EPFL

Thèse n° 8120

Adaptation in Stochastic Algorithms: From Nonsmooth Optimization to Min-Max Problems and Beyond

Présentée le 25 août 2021

Faculté des sciences et techniques de l'ingénieur
Laboratoire de systèmes d'information et d'inférence
Programme doctoral en informatique et communications

pour l'obtention du grade de Docteur ès Sciences

par

Ahmet ALACAOGLU

Acceptée sur proposition du jury

Prof. A. H. Sayed, président du jury
Prof. V. Cevher, directeur de thèse
Prof. S. Wright, rapporteur
Prof. O. Fercoq, rapporteur
Prof. M. Jaggi, rapporteur

École
polytechnique
fédérale
de Lausanne

2021

Young man, in mathematics
you don’t understand things.
You just get used to them.
— John von Neumann

Anneme…

Acknowledgements

First and foremost, I would like to express my gratitude to my advisor Volkan Cevher. Volkan introduced me to the field of optimization and supported me throughout the way. I am especially thankful to him for being so generous with his time for many research discussions, ideas and also non-technical discussions about life. I have greatly benefited from the collaborative environment Volkan built in LIONS and the freedom and guidance that he gave me while pursuing the research directions I valued. I am grateful to him for his belief in me right from the start and his optimism that helped me greatly for the last five years.

I would like to thank Stephen Wright, Olivier Fercoq, Martin Jaggi and Ali Sayed for being in my jury and their valuable constructive criticism which helped to improve this thesis. I also wish to thank Olivier, Quoc Tran-Dinh and Volkan for arranging my academic visits to Paris and Chapel Hill; and Olivier and Quoc for being such great hosts. I am so grateful to Stephen for his hospitality even before my arrival in Madison and I am looking forward to joining his lab as a postdoc.

This thesis would not have been possible without the help and support of my amazing collaborators Olivier, Yura Malitsky, Quoc, Panayotis Mertikopoulos, Niao He and Volkan. I was extremely fortunate to work with and learn from Olivier since the beginning of my PhD. I learned a great deal from him and his papers which I continue to use in my work every day. Each conversation I had with him has been full of inspiration and he was always too kind and patient with me from the beginning. Later in my PhD, I was so lucky to have met Yura who not only taught me a lot about mathematics but also about life. I am especially thankful to him for answering to my easy mathematical questions that I am ashamed to ask anyone else. Panayotis, Nadav Hallak, Quoc and Niao were always too generous with their time either for mathematical discussions, career advices or general advices in life.

I would like to thank Ali Kavis, Yura, Ya-Ping Hsieh and Alp Yurtsever for their feedback on my thesis and to Leello Dadi for writing the French abstract. Our lab admin Gosia Baltaian was incredibly helpful to me on so many different occasions.

I was incredibly fortunate to have met Emre Telatar before arriving in Lausanne, who was my mentor during my PhD. He was always there for me in difficult times and his advices were invaluable. He continues to be a role model for me both academically and personally.

My undergraduate research advisor in Bilkent was Emine Ulku Saritas whose advices have been instrumental since the beginning of my academic career, for which I am incredibly

thankful. I am also grateful to my other professors at Bilkent for inspiring me to start a PhD.

I have met some amazing friends and colleagues in Lausanne who made the last five years full of fun and laughter. This thesis would not have been possible without their support in difficult times and without their company in the joyful times. I am also indebted to my friends from high school and university who have supported me in my journey.

Last but not the least, I wish to express my gratitude to my parents Dilek and Muharrem, and my sister Merve for constant support and encouragement in all my choices: Without you by my side, I could not start a PhD, let alone finish it.

Lausanne, August 25, 2021 A. A.

Abstract

Stochastic gradient descent (SGD) and randomized coordinate descent (RCD) are two of the workhorses for training modern automated decision systems. Intriguingly, convergence properties of these methods are not well-established as we move away from the specific case of smooth minimization. In this dissertation, we focus on related problems of nonsmooth optimization and min-max optimization to improve the theoretical understanding of stochastic algorithms.

First, we study SGD-based adaptive algorithms and propose a regret analysis framework overcoming the limitations of the existing ones in the convex case. In the nonconvex case, we prove convergence of an adaptive gradient algoritm for solving constrained weakly convex optimization, generalizing the previously known results on unconstrained smooth optimization. We also propose an algorithm combining Nesterov's smoothing with SGD to solve convex problems with infinitely many linear constraints, with optimal rates.

Then, we move on to convex-concave min-max problems with bilinear coupling and analyze primal-dual coordinate descent (PDCD) algorithms. We obtain the first PDCD methods with the optimal $\mathcal{O}(1/k)$ rate on the the standard optimality measure expected primal-dual gap, which was an open question since 2014. Our analysis also aims to explain the practical behavior of these algorithms by showing that the last iterate enjoys adaptive linear convergence without altering the parameters, depending on a certain error bound condition. Furthermore, we propose an algorithm combining the favorable properties of two branches of PDCD methods: the new method uses large step sizes with dense data and its per-iteration cost depends on the number of nonzeros of the data matrix. Thanks to these unique properties, this method enjoys compelling practical performance complementing its rigorous theoretical guarantees.

Next, we consider monotone variational inequalities that generalize convex-concave min-max problems with nonbilinear coupling. We introduce variance reduced algorithms that converge under the same set of assumptions as their deterministic counterparts and improve the best-known complexities for solving convex-concave min-max problems with finite-sum structure. Optimality of our algorithms for this problem class is established in a recent work via matching lower bounds.

Finally, we show our preliminary results on policy optimization methods for solving two player zero-sum Markov games for competitive reinforcement learning (RL). Even though this is a nonconvex-nonconcave min-max problem in general, thanks to the special structure, it is tractable to find an approximate Nash equilibrium. We introduce an algorithm that improves the best-known sample complexity of policy gradient methods. This development combines tools from RL and stochastic primal-dual optimization, showing the importance of techniques from convex-concave optimization.

Key words: randomized primal-dual methods, coordinate descent, variance reduction, adaptive gradient algorithms, min-max optimization, linearly constrained optimization, variational inequalities.

Résumé

L'algorithme du gradient stochastique et la méthode de descente par coordonnée randomisée sont deux des principales approches utilisées pour l'entraînement des systémes de décision modernes. Pourtant, leurs propriétés de convergence restent largement inexplorées lorsqu'il s'agit d'optimiser des fonctions non lisses. Dans ce manuscrit, nous étudions des problémes d'optimisation non lisses, ainsi que des problémes d'optimisation min-max dans le but d'améliorer notre compréhension des ces méthodes stochastiques.

Nous commencons par l'étude des algorithmes de gradient stochastique adaptatifs et proposons un cadre d'analyse qui permet de dépasser les limites des cadres existant pour les fonctions convexes. Dans le cas non-convexe, nous démontrons la convergence d'un algorithme adaptatif pour résoudre des problémes [weakly convex constrained] en généralisant les résultats connus pour les problémes non-contraint et lisses. Nous proposons également une méthode combinant le "smoothing" de Nesterov avec l'algorithme du gradient stochastique qui permet de résoudre des problémes convexes avec une infinité de contraintes linéaires en atteignant un taux de convergence optimal.

Nous passons ensuite á l'étude des problémes min-max convexe-concave ayant un couplage bilinéaire et nous analysons les algorithmes primal-dual de descente par coordonnée (PDCD). Nous obtenons la premiére méthode PDCD atteignant le taux de convergence optimal de O(1/k) mesuré en termes de l'écart primal-dual, résolvant ainsi un probléme resté ouvert depuis 2014. Notre analyse cherche également á expliquer le comportement observé en pratique de ces algorithmes en montrant que, sous une condition simple, le dernier itéré admet un taux de convergence linéaire adaptatif sans nécessiter une modification des paramètres. En outre, nous proposons un algorithme qui mélange les propriétés favorables des deux branches des méthodes PDCD : notre méthode utilise des pas larges pour des données denses et son coût par itération dépend du nombre de coefficients non nuls de la matrice des données. Ces propriétés uniques permettent á notre méthode d'avoir de bonnes performances en pratique qui complimentent bien ses rigoureuses garanties théoriques.

Nous continuons par considérer les inégalités variationnelles monotones qui généralisent les problémes min-max convexe-concave avec couplage non bilinéaire. Nous introduisons des algorithmes avec réduction de variance qui convergent sous les mêmes hypothéses que leurs équivalents déterministes et qui améliorent la complexité de résolution problémes min-max convexe-concave ayant une structure de somme finie. L'optimalité de notre méthode a été démontrée indépendamment par d'autres auteurs par l'établissement d'une borne inférieure.

Enfin, nous montrons quelques résultats préliminaires sur l'optimisation de politique dans les jeux markoviens à somme nulle à deux joueurs dans le cadre de l'apprentissage par renforcement compétitif. Bien que, en general, ces problémes soit des problemes min-max non-convexe non-concave, la structure spéciale dont ils jouissent rend possible la détermination d'un équilibre de Nash approché. Nous introduisons un algorithme qui atteint la meilleure complexité en termes d'échantillons connue pour les méthodes dites de "policy gradient". L'élaboration de cette méthode a fait appel à des outils de l'apprentissage par renforcement et à des outils de l'optimisation primale-duale stochastique, ce qui nous montre l'importance des techniques issues de l'optimisation convexe-concave.

Key words : algorithmes primal-dual randomisée, méthode de descente par coordonnée, réduction de variance, algorithmes de gradient adaptatifs, des problémes min-max, des problémes avec de contraintes linéaires, inégalités variationnelles.

Bibliographic Note

This dissertation is based on the following publications:

- Ahmet Alacaoglu, Quoc Tran-Dinh, Olivier Fercoq and Volkan Cevher. “Smooth Primal-Dual Coordinate Descent Algorithms for Nonsmooth Convex Optimization.” Conference on Neural Information Processing Systems (NeurIPS), 2017
- Olivier Fercoq, Ahmet Alacaoglu, Ion Necoara and Volkan Cevher. “Almost surely constrained convex optimization.” International Conference on Machine Learning (ICML), 2019
- Ahmet Alacaoglu, Yura Malitsky, Panayotis Mertikopoulos and Volkan Cevher. “A new regret analysis for Adam-type algorithms.” International Conference on Machine Learning (ICML), 2020
- Ahmet Alacaoglu, Olivier Fercoq and Volkan Cevher. “Random extrapolation for primal-dual coordinate descent.” International Conference on Machine Learning (ICML), 2020
- Ahmet Alacaoglu, Olivier Fercoq and Volkan Cevher. “On the convergence of stochastic primal-dual hybrid gradient.” SIAM Journal on Optimization, to appear, 2021
- Ahmet Alacaoglu, Yura Malitsky and Volkan Cevher. “Forward-reflected-backward method with variance reduction.” Computational Optimization and Applications, 2021
- Ahmet Alacaoglu, Yura Malitsky and Volkan Cevher. “Convergence of adaptive algorithms for constrained weakly convex optimization.” arXiv:2006.06650
- Ahmet Alacaoglu and Yura Malitsky. “Stochastic Variance Reduction for Variational Inequality Methods.” arXiv: 2102.08352
- Ahmet Alacaoglu, Niao He, Luca Viano and Volkan Cevher. “Sample-efficient actor-critic methods for solving zero-sum Markov game.” Manuscript, 2021

At the end of some chapters we include a “Bibliographic Note” section to distinguish the contributions of the author of this dissertation in the abovementioned publications. For the chapters without a “Bibliographic Note”, the author of the dissertation contributed to all the results within the chapter.

My other publications on the same research line, but not included in the dissertation:

- Quoc Tran-Dinh, Ahmet Alacaoglu, Olivier Fercoq and Volkan Cevher. "An Adaptive Primal-Dual Framework for Nonsmooth Convex Minimization." Mathematical Programming Computation, 2019
- Mehmet Fatih Sahin, Armin Eftekhari, Ahmet Alacaoglu, Fabian Latorre and Volkan Cevher. "An Inexact Augmented Lagrangian Framework for Nonconvex Optimization with Nonlinear Constraints." Conference on Neural Information Processing Systems (NeurIPS), 2019
- Maria-Luiza Vladarean, Ahmet Alacaoglu, Ya-Ping Hsieh and Volkan Cevher. "Conditional gradient methods for stochastically constrained convex minimization". International Conference on Machine Learning (ICML), 2020

Contents

1 Introduction

Two of the most widely-used optimization algorithms nowadays are stochastic gradient descent (SGD) and coordinate descent (CD), the ideas of which have been around for more than 70 years [RM51, Hil57, Kar37]. Despite their long histories, the research activity around these methods have witnesses a surge of interest in the last two decades owing to the so-called *big data*. Simple update rules with cheap per-iteration costs make them suitable in this era.

In 1990s, the focus of continuous optimization was on interior point methods which, in contrast to stochastic methods, had computationally intensive iterations. While the main tools of optimization are significantly different compared to 1990s, the guiding principle remains: *take advantage of structure when present.* In the last decade, the particular revelation surrounding SGD and CD was that randomization can be used as a *technique* for carefully harnessing the structure of optimization problems, to design faster algorithms. We refer to the class of algorithms based on SGD and randomized CD as *stochastic algorithms*.

Two concerns while deploying these methods in practice are *reliability* and *adaptivity*. The former is related to the theoretical convergence guarantees of algorithms. The latter defines the ability of an algorithm to enjoy fast convergence when favorable structures are present, without the need to modify the algorithm. Surprisingly, as we move away from the standard setting of *smooth minimization*, these natural requirements are often not satisfied for stochastic algorithms.

This dissertation focuses on improving the understanding of stochastic algorithms for solving structured *nonsmooth optimization* and *min-max* problems. We now make a brief technical excursion and introduce these concepts more concretely to facilitate our discussion and introduce the contributions.

1.1 Problem description

The basic continuous optimization template is

$$\min_{x\in\mathbb{R}^d} \ell(x), \tag{1.1}$$

where $\ell\colon \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$. Depending on the structure of ℓ, an optimization problem can be *smooth* or *nonsmooth, (strongly) convex* or *non-(strongly) convex.* In particular, we say that ℓ is

▷ L-smooth if its gradient is L-Lipschitz continuous:

$$\|\nabla\ell(x_1)-\nabla\ell(x_2)\| \le L\|x_1-x_2\|, \quad \text{for all } \; x_1, x_2 \in \mathbb{R}^d,$$

▷ convex if

$$\ell(\alpha x_1+(1-\alpha)x_2) \le \alpha\ell(x_1)+(1-\alpha)\ell(x_2), \;\; \text{for any } \; \alpha\in[0,1], x_1, x_2\in\mathbb{R}^d,$$

▷ μ-strongly convex if there exists $\mu > 0$ such that

$$\ell(\alpha x_1+(1-\alpha)x_2) \le \alpha\ell(x_1)+(1-\alpha)\ell(x_2)-\alpha(1-\alpha)\frac{\mu}{2}\|x_1-x_2\|^2, \;\; \text{for all } \; x_1, x_2\in\mathbb{R}^d.$$

The most favorable case is when ℓ is smooth and strongly convex. In this case, stochastic algorithms and their deterministic counterparts attain fast linear convergence [Nes03].

1.1.1 Lack of smoothness

Even though smooth optimization problems are common in practice, nonsmooth problems provide a much more powerful framework. Even though nonsmooth problems are difficult in full generality [NY83], empirical observations show that stochastic algorithms can still enjoy fast convergence for solving them. Nesterov showed in [Nes05] that by opening the *black box* and studying structured nonsmooth problems, we can avoid the worst-case lower bounds in [NY83].

One class of structured nonsmooth problems is constrained optimization. We define constraints via indicator functions $\delta_{\mathcal{K}}$ which is equal to 0 for admissible values ($x\in\mathcal{K}$) and $+\infty$ otherwise.

Problems with a smooth objective and a constraint can be solved as if they are fully smooth when the constraint set admits an efficient projection operator. On the other hand, when the constraint set is defined with a linear equality, projection requires solving a linear system. With large scale problems, algorithms requiring such projection steps are not feasible.

We formalize linearly constrained optimization as

$$\min_{x\in\mathbb{R}^d} \ell(x), \qquad \ell(x) = f(x)+\delta_{\mathcal{K}}(Ax)+\delta_{\mathcal{B}}(x) \qquad \Longrightarrow \qquad \min_{x\in\mathcal{B}} f(x) : Ax\in\mathcal{K}, \tag{1.2}$$

where A is a linear operator, $\mathcal{K}, \mathcal{B}$ are convex sets and f is a convex function.

This template covers classical problems such as quadratically constrained quadratic programs and it also naturally arises for representing the communication graph in distributed optimization. Despite being classical, these fundamental problems continue to emerge in important applications in machine learning (ML) and reinforcement learning (RL) [Wan20, AAF+20, BRS18, JOV09, TSR+05].

The second example, also intimately connected to the first, is min-max optimization:

$$\min_{x\in\mathbb{R}^d} \ell(x), \qquad \ell(x) = \delta_{\mathcal{X}}(x) + \max_{y\in\mathcal{Y}} \Phi(x,y) \qquad \Longrightarrow \qquad \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \Phi(x,y), \tag{1.3}$$

where Φ is a coupling function describing the interaction between variables x, y. This problem is central in game theory [vN28] and it is also a convenient way to solve constrained problems via Lagrangian duality [Ber99]. Recently, the problem (1.3) is getting increasingly more popular thanks to many applications in adversarially robust, safe, and fair formulations of ML [MMS+18, BTSK17, HSNL18, GPAM+14], competitive RL [DFG20, WLZL21], and many others.

In this dissertation, we will be mostly focusing on the case when Φ is convex-concave.

1.1.2 Presence of favorable structure

As mentioned before, optimization algorithms can attain a fast rate of convergence for smooth and strongly convex problems. However, for the problems described in the previous section, none of these assumptions hold in general. On the other hand, their specific structures make them easier compared to generic nonsmooth problems. Next, we give three representing examples of the favorable structures that can be used by stochastic algorithms.

◦ When the data matrix A in (1.2) is sparse, we can design CD-based algorithms to have a per-iteration cost depending on the nonzeros of the associated matrix rather than its dimensions.

◦ When the coupling function in (1.3) is given as a a finite-sum of component functions, we can design SGD-based methods to obtain stochastic estimates with a reduced variance.

◦ Even though the functions in eqs. (1.2) and (1.3) are not smooth and not strongly convex in general, they might satisfy error bound conditions, such as having piecewise linear-quadratic structure. For such problems, we can prove linear rate of convergence.

1.2 Contributions and Organization

In this dissertation, we have two research goals for solving nonsmooth optimization and min-max problems: *(i)* providing a rigorous understanding of some existing stochastic algorithms that have seen empirical success, *(ii)* introducing new provable algorithms when the existing ones are insufficient.

Chapter 2: *Convergence of adaptive gradient algorithms for nonsmooth optimization with convex and nonconvex objectives (based on works [AMMC20, AMC20]).*

In this chapter, we focus on nonsmooth stochastic optimization problem

$$\min_{x\in\mathcal{K}} f(x) = \mathbb{E}_\xi[f_\xi(x)],$$

where $\mathcal{K}$ is convex, closed and f is either convex or weakly convex.

The class of Adam-type algorithms, (also referred to as adaptive gradient algorithms), are extremely popular in deep learning applications. These methods incorporate exponential moving averaging (EMA) for past gradient vectors and their element-wise squares, with the EMA parameters denoted by β_1 and β_2. In practice, these parameters are chosen to be constant values close to 1 (default values in Tensorflow and Pytorch are $\beta_1 = 0.9$).

The work [RKK18] identified that the convergence analysis in [KB15] for Adam is incorrect and proposed new Adam-type methods with small modifications to ensure convergence. On the other hand, even the corrected regret analysis by [RKK18] follows the analysis of [KB15] closely, therefore requires a linearly diminishing β_1 schedule, which is opposite to what is used practice. Moreover, these methods are being used with increasingly complicated neural network structures, however, their convergence is only known for nonconvex problems which are smooth and unconstrained.

- We show that the requirement of β_1 in the convex analysis is a mere artifact of the proof, and we propose a new regret analysis framework that allows a constant β_1 parameter, with provably better bounds compared to previous works. We show the generality of our technique by applying it to most of the existing convergent Adam-type methods and equip all of their guarantees with a constant β_1 choice.

- Using our framework, we propose a convergence analysis for an adaptive gradient algorithm for solving a class of nonsmooth nonconvex optimization problems. The problem class we consider is constrained weakly convex problems which contain smooth nonconvex optimization with convex constraints and also potentially nonsmooth problems. This result generalizes the class of nonconvex problems where adaptive gradient algorithms have convergence guarantees.

Chapter 3: *Smoothing and stochastic algorithms for linearly constrained problems (based on works [ADFC17, FANC19])*

In this chapter, we focus on the linearly constrained optimization problem,

$$\min_{x\in\mathbb{R}^d} h(x) + g(x), \text{ s.t. } x \in \cap_{i=1}^N \mathcal{K}_i, \text{ where } \mathcal{K}_i = \{x\colon A_i x \in b_i\}.$$

with convex functions h, g, matrix A_i and set b_i.

A useful structure in this problem is the separability of the constraints, which can be used by coordinate methods (CD) to decrease per iteration cost. Moreover, these methods use coordinate-wise Lipschitz constants for h and norms of blocks of the matrix A rather than their global counterparts. These features help CD methods show fast convergence in practice. Prior to our work [ADFC17] we were not aware of any CD algorithm with rate guarantees for linearly constrained problem.

• We first focus on the case where N can be potentially infinite to solve problems with infinite number of linear inclusion constraints. Using Nesterov's smoothing along with SGD [RM51], we obtain an algorithm with the rate $\tilde{\mathcal{O}}(1/\sqrt{k})$ when $h+g$ is only convex and $\tilde{\mathcal{O}}(1/k)$ when $h+g$ is restricted strongly convex, both of which are optimal (up to log factors) even in the unconstrained setting [AWBR09].

• Unlike the previous approach for this problem based on alternating projections, our work does not assume projectability of $\mathcal{K}_i$ which, depending on the dimensions of A_i, can be prohibitive.

• When N is finite, by combining Nesterov's smoothing technique [Nes05] and accelerated proximal coordinate descent [FR15], we obtain a CD algorithm with $\mathcal{O}(1/k)$ rate which is optimal in its dependence on k [Nes05].

• Our algorithm uses coordinatewise Lipschitz constants of h and norms of blocks of A, rather than their global counterparts.

Chapter 4: *Convergence of primal-dual coordinate descent (PDCD) methods and adaptivity to functional structures (based on work [AFC21])*

Primal-dual hybrid gradient method (PDHG) [CP11] is a classical algorithm for solving convex-concave min-max problems with bilinear coupling function:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} g(x) + \sum_{i=1}^{n} \langle A_i x, y^{(i)}\rangle - f_i^*(y^{(i)}) \tag{1.4}$$

Two fundamental results on the convergence of PDHG are *(i)* asymptotic convergence of the sequence to a solution, *(ii)* $\mathcal{O}(1/k)$ convergence rate on the primal-dual gap function, which is the standard optimality measure. Since 2014, different randomized variants of PDHG are proposed [DL14, ZL15, FB19, LFP19]. Recently, Chambolle et al. [CERS18] introduced stochastic PDHG (SPDHG) which have been popular especially in computational imaging due to its practical performance, with implementations in different software packages [EMC$^+$17, PAD$^+$21, LL19, KPB$^+$19]. Despite the practical interest, abovementioned standard theoretical results regarding the convergence of SPDHG remained open. In fact, in the work that introduced one of the first randomized PDHG algorithm in 2014 [DL14], the difficulty of deriving the $\mathcal{O}(1/k)$ guarantee for the expected primal-dual gap was pointed out and this question remained open ever since.

In this chapter, we analyze SPDHG and prove under convexity assumption:

• Almost sure convergence of the sequence to a solution.

- $\mathcal{O}(1/k)$ convergence rate for the expected primal-dual gap.

- For the latter result, we introduce a new analysis technique for PDCD methods, inspired by [NJLS09], which is of independent interest.

One reason for popularity of PDHG/SPDHG in practice is the linear rate of convergence often observed in empirical studies. On the other hand, the only case where we knew linear convergence of SPDHG was when f, g_i are strongly convex and the step sizes are set accordingly [CERS18]. To derive a more general result on linear rate of convergence, we use an error bound condition [DR09]. This condition not only holds for restricted strongly convex functions, but also for problems with piecewise linear quadratic (PLQ) objective and linear constraints, including quadratic programming, Lasso, support vector machines, etc.

- We show that without any modification on the algorithmic parameters, SPDHG converges linearly with the error bound condition, which is a first step towards explaining its favorable adaptive linear convergence in practice.

Chapter 5: *Adapting to sparsity of the data via PDCD methods (based on work [AFC20])*

One drawback of SPDHG [CERS18] is that its per iteration cost depends on one of the dimensions of the data matrix, and does not decrease with sparse data. An alternative of SPDHG in the prior literature was due to [FB19] where the per iteration cost depends on the number of nonzeros of the data matrix, adapting to sparsity when it is encountered. The drawback of the method of [FB19] however, was the restriction to smaller step sizes than SPDHG with dense data. In this chapter, we design an algorithm that achieves the best of both worlds, for solving (1.4).

- Our algorithm is the first to simultaneously use large step sizes with dense data and have cheap per-iteration cost with sparse data.

- By assuming convexity, we prove almost sure convergence of the sequence to a solution.

- By assuming convexity, $\mathcal{O}(1/k)$ rate for expected duality gap.

- We prove adaptive linear convergence with an error bound condition (as in the result proven for SPDHG in the previous chapter).

- Our algorithm can also handle an additional smooth term on top of (1.4), with step size rule depending on coordinate-wise Lipschitz constants.

We conduct experiments on sparse, moderately sparse and dense datasets, to illustrate the adaptation of our method to sparsity.

- As predicted by theory, our method enjoys the best performance in all the sparsity configurations, compared to SPDHG [CERS18] and the method of [FB19].

We also compared our algorithm with stochastic variance reduction methods from ML literature [JZ13] that are designed specifically for strongly convex strongly concave problems and our method was competitive despite its generality.

Chapter 6: *Variance reduction for provably faster min-max optimization (based on works [AMC21, AM21])*

In this chapter, we solve potentially nonbilinear convex-concave min-max problems with finite sum structure.

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\Phi(x,y). \quad \text{Let} \quad (x,y)=z, \quad F(z)=[\nabla_x\Phi(x,y),-\nabla_y\Phi(x,y)], \qquad F(z)=\sum_{i=1}^{N}F_i(z). \tag{1.5}$$

This structure is present for example in matrix games, Lagrangian formulation of constrained optimization, empirical risk minimization. Prior to our work, Carmon et al. [CJST19] focused on matrix games and showed that variance reduction can improve the complexity of Mirror-Prox. On the other hand, [CJST19] required additional strong assumptions on top of Mirror-Prox, such as boundedness of the domain or convexity of the map $z\mapsto\langle F(z),z-u\rangle,\forall u\in\mathcal{Z}$. These assumptions are satisfied for matrix games, but can be violated even for slightly more general linearly constrained optimization problems. Moreover, the analysis in [CJST19] required conservative step sizes, and for the nonbilinear problem, a complicated three-loop algorithm was used with suboptimal complexity. In addition, almost sure convergence of the sequence was not proven.

• We design a variance reduction framework for min-max problems, and apply to extragradient, forward-reflected-backward, forward-backward-forward methods, with Euclidean and Bregman setups.

• For problem (1.5), our algorithms converge under the same set of assumptions as deterministic methods, unlike previous work with spurious assumptions.

• Our results match the best-known complexity for bilinear case, and improve the best-known complexity in the nonbilinear case.

• Our algorithms enjoy more freedom in choosing the step size, resulting in better practical performance compared to [CJST19], even in the bilinear case.

• We show that our algorithms can converge linearly while staying agnostic to the strong convexity (strong monotonicity) parameter.

• We also prove almost sure convergence of the sequence for monotone variational inequalities and monotone inclusions.

• A recent independent work [HXZ21] proved matching lower bounds in this setting, establishing the optimality of our algorithms.

Chapter 7: *Improving sample complexity of policy optimization for competitive reinforcement learning*

In this chapter, we present our preliminary results for solving two player zero-sum Markov games (also known as stochastic games). This problem template generalizes matrix games and Markov Decision Processes (MDP) and used in competitive RL. Even without function approximation, this problem is nonconvex nonconcave in general. We focus on a class of policy optimization methods, which are also known as actor-critic algorithms or policy gradient algorithms. These methods jointly learn the optimal policy and the optimal value function.

By building on the recent advances on policy gradient methods for single agent RL and stochastic primal-dual optimization, we prove sample complexity results for solving two player zero-sum Markov games, for reaching to an approximate Nash equilibrium. Our results improve the best-known sample complexity of policy gradient methods in the literature.

1.3 Notation and definitions

We introduce the basic notation, definitions and properties used throughout the dissertation. Chapter-specific notations and definitions are included in the corresponding chapters.

Notation. Let $\mathcal{X}$ and $\mathcal{Y}$ be Euclidean spaces with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$. We also define $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$ and $z=(x,y)\in\mathcal{Z}$. For a positive definite matrix Q, we use $\langle x,y\rangle_Q=\langle Qx,y\rangle$ to denote weighted inner product and $\|x\|_Q^2=\langle Qx,x\rangle$ to denote weighted Euclidean norm. For a set $\mathcal{C}$, and positive definite Q, distance of a point x to $\mathcal{C}$, measured in weighted norm is defined as $\operatorname{dist}_Q^2(x,\mathcal{C})=\min_{y\in\mathcal{C}}\|x-y\|_Q^2=\|x-\mathcal{P}_{\mathcal{C}}^Q(x)\|_Q$, where we have defined the corresponding projection operator $\mathcal{P}$ implicitly. When $Q=I$, we drop the subscript and write $\operatorname{dist}(x,\mathcal{C})$.

We define for $\sigma\in\mathbb{R}^n$, the diagonal matrix $D(\sigma)=\operatorname{diag}(\sigma_1,\ldots,\sigma_n)$. We define the support function of the set $\mathcal{K}$ as $\operatorname{supp}_{\mathcal{K}}(x)=\sup_{y\in\mathcal{K}}\langle x,y\rangle$.

Given a vector x, we access i-th element as $x^{(i)}$. We define e_i as the i-th unit vector and $E(i)=e_ie_i^\top$. Unless used with a subscript, 1 in Kronecker products denotes the all-ones vector $1_n\in\mathbb{R}^n$. The notation $[N]$ represents the set $\{1,\ldots,N\}$.

Properties of convex functions. Unless indicated otherwise, we consider a proper lower semicontinuous (l.s.c.) convex function $h\colon\mathcal{X}\to\mathbb{R}\cup\{+\infty\}$.

Domain of a function h is denoted as $\operatorname{dom} h$. We encode equality constraints using the indicator function $\delta_{\{b\}}$ such that $\delta_{\{b\}}(x)=0$ if $x=b$ and $\delta_{\{b\}}(x)=+\infty$ if $x\neq b$.

The proximal operator of h with weighting matrix τ, is defined as

$$\operatorname{prox}_{\tau,h}(x)=\arg\min_{u\in\mathcal{X}} h(u)+\frac{1}{2}\|u-x\|_{\tau^{-1}}^2. \tag{1.6}$$

When τ is a scalar, we write $\operatorname{prox}_{\tau h}$ instead of $\operatorname{prox}_{\tau,h}$. We sometimes say *proximable* to mean that the proximal operator of h can be computed efficiently.

A standard identity is

$$\bar{x}=\operatorname{prox}_h(\hat{x}) \iff \langle\bar{x}-\hat{x},x-\bar{x}\rangle\geq h(\bar{x})-h(x)\quad\forall x\in\mathcal{X}. \tag{1.7}$$

The Fenchel conjugate of h is defined as

$$h^*(y) = \sup_{z\in\mathcal{X}} \langle z, y\rangle - h(z).$$

We say that h is:

∘ μ-strongly convex if $h(x) - \frac{\mu}{2}\|x\|^2$ is convex for all $x \in \mathcal{X}$.

∘ μ-weakly convex if $h(x) + \frac{\mu}{2}\|x\|^2$ is convex for all $x \in \mathcal{X}$.

∘ L-Lipschitz (or L-Lipschitz continuous) if $|h(x_1) - h(x_2)| \le L\|x_1 - x_2\|$ for all $x_1, x_2 \in \mathcal{X}$.

∘ L-smooth if h is differentiable and its gradient is Lipschitz continuous: $\|\nabla h(x_1) - \nabla h(x_2)\| \le L\|x_1 - x_2\|$, for all $x_1, x_2 \in \mathcal{X}$.

These definitions can be written with other norms that we omit for simplicity.

Lagrangian duality. Given a *primal* optimization problem

$$\min_{x\in\mathcal{X}} h(x) + g(x) + f(Ax), \tag{1.8}$$

where all functions are proper l.s.c. convex, $h\colon \mathcal{X} \to \mathbb{R}$ is smooth, $g\colon \mathcal{X} \to \mathbb{R}\cup\{+\infty\}$, $f\colon \mathcal{Y} \to \mathbb{R}\cup\{+\infty\}$, $A\colon \mathcal{X} \to \mathcal{Y}$. We can write the Lagrangian as

$$\mathcal{L}(x, y) = h(x) + g(x) + \langle Ax, y\rangle - f^*(y),$$

and the dual problem as

$$\max_{y\in\mathcal{Y}} \left\{ \min_{x\in\mathcal{X}} L(x, y) \right\}.$$

Unless stated otherwise, we assume a *primal-dual solution pair* $z_\star = (x_\star, y_\star)$ exists. Then, we have for any $x \in \mathcal{X}$, $y \in \mathcal{Y}$,

$$\mathcal{L}(x_\star, y) \le \mathcal{L}(x_\star, y_\star) \le \mathcal{L}(x, y_\star).$$

Optimality measure. Primal-dual gap is defined as:

$$\mathrm{Gap}(\bar{x}, \bar{y}) = \sup_{x,y} \mathcal{L}(\bar{x}, y) - \mathcal{L}(x, \bar{y}).$$

It is easy to see that $\mathrm{Gap}(\bar{x}, \bar{y}) = 0$ if and only if $(\bar{x}, \bar{y})$ is a primal-dual solution. Consequently, for stochastic algorithms, the optimality measure is $\mathbb{E}[\mathrm{Gap}(\bar{x}, \bar{y})]$.

Operators. In the last two chapters, we work with operator notation. We say that $F\colon \mathcal{Z} \to \mathcal{Z}$ is

∘ monotone if: $\langle F(z_1) - F(z_2), z_1 - z_2\rangle \ge 0 \quad \forall z_1, z_2 \in \mathcal{Z}$.

∘ μ-strongly monotone if: $\langle F(z_1) - F(z_2), z_1 - z_2\rangle \ge \frac{\mu}{2}\|z_1 - z_2\|^2 \quad \forall z_1, z_2 \in \mathcal{Z}$, with $\mu > 0$.

∘ L-Lipschitz if: $\|F(z_1) - F(z_2)\| \le L\|z_1 - z_2\| \quad \forall z_1, z_2 \in \mathcal{Z}$.

Constraint qualification. Let us consider the problem (1.8). Slater's condition is a sufficient condition for strong duality that states $0 \in \operatorname{ri}(A \operatorname{dom} g - \operatorname{dom} f)$, where ri denotes relative interior [BC11]. For existence of primal-dual solution, it is standard to assume that a solution exists for primal problem (1.8) and Slater's condition holds.

2 Convergence of adaptive gradient algorithms for nonsmooth problems

In this section we focus on the convergence properties of adaptive gradient algorithms that are variants of the popular ADAM (Adaptive Moment estimation) algorithm, in the convex and nonconvex settings. We first identify a limitation in the existing convex regret analyses which requires parameter choices inconsistent with practice. We propose a new regret analysis framework to overcome this issue. Using our framework, we also prove the first convergence result of Adam-type adaptive gradient algorithms for solving a class of nonsmooth nonconvex optimization problems.

This chapter is based on joint works with Yura Malitsky, Panayotis Mertikopoulos and Volkan Cevher [AMMC20, AMC20].

2.1 Introduction

One of the most popular optimization algorithms for training neural networks is ADAM [KB15], which is a variant of the general class of adaptive gradient algorithms [DHS11]. The main novelty of ADAM is to apply an exponential moving average (EMA) to gradient estimate (first-order) and to element-wise square-of-gradients (second-order), with parameters β_1 and β_2. In practice, constant β_1 and β_2 values are used (the default parameters in PYTORCH and TENSORFLOW are $\beta_1 = 0.9$ and $\beta_2 = 0.999$). However, the regret analysis in [KB15] requires $\beta_1 \to 0$ with a linear rate, causing a clear discrepancy between theory and practice.

Recently, [RKK18] showed that the regret analysis of ADAM for online convex optimization (OCO) in [KB15] contains a mistake and proposed AMSGRAD and ADAMNC as convergent alternatives, along with proofs for nonconvergence of ADAM. Following this discovery, many variants of ADAM are proposed with regret guarantees [RKK18, CZT+20, HWD19]. Unfortunately, in all these analyses, the requirement $\beta_1 \to 0$ is inherited from [KB15] and is needed to derive the optimal $\mathcal{O}(\sqrt{T})$ regret. In contrast, for favorable practical performance, methods continue to use constant β_1 in experiments.

Given the nonconvergence issues surrounding these methods, one can wonder whether there is an inherent obstacle – in the proposed methods or the setting – which prohibits optimal regret bounds with a constant β_1? In this chapter, we show that this specific discrepancy between the theory and practice is only an artifact of the previous analyses. We point out the shortcomings responsible for this artifact, and then introduce a new analysis framework that attains optimal regret bounds for OCO with constant β_1 at no additional cost (and with better constants in the obtained bounds).

Given that the main applications of Adam-type methods is on training neural networks, their behavior in nonconvex problems is of paramount importance. Many recent works made progress on this direction [CLSH19, CZT$^+$20, ZSJ$^+$19, WWB19, LO19, DBBU20]. These works focus on unconstrained smooth stochastic optimization, where the standard analysis framework of the stochastic gradient descent (SGD) [GL13] can be used. Convergence of adaptive methods for the more general setting of constrained and/or nonsmooth stochastic nonconvex optimization has remained unexplored, while these settings have broad practical applications [VPG19, MNSF17, MMS$^+$18, IEAL18, DD19, DP19].

The difficulty of handling the combination of nonconvexity, adaptive step sizes, momentum and constraints is mentioned in [CLX$^+$19, Section 4.3]. In particular, in terms of our analysis, *(i)* adaptivity introduces coupling between the step sizes and iterates, *(ii)* time-dependent diagonal step size requires an analysis framework based on variable metrics, *(iii)* using a constant β_1 requires the new analysis framework we develop in this chapter. For details, please see the discussions around Lemmas 2.10 and 2.11.

2.1.1 Contributions

▷ In the convex setting, our technique obtains data-dependent $\mathcal{O}(\sqrt{T})$ regret bounds for AMSGRAD and ADAMNC [RKK18].

▷ We apply our technique to a strongly convex variant of ADAMNC, known as SADAM [WLC$^+$20], yielding data-dependent logarithmic regret with constant β_1.

To the best of our knowledge, these are the first optimal regret bounds with constant β_1.

▷ Finally, we apply our framework to derive a convergence guarantee for AMSGRAD for solving constrained weakly convex optimization. This is the first result for adaptive gradient methods for solving nonconvex problems beyond the simplest unconstrained smooth template.

It is worth noting that even though our analysis is more flexible and it provides better bounds than prior works, it is not sufficient to explain why nonzero β_1 helps in practice. This is an interesting question requiring further investigation and is outside the scope of this chapter.

Organization. In the first part of the chapter, we analyze Adam-type algorithms for online convex optimization (OCO). We introduce a new regret analysis framework that enables optimal bounds with constant β_1 parameter. In the second part, we move on to the nonconvex setting to study constrained stochastic optimization with weakly convex objectives. We provide

	f	Constraints	β_1	minibatch size	Diagonal	Adaptive
[CLSH19]	L-smooth	×	const.	1	✓	✓
[CLX+19]	L-smooth	✓	0	$\sim \sqrt{t}$	✓	✓
[DD19]	ρ-weak. cvx.	✓	0	1	×	×
[MJ20]	ρ-weak. cvx.	✓	const.	1	×	×
This chapter	ρ-weak. cvx.	✓	const.	1	✓	✓

Table 2.1 – Comparison with adaptive methods for smooth nonconvex optimization and SGD-based methods for weakly convex optimization. Column "diagonal" refers to coordinate-wise step sizes and "adaptive" refers to step sizes depending on observed gradients á la AdaGrad.

a brief comparison with the existing nonconvex results in Table 2.1. A comprehensive literature review for both parts is given in Section 2.4.

2.2 Regret analysis for Online Convex Optimization

Problem Setup. In OCO, a loss function $f_t \colon \mathcal{K} \to \mathbb{R}^d$ is revealed, after a decision vector $x_t \in \mathcal{K}$ is picked by the algorithm. We then minimize the regret defined as

$$R(T) = \sum_{t=1}^{T} f_t(x_t) - \min_{x \in \mathcal{K}} \sum_{t=1}^{T} f_t(x). \tag{2.1}$$

Our assumptions below are standard for OCO [Haz16] and are the same as [RKK18].

Assumption 2.1.
▷ $\mathcal{K} \subset \mathbb{R}^d$ is a compact convex set.
▷ $f_t \colon \mathcal{K} \to \mathbb{R}$ is a convex lsc function, $g_t \in \partial f_t(x_t)$.
▷ $D = \max_{x,y \in \mathcal{K}} \|x - y\|_\infty$, $G = \max_t \|g_t\|_\infty$.

Notation. For vectors $a, b \in \mathbb{R}^d$ standard operations ab, a^2, a/b, $a^{1/2}$, $1/a$, $\max\{a, b\}$ are supposed to be coordinate-wise. For a given $a_t \in \mathbb{R}^d$, we denote the i-th coordinate as $a_t^{(i)}$. We use $\mathbf{1}$ for the vector of all ones. For lighter notation, throughout this chapter we use the following notation for weighted projection: $\mathcal{P}^v = \mathcal{P}^{D(v)}$, where $v \in \mathbb{R}^d$, $v^{(i)} > 0$, for all i.

2.2.1 Dissection of the standard analysis

For the discussion, we use AMSGRAD in Algorithm 2.1, proposed by [RKK18] as a fix to ADAM. Compared to ADAM, it has an extra step to enforce monotonicity of the second moment estimator $\hat{v}_t$.

We first describe the shortcoming of the previous approaches in [RKK18, WLC+20], dating back to [KB15]. Then we explain the mechanism that allows us to obtain regret bounds with

Algorithm 2.1 AMSGrad[RKK18]

```
**Input:** $x_1 \in \mathcal{K}$, $\alpha_t = \frac{\alpha}{\sqrt{t}}$, $\alpha > 0$, $\beta_1 < 1$, $\beta_2 < 1$,
    $m_0 = v_0 = 0$, $\hat{v}_0 = \epsilon \mathbf{1}$, $\epsilon \geq 0$
**for** $t = 1, 2 \dots$ **do**
    $g_t \in \partial f_t(x_t)$
    $m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$
    $v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$
    $\hat{v}_t = \max(\hat{v}_{t-1}, v_t)$
    $x_{t+1} = \mathcal{P}_{\mathcal{K}}^{\hat{v}_t^{1/2}}(x_t - \alpha_t \hat{v}_t^{-1/2} m_t)$
**end for**
```

constant β_1. In this subsection, for full generality, consider m_t being updated with β_{1t}, as in [RKK18, KB15]:

$$m_t = \beta_{1t} m_{t-1} + (1-\beta_{1t}) g_t. \tag{2.2}$$

The standard way to analyze Adam-type algorithms is to start by nonexpansiveness to write

$$\|x_{t+1} - x\|^2_{\hat{v}_t^{1/2}} \leq \|x_t - x\|^2_{\hat{v}_t^{1/2}} - 2\alpha_t \langle m_t, x_t - x\rangle + \alpha_t^2 \|m_t\|^2_{\hat{v}_t^{-1/2}}.$$

Then using (2.2), one can deduce

$$(1-\beta_{1t})\langle g_t, x_t - x\rangle \leq -\beta_{1t}\langle m_{t-1}, x_t - x\rangle + \frac{\alpha_t}{2}\|m_t\|^2_{\hat{v}_t^{-1/2}} + \frac{1}{2\alpha_t}\left(\|x_t - x\|^2_{\hat{v}_t^{1/2}} - \|x_{t+1} - x\|^2_{\hat{v}_t^{1/2}}\right).$$

Let us analyze the above inequality. Its left-hand side is exactly what we want to bound, since by convexity $R(T) \leq \sum_{t=1}^T \langle g_t, x_t - x\rangle$. The last two terms in the right-hand side are easy to analyze, all of them can be bounded in a standard way using just definitions of $\hat{v}_t$, m_t, and α_t.

What can we do with the term $-\beta_{1t}\langle m_{t-1}, x_t - x\rangle$? Analysis in [RKK18] uses Young's inequality

$$-\beta_{1t}\langle m_{t-1}, x_t - x\rangle \leq \frac{\beta_{1t}}{2\alpha_t}\|x_t - x\|^2_{\hat{v}_t^{1/2}} + \frac{\beta_{1t}\alpha_t}{2}\|m_{t-1}\|^2_{\hat{v}_t^{-1/2}}.$$

The term $\frac{\beta_{1t}}{2\alpha_t}\|x_t - x\|^2_{\hat{v}_t^{1/2}}$ is precisely what leads to the second term in the regret bound in [RKK18, Theorem 4]. Since $\alpha_t = \frac{\alpha}{\sqrt{t}}$, one must require $\beta_{1t} \to 0$.

Note that the update for x_{t+1} has a projection. This is important, since otherwise a solution must lie in the interior of $\mathcal{X}$, which is not the case in general for problems with a compact domain. However, let us assume for a moment that the update for x_{t+1} does not have any projection. In this simplified setting, applying the following trick will work.

Recall that $x_t = x_{t-1} - \alpha_{t-1}\hat{v}_{t-1}^{-1/2} m_{t-1}$, or equivalently $m_{t-1} = \frac{1}{\alpha_{t-1}}\hat{v}_{t-1}^{1/2}(x_{t-1} - x_t)$. Plugging it into the error term $\langle m_{t-1}, x_t - x\rangle$ yields

$$-\langle m_{t-1}, x_t - x\rangle = -\frac{1}{\alpha_{t-1}}\langle \hat{v}_{t-1}^{1/2}(x_{t-1} - x_t), x_t - x\rangle$$

$$
\begin{aligned}
&= \frac{1}{2\alpha_{t-1}}\left[\|x_t - x_{t-1}\|^2_{\hat{v}^{1/2}_{t-1}} + \|x_t - x\|^2_{\hat{v}^{1/2}_{t-1}} - \|x_{t-1} - x\|^2_{\hat{v}^{1/2}_{t-1}}\right] \\
&\le \frac{1}{2}\alpha_{t-1}\|m_{t-1}\|^2_{\hat{v}^{-1/2}_{t-1}} + \frac{1}{2}\|x_t - x\|^2_{\hat{v}^{1/2}_{t}/\alpha_t} - \frac{1}{2}\|x_{t-1} - x\|^2_{\hat{v}^{1/2}_{t-1}/\alpha_{t-1}},
\end{aligned}
$$

where the second equality follows from the Cosine Law and the first inequality is from $x_t - x_{t-1} = -\alpha_{t-1}\hat{v}^{-1/2}_{t-1}m_{t-1}$ and $\hat{v}^{1/2}_t/\alpha_t \ge \hat{v}^{1/2}_{t-1}/\alpha_{t-1}$. We now compare this bound with the previous one. The term $\alpha_{t-1}\|m_{t-1}\|^2_{\hat{v}^{-1/2}_{t-1}}$, as we mentioned, is good for summation. Other two terms are going to cancel after summation over t. Hence, it is easy to finish the analysis to conclude $\mathcal{O}(\sqrt{T})$ regret with a fixed $\beta_{1t} = \beta_1$.

Unfortunately, x_{t+1} update has a projection, since otherwise the bounded domain assumption is very restrictive. This prevents us from using the above trick. Its message, however, is that one can expect a good bound with a fixed β_{1t}.

For having a more general technique to handle β_1, we will take a different route in the very beginning — we will analyze the term $\langle g_t, x_t - x\rangle$ in a completely different way, without resorting to crude Young's inequality as in [RKK18]. Basically, this idea can be applied to any framework with a similar update for the moment m_t, as we will show for AdamNc and SAdam.

2.2.2 A key lemma

As we understood above, the presence of the projection complicates handling $\langle m_{t-1}, x_t - x\rangle$. A high level explanation for the cause of the issue is that the standard analysis does not leave much flexibility, since it uses nonexpansiveness in the very beginning.

Lemma 2.1. *Under the definition $m_t = \beta_1 m_{t-1} + (1-\beta_1)g_t$, it follows that*

$$
\langle g_t, x_t - x\rangle = \langle m_{t-1}, x_{t-1} - x\rangle - \frac{\beta_1}{1-\beta_1}\langle m_{t-1}, x_t - x_{t-1}\rangle + \frac{1}{1-\beta_1}\left(\langle m_t, x_t - x\rangle - \langle m_{t-1}, x_{t-1} - x\rangle\right).
$$

The main message of Lemma 2.1 is that the decomposition of m_t, in the second part of the analysis in Section 2.2.1 is now done before using nonexpansiveness, therefore there would be no need for using Young's inequality which is the main shortcoming of the previous analysis.

Upon inspection on the bound, we see that the last two terms will telescope. The second term can be shown to be of the order $\alpha_t\|m_t\|^2_{\hat{v}^{-1/2}_t}$, and as we mentioned before, summing this term will give $\mathcal{O}\left(\sqrt{T}\right)$. To see that the first term is benign, a high level explanation is to notice that m_{t-1} is the gradient estimate used in the update $x_t = x_{t-1} - \alpha_{t-1}\hat{v}^{-1/2}_{t-1}m_{t-1}$, therefore it can be analyzed in the classical way.

Proof of Lemma 2.1. By definition of m_t, $g_t = \frac{1}{1-\beta_1}m_t - \frac{\beta_1}{1-\beta_1}m_{t-1}$. Thus, we have

$$
\langle g_t, x_t - x\rangle = \frac{1}{1-\beta_1}\langle m_t, x_t - x\rangle - \frac{\beta_1}{1-\beta_1}\langle m_{t-1}, x_t - x\rangle
$$

$$
\begin{aligned}
&= \frac{1}{1-\beta_1}\langle m_t, x_t - x\rangle - \frac{\beta_1}{1-\beta_1}\langle m_{t-1}, x_{t-1} - x\rangle - \frac{\beta_1}{1-\beta_1}\langle m_{t-1}, x_t - x_{t-1}\rangle \\
&= \frac{1}{1-\beta_1}\big(\langle m_t, x_t - x\rangle - \langle m_{t-1}, x_{t-1} - x\rangle\big) + \langle m_{t-1}, x_{t-1} - x\rangle - \frac{\beta_1}{1-\beta_1}\langle m_{t-1}, x_t - x_{t-1}\rangle.
\end{aligned}
$$

■

This simple proof of the decomposition in Lemma 2.1 enables the new analysis without the previous restrictions.

2.2.3 AMSGrad

The regret bound for AMSGrad in [RKK18, Theorem 4, Corollary 1] requires a decreasing β_1 at least at the order of $1/t$ to obtain $\mathcal{O}(\sqrt{T})$ worst case regret. Moreover, a constant β_1 results in $\mathcal{O}(T\sqrt{T})$ regret in [RKK18, Theorem 4]. We now continue with the theorem showing that the same $\mathcal{O}(\sqrt{T})$ can be obtained by AMSGrad under the same assumptions as [RKK18].

Theorem 2.2. *Under Assumption 2.1, $\beta_1 < 1$, $\beta_2 < 1$, $\gamma = \frac{\beta_1^2}{\beta_2} < 1$, and $\epsilon > 0$,* AMSGrad *has the regret*

$$
R(T) \le \frac{D^2\sqrt{T}}{2\alpha(1-\beta_1)} \sum_{i=1}^{d} \sqrt{\hat{v}_T^{(i)}} + \frac{\alpha\sqrt{1+\log T}}{\sqrt{(1-\beta_2)(1-\gamma)}} \sum_{i=1}^{d} \sqrt{\sum_{t=1}^{T} (g_t^{(i)})^2}. \tag{2.3}
$$

Our bound for $R(T)$ is also better than the one in [RKK18] in term of constants. We have two terms in contrast to three in [RKK18] and each of them is strictly smaller than their counterparts in [RKK18]. The reason is that we used i) new way of decomposition $\langle g_t, x_t - x\rangle$ as in Lemma 2.1, ii) wider admissible range for β_1, β_2, iii) more refined estimates for analyzing terms. For example, the standard analysis to estimate $\|m_t\|^2_{\hat{v}_t^{-1/2}}$ uses several Cauchy-Schwarz inequalities. We instead give a better bound by applying generalized Hölder inequality [BB61].

Another observation is that having a constant β_1 explicitly improves the last term in the regret bound. With a non-decreasing β_1, instead of constant β_1, this term would have an additional multiple of $\frac{1}{(1-\beta_1)^2}$. Since in general one chooses β_1 close to 1, this factor is significant.

Remark 2.3. Notice that Theorem 2.2 requires $\epsilon > 0$ in order to have the weighted projection well-defined. Such a requirement is common in the literature for theoretical analysis, see [DHS11, Theorem 5]. In practice, however, one can set $\epsilon = 0$.

Proof sketch of Theorem 2.2. We sum $\langle g_t, x_t - x\rangle$ from Lemma 2.1 over t, use $m_0 = 0$ to get

$$
\sum_{t=1}^{T} \langle g_t, x_t - x\rangle \le \underbrace{\sum_{t=1}^{T} \langle m_t, x_t - x\rangle}_{S_1} + \frac{\beta_1}{1-\beta_1} \underbrace{\sum_{t=1}^{T} \langle m_{t-1}, x_{t-1} - x_t\rangle}_{S_2} + \frac{\beta_1}{1-\beta_1} \underbrace{\langle m_T, x_T - x\rangle}_{S_3}.
$$

By Hölder inequality, we can show that

$$S_2 \leq \sum_{t=1}^{T-1} \alpha_t \|m_t\|^2_{\hat{v}_t^{-1/2}}.$$

By using the fact that $\hat{v}_t^{(i)} \geq \hat{v}_{t-1}^{(i)}$, and the same estimation as deriving S_2,

$$S_1 \leq \frac{D^2}{2\alpha_T} \sum_{i=1}^{d} \sqrt{\hat{v}_T^{(i)}} + \sum_{t=1}^{T} \frac{\alpha_t}{2} \|m_t\|^2_{\hat{v}_t^{-1/2}}.$$

By Hölder and Young's inequalities, we can bound S_3 as

$$S_3 \leq \alpha_T \|m_T\|^2_{\hat{v}_T^{-1/2}} + \frac{D^2}{4\alpha_T} \sum_{i=1}^{d} \sqrt{\hat{v}_T^{(i)}}.$$

We see that $\alpha_t \|m_t\|^2_{\hat{v}_t^{-1/2}}$ is common in all terms and it is well known that this term is good for summation

$$\sum_{t=1}^{T} \alpha_t \|m_t\|^2_{\hat{v}_t^{-1/2}} \leq \frac{(1-\beta_1)\alpha\sqrt{1+\log T}}{\sqrt{(1-\beta_2)(1-\gamma)}} \sum_{i=1}^{d} \sqrt{\sum_{t=1}^{T} (g_t^{(i)})^2}.$$

Combining the terms gives the final bound. ■

Finally, if we are interested in the worst case scenario, it is clear that Theorem 2.2 gives regret $R(T) = \mathcal{O}(\sqrt{\log(T)T})$. A quick look into the calculations yields that if one uses the worst case bound $g_t^{(i)} \leq G$, then the bound will not include a logarithmic term. However, then the data-dependence of the bound will be lost. It is not clear if one can obtain a data-dependent $\mathcal{O}(\sqrt{T})$ regret bound. In the following corollary, we give a partial answer to this question.

Corollary 2.4. *Under Assumption 2.1, $\beta_1 < 1$, $\beta_2 < 1$, $\gamma = \frac{\beta_1^2}{\beta_2} < 1$, and $\epsilon > 0$,* AMSGRAD *achieves*

$$R(T) \leq \frac{D^2\sqrt{T}}{2\alpha(1-\beta_1)} \sum_{i=1}^{d} \sqrt{\hat{v}_T^{(i)}} + \frac{\alpha\sqrt{G}}{\sqrt{1-\beta_2}(1-\gamma)} \sum_{i=1}^{d} \sqrt{\sum_{t=1}^{T} |g_t^{(i)}|}. \tag{2.4}$$

This bound has no $\log(T)$ term, thus it is better in the worst-case. However its data-dependence is worse than the bound in Theorem 2.2. Bound in Theorem 2.2 contains $(g_t^{(i)})^2$ whereas bound above contains $|g_t^{(i)}|$. With small $g_t^{(i)}$, the bound with $\log T$ can be better. We leave it as an open question to have a $\sqrt{T}$ bound with the same data-dependence as Theorem 2.2.

2.2.4 ADAMNC

Another variant that is proposed by [RKK18] as a fix to ADAM is ADAMNC which features an increasing schedule for β_{2t}. In particular, one sets $\beta_{2t} = 1 - \frac{1}{t}$ in

$$v_t = \beta_{2t} v_{t-1} + (1-\beta_{2t}) g_t^2,$$

that results in the expression $v_t = \frac{1}{t}\sum_{j=1}^t g_j^2$, which is a reminiscent of ADAGRAD [DHS11]. In fact, to ensure that $\mathcal{P}_{\mathcal{X}}^{v_t^{1/2}}$ is well-defined, one needs to consider the more general update $v_t = \frac{1}{t}\left(\sum_{j=1}^t g_j^2 + \epsilon\mathbf{1}\right)$ similar to the previous case with AMSGRAD.

Algorithm 2.2 ADAMNC[RKK18]

```
Input: $x_1 \in \mathcal{K}$, $\alpha_t = \frac{\alpha}{\sqrt{t}}$, $\alpha > 0$, $\beta_1 < 1$, $\epsilon \ge 0$, $m_0 = 0$.
for $t = 1, 2 \dots$ do
    $g_t \in \partial f_t(x_t)$
    $m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$
    $v_t = \frac{1}{t}\left(\sum_{j=1}^t g_j^2 + \epsilon\mathbf{1}\right)$
    $x_{t+1} = \mathcal{P}_{\mathcal{K}}^{v_t^{1/2}}(x_t - \alpha_t {v_t}^{-1/2} m_t)$
end for
```

ADAMNC is analyzed in [RKK18, Theorem 5, Corollary 2] and similar to AMSGRAD it has been shown to exhibit $\mathcal{O}(\sqrt{T})$ worst case regret only when β_1 decreases to 0. We show in the following theorem that the same regret can be obtained with a constant β_1.

Theorem 2.5. *Under Assumption 2.1, $\beta_1 < 1$, and $\epsilon > 0$, ADAMNC has the regret*

$$R(T) \le \frac{D^2\sqrt{T}}{2\alpha(1-\beta_1)}\sum_{i=1}^d \sqrt{v_T^{(i)}} + \frac{2\alpha}{1-\beta_1}\sum_{i=1}^d \sqrt{\sum_{t=1}^T (g_t^{(i)})^2}.$$

We skip the proof sketch of this theorem as it will have the same steps as AMSGRAD, just different estimation for $\alpha_t \|m_t\|^2_{v_t^{-1/2}}$, due to different v_t. The full proof is given in the appendix. As before, compared to the bound from [RKK18, Corollary 2], constant β_1 not only removes the middle term of [RKK18, Corollary 2] but improves the last term of the bound by $(1-\beta_1)^2$.

2.2.5 SADAM

We know that ADAGRAD obtains logarithmic regret [DHS10], when the loss functions are μ-strongly convex. A variant of ADAMNC for this setting is proposed in [WLC+20] and is shown to obtain logarithmic regret when β_1 decreases linearly to 0 [WLC+20, Theorem 1].

Algorithm 2.3 SADAM [WLC+20]

```
Input: $x_1 \in \mathcal{K}$, $\alpha_t = \frac{\alpha}{t}$, $\alpha > 0$, $\beta_1 < 1$, $m_0 = 0$, $\epsilon \ge 0$, $\beta_{2t} = 1 - 1/t$.
for $t = 1, 2 \dots$ do
    $g_t \in \partial f_t(x_t)$
    $m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$
    $v_t = \beta_{2t} v_{t-1} + (1-\beta_{2t}) g_t^2$
    $\hat{v}_t = v_t + \frac{\epsilon\mathbf{1}}{t}$
    $x_{t+1} = \mathcal{P}_{\mathcal{K}}^{\hat{v}_t}(x_t - \alpha_t \hat{v}_t^{-1} m_t)$
end for
```

Similar to AMSGRAD and ADAMNC, our new technique applies to SADAM to show logarithmic regret with a constant β_1 under the same assumptions as [WLC+20].

Theorem 2.6. *Let Assumption 2.1 hold and f_t be μ-strongly convex, $\forall t$. Then, if $\beta_1 < 1, \epsilon > 0$, and $\alpha \geq \frac{G^2}{\mu}$,* SADAM *has the regret*

$$R(T) \leq \frac{\beta_1 dGD}{1-\beta_1} + \frac{\alpha}{1-\beta_1}\sum_{i=1}^{d}\log\left(\frac{\sum_{t=1}^{T}(g_t^{(i)})^2}{\epsilon}+1\right).$$

Consistent with the standard literature of OGD [HAK07], to obtain the logarithmic regret, first step size α has a lower bound that depends on μ. Compared with the requirement of [WLC+20] for $\alpha \geq \frac{G^2}{\mu(1-\beta_1)}$, our requirement is strictly milder as $1-\beta_1 \leq 1$ and in practice since β_1 is near 1, it is much milder. Our bound is also strictly better than [WLC+20]. Moreover, we remove a factor of $\frac{1}{(1-\beta_1)^2}$ from the last term of the bound, compared to [WLC+20, Theorem 1].

2.3 Weakly convex optimization

Problem setup. We will prove the convergence of AMSGRAD for solving the problem

$$\min_{x\in\mathcal{K}}\left\{f(x) = \mathbb{E}_{\xi}\left[f(x;\xi)\right]\right\}, \tag{2.5}$$

where $f\colon \mathbb{R}^d \to \mathbb{R}$ is ρ-weakly convex, $\mathcal{K}$ is closed convex, ξ is a r.v. following a fixed unknown distribution. Problem (2.5) generalizes the previous analyses when f is L-smooth, as this implies L-weak convexity, and $\mathcal{K} = \mathbb{R}^d$. However, there are many applications when $\mathcal{K} \neq \mathbb{R}^d$ [VPG19, MNSF17, MMS+18] or when f is not L- smooth [DD19, Section 2.1],[DR18, DP19].

Constrained stochastic minimization with nonconvexity presents challenges not met in the convex setting [GLZ16, CLX+19]. In particular, until the recent work [DD19], even for SGD, increasing mini-batch sizes were required for convergence in constrained nonconvex optimization. To analyze AMSGRAD for solving (2.5), we build on the analysis framework of [DD19].

We show that AMSGrad achieves $\mathcal{O}(\log(T)/\sqrt{T})$ rate for near-stationarity (see (2.8)) for solving (2.5). Key specifications for this result are the following:

◦ We can use a mini-batch size of 1.

◦ We use constant parameters β_1, β_2 used in practice [KB15, RKK18, CLSH19, AMMC20].

◦ We do not assume boundedness of the domain $\mathcal{X}$.

We particularize our results for constrained optimization with L-smooth objectives and for a variant of RMSprop. Comparison of our results with state-of-the-art is given in Table 2.1.

Finally, in a numerical experiment for robust phase retrieval, we show that AMSGRAD is empirically more robust to variation of initial step sizes, than SGD and SGD with momentum.

Examples of weakly convex problems. The class of problems we consider in this chapter include constrained problems with L-smooth objectives which are, for example, studied in [CLX$^+$19] in the context of adversarial attacks. Other important examples with weak convexity are composite objectives $h(c(x))$, where h is a convex Lipschitz continuous function and c is a smooth map with Lipschitz continuous Jacobian. Concrete examples of weakly convex problems are listed in [DD19, Section 2.1], which include robust phase retrieval, sparse dictionary learning, Conditional Value-at-Risk, to name a few.

Notation. Due to nonconvexity, we cannot use standard definition of subgradients to form a global under-estimator. *Regular subdifferential*, denoted as ∂f, for nonconvex functions [RW09, Ch. 8] is defined as the set of vectors $q \in \mathbb{R}^d$ such that, $\forall x, y \in \mathbb{R}^d$, $q \in \partial f(x)$ if

$$f(y) \geq f(x) + \langle y - x, q\rangle + o(\|y-x\|), \quad \text{as } y \to x. \tag{2.6}$$

When f is convex, this reduces to standard definition of a subdifferential and when f is differentiable, this set coincides with $\{\nabla f(x)\}$.

Given random iterates $x_1, \dots, x_t$, we denote the filtration generated by these realizations as $\mathcal{F}_t = \sigma(x_1, \dots, x_t)$, and the corresponding conditional expectation as $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_t]$.

Assumption 2.2.

- $f\colon \mathbb{R}^d \to \mathbb{R}$ is ρ-weakly convex with respect to norm $\|\cdot\|$.
- The set $\mathcal{K} \subseteq \mathbb{R}^d$ is convex and closed.
- There exists g_t such that $\mathbb{E}[g_t] \in \partial f(x_t, \xi_t)$ and $\|g_t\|_\infty \leq G, \forall t$.
- f is lower bounded: $f^\star \leq f(x), \forall x \in \mathcal{K}$.

Remark 2.7. We note that when f is ρ-weakly convex w.r.t. $\|\cdot\|$, then it is $\frac{\rho}{\sqrt{\epsilon}}$-weakly convex w.r.t. $\|\cdot\|_{\hat{v}_t^{1/2}}, \forall t$, since $\hat{v}_t^{(i)} \geq \epsilon > 0$ (see Algorithm 2.1). We denote $\hat{\rho} = \frac{\rho}{\sqrt{\epsilon}}$.

It is easy to verify this remark by noticing that $x \mapsto f(x) + \frac{\rho}{2}\|x\|^2$ is convex and $\frac{\hat{\rho}}{2}\|x\|^2_{\hat{v}_t^{1/2}} \geq \frac{\rho}{2}\|x\|^2$.

A few remarks are in order for Assumption 2.2. First, we do not require boundedness of the domain $\mathcal{K}$. Second, weak convexity assumption is weaker than smoothness assumption on f and the assumption of bounded gradients is standard [CZT$^+$20, CLSH19, DBBU20]. In principle, it is possible to relax the bounded gradient assumption to the weaker requirement $\mathbb{E}\|g_t\|^2 \leq G$ as in [ZSJ$^+$19, Remark 6. (ii)] with a slightly worse and complicated convergence rate. For simplicity, we use Assumption 2.2.

Algorithm. We analyze AMSGRAD (see Algorithm 2.1) proposed in [RKK18]. As standard in stochastic nonconvex optimization, we output a randomly selected iterate [DD19, GL13, GLZ16]. We next define the composite objective

$$\varphi(x) = f(x) + \delta_{\mathcal{K}}(x).$$

For nonsmooth problems, the standard stationarity measures such as the norm of (sub)gradients are no longer applicable, see [DD19, MJ20] and [DP19, Section 4]. This motivates the following definitions that, as we show below, relate to a relaxed form of stationarity. Based on φ and a parameter $\bar\rho > 0$, we define the proximal point of x_t and the Moreau envelope

$$\hat{x}_t = \operatorname{prox}_{\varphi/\bar\rho}^{\hat{v}_t^{1/2}}(x_t) = \operatorname*{argmin}_y \left\{ \varphi(y) + \frac{\bar\rho}{2} \| y - x_t \|^2_{\hat{v}_t^{1/2}} \right\} \tag{2.7}$$

$$\varphi^t_{1/\bar\rho}(x_t) = \min_y \left\{ \varphi(y) + \frac{\bar\rho}{2} \| y - x_t \|^2_{\hat{v}_t^{1/2}} \right\}.$$

We compare the definitions with that of [DD19]. Due to the use of variable metric $\hat{v}_t$ in adaptive methods, we have a time dependent Moreau envelope, where the corresponding vector $\hat{v}_t$ is used for defining the norm. Important considerations for these quantities are the uniqueness of $\hat{x}_t$ and the smoothness of $\varphi^t_{1/\bar\rho}$. As we see now, choice of $\bar\rho$ is critical for ensuring these. In light of Remark 2.7, selecting $\bar\rho > \hat\rho = \frac{\rho}{\sqrt{\epsilon}}$, and by using similar arguments as [DD19, Lemma 2.2], it follows $\hat{x}_t$ is unique and $\varphi^t_{1/\bar\rho}$ is smooth with the gradient

$$\nabla \varphi^t_{1/\bar\rho}(x_t) = \bar\rho \hat{v}_t^{1/2}(x_t - \hat{x}_t).$$

Near stationarity. Near-stationarity conditions follow from the optimality condition of $\hat{x}_t$: $0 \in \partial\varphi(\hat{x}_t) + \bar\rho \hat{v}_t^{1/2}(\hat{x}_t - x_t)$, where we have used $\hat{v}_{t,i} \le G^2$:

$$\begin{cases} \|x_t - \hat{x}_t\|^2_{\hat{v}_t^{1/2}} = \frac{1}{\bar\rho^2} \|\nabla \varphi^t_{1/\bar\rho}(x_t)\|^2_{\hat{v}_t^{-1/2}} \\ \operatorname{dist}^2(0, \partial\varphi(\hat{x}_t)) \le G \|\nabla \varphi^t_{1/\bar\rho}(x_t)\|^2_{\hat{v}_t^{-1/2}} \\ \varphi(\hat{x}_t) \le \varphi(x_t). \end{cases} \tag{2.8}$$

Consistent with previous literature for weakly convex optimization [DD19, MJ20], we state the guarantees in terms of the norm of the gradient of Moreau envelope. Given (2.8), this means that x_t is near stationary: it is close to its proximal point $\hat{x}_t$ and $\hat{x}_t$ is approximately stationary.

2.3.1 Convergence

We start with our main theorem that shows that the norm of the gradient of Moreau envelope converges to 0 at the claimed rate, resulting in near-stationarity of x_{t^*}, as in (2.8).

Theorem 2.8. *Let Assumption 2.2 hold. Let $\beta_1 < 1$, $\beta_2 < 1$, $\gamma = \frac{\beta_1^2}{\beta_2} < 1$, $\bar\rho = 2\hat\rho$, $\epsilon > 0$ and t^* selected randomly from $[T]$. For, x_{t^*} as output of Algorithm 2.1, it follows that*

$$\mathbb{E}\|\nabla \varphi^{t^*}_{1/\bar\rho}(x_{t^*})\|^2_{\hat{v}_{t^*}^{-1/2}} \le \frac{2dG}{\alpha\sqrt{\epsilon T}(1-\beta_1)} \left[C_1 + (1 + \log T) C_2 + C_3 \right],$$

where $C_1 = 8\rho\alpha G + \frac{1}{dG}\left(\varphi^1_{1/\bar\rho}(x_1) - f^\star\right)$, $C_2 = \frac{2\rho}{\sqrt{(1-\beta_2)(1-\gamma)}}\left(4 + \frac{6G}{\sqrt{\epsilon}}\right)$, $C_3 = \frac{8G}{\rho}\sum_{i=1}^d \mathbb{E}(\hat{v}^{(i)}_{T+1})^{1/2}$.

The bound in Theorem 2.8 has complicated constants as it is usual for adaptive algorithms in nonconvex case [CLSH19, CZT$^+$20]. These constants are slightly simplified and the proof of Theorem 2.8 in Section 2.5.4 includes the non-simplified version. Next, we explain and interpret the bound in terms of dependence to key parameters.

Discussion on Theorem 2.8 In the context of near-stationarity (2.8), Theorem 2.8 states that to have x_{t^*} such that $\|\nabla\varphi^{t^*}_{1/\bar{\rho}}(x_{t^*})\|_{\hat{v}_t^{-1/2}} \le \varepsilon$, we require $\tilde{\mathcal{O}}(\varepsilon^4)$ iterations. This matches the known complexities for adaptive methods in unconstrained smooth stochastic optimization [AMMC20, DBBU20, WWB19, ZSJ$^+$18, CLSH19, CZT$^+$20, LO19, ZSJ$^+$19], and SGD-type methods in weakly convex optimization [MJ20, DD19].

Next remark is about the metric of the norm used for the gradient of Moreau envelope in Theorem 2.8. We then discuss the dependence of our bound w.r.t. important quantities.

Remark 2.9. By (2.8), one has $\|\nabla\varphi^{t^*}_{1/\bar{\rho}}(x_{t^*})\|^2_{\hat{v}_{t^*}^{-1/2}} = \bar{\rho}^2\|x_{t^*} - \hat{x}_{t^*}\|^2_{\hat{v}_{t^*}^{1/2}}$. Next, $\|x_{t^*} - \hat{x}_{t^*}\|^2_{\hat{v}_{t^*}^{1/2}} \ge \sqrt{\epsilon}\|x_{t^*} - \hat{x}_{t^*}\|^2$ as $\hat{v}_{t,i} \ge \epsilon$. It also holds that $\hat{v}_t^{(i)} \le G^2$. Therefore, one can convert our guarantees to $\|x_{t^*} - \hat{x}_{t^*}\|^2$ or $\|\nabla\varphi^{t^*}(x_{t^*})\|$ by multiplying the right hand side by appropriate quantities depending on δ or G. We leave the result with the metric, as δ and G are the worst case bounds.

Knowledge of ρ. To run the algorithm, one does not need to know the weak convexity parameter ρ. The parameters $\bar{\rho}$ and $\hat{\rho}$ are merely for analysis purposes [DDKL20, MJ20], and the convergence rate holds for any choice of step size α_t, independent of ρ.

Dependence w.r.t. β_1. Comparing with the previous work, the scaling of our bound in terms of β_1 is $(1-\beta_1)^{-1}$ matching the dependence for the unconstrained setting [AMMC20, DBBU20].

Dependence w.r.t. d. Standard dependence in the convergence rates of Adam-type algorithms for unconstrained case is $d/\sqrt{T}$ [AMMC20, DBBU20].[1] Even though in Theorem 2.8, C_3 has worst case dependence d^2, this is merely due to assumptions. The main reason is that we do not assume boundedness of the sequence x_t, instead we prove the necessary result for the analysis in Lemma 2.10. However, this result gives a bound for $\|x_t - \hat{x}_t\|$, which is naturally dimension dependent. We used this bound in (2.45), where we need to bound $\|x_t - \hat{x}_t\|_\infty$.

In particular, if we had assumed a bound for $\|x_t - \hat{x}_t\|_\infty$, then in (2.45) we could have used it instead of Lemma 2.10 to have standard $d/\sqrt{T}$ in C_3. Boundedness assumption also would remove a factor of $\frac{1}{\sqrt{\epsilon}}$ in the bound, as those terms appear in the steps where we avoid boundedness assumption. However, for generality, we do not assume boundedness.

Dependence w.r.t. ϵ. Our bound has a polynomial dependence of $1/\epsilon$ similar to [AMMC20, CZT$^+$20, CLSH19]. In [DBBU20], a more refined technique from [WWB19] is used to have a logarithmic dependence of $1/\epsilon$. This technique, used on the case of smooth unconstrained problems in these works, did not seem to apply to our setting.

In this section, we will flesh out the main ideas of our proof with three lemmas. The proof

[1] In [CZT$^+$20] better dependence is obtained by using step sizes in the order of $\frac{1}{\sqrt{d}}$, which we do not consider.

of Theorem 2.8 is then a careful combination of these results and given in Section 2.5.4.

Useful lemmas. We start with a result showing that under Assumption 2.2, the quantity $\|x_t - \hat{x}_t\|$ from (2.8) stays bounded. Third term on RHS in (2.44) arises as a spurious term due to time-dependent diagonal step sizes, which was not the case in previous works on weakly convex optimization with scalar step sizes [MJ20]. Next lemma is the main tool for us to avoid assuming boundedness of $\mathcal{X}$. The proof of this lemma given in Section 2.5.4 combines the definition of $\hat{x}_t$ with weak convexity to reach the result.

Lemma 2.10. *Let Assumption 2.2 hold. Let $\bar{\rho} > \hat{\rho}$, and $\hat{v}_t \geq \delta > 0$. It follows that*

$$\|x_t - \hat{x}_t\|^2 \leq \hat{D}^2 := \frac{4dG^2}{\epsilon(\bar{\rho} - \hat{\rho})^2}.$$

A key aspect in the analysis of adaptive algorithms is the dependence of $\hat{v}_t$ and g_t that couples $\hat{x}_t$ and g_t (see (2.7)), preventing taking expectation of $\langle x_t - \hat{x}_t, g_t \rangle$ that we use for obtaining the stationarity measure in the proof. Since this was not the case in prior works [DD19, MJ20], we need a more refined analysis.

Lemma 2.11. *Let Assumption 2.2 hold. Let $q_t = \mathbb{E}_t[g_t] \in \partial f(x_t)$, then it follows that*

$$\begin{aligned}\alpha_t \mathbb{E}_t \langle x_t - \hat{x}_t, g_t \rangle \geq \alpha_t(\bar{\rho} - \hat{\rho}) \mathbb{E}_t \|x_t - \hat{x}_t\|^2_{\hat{v}_t^{1/2}} - (\alpha_{t-1} - \alpha_t)\sqrt{d}\hat{D}G - \frac{\bar{\rho} - \hat{\rho}}{4\bar{\rho}} \mathbb{E}_t \|\hat{x}_t - \hat{x}_{t-1}\|^2_{\hat{v}_{t-1}^{1/2}} \\ - \frac{\alpha_{t-1}}{2} \mathbb{E}_t \|m_{t-1}\|^2_{\hat{v}_{t-1}^{-1/2}} - \left(\frac{1}{2} + \frac{\bar{\rho}}{\bar{\rho} - \hat{\rho}}\right) \frac{\alpha_{t-1}^2}{\sqrt{\epsilon}} \mathbb{E}_t \|g_t\|^2.\end{aligned}$$

Interpreting Lemma 2.11. We review the terms in this bound to gain some intuition. The first term in the RHS is the stationarity measure (see (2.8)), second term will sum to a constant, fourth and fifth terms will sum to $\log(T)$ by Lemma 2.12. Handling the third term in RHS is not as obvious, but we can show that we can cancel it using the contribution from another part of the analysis that we detail in the full proof (see (2.44)).

One critical issue for Adam-type algorithms is to obtain results with constant β_1 parameter. As we show in Section 2.2, Lemma 2.1 is critical for this. Without Lemma 2.1, we would require decreasing β_1, especially for constrained problems, which we would like to avoid.

Next lemma is a standard estimation used for the analysis of Adam-based methods, since [KB15]. We used such estimations for Section 2.2 (see Lemma 2.15). As mentioned before, we get a tighter bound than previous works, due to using a constant β_1. In words, we bound the sum of the norms of first moment vectors multiplied by the step size.

Lemma 2.12. *Let $\beta_1 < 1$, $\beta_2 < 1$, $\gamma = \frac{\beta_1^2}{\beta_2} < 1$, then it holds that*

$$\sum_{t=1}^{T} \alpha_t^2 \|m_t\|^2_{\hat{v}_t^{-1/2}} \leq \frac{(1-\beta_1)\alpha^2}{\sqrt{(1-\beta_2)(1-\gamma)}} dG(1 + \log T).$$

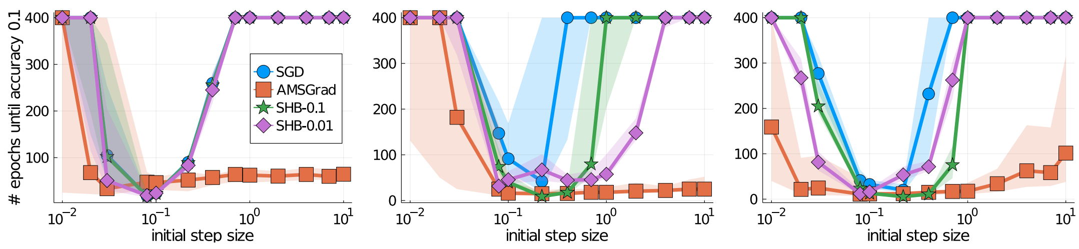

Figure 2.1 – left to right: $\kappa = \{1, 10, 100\}$. Number of epochs to reach $f(x) - f^\star \le 0.1$ vs. initial step size[2]. Each experiment is run 50 times, lines show the median and shaded areas cover between 20th and 80th percentiles. We denote SGD with momentum by SHB.

2.3.2 A numerical experiment

This section illustrates the potential advantages of adaptive algorithms, in particular AMSGrad, for solving a prototypical setting of a weakly convex problem, compared to SGD and SGD with momentum [MJ20]. As popular in the literature of weakly convex stochastic optimization [MJ20, DG19, DDKL20], we compare the algorithms in terms of their robustness to initial step sizes. *"Robustness to tuning"* of algorithms is also investigated in the context of deep learning in the literature and the advantage of adaptive algorithms such as Adam/AMSGrad is observed [SMV⁺20, CSN⁺19].

We solve the robust phase retrieval problem [EM14, DDP20, DR19],

$$\min_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} |\langle a_i, x \rangle^2 - b_i|,$$

where $A = [a_1, \ldots, a_n]^\top \in \mathbb{R}^{n \times d}$, $n = 300$, $d = 50$. Weak convexity of this problem is well-known [DD19, DDP20]. We recall the setup from [MJ20] that considered SGD with momentum for solving this problem. The data is generated as $A = QD$, with a standard normal distributed $Q \in \mathbb{R}^{n \times d}$ and $D = \text{linspace}(1/\kappa, 1, d)$, where $\kappa \ge 1$ controls the conditioning. We generate $x^\star$ as a standard normal random vector with unit norm. Then, $b = Ax^\star + \delta\eta$ where elements of $\eta \in \mathbb{R}^n$ have distribution $\mathcal{N}(0, 25)$ and $\delta = \text{diag}(\delta_1, \ldots, \delta_d)$ is such that $\frac{|\{i \in [n] \colon \delta_i = 1\}|}{n} = 0.2$, meaning that 20% of the observations are corrupted.

With this setup, it is proven in [DDP20, Lemma B.12] that only solutions of the problem are $\{x^\star, -x^\star\}$. Therefore, for the algorithms, we will use $f(x_k) - f(x^\star) \le \varepsilon$ as the stopping criterion.

We run stochastic subgradient method (SGD) [DD19], momentum SGD (SHB) [MJ20] and AMSGrad that we analyzed. For all algorithms, the step size is chosen as $\alpha_k = \frac{\alpha_0}{\sqrt{k}}$. We varied the initial step size[2] between 0.01 and 10 for all algorithms, and we plotted the number of epochs to reach $f(x) - f(x^\star) \le 0.1$. In terms of other parameters, we use both $\beta = 0.1$ and

[2] We make the "effective initial step sizes" of algorithms equal. In particular we pick $\alpha_0^{\text{SGD}} = \alpha_0^{\text{MSGD}} = \frac{\alpha_0^{\text{AMS}}}{\beta_2 \sqrt{\max_i (g_{1,i}^2)}}$, since the initial step size of AMSGrad is $\frac{\alpha_0}{\sqrt{\hat{v}_1^2}}$.

$\beta = 0.01$ for SHB, as recommended in [MJ20] and $\beta_1 = \beta_2 = 0.99$ as popular, for AMSGrad.

We present the results in Figure 2.1 for varying values of $\kappa = \{1, 10, 100\}$, where each setup is run for 50 times, medians are drawn as lines and the region between 20th and 80th percentiles is shaded. [MJ20] observed that SHB improves the robustness of SGD to initial step sizes. We observe in Figure 2.1 that AMSGrad shows a more robust behavior compared to both algorithms. Our observations support the potential of AMSGrad and adaptive methods for weakly convex optimization. Moreover, our findings about robustness of adaptive algorithms to tuning is consistent with the findings from deep learning literature [SMV⁺20, CSN⁺19].

2.4 Related work

Convex world. AMSGRAD and ADAMNC were proposed by [RKK18] to fix the nonconvergence issue for ADAM [KB15]. However, as the proof template of [RKK18] follows very closely the proof of [KB15], the requirement for $\beta_1 \to 0$ remains in all the regret guarantees of these algorithms. In particular, as noted by [RKK18, Corollary 1, 2], a schedule of $\beta_{1t} = \beta_1 \lambda^{t-1}$ is needed for obtaining optimal regret. [RKK18] also noted that regret bounds of the same order can be obtained by setting $\beta_{1t} = \beta_1 / t$. On the other hand, in the numerical experiments, a constant value $\beta_{1t} = \beta_1$ is used consistent with the huge literature following [KB15].

Following [RKK18], there has been a surge of interest in proposing new variants of ADAM with good practical properties; to name a few, PADAM by [CZT⁺20], ADABOUND and AMSBOUND by [LXL19, Sav19], NOSTALGIC ADAM by [HWD19]. As the regret analyses of these methods follow very closely the analysis of [RKK18], the resulting bounds inherited the same shortcomings explained in the previous paragraph. The experimental results reported on these algorithms use a constant value of β_1 in practice in order to obtain better performance.

Similar issues are present in other problem settings. For strongly convex setting, [WLC⁺20] proposed SADAM as a variant of ADAMNC, which exploits strong convexity to obtain $\mathcal{O}\big(\log T\big)$ regret. SADAMwas shown to exhibit favorable practical performance in the experimental results of [WLC⁺20]. However, the same discrepancy exists as previous ADAM variants: a linearly decreasing β_{1t} is required in theory but a constant $\beta_{1t} = \beta_1$ is used in practice.

One work that tried to address this issue is that of [FK19], where the authors focused on OCO with strongly convex loss functions and derived an $\mathcal{O}(\sqrt{T})$ regret bound with a constant value of $\beta_1 \le \frac{\mu\alpha}{1+\mu\alpha}$, where μ is the strong convexity constant and α is the step size that is set as $\alpha_1 / \sqrt{T}$. [FK19, Theorem 2]. However, this result is not satisfactory, since the obtained bound for β_1 is weak: both strong convexity μ and the step size $\frac{\alpha_1}{\sqrt{T}}$ are small. This does not allow for the standard choices of $\beta_1 \in (0.9, 0.99)$ and the regret is suboptimal with strong convexity.

Moreover, a quick look into the proof of [FK19, Theorem 2] reveals that the proof in fact follows the same lines as [RKK18] with the difference of using the contribution of strong convexity to get rid of the spurious terms that require $\beta_1 \to 0$. Therefore, it is not surprising that the theoretical bound for β_1 depends on μ and α and can only take values close to 0. Second, in addition to the standard Assumption 2.1, [FK19] also assumes strong convexity, which is a quite stringent assumption by itself. In contrast, our approach does not follow the lines of [RKK18], but is an alternative way that does not encounter the same roadblocks.

Nonconvex world. When convexity is removed, the standard setting in which the algorithms are analyzed is stochastic optimization with a smooth loss function and no constraints [CLSH19, ZTY$^+$18, ZSJ$^+$19]. As a result, these algorithms, compared to the convex counterparts, do not perform projections in the update step of x_{t+1} (*cf.*, Algorithm 2.1). The standard results bound the minimum gradient norm across all iterations.

An interesting phenomenon in this line of work is that a constant $\beta_1 < 1$ is permitted for the theoretical results, which may seem like weakening our claims. However, it is worth noting that these results do not imply a guarantee for regret in OCO. Adding the convexity assumption to these analyses for unconstrained, smooth stochastic optimization, does not give a guarantee in the objective value, unless more stringent Polyak-Lojasiewicz or strong convexity requirements are added in the mix.

Moreover, in the OCO setting that we analyze, loss functions are nonsmooth, and the algorithm performs projections to the constraint set (which as we see is the main difficulty for constant β_1 analysis). Finally, online optimization includes stochastic optimization as a special case. Given the difference of assumptions, the analyses in [CLSH19, ZTY$^+$18, ZSJ$^+$19] do not help obtaining a regret guarantee for standard OCO.

A good example demonstrating this difference on the set of assumptions is the work [CLX$^+$19], where a variant of AMSGRAD is proposed for zeroth order optimization and it is analyzed in the convex and nonconvex settings. Consistent with the previous literature in both, convergence result for the nonconvex setting allows a constant $\beta_1 < 1$ [CLX$^+$19, Theorem 1]. However, the result in the convex setting requires a decreasing schedule $\beta_{1t} = \frac{\beta_1}{t}$ [CLX$^+$19, Proposition 4]. Moreover, this result applies for the specific case of $\beta_1 = 0$ which corresponds to a variant of RMSprop [TH12, RKK18]. More importantly, since its analysis follows the one of [GLZ16], increasing mini-batch sizes of the order $\sqrt{t}$ are required [CLX$^+$19, Theorem 2].

As we highlighted above, the analyses in convex/nonconvex settings follow different paths and the results or techniques are not transferrable to each other. Thus, our main aim in this chapter is to bridge the gap in the understanding of regret analysis for OCO and propose a new analytic framework. As we see in the sequel, our analysis not only gives the first results in OCO setting, it is also general enough to apply to the nonconvex case to derive guarantees for constrained problems, generalizing the previous nonconvex analyses [LO19, WWB19, ZSJ$^+$18, CZT$^+$20, CLSH19, ZSJ$^+$19, DBBU20, BB20].

Weakly convex optimization is well studied with SGD based methods [DR18, DG19, DD19]. A recent work by [MJ20], considers momentum SGD for solving (2.5). However, this algorithm ***(i)*** does not use momentum with β_2 and ***(ii)*** uses non-adaptive, scalar, fixed step size: in the notation of Algorithm 2.1, $\hat{v}_t = 1$, $\alpha_t = \alpha/\sqrt{T}$. These make the algorithm less practical, while simpler for analysis.

2.5 Proofs

2.5.1 Proofs for Section 2.2.3

Lemma 2.13 (Generalized Hölder inequality, BB61, Chap. 1.18)**.** *For $x, y, z \in \mathbb{R}_+^n$ and positive p, q, r such that $\frac{1}{p} + \frac{1}{q} + \frac{1}{r} = 1$, we have: $\sum_{j=1}^n x_j y_j z_j \le \|x\|_p \|y\|_q \|z\|_r$.*

Above lemma is used to obtain a slightly tighter bound for $\|m_t\|^2_{\hat{v}_t^{-1/2}}$, than standard analysis.

Lemma 2.14 (Bound for $\|m_t\|^2_{\hat{v}_t^{-1/2}}$)**.** *Under Assumption 2.1, $\beta_1 < 1$, $\beta_2 < 1$, $\gamma = \frac{\beta_1^2}{\beta_2} < 1$, $\epsilon > 0$, and the definitions of α_t, m_t, v_t, $\hat{v}_t$ in* AMSGrad*, it holds that*

$$\|m_t\|^2_{\hat{v}_t^{-1/2}} \le \frac{(1-\beta_1)^2}{\sqrt{(1-\beta_2)(1-\gamma)}} \sum_{i=1}^d \sum_{j=1}^t \beta_1^{t-j} |g_j^{(i)}|. \tag{2.9}$$

Proof. From the definition of m_t and v_t, it follows that

$$m_t = (1-\beta_1) \sum_{j=1}^t \beta_1^{t-j} g_j, \qquad v_t = (1-\beta_2) \sum_{j=1}^t \beta_2^{t-j} g_j^2. \tag{2.10}$$

Then we have

$$\begin{aligned}
\|m_t\|^2_{\hat{v}_t^{-1/2}} &\le \|m_t\|^2_{v_t^{-1/2}} = \sum_{i=1}^d \frac{(m_t^{(i)})^2}{(v_t^{(i)})^{1/2}} = \sum_{i=1}^d \frac{\left(\sum_{j=1}^t (1-\beta_1)\beta_1^{t-j} g_j^{(i)}\right)^2}{\sqrt{\sum_{j=1}^t (1-\beta_2)\beta_2^{t-j} (g_j^{(i)})^2}} \\
&= \frac{(1-\beta_1)^2}{\sqrt{1-\beta_2}} \sum_{i=1}^d \frac{\left(\sum_{j=1}^t \beta_1^{t-j} g_{j,i}\right)^2}{\sqrt{\sum_{j=1}^t \beta_2^{t-j} g_{j,i}^2}} \\
&\le \frac{(1-\beta_1)^2}{\sqrt{1-\beta_2}} \sum_{i=1}^d \frac{\left[\left(\sum_{j=1}^t (\beta_2^{\frac{t-j}{4}} |g_j^{(i)}|^{\frac{1}{2}})^4\right)^{\frac{1}{4}} \left(\sum_{j=1}^t (\beta_1^{1/2}\beta_2^{-1/4})^{4(t-j)}\right)^{\frac{1}{4}} \left(\sum_{j=1}^t (\beta_1^{t-j} |g_j^{(i)}|)^{\frac{1}{2}\cdot 2}\right)^{\frac{1}{2}}\right]^2}{\sqrt{\sum_{j=1}^t \beta_2^{t-j} (g_j^{(i)})^2}} \\
&= \frac{(1-\beta_1)^2}{\sqrt{1-\beta_2}} \sum_{i=1}^d \left(\sum_{j=1}^t \gamma^{t-j}\right)^{\frac{1}{2}} \sum_{j=1}^t \beta_1^{t-j} |g_j^{(i)}| \le \frac{(1-\beta_1)^2}{\sqrt{(1-\beta_2)(1-\gamma)}} \sum_{i=1}^d \sum_{j=1}^t \beta_1^{t-j} |g_j^{(i)}|,
\end{aligned}$$

where the first inequality is by $(\hat{v}_t^{(i)})^{1/2} \ge (v_t^{(i)})^{1/2}$, the second one follows from Lemma 2.13 with

$$x_j = \beta_2^{\frac{t-j}{4}} |g_j^{(i)}|^{\frac{1}{2}}, \quad y_j = (\beta_1 \beta_2^{-1/2})^{\frac{t-j}{2}}, \quad z_j = (\beta_1^{t-j} |g_j^{(i)}|)^{\frac{1}{2}} \quad \text{and} \quad p = q = 4, \quad r = 2,$$

and the third one follows from the sum of geometric series and the assumption $\gamma = \frac{\beta_1^2}{\beta_2} < 1$.

We comment on the possibility of observing many zero gradients in the beginning, causing $v_t = 0$ until some t, which would cause the appearance of indeterminate form $\frac{0}{0}$ in the upper bound derived above, specifically in $\frac{(m_t^{(i)})^2}{(v_t^{(i)})^{1/2}}$. For this, we use the convention $\frac{0}{0} = 0$, in which case the above derivations are always well-defined. For this, we argue as follows: recall first that $v_t^{(i)} = 0$ iff $g_j^{(i)} = 0$ for all $j = 1, \dots, t$. This being the case, we also get $m_t^{(i)} = 0$, and hence, $\frac{(m_t^{(i)})^2}{(v_t^{(i)})^{1/2}} = 0$. In fact, this was done only for convenience, since $\hat{v}_t^{(i)} \geq \epsilon$ and we can always exclude zero terms from $\|m_t\|^2_{\hat{v}_t^{-1/2}}$, before using the first line in the above chain of inequalities. ■

Lemma 2.15 (Bound for $\sum_{t=1}^T \alpha_t \|m_t\|^2_{\hat{v}_t^{-1/2}}$). *Under Assumption 2.1, $\beta_1 < 1$, $\beta_2 < 1$, $\gamma = \frac{\beta_1^2}{\beta_2} < 1$, $\epsilon > 0$, and the definitions of α_t, m_t, v_t, $\hat{v}_t$ in* AMSGRAD, *we have*

$$\sum_{t=1}^T \alpha_t \|m_t\|^2_{\hat{v}_t^{-1/2}} \leq \frac{(1-\beta_1)\alpha\sqrt{1+\log T}}{\sqrt{(1-\beta_2)(1-\gamma)}} \sum_{i=1}^d \sqrt{\sum_{t=1}^T (g_t^{(i)})^2}. \tag{2.11}$$

Proof. We have

$$\begin{aligned}
\sum_{t=1}^T \alpha_t \|m_t\|^2_{\hat{v}_t^{-1/2}} &\leq \frac{(1-\beta_1)^2}{\sqrt{(1-\beta_2)(1-\gamma)}} \sum_{i=1}^d \sum_{t=1}^T \alpha_t \sum_{j=1}^t \beta_1^{t-j} |g_j^{(i)}| && \text{(Equation (2.9))} \\
&= \frac{(1-\beta_1)^2}{\sqrt{(1-\beta_2)(1-\gamma)}} \sum_{i=1}^d \sum_{j=1}^T \sum_{t=j}^T \alpha_t \beta_1^{t-j} |g_j^{(i)}| && \text{(Changing order of summation)} \\
&\leq \frac{(1-\beta_1)}{\sqrt{(1-\beta_2)(1-\gamma)}} \sum_{i=1}^d \sum_{j=1}^T \alpha_j |g_j^{(i)}| && \Big(\text{Using} \sum_{t=j}^T \alpha_t \beta_1^{t-j} \leq \frac{\alpha_j}{1-\beta_1}\Big) \\
&\leq \frac{1-\beta_1}{\sqrt{(1-\beta_2)(1-\gamma)}} \sum_{i=1}^d \sqrt{\sum_{j=1}^T \alpha_j^2} \sqrt{\sum_{j=1}^T (g_j^{(i)})^2} && \text{(Cauchy-Schwarz)} \\
&\leq \frac{(1-\beta_1)\alpha\sqrt{1+\log T}}{\sqrt{(1-\beta_2)(1-\gamma)}} \sum_{i=1}^d \sqrt{\sum_{t=1}^T (g_t^{(i)})^2} && \Big(\text{Using } \sum_{j=1}^T \frac{1}{j} \leq 1 + \log T\Big).
\end{aligned}$$

■

We now continue with the proof of Theorem 2.2.

Proof of Theorem 2.2. Let $x \in \operatorname{argmin}_{y \in \mathcal{K}} \sum_{t=1}^T f_t(y)$. Then by convexity, we immediately have

$$R(T) \leq \sum_{t=1}^T \langle g_t, x_t - x \rangle.$$

Hence, our goal is to bound the latter expression. If we sum the inequality from Lemma 2.1

over $t = 1, \ldots, T$ and use the fact that $m_0 = 0$, we obtain

$$\begin{aligned}\sum_{t=1}^{T}\langle g_t, x_t - x\rangle &= \frac{1}{1-\beta_1}\big(\langle m_T, x_T - x\rangle - \langle m_0, x_0 - x\rangle\big) + \langle m_0, x_0 - x\rangle + \sum_{t=1}^{T-1}\langle m_t, x_t - x\rangle \\ &\qquad + \frac{\beta_1}{1-\beta_1}\sum_{t=1}^{T}\langle m_{t-1}, x_{t-1} - x_t\rangle \\ &= \frac{\beta_1}{1-\beta_1}\langle m_T, x_T - x\rangle + \sum_{t=1}^{T}\langle m_t, x_t - x\rangle + \frac{\beta_1}{1-\beta_1}\sum_{t=1}^{T}\langle m_{t-1}, x_{t-1} - x_t\rangle. \qquad (2.12)\end{aligned}$$

We will separately bound each term in the right-hand side of (2.12) and then combine these bounds together.

- *Bound for* $\sum_{t=1}^{T}\langle m_t, x_t - x\rangle$: As $x \in \mathcal{K}$, by nonexpansiveness, we get

$$\begin{aligned}\|x_{t+1} - x\|^2_{\hat{v}_t^{1/2}} &= \|\mathcal{P}_{\mathcal{K}}^{\hat{v}_t^{1/2}}\left(x_t - \alpha_t \hat{v}_t^{-1/2} m_t\right) - x\|^2_{\hat{v}_t^{1/2}} \le \|x_t - \alpha_t \hat{v}_t^{-1/2} m_t - x\|^2_{\hat{v}_t^{1/2}} \\ &= \|x_t - x\|^2_{\hat{v}_t^{1/2}} - 2\alpha_t\langle m_t, x_t - x\rangle + \|\alpha_t \hat{v}_t^{-1/2} m_t\|^2_{\hat{v}_t^{1/2}} \\ &= \|x_t - x\|^2_{\hat{v}_t^{1/2}} - 2\alpha_t\langle m_t, x_t - x\rangle + \alpha_t^2\|m_t\|^2_{\hat{v}_t^{-1/2}}. \qquad (2.13)\end{aligned}$$

We rearrange and divide both sides of (2.13) by $2\alpha_t$ to get

$$\begin{aligned}\langle m_t, x_t - x\rangle &\le \frac{1}{2\alpha_t}\|x_t - x\|^2_{\hat{v}_t^{1/2}} - \frac{1}{2\alpha_t}\|x_{t+1} - x\|^2_{\hat{v}_t^{1/2}} + \frac{\alpha_t}{2}\|m_t\|^2_{\hat{v}_t^{-1/2}} \\ &= \frac{1}{2\alpha_{t-1}}\|x_t - x\|^2_{\hat{v}_{t-1}^{1/2}} - \frac{1}{2\alpha_t}\|x_{t+1} - x\|^2_{\hat{v}_t^{1/2}} + \frac{1}{2}\sum_{i=1}^{d}\left(\frac{(\hat{v}_t^{(i)})^{1/2}}{\alpha_t} - \frac{(\hat{v}_{t-1}^{(i)})^{1/2}}{\alpha_{t-1}}\right)(x_t^{(i)} - x^{(i)})^2 \\ &\qquad + \frac{\alpha_t}{2}\|m_t\|^2_{\hat{v}_t^{-1/2}} \\ &\le \frac{1}{2\alpha_{t-1}}\|x_t - x\|^2_{\hat{v}_{t-1}^{1/2}} - \frac{1}{2\alpha_t}\|x_{t+1} - x\|^2_{\hat{v}_t^{1/2}} + \frac{D^2}{2}\sum_{i=1}^{d}\left(\frac{(\hat{v}_t^{(i)})^{1/2}}{\alpha_t} - \frac{(\hat{v}_{t-1}^{(i)})^{1/2}}{\alpha_{t-1}}\right) + \frac{\alpha_t}{2}\|m_t\|^2_{\hat{v}_t^{-1/2}}, \\ &\qquad (2.14)\end{aligned}$$

where the last inequality is due to the fact that $\hat{v}_t^{(i)} \ge \hat{v}_{t-1}^{(i)}$, $\frac{1}{\alpha_t} \ge \frac{1}{\alpha_{t-1}}$, and the definition of D.[2]

Summing (2.14) over $t = 1, \ldots T$ and using that $\frac{1}{2\alpha_0}\|x_1 - x\|^2_{\hat{v}_0^{1/2}} = 0$ yields

$$\sum_{t=1}^{T}\langle m_t, x_t - x\rangle \le \frac{D^2}{2\alpha_T}\sum_{i=1}^{d}(\hat{v}_T^{(i)})^{1/2} + \frac{1}{2}\sum_{t=1}^{T}\alpha_t\|m_t\|^2_{\hat{v}_t^{-1/2}}. \qquad (2.15)$$

- *Bound for* $\sum_{t=1}^{T}\langle m_{t-1}, x_{t-1} - x_t\rangle$: Let us bound the last term in (2.12).

$$\sum_{t=1}^{T}\langle m_{t-1}, x_{t-1} - x_t\rangle = \sum_{t=2}^{T}\langle m_{t-1}, x_{t-1} - x_t\rangle = \sum_{t=1}^{T-1}\langle m_t, x_t - x_{t+1}\rangle \quad (\text{Using } m_0 = 0)$$

[2] Note that for $t = 1$ we suppose that $\frac{1}{\alpha_0} = 0$; this makes the above derivation still valid, as α_0 is not used in the algorithm, and this is only for convenience.

$$
\begin{aligned}
&\le \sum_{t=1}^{T-1} \|m_t\|_{\hat{v}_t^{-1/2}} \|x_{t+1} - x_t\|_{\hat{v}_t^{1/2}} && \text{(Hölder inequality)}\\
&= \sum_{t=1}^{T-1} \|m_t\|_{\hat{v}_t^{-1/2}} \left\| \mathcal{P}_{\mathcal{K}}^{\hat{v}_t^{1/2}} \left(x_t - \alpha_t \hat{v}_t^{-1/2} m_t\right) - \mathcal{P}_{\mathcal{K}}^{\hat{v}_t^{1/2}}(x_t)\right\|_{\hat{v}_t^{1/2}} && \text{(Using } x_t \in \mathcal{X}\text{)}\\
&\le \sum_{t=1}^{T-1} \alpha_t \|m_t\|_{\hat{v}_t^{-1/2}} \|\hat{v}_t^{-1/2} m_t\|_{\hat{v}_t^{1/2}} && \text{(Nonexpansiveness)}\\
&= \sum_{t=1}^{T-1} \alpha_t \|m_t\|^2_{\hat{v}_t^{-1/2}} && (\|u^{-1}x\|_u = \|x\|_{u^{-1}}). \quad (2.16)
\end{aligned}
$$

At this point, we could use eq. (2.11) to obtain a final bound for $\sum_{t=1}^{T} \langle m_{t-1}, x_{t-1} - x_t\rangle$. However, we postpone it to combine it with the term $\langle m_T, x_T - x\rangle$ in (2.12) to have a shorter expression.

- *Bound for* $\langle m_T, x_T - x\rangle$: This term is the easiest for estimation:

$$
\begin{aligned}
\langle m_T, x_T - x\rangle &\le \|m_T\|_{\hat{v}_T^{-1/2}} \|x_T - x\|_{\hat{v}_T^{1/2}} && \text{(Hölder's inequality)}\\
&\le \alpha_T \|m_T\|^2_{\hat{v}_T^{-1/2}} + \frac{1}{4\alpha_T} \|x_T - x\|^2_{\hat{v}_T^{1/2}} && \text{(Young's inequality)}\\
&\le \alpha_T \|m_T\|^2_{\hat{v}_T^{-1/2}} + \frac{D^2}{4\alpha_T} \sum_{i=1}^{d} (\hat{v}_T^{(i)})^{1/2} && \text{(Definition of } D\text{)} \quad (2.17)
\end{aligned}
$$

We now have all the ingredients required to bound the right-hand side of (2.12). To that end, after all substitutions and some straightforward algebra, we obtain

$$
\begin{aligned}
\text{RHS of (2.12)} &= \frac{\beta_1}{1-\beta_1}\left(\langle m_T, x_T - x\rangle + \sum_{t=1}^{T} \langle m_{t-1}, x_{t-1} - x_t\rangle \right) + \sum_{t=1}^{T} \langle m_t, x_t - x\rangle\\
&\le \frac{\beta_1}{1-\beta_1}\left(\frac{D^2}{4\alpha_T} \sum_{i=1}^{d} (\hat{v}_T^{(i)})^{1/2} + \sum_{t=1}^{T} \alpha_t \|m_t\|^2_{\hat{v}_t^{-1/2}} \right) + \frac{D^2}{2\alpha_T} \sum_{i=1}^{d} (\hat{v}_T^{(i)})^{1/2} + \frac{1}{2} \sum_{t=1}^{T} \alpha_t \|m_t\|^2_{\hat{v}_t^{-1/2}}\\
&= \frac{(2-\beta_1)D^2}{4\alpha_T(1-\beta_1)} \sum_{i=1}^{d} (\hat{v}_T^{(i)})^{1/2} + \frac{1+\beta_1}{2(1-\beta_1)} \sum_{t=1}^{T} \alpha_t \|m_t\|^2_{\hat{v}_t^{-1/2}}\\
&\le \frac{D^2\sqrt{T}}{2\alpha(1-\beta_1)} \sum_{i=1}^{d} (\hat{v}_T^{(i)})^{1/2} + \frac{1}{1-\beta_1} \sum_{t=1}^{T} \alpha_t \|m_t\|^2_{\hat{v}_t^{-1/2}}\\
&\le \frac{D^2\sqrt{T}}{2\alpha(1-\beta_1)} \sum_{i=1}^{d} (\hat{v}_T^{(i)})^{1/2} + \frac{\alpha\sqrt{1+\log T}}{\sqrt{(1-\beta_2)(1-\gamma)}} \sum_{i=1}^{d} \sqrt{\sum_{t=1}^{T} (g_t^{(i)})^2}, \quad (2.18)
\end{aligned}
$$

where the second inequality follows from the assumption $\frac{2-\beta_1}{4} \le \frac{1}{2}$, $\frac{1+\beta_1}{2} \le 1$, and $\alpha_T = \frac{\alpha}{\sqrt{T}}$, and the last follows by Lemma 2.15. ■

2.5.2 Proofs for Section 2.2.4

We give analogous results to Lemmas 2.14 and 2.15, which are mostly standard and simplified thanks to a constant β_1.

Lemma 2.16 (Bound for $\|m_t\|^2_{v_t^{-1/2}}$). *Under Assumption 2.1,* $\beta_1 < 1, \epsilon > 0$, *and the definitions of* α_t, m_t, v_t *in* ADAMNC, *it holds that*

$$\|m_t\|^2_{v_t^{-1/2}} \le \sqrt{t}(1-\beta_1)\sum_{i=1}^{d}\sum_{j=1}^{t}\frac{\beta_1^{t-j}(g_j^{(i)})^2}{\sqrt{\sum_{k=1}^{j}(g_k^{(i)})^2}}.$$

Proof. Using the expression (2.10) for m_t and $v_t^{(i)} = \frac{1}{t}\left(\sum_{j=1}^{t}(g_j^{(i)})^2+\epsilon\right)$, we obtain:[3]

$$\begin{aligned}\|m_t\|^2_{v_t^{-1/2}} &= \sum_{i=1}^{d}\frac{(m_t^{(i)})^2}{(v_t^{(i)})^{1/2}} = \sum_{i=1}^{d}\frac{\left(\sum_{j=1}^{t}(1-\beta_1)\beta_1^{t-j}g_j^{(i)}\right)^2}{\sqrt{\frac{1}{t}\left(\epsilon+\sum_{k=1}^{t}(g_k^{(i)})^2\right)}} \le \sqrt{t}(1-\beta_1)^2\sum_{i=1}^{d}\frac{\left(\sum_{j=1}^{t}\beta_1^{t-j}g_j^{(i)}\right)^2}{\sqrt{\sum_{k=1}^{t}(g_k^{(i)})^2}}\\ &\le \sqrt{t}(1-\beta_1)^2\sum_{i=1}^{d}\frac{\left(\sum_{j=1}^{t}\beta_1^{t-j}(g_j^{(i)})^2\right)\left(\sum_{j=1}^{t}\beta_1^{t-j}\right)}{\sqrt{\sum_{k=1}^{t}(g_k^{(i)})^2}} \le \sqrt{t}(1-\beta_1)\sum_{i=1}^{d}\frac{\sum_{j=1}^{t}\beta_1^{t-j}(g_j^{(i)})^2}{\sqrt{\sum_{k=1}^{t}(g_k^{(i)})^2}}\\ &\le \sqrt{t}(1-\beta_1)\sum_{i=1}^{d}\sum_{j=1}^{t}\frac{\beta_1^{t-j}(g_j^{(i)})^2}{\sqrt{\sum_{k=1}^{j}(g_k^{(i)})^2}},\end{aligned} \tag{2.19}$$

where the first inequality is due to $\epsilon > 0$, second inequality is by Cauchy-Schwarz, the third one by the sum of geometric series, and the final one is by $j \le t$. ■

Lemma 2.17 (Bound for $\sum_{t=1}^{T}\alpha_t\|m_t\|^2_{v_t^{-1/2}}$). *Under Assumption 2.1,* $\beta_1 < 1$, $\epsilon > 0$, *and the definitions of* α_t, m_t, v_t *in* ADAMNC, *it holds that*

$$\sum_{t=1}^{T}\alpha_t\|m_t\|^2_{v_t^{-1/2}} \le 2\alpha\sum_{i=1}^{d}\sqrt{\sum_{t=1}^{T}(g_t^{(i)})^2}. \tag{2.20}$$

Proof. We have, by using Lemma 2.16

$$\begin{aligned}\sum_{t=1}^{T}\alpha_t\|m_t\|^2_{v_t^{-1/2}} &= \sum_{t=1}^{T}\alpha_t\sqrt{t}(1-\beta_1)\sum_{i=1}^{d}\sum_{j=1}^{t}\frac{\beta_1^{t-j}(g_j^{(i)})^2}{\sqrt{\sum_{k=1}^{j}(g_k^{(i)})^2}}\\ &= \alpha(1-\beta_1)\sum_{i=1}^{d}\sum_{t=1}^{T}\sum_{j=1}^{t}\frac{\beta_1^{t-j}(g_j^{(i)})^2}{\sqrt{\sum_{k=1}^{j}(g_k^{(i)})^2}} = \alpha(1-\beta_1)\sum_{i=1}^{d}\sum_{j=1}^{T}\sum_{t=j}^{T}\frac{\beta_1^{t-j}(g_j^{(i)})^2}{\sqrt{\sum_{k=1}^{j}(g_k^{(i)})^2}}\\ &\le \alpha\sum_{i=1}^{d}\sum_{j=1}^{T}\frac{(g_j^{(i)})^2}{\sqrt{\sum_{k=1}^{j}(g_k^{(i)})^2}} \le 2\alpha\sum_{i=1}^{d}\sqrt{\sum_{j=1}^{T}(g_j^{(i)})^2},\end{aligned}$$

where the second equality is due to $\alpha_t = \frac{\alpha}{\sqrt{t}}$, third equality is by changing the order of summation, first inequality by summation of the geometric series. For the last inequality, we use a

[3] In the sequel, the same comments about the indeterminate form $\frac{0}{0}$ apply here as in Lemma 2.14.

standard inequality for numerical sequences, see for example [ACBG02, Lemma 3.5]

$$\sum_{j=1}^{T} \frac{a_j}{\sqrt{\sum_{k=1}^{j} a_k}} \leq 2\sqrt{\sum_{j=1}^{T} a_j} \quad \text{for all } a_1, \ldots, a_T \geq 0. \qquad \blacksquare$$

Proof of Theorem 2.5. We will follow the proof structure of Theorem 2.2. First, we start from (2.12) which applies to ADAMNCas the update of m_t is the same as AMSGRAD

$$R(T) \leq \sum_{t=1}^{T} \langle g_t, x_t - x \rangle = \frac{\beta_1}{1-\beta_1} \langle m_T, x_T - x \rangle + \sum_{t=1}^{T} \langle m_t, x_t - x \rangle + \frac{\beta_1}{1-\beta_1} \sum_{t=1}^{T} \langle m_{t-1}, x_{t-1} - x_t \rangle. \tag{2.21}$$

Then we again bound each term in the right-hand side seperately.

• *Bound for* $\sum_{t=1}^{T} \langle m_t, x_t - x \rangle$: We proceed similarly to the derivations in (2.13) and (2.14), the main change being that we now have v_t instead of $\hat{v}_t$. We have:

$$\begin{aligned} \langle m_t, x_t - x \rangle &\leq \frac{1}{2\alpha_{t-1}} \|x_t - x\|^2_{v_{t-1}^{1/2}} - \frac{1}{2\alpha_t} \|x_{t+1} - x\|^2_{v_t^{1/2}} + \frac{1}{2} \sum_{i=1}^{d} \left(\frac{(v_t^{(i)})^{1/2}}{\alpha_t} - \frac{(v_{t-1}^{(i)})^{1/2}}{\alpha_{t-1}} \right) (x_t^{(i)} - x^{(i)})^2 \\ &\quad + \frac{\alpha_t}{2} \|m_t\|^2_{v_t^{-1/2}} \\ &\leq \frac{1}{2\alpha_{t-1}} \|x_t - x\|^2_{v_{t-1}^{1/2}} - \frac{1}{2\alpha_t} \|x_{t+1} - x\|^2_{v_t^{1/2}} + \frac{D^2}{2} \sum_{i=1}^{d} \left(\frac{(v_t^{(i)})^{1/2}}{\alpha_t} - \frac{(v_{t-1}^{(i)})^{1/2}}{\alpha_{t-1}} \right) + \frac{\alpha_t}{2} \|m_t\|^2_{v_t^{-1/2}}, \end{aligned}$$

where the last inequality is due to $\frac{(v_t^{(i)})^{1/2}}{\alpha_t} \geq \frac{(v_{t-1}^{(i)})^{1/2}}{\alpha_{t-1}}$, since by definition $v_t^{(i)} = \frac{1}{t} \sum_{j=1}^{t} (g_j^{(i)})^2$ and $\alpha_t = \frac{\alpha}{\sqrt{t}}$. We now proceed to telescope this inequality, assuming as before that $\frac{1}{\alpha_0} = 0$. Doing so, we obtain:

$$\sum_{t=1}^{T} \langle m_t, x_t - x \rangle \leq \frac{D^2}{2} \sum_{i=1}^{d} \frac{(v_T^{(i)})^{1/2}}{\alpha_T} + \frac{1}{2} \sum_{t=1}^{T} \alpha_t \|m_t\|^2_{v_t^{-1/2}}. \tag{2.22}$$

• *Bounds for* $\langle m_T, x_T - x \rangle$ *and* $\sum_{t=1}^{T} \langle m_{t-1}, x_{t-1} - x_t \rangle$: These bounds will be similar as in the proof of Theorem 2.2. Again, the only change in calculations in (5.38) and (2.17) is that now we have v_t instead of $\hat{v}_t$

$$\sum_{t=1}^{T} \langle m_{t-1}, x_{t-1} - x_t \rangle \leq \sum_{t=1}^{T-1} \alpha_t \|m_t\|^2_{v_t^{-1/2}}, \tag{2.23}$$

and

$$\langle m_T, x_T - x \rangle \leq \alpha_T \|m_T\|^2_{v_T^{-1/2}} + \frac{D^2}{4\alpha_T} \sum_{i=1}^{d} (v_T^{(i)})^{1/2}. \tag{2.24}$$

We now combine (2.22), (2.23), and (2.24) in (2.21), estimate using the same steps in (2.18), and use the bound for $\sum_{t=1}^T \alpha_t \|m_t\|^2_{v_t^{-1/2}}$ from Lemma 2.17 to conclude:

$$\begin{aligned}\sum_{t=1}^T \langle g_t, x_t - x\rangle &= \frac{\beta_1}{1-\beta_1}\langle m_T, x_T - x\rangle + \sum_{t=1}^T \langle m_t, x_t - x\rangle + \frac{\beta_1}{1-\beta_1}\sum_{t=1}^T \langle m_{t-1}, x_{t-1} - x_t\rangle \\ &\le \left(\frac{D^2}{2} + \frac{\beta_1 D^2}{4(1-\beta_1)}\right)\sum_{i=1}^d \frac{(v_T^{(i)})^{1/2}}{\alpha_T} + \left(\frac{1}{2} + \frac{\beta_1}{1-\beta_1}\right)\sum_{t=1}^T \alpha_t \|m_t\|^2_{v_t^{-1/2}} \\ &\le \frac{D^2\sqrt{T}}{2\alpha(1-\beta_1)}\sum_{i=1}^d (v_T^{(i)})^{1/2} + \frac{2\alpha}{1-\beta_1}\sum_{i=1}^d \sqrt{\sum_{t=1}^T (g_t^{(i)})^2}.\end{aligned}$$

■

2.5.3 Proofs for Section 2.2.5

Lemma 2.18 (Bound for $\|m_t\|^2_{\hat{v}_t^{-1}}$)**.** *Under Assumption 2.1, $\beta_1 < 1, \epsilon > 0$, and the definitions of $\alpha_t, m_t, v_t, \hat{v}_t$ in* SADAM, *it holds that*

$$\|m_t\|^2_{\hat{v}_t^{-1}} \le t(1-\beta_1)\sum_{i=1}^d \sum_{j=1}^t \frac{\beta_1^{t-j}(g_j^{(i)})^2}{\sum_{k=1}^j (g_k^{(i)})^2 + \epsilon}. \tag{2.25}$$

Proof. We have

$$\begin{aligned}\|m_t\|^2_{\hat{v}_t^{-1}} &= \sum_{i=1}^d \frac{(m_t^{(i)})^2}{\hat{v}_t^{(i)}} = \sum_{i=1}^d \frac{(m_t^{(i)})^2}{v_t^{(i)} + \frac{\epsilon}{t}} = t(1-\beta_1)^2 \sum_{i=1}^d \frac{\left(\sum_{j=1}^t \beta_1^{t-j} g_j^{(i)}\right)^2}{\sum_{k=1}^t (g_k^{(i)})^2 + \epsilon} \\ &\le t(1-\beta_1)\sum_{i=1}^d \frac{\sum_{j=1}^t \beta_1^{t-j}(g_j^{(i)})^2}{\sum_{k=1}^t (g_k^{(i)})^2 + \epsilon} \le t(1-\beta_1)\sum_{i=1}^d\sum_{j=1}^t \frac{\beta_1^{t-j}(g_j^{(i)})^2}{\sum_{k=1}^j (g_k^{(i)})^2 + \epsilon},\end{aligned} \tag{2.26}$$

where we used $\hat{v}_t^{(i)} = \frac{1}{t}\sum_{k=1}^t (g_k^{(i)})^2 + \frac{\epsilon}{t}$ and expression for m_t from (2.10) in the first line. First inequality is by Cauchy-Schwarz and sum of geometric series; the last inequality is by $j \le t$. ■

Lemma 2.19 (Bound for $\sum_{t=1}^T \alpha_t \|m_t\|^2_{\hat{v}_t^{-1}}$)**.** *Under Assumption 2.1, $\beta_1 < 1, \epsilon > 0$, and the definitions of $\alpha_t, m_t, v_t, \hat{v}_t$ in* SADAM, *it holds that*

$$\sum_{t=1}^T \alpha_t \|m_t\|^2_{\hat{v}_t^{-1}} \le \alpha \sum_{i=1}^d \log\left(\frac{\sum_{t=1}^T (g_t^{(i)})^2}{\epsilon} + 1\right). \tag{2.27}$$

Proof. We have, by Lemma 2.18

$$\begin{aligned}\sum_{t=1}^T \alpha_t \|m_t\|^2_{\hat{v}_t^{-1}} &= \sum_{t=1}^T \alpha_t t(1-\beta)\sum_{i=1}^d\sum_{j=1}^t \frac{\beta_1^{t-j}(g_j^{(i)})^2}{\sum_{k=1}^j (g_k^{(i)})^2 + \epsilon} \\ &= \alpha(1-\beta)\sum_{i=1}^d\sum_{t=1}^T\sum_{j=1}^t \frac{\beta_1^{t-j}(g_j^{(i)})^2}{\sum_{k=1}^j (g_k^{(i)})^2 + \epsilon} = \alpha(1-\beta)\sum_{i=1}^d\sum_{j=1}^T\sum_{t=j}^T \frac{\beta_1^{t-j}(g_j^{(i)})^2}{\sum_{k=1}^j (g_k^{(i)})^2 + \epsilon}\end{aligned}$$

$$\leq \alpha \sum_{i=1}^{d} \sum_{j=1}^{T} \frac{(g_j^{(i)})^2}{\sum_{k=1}^{j} (g_k^{(i)})^2 + \epsilon} \leq \alpha \sum_{i=1}^{d} \log\left(\frac{\sum_{t=1}^{T} (g_t^{(i)})^2}{\epsilon} + 1 \right), \tag{2.28}$$

where the second equality is by the definition of α_t and the third equality is by changing the order of summation. First inequality is by the sum of geometric series and the last inequality is by

$$\sum_{j=1}^{T} \frac{a_j}{\sum_{k=1}^{j} a_k + \epsilon} \leq \log\left(\frac{\sum_{j=1}^{T} a_j}{\epsilon} + 1 \right), \tag{2.29}$$

for nonnegative $a_1, \ldots, a_T$ and $\epsilon > 0$ – see e.g., [DHS10, Lemma 12],[HAK07, Lemma 11]. ■

Proof of Theorem 2.6. Let $x = \operatorname{argmin}_{y \in \mathcal{K}} \sum_{t=1}^{T} f_t(y)$. In Theorem 2.2 we used convexity only once: going from $R(T)$ to $\sum_{t=1}^{T} \langle g_t, x_t - x \rangle$. Instead, strong convexity gives us $f_t(x) \geq f_t(x_t) + \langle g_t, x - x_t \rangle + \frac{\mu}{2} \|x_t - x\|^2$, which combined for all t yields

$$R(T) = \sum_{t=1}^{T} f_t(x_t) - f_t(x) \leq \sum_{t=1}^{T} \langle g_t, x_t - x \rangle - \frac{\mu}{2} \sum_{t=1}^{T} \|x_t - x\|^2. \tag{2.30}$$

We want to estimate $\sum_{t=1}^{T} \langle g_t, x_t - x \rangle$. Similarly to (2.12), we have

$$\sum_{t=1}^{T} \langle g_t, x_t - x \rangle \leq \frac{\beta_1}{1-\beta_1} \langle m_T, x_T - x \rangle + \sum_{t=1}^{T} \langle m_t, x_t - x \rangle + \frac{\beta_1}{1-\beta_1} \sum_{t=1}^{T} \langle m_{t-1}, x_{t-1} - x_t \rangle. \tag{2.31}$$

• *Bound for* $\sum_{t=1}^{T} \langle m_t, x_t - x \rangle$: We proceed similar to (2.13) and (2.14). The only change is that now we have $\hat{v}_t$ instead of $\hat{v}_t^{1/2}$

$$\langle m_t, x_t - x \rangle \leq \frac{1}{2\alpha_{t-1}} \|x_t - x\|^2_{\hat{v}_{t-1}} - \frac{1}{2\alpha_t} \|x_{t+1} - x\|^2_{\hat{v}_t} + \frac{1}{2} \sum_{i=1}^{d} \left(\frac{\hat{v}_t^{(i)}}{\alpha_t} - \frac{\hat{v}_{t-1}^{(i)}}{\alpha_{t-1}} \right) (x_t^{(i)} - x^{(i)})^2 + \frac{\alpha_t}{2} \|m_t\|^2_{\hat{v}_t^{-1}}.$$

We sum the above inequality and use the fact that $\frac{1}{\alpha_0} \|x_1 - x\|^2_{\hat{v}_0} = 0$ to obtain

$$\sum_{t=1}^{T} \langle m_t, x_t - x \rangle \leq \sum_{t=1}^{T} \sum_{i=1}^{d} \left(\frac{\hat{v}_t^{(i)}}{2\alpha_t} - \frac{\hat{v}_{t-1}^{(i)}}{2\alpha_{t-1}} \right) (x_t^{(i)} - x^{(i)})^2 + \sum_{t=1}^{T} \frac{\alpha_t}{2} \|m_t\|^2_{\hat{v}_t^{-1}}. \tag{2.32}$$

• *Bound for* $\sum_{t=1}^{T} \langle m_{t-1}, x_{t-1} - x_t \rangle$: This bound will be similar to the one we derived for Theorem 2.2. The main change in the calculations of (5.38) is that we will have $\hat{v}_t$ instead of $\hat{v}_t^{1/2}$ for using Hölder's inequality and nonexpansiveness

$$\sum_{t=1}^{T} \langle m_{t-1}, x_{t-1} - x_t \rangle \leq \sum_{t=1}^{T-1} \alpha_t \|m_t\|^2_{\hat{v}_t^{-1}} \leq \sum_{t=1}^{T} \alpha_t \|m_t\|^2_{\hat{v}_t^{-1}}. \tag{2.33}$$

We collect these estimations in (2.31) and (2.30) to derive

$$
\begin{aligned}
R(T) = \sum_{t=1}^{T} f_t(x_t) - f_t(x) \le \frac{\beta_1}{1-\beta_1}\langle m_T, x_T - x\rangle + \frac{1+\beta_1}{2(1-\beta_1)}\sum_{t=1}^{T}\alpha_t\|m_t\|^2_{\hat{v}_t^{-1}} \\
+ \sum_{t=1}^{T}\sum_{i=1}^{d}\left(\frac{\hat{v}_t^{(i)}}{2\alpha_t} - \frac{\hat{v}_{t-1}^{(i)}}{2\alpha_{t-1}}\right)(x_t^{(i)} - x^{(i)})^2 - \sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\mu}{2}(x_t^{(i)} - x^{(i)})^2. \qquad (2.34)
\end{aligned}
$$

We collect the last two terms and use the assumption on the step size $\alpha \ge \frac{G^2}{\mu}$ and the definition $\hat{v}_t^{(i)} = \frac{1}{t}\sum_{j=1}^{t}(g_j^{(i)})^2 + \frac{\epsilon}{t}$ to derive

$$
\frac{\hat{v}_t^{(i)}}{2\alpha_t} - \frac{\hat{v}_{t-1}^{(i)}}{2\alpha_{t-1}} - \frac{\mu}{2} = \frac{(g_t^{(i)})^2}{2\alpha} - \frac{\mu}{2} \le 0.
$$

Thus, (2.34) becomes

$$
\sum_{t=1}^{T} f_t(x_t) - f_t(x) \le \frac{\beta_1}{1-\beta_1}\langle m_T, x_T - x\rangle + \frac{1+\beta_1}{2(1-\beta_1)}\sum_{t=1}^{T}\alpha_t\|m_t\|^2_{\hat{v}_t^{-1}}.
$$

We finalize by using $\frac{1+\beta_1}{2} \le 1$, Lemma 2.19 for the last term, and $\|m_t\|_\infty \le G$, $\|x_t - x\|_\infty \le D$ for the first term

$$
\sum_{t=1}^{T} f_t(x_t) - f_t(x) \le \frac{\beta_1 dGD}{1-\beta_1} + \frac{\alpha}{1-\beta_1}\sum_{i=1}^{d}\log\left(\frac{\sum_{t=1}^{T}(g_t^{(i)})^2}{\epsilon} + 1\right).
$$

■

2.5.4 Proofs for Section 2.3

Proof of Lemma 2.10. By the definition of $\hat{x}_t$ in (2.7), it follows that

$$
\varphi(\hat{x}_t) + \frac{\bar{\rho}}{2}\|x_t - \hat{x}_t\|^2_{\hat{v}_t^{1/2}} \le \varphi(x_t) + \frac{\bar{\rho}}{2}\|x_t - x_t\|^2_{\hat{v}_t^{1/2}} = \varphi(x_t).
$$

Next, we use $\hat{\rho}$-weak convexity of φ with respect to norm $\|\cdot\|_{\hat{v}_t^{1/2}}$ from Remark 2.7, and the fact that $x_t, \hat{x}_t \in \mathcal{X}$ to get for any vector q_t such that $q_t \in \partial f(x_t)$,

$$
\varphi(x_t) - \varphi(\hat{x}_t) \le \langle x_t - \hat{x}_t, q_t\rangle + \frac{\hat{\rho}}{2}\|x_t - \hat{x}_t\|^2_{\hat{v}_t^{1/2}}.
$$

We sum two inequalities and apply Cauchy-Schwarz inequality

$$
\frac{\bar{\rho} - \hat{\rho}}{2}\|x_t - \hat{x}_t\|^2_{\hat{v}_t^{1/2}} \le \langle x_t - \hat{x}_t, g_t\rangle \le \|q_t\|_{\hat{v}_t^{-1/2}}\|x_t - \hat{x}_t\|_{\hat{v}_t^{1/2}},
$$

which yields

$$
\frac{\bar{\rho} - \hat{\rho}}{2}\|x_t - \hat{x}_t\|_{\hat{v}_t^{1/2}} \le \|q_t\|_{\hat{v}_t^{-1/2}}.
$$

As $\hat{v}_{t,i} \geq \epsilon$ and for q_t such that $\mathbb{E}g_t = q_t$, $\|q_t\|^2 = \|\mathbb{E}g_t\|^2 \leq \mathbb{E}\|g_t\|^2 \leq dG^2$ by Assumption 2.2, we have

$$\|q_t\|^2_{\hat{v}_t^{-1/2}} \leq \frac{dG^2}{\sqrt{\epsilon}}$$

and the final bound follows immediately. ■

Proof of Lemma 2.11. We first decompose the LHS

$$\begin{aligned}
\alpha_t\langle x_t - \hat{x}_t, g_t\rangle &= \alpha_t\langle x_t - \hat{x}_t, q_t\rangle + \alpha_t\langle x_t - \hat{x}_t, g_t - q_t\rangle \\
&= \alpha_t\langle x_t - \hat{x}_t, q_t\rangle + \langle \alpha_t(x_t - \hat{x}_t) - \alpha_{t-1}(x_{t-1} - \hat{x}_{t-1}), g_t - q_t\rangle \\
&+ \langle \alpha_{t-1}(x_{t-1} - \hat{x}_{t-1}), g_t - q_t\rangle
\end{aligned} \tag{2.35}$$

In this bound, the last term will be 0 after taking conditional expectation $\mathbb{E}_t$ as $\hat{x}_{t-1}$ depends on $\hat{v}_{t-1}$, which, in turn, depends only on $g_1, \dots, g_{t-1}$, thus, independent of g_t.

For the first term in (2.35), we recall that $\hat{x}_t \in \mathcal{K}$, $x_t \in \mathcal{K}$, $q_t \in \partial f(x_t)$. Then we use $\hat{\rho}$-weak convexity of f with respect to $\|\cdot\|_{\hat{v}_t^{1/2}}$,

$$\begin{aligned}
\langle x_t - \hat{x}_t, q_t\rangle &\geq f(x_t) - f(\hat{x}_t) - \frac{\hat{\rho}}{2}\|x_t - \hat{x}_t\|^2_{\hat{v}_t^{1/2}} \\
&= \left(f(x_t) + \frac{\bar{\rho}}{2}\|x_t - x_t\|^2_{\hat{v}_t^{1/2}}\right) - \left(f(\hat{x}_t) + \frac{\bar{\rho}}{2}\|x_t - \hat{x}_t\|^2_{\hat{v}_t^{1/2}}\right) + \frac{\bar{\rho} - \hat{\rho}}{2}\|x_t - \hat{x}_t\|^2_{\hat{v}_t^{1/2}} \\
&\geq (\bar{\rho} - \hat{\rho})\|x_t - \hat{x}_t\|^2_{\hat{v}_t^{1/2}},
\end{aligned} \tag{2.36}$$

where the last step is due to $x \mapsto f(x) + \delta_{\mathcal{K}}(x) + \frac{\bar{\rho}}{2}\|x - x_t\|^2_{\hat{v}_t^{1/2}}$ being $\bar{\rho} - \hat{\rho}$ strongly convex w.r.t. $\|\cdot\|_{\hat{v}_t^{1/2}}$, with the minimizer $\hat{x}_t$, and $x_t, \hat{x}_t \in \mathcal{K}$.

Next, we need to lower bound the second term in (2.35). For this we upper bound the term

$$\begin{aligned}
\langle \alpha_{t-1}(x_{t-1} - \hat{x}_{t-1}) - \alpha_t(x_t - \hat{x}_t), g_t - q_t\rangle &= (\alpha_{t-1} - \alpha_t)\langle x_t - \hat{x}_t, g_t - q_t\rangle \\
&+ \alpha_{t-1}\langle x_{t-1} - x_t, g_t - q_t\rangle + \alpha_{t-1}\langle \hat{x}_t - \hat{x}_{t-1}, g_t - q_t\rangle.
\end{aligned} \tag{2.37}$$

We proceed with bounding the first term in the RHS of (2.37), using $\alpha_t \leq \alpha_{t-1}$,

$$\begin{aligned}
\mathbb{E}_t(\alpha_{t-1} - \alpha_t)\langle x_t - \hat{x}_t, g_t - q_t\rangle &\leq (\alpha_{t-1} - \alpha_t)\mathbb{E}_t\|x_t - \hat{x}_t\|\|g_t - q_t\| \\
&\leq (\alpha_{t-1} - \alpha_t)\hat{D}\mathbb{E}_t\|g_t - q_t\| \\
&\leq (\alpha_{t-1} - \alpha_t)\hat{D}\sqrt{\mathbb{E}_t\|g_t\|^2} \\
&\leq (\alpha_{t-1} - \alpha_t)\hat{D}\sqrt{d}G,
\end{aligned}$$

where the second inequality follows from Lemma 2.10 and third inequality follows from Jensen's inequality and $\mathbb{E}_t\|g_t - \mathbb{E}_t g_t\|^2 \leq \mathbb{E}_t\|g_t\|^2$.

For the second term in the RHS of (2.37) we use Cauchy-Schwarz and Young's inequalities and nonexpansiveness of weighted projection to get

$$\begin{aligned}
\mathbb{E}_t\alpha_{t-1}\langle x_{t-1}-x_t, g_t-q_t\rangle &\le \frac{1}{2}\mathbb{E}_t\|x_t-x_{t-1}\|^2_{\hat{v}_{t-1}^{1/2}} + \frac{\alpha_{t-1}^2}{2}\mathbb{E}_t\|g_t-q_t\|^2_{\hat{v}_{t-1}^{-1/2}} \\
&\le \frac{\alpha_{t-1}^2}{2}\mathbb{E}_t\|m_{t-1}\|^2_{\hat{v}_{t-1}^{-1/2}} + \frac{\alpha_{t-1}^2}{2\sqrt{\epsilon}}\mathbb{E}_t\|g_t-q_t\|^2 \\
&\le \frac{\alpha_{t-1}^2}{2}\mathbb{E}_t\|m_{t-1}\|^2_{\hat{v}_{t-1}^{-1/2}} + \frac{\alpha_{t-1}^2}{2\sqrt{\epsilon}}\mathbb{E}_t\|g_t\|^2.
\end{aligned}$$

Similarly, we estimate the third term in the RHS of (2.37)

$$\begin{aligned}
\mathbb{E}_t\alpha_{t-1}\langle \hat{x}_t-\hat{x}_{t-1}, g_t-q_t\rangle &\le \frac{\bar{\rho}-\hat{\rho}}{4\bar{\rho}}\|\hat{x}_t-\hat{x}_{t-1}\|^2_{\hat{v}_t^{1/2}} + \frac{\alpha_{t-1}^2\bar{\rho}}{\bar{\rho}-\hat{\rho}}\mathbb{E}_t\|g_t-q_t\|^2_{\hat{v}_t^{-1/2}} \\
&\le \frac{\bar{\rho}-\hat{\rho}}{4\bar{\rho}}\|\hat{x}_t-\hat{x}_{t-1}\|^2_{\hat{v}_t^{1/2}} + \frac{\alpha_{t-1}^2\bar{\rho}}{(\bar{\rho}-\hat{\rho})\sqrt{\epsilon}}\mathbb{E}_t\|g_t\|^2.
\end{aligned}$$

Combining all the bounds gives the result. ■

Proof of Lemma 2.12. We start with the result of Lemma 2.14

$$\|m_t\|^2_{\hat{v}_t^{-1/2}} \le \frac{(1-\beta_1)^2}{\sqrt{(1-\beta_2)(1-\gamma)}}\sum_{i=1}^d\sum_{j=1}^t \beta_1^{t-j}|g_{j,i}|.$$

We will proceed similar to Lemma 2.15 with the only change of having α_t^2 instead of α_t

$$\begin{aligned}
\sum_{t=1}^T \alpha_t^2\|m_t\|^2_{\hat{v}_t^{-1/2}} &\le \frac{(1-\beta_1)^2}{\sqrt{(1-\beta_2)(1-\gamma)}}\sum_{i=1}^d\sum_{t=1}^T\alpha_t^2\sum_{j=1}^t\beta_1^{t-j}|g_{j,i}| \\
&= \frac{(1-\beta_1)^2}{\sqrt{(1-\beta_2)(1-\gamma)}}\sum_{i=1}^d\sum_{j=1}^T\sum_{t=j}^T\alpha_t^2\beta_1^{t-j}|g_{j,i}| \\
&\le \frac{1-\beta_1}{\sqrt{(1-\beta_2)(1-\gamma)}}\sum_{i=1}^d\sum_{j=1}^T\alpha_j^2|g_{j,i}| \\
&\le \frac{1-\beta_1}{\sqrt{(1-\beta_2)(1-\gamma)}}\alpha^2 dG(1+\log T).
\end{aligned}$$

■

Theorem 2.8. Let Assumption 2.2 hold. Let $\beta_1 < 1$, $\beta_2 < 1$, $\gamma = \frac{\beta_1^2}{\beta_2} < 1$, $\bar{\rho} = 2\hat{\rho}$, $\epsilon > 0$ and t^* selected randomly from $[T]$. Then, for iterate x_{t^*} generated by Algorithm 2.1, it follows that

$$\mathbb{E}\|\nabla\varphi_{1/\bar{\rho}}^{t^*}(x_{t^*})\|^2_{\hat{v}_{t^*}^{-1/2}} \le \frac{2}{\alpha\sqrt{T}}\left[C_1 + (1+\log T)C_2 + C_3\right],$$

with $C_1 = \frac{4\rho\beta_1\alpha}{\sqrt{\epsilon}(1-\beta_1)}\sqrt{d}\hat{D}G + \varphi_{1/\bar{\rho}}^1(x_1) - f^\star$, $C_2 = \frac{5\rho}{\epsilon}dG^2 + \frac{2\rho}{\sqrt{\epsilon}}\left(\frac{2G}{\sqrt{\epsilon}} + \frac{\beta_1}{1-\beta_1} + \frac{2\beta_1^2}{(1-\beta_1)^2}\right)\frac{1-\beta_1}{\sqrt{(1-\beta_2)(1-\gamma)}}dG$, $C_3 = \bar{\rho}\hat{D}^2\sum_{i=1}^d \mathbb{E}\hat{(}v_{T+1}^{(i)})^{1/2}$, and $\hat{D} := \frac{2\sqrt{d}G}{\rho}$.

Proof. We sum the result of Lemma 2.1 by using A_t instead of $x_t - x$, and use $A_1 = A_0$. with $m_0 = 0$. We note that we have $A_t = \bar\rho\alpha_t(x_t - \hat x_t)$, for $t \geq 1$.

$$\sum_{t=1}^{T}\langle A_t, g_t\rangle = \frac{\beta_1}{1-\beta_1}\langle A_T, m_T\rangle + \sum_{t=1}^{T}\langle A_t, m_t\rangle + \frac{\beta_1}{1-\beta_1}\sum_{t=1}^{T-1}\langle A_t - A_{t+1}, m_t\rangle. \tag{2.38}$$

After plugging in the value of A_t, (2.38) becomes

$$\begin{aligned}\sum_{t=1}^{T}\bar\rho\alpha_t\langle x_t - \hat x_t, g_t\rangle \leq \frac{\beta_1\bar\rho\alpha_T}{1-\beta_1}\langle x_T - \hat x_T, m_T\rangle + \sum_{t=1}^{T}\bar\rho\alpha_t\langle x_t - \hat x_t, m_t\rangle \\ + \frac{\beta_1\bar\rho}{1-\beta_1}\sum_{t=1}^{T-1}\langle \alpha_t(x_t - \hat x_t) - \alpha_{t+1}(x_{t+1} - \hat x_{t+1}), m_t\rangle.\end{aligned} \tag{2.39}$$

LHS of this bound is suitable for applying Lemma 2.11 to obtain the stationarity measure. We have to estimate the three terms on the RHS.

- Bound for $\frac{\beta_1\bar\rho\alpha_T}{1-\beta_1}\langle x_T - \hat x_T, m_T\rangle$ in (2.39): We bound this term by Cauchy-Schwarz inequality, Lemma 2.10, and $\|m_t\|_\infty \leq G$:

$$\langle x_T - \hat x_T, m_T\rangle \leq \|x_T - \hat x_T\|\|m_T\| \leq \hat D\sqrt{d}G. \tag{2.40}$$

- Bound for $\frac{\beta_1\bar\rho}{1-\beta_1}\sum_{t=1}^{T-1}\langle \alpha_t(x_t - \hat x_t) - \alpha_{t+1}(x_{t+1} - \hat x_{t+1}), m_t\rangle$ in (2.39): We have

$$\begin{aligned}\langle \alpha_t(x_t - \hat x_t) - \alpha_{t+1}(x_{t+1} - \hat x_{t+1}), m_t\rangle = (\alpha_t - \alpha_{t+1})\langle x_{t+1} - \hat x_{t+1}, m_t\rangle + \alpha_t\langle x_t - x_{t+1}, m_t\rangle \\ + \alpha_t\langle \hat x_{t+1} - \hat x_t, m_t\rangle.\end{aligned} \tag{2.41}$$

For the first term in (2.41), we use that $\alpha_t \geq \alpha_{t+1}$, Lemma 2.10, Cauchy-Schwarz inequality and $\|m_t\|_\infty \leq G$ to obtain

$$\sum_{t=1}^{T-1}(\alpha_t - \alpha_{t+1})\langle x_{t+1} - \hat x_{t+1}, m_t\rangle \leq \sum_{t=1}^{T-1}(\alpha_t - \alpha_{t+1})\hat D\sqrt{d}G \leq \alpha_1\hat D\sqrt{d}G.$$

For the second term of (2.41), using nonexpansiveness of weighted projection, we deduce

$$\begin{aligned}\alpha_t\langle x_t - x_{t+1}, m_t\rangle \leq \alpha_t\|x_t - x_{t+1}\|_{\hat v_t^{1/2}}\|m_t\|_{\hat v_t^{-1/2}} = \alpha_t\|x_t - \mathcal{P}_{\mathcal{K}}^{\hat v_t^{1/2}}(x_t - \alpha_t\hat v_t^{-1/2}m_t)\|_{\hat v_t^{1/2}}\|m_t\|_{\hat v_t^{-1/2}} \\ \leq \alpha_t^2\|m_t\|^2_{\hat v_t^{-1/2}}.\end{aligned}$$

First, summing (2.41), multiplying both sides of the inequality by $\frac{\beta_1\bar\rho}{1-\beta_1}$, and then plugging the last two bounds, we have

$$\begin{aligned}&\frac{\beta_1\bar\rho}{1-\beta_1}\sum_{t=1}^{T-1}\langle \alpha_t(x_t - \hat x_t) - \alpha_{t+1}(x_{t+1} - \hat x_{t+1}), m_t\rangle \\ &\qquad \leq \frac{\beta_1\bar\rho}{1-\beta_1}\alpha_1\hat D\sqrt{d}G + \sum_{t=1}^{T}\frac{\beta_1\bar\rho\alpha_t^2}{1-\beta_1}\|m_t\|^2_{\hat v_t^{-1/2}} + \sum_{t=1}^{T-1}\frac{\beta_1\bar\rho\alpha_t}{1-\beta_1}\langle \hat x_{t+1} - \hat x_t, m_t\rangle\end{aligned}$$

$$\begin{aligned}&\le \frac{\beta_1\bar{\rho}}{1-\beta_1}\alpha_1\hat{D}\sqrt{d}G + \sum_{t=1}^{T}\frac{\beta_1\bar{\rho}\alpha_t^2}{1-\beta_1}\|m_t\|^2_{\hat{v}_t^{-1/2}} + \sum_{t=1}^{T}\frac{\bar{\rho}-\hat{\rho}}{4}\|\hat{x}_{t+1}-\hat{x}_t\|^2_{\hat{v}_t^{1/2}}\\ &\quad + \frac{\bar{\rho}^2}{(\bar{\rho}-\hat{\rho})}\frac{\beta_1^2}{(1-\beta_1)^2}\sum_{t=1}^{T}\alpha_t^2\|m_t\|^2_{\hat{v}_t^{-1/2}},\end{aligned}\tag{2.42}$$

where we used Young's inequality in the last step.

• Bound for $\sum_{t=1}^{T}\bar{\rho}\alpha_t\langle x_t-\hat{x}_t, m_t\rangle$ in (2.39): We proceed as in eq. (3.6) to (3.8) in [DD19], but with a tighter bound in the beginning, where we use $x\mapsto f(x)+\delta_{\mathcal{K}}(x)+\frac{\bar{\rho}}{2}\|x-x_{t+1}\|^2_{\hat{v}_{t+1}^{1/2}}$ being $\bar{\rho}-\hat{\rho}$ strongly convex w.r.t. $\|\cdot\|_{\hat{v}_{t+1}^{1/2}}$, with the minimizer $\hat{x}_{t+1}$

$$\begin{aligned}\varphi_{1/\bar{\rho}}^{t+1}(x_{t+1}) &\le f(\hat{x}_t)+\frac{\bar{\rho}}{2}\|\hat{x}_t-x_{t+1}\|^2_{\hat{v}_{t+1}^{1/2}}-\frac{\bar{\rho}-\hat{\rho}}{2}\|\hat{x}_t-\hat{x}_{t+1}\|^2_{\hat{v}_{t+1}^{1/2}}\\ &= f(\hat{x}_t)+\frac{\bar{\rho}}{2}\|\hat{x}_t-x_{t+1}\|^2_{\hat{v}_t^{1/2}}+\frac{\bar{\rho}}{2}\|\hat{x}_t-x_{t+1}\|^2_{\hat{v}_{t+1}^{1/2}-\hat{v}_t^{1/2}}-\frac{\bar{\rho}-\hat{\rho}}{2}\|\hat{x}_t-\hat{x}_{t+1}\|^2_{\hat{v}_{t+1}^{1/2}}.\end{aligned}\tag{2.43}$$

We estimate the second term in the RHS of (2.43) by the definition of x_{t+1}, then using $\hat{x}_t\in\mathcal{K}$ and nonexpansiveness of the weighted projection in the weighted norm

$$\begin{aligned}\frac{\bar{\rho}}{2}\|\hat{x}_t-x_{t+1}\|^2_{\hat{v}_t^{1/2}} &= \frac{\bar{\rho}}{2}\|\mathcal{P}_{\mathcal{K}}^{\hat{v}_t^{1/2}}(x_t-\alpha_t\hat{v}_t^{-1/2}m_t)-\hat{x}_t\|^2_{\hat{v}_t^{1/2}}\\ &= \frac{\bar{\rho}}{2}\|\mathcal{P}_{\mathcal{K}}^{\hat{v}_t^{1/2}}(x_t-\alpha_t\hat{v}_t^{-1/2}m_t)-\mathcal{P}_{\mathcal{K}}^{\hat{v}_t^{1/2}}(\hat{x}_t)\|^2_{\hat{v}_t^{1/2}}\\ &\le \frac{\bar{\rho}}{2}\|x_t-\alpha_t\hat{v}_t^{-1/2}m_t-\hat{x}_t\|^2_{\hat{v}_t^{1/2}}\\ &= \frac{\bar{\rho}}{2}\|x_t-\hat{x}_t\|^2_{\hat{v}_t^{1/2}}+\bar{\rho}\langle\hat{x}_t-x_t,\alpha_t m_t\rangle+\frac{\bar{\rho}}{2}\alpha_t^2\|m_t\|^2_{\hat{v}_t^{-1/2}}.\end{aligned}$$

We insert this estimate into (2.43) and use the definition of $\varphi_{1/\bar{\rho}}^{t}(x_t)$ to obtain

$$\begin{aligned}\varphi_{1/\bar{\rho}}^{t+1}(x_{t+1}) \le \varphi_{1/\bar{\rho}}^{t}(x_t)+\bar{\rho}\alpha_t\langle\hat{x}_t-x_t,m_t\rangle+\frac{\bar{\rho}}{2}\alpha_t^2\|m_t\|^2_{\hat{v}_t^{-1/2}}+\frac{\bar{\rho}}{2}\|\hat{x}_t-x_{t+1}\|^2_{\hat{v}_{t+1}^{1/2}-\hat{v}_t^{1/2}}\\ -\frac{\bar{\rho}-\hat{\rho}}{2}\|\hat{x}_t-\hat{x}_{t+1}\|^2_{\hat{v}_{t+1}^{1/2}}.\end{aligned}\tag{2.44}$$

We manipulate the second to last term, by $\|a+b\|^2\le 2\|a\|^2+2\|b\|^2$, $\hat{v}_{t+1}^{(i)}\ge\hat{v}_t^{(i)}$, and Lemma 2.10

$$\begin{aligned}\frac{\bar{\rho}}{2}\|\hat{x}_t-x_{t+1}\|^2_{\hat{v}_{t+1}^{1/2}-\hat{v}_t^{1/2}} &\le \bar{\rho}\|\hat{x}_t-x_t\|^2_{\hat{v}_{t+1}^{1/2}-\hat{v}_t^{1/2}}+\frac{G\bar{\rho}}{\sqrt{\epsilon}}\|x_t-x_{t+1}\|^2_{\hat{v}_t^{1/2}}\\ &\le \bar{\rho}\hat{D}^2\sum_{i=1}^{d}((\hat{v}_{t+1}^{(i)})^{1/2}-(\hat{v}_t^{(i)})^{1/2})+\frac{G\bar{\rho}}{\sqrt{\epsilon}}\alpha_t^2\|m_t\|^2_{\hat{v}_t^{-1/2}}.\end{aligned}\tag{2.45}$$

We use this estimate in (2.44) and sum the inequality to get

$$\bar{\rho}\alpha_t\sum_{t=1}^{T}\langle x_t-\hat{x}_t,m_t\rangle \le \varphi_{1/\bar{\rho}}^{1}(x_1)-\varphi_{1/\bar{\rho}}^{T+1}(x_{T+1})+\sum_{t=1}^{T}\left(\frac{1}{2}+\frac{G}{\sqrt{\epsilon}}\right)\bar{\rho}\alpha_t^2\|m_t\|^2_{\hat{v}_t^{-1/2}}$$

$$+\bar\rho\hat D^2\sum_{i=1}^{d}(\hat v_{T+1}^{(i)})^{1/2}-\sum_{t=1}^{T}\frac{\bar\rho-\hat\rho}{2}\|\hat x_t-\hat x_{t+1}\|^2_{\hat v_{t+1}^{1/2}}. \tag{2.46}$$

Combining estimates into (2.39). We now plug in (2.40), (2.42), (2.46) into (2.39) and use $\alpha_T \le \alpha$, $\hat v_{t+1}^{1/2} \ge \hat v_t^{1/2}$ to get

$$\begin{aligned}\sum_{t=1}^{T}\bar\rho\alpha_t\langle x_t-\hat x_t, g_t\rangle &\le \frac{2\beta_1\bar\rho\alpha}{(1-\beta_1)}\hat D\sqrt{d}G+\varphi^1_{1/\bar\rho}(x_1)-\varphi^{T+1}_{1/\bar\rho}(x_{T+1})+\bar\rho\hat D^2\sum_{i=1}^{d}(\hat v_{T+1}^{(i)})^{1/2}\\ &+\sum_{t=1}^{T}\left(\frac{1}{2}+\frac{G}{\sqrt{\epsilon}}+\frac{\beta_1}{1-\beta_1}+\frac{\bar\rho}{\bar\rho-\hat\rho}\frac{\beta_1^2}{(1-\beta_1)^2}\right)\bar\rho\alpha_t^2\|m_t\|^2_{\hat v_t^{-1/2}}-\sum_{t=1}^{T}\frac{\bar\rho-\hat\rho}{4}\|\hat x_t-\hat x_{t+1}\|^2_{\hat v_{t+1}^{1/2}}.\end{aligned} \tag{2.47}$$

At this point, due to the coupling between $\hat x_t$, $\hat v_t$, and g_t, we cannot directly take expectations, so we use the estimations of Lemma 2.11. First we sum the result of Lemma 2.11 which gives

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}_t\left[\alpha_t\langle x_t-\hat x_t, g_t\rangle\right] &\ge \sum_{t=1}^{T}\mathbb{E}_t(\bar\rho-\hat\rho)\alpha_t\|x_t-\hat x_t\|^2_{\hat v_t^{1/2}}-(\alpha_0)\sqrt{d}\hat D G\\ &-\sum_{t=1}^{T}\frac{\bar\rho-\hat\rho}{4\bar\rho}\mathbb{E}_t\|\hat x_t-\hat x_{t-1}\|^2_{\hat v_{t-1}^{1/2}}-\sum_{t=1}^{T}\frac{\alpha_{t-1}}{2}\mathbb{E}_t\|m_{t-1}\|^2_{\hat v_{t-1}^{-1/2}}-\sum_{t=1}^{T}\left(\frac{1}{2}+\frac{\bar\rho}{\bar\rho-\hat\rho}\right)\frac{\alpha_{t-1}^2}{\sqrt{\epsilon}}\mathbb{E}_t\|g_t\|^2.\end{aligned}$$

We use here the assignments used for convenience: $\alpha_0 = 0$ and $\hat x_0 = \hat x_1$ and recall that $m_0 = 0$. We plug this estimation after taking full expectation in (2.47) and use $\hat v_{t-1}^{1/2} \le \hat v_t^{1/2}$ to obtain

$$\begin{aligned}\bar\rho(\bar\rho-\hat\rho)\sum_{t=1}^{T}\alpha_t\mathbb{E}\|x_t-\hat x_t\|^2_{\hat v_t^{1/2}} &\le \frac{2\beta_1\bar\rho\alpha}{(1-\beta_1)}\hat D\sqrt{d}G+\varphi^1_{1/\bar\rho}(x_1)-\mathbb{E}\varphi^{T+1}_{1/\bar\rho}(x_{T+1})+\bar\rho\hat D^2\sum_{i=1}^{d}\mathbb{E}(\hat v_{T+1}^{(i)})^{1/2}\\ &+\sum_{t=1}^{T}\left(1+\frac{G}{\sqrt{\epsilon}}+\frac{\beta_1}{1-\beta_1}+\frac{\bar\rho}{\bar\rho-\hat\rho}\frac{\beta_1^2}{(1-\beta_1)^2}\right)\bar\rho\alpha_t^2\mathbb{E}\|m_t\|^2_{\hat v_t^{-1/2}}+\sum_{t=1}^{T}\left(\frac{1}{2\sqrt{\epsilon}}+\frac{\bar\rho}{(\bar\rho-\hat\rho)\sqrt{\epsilon}}\right)\bar\rho\alpha_{t-1}^2\mathbb{E}\|g_t\|^2.\end{aligned}$$

The only quantities left to estimate are $\sum_{t=1}^{T}\alpha_{t-1}^2\|g_t\|^2$ and $\sum_{t=1}^{T}\alpha_t^2\|m_t\|^2_{\hat v_t^{-1/2}}$. Using Lemma 2.12 and $\alpha_0 = 0$ shows that both these quantities are bounded by $\mathcal{O}(\log T)$:

$$\sum_{t=1}^{T}\alpha_t^2\|m_t\|^2_{\hat v_t^{1/2}} \le \frac{1-\beta_1}{\sqrt{(1-\beta_2)(1-\gamma)}}dG(1+\log T).$$

$$\sum_{t=1}^{T}\alpha_{t-1}^2\|g_t\|^2=\sum_{t=2}^{T}\alpha_{t-1}^2\|g_t\|^2 \le dG^2(1+\log T).$$

The proof then follows by using (2.8), $f^\star \le f(x), \forall x \in \mathcal{X}$, picking $\bar\rho = 2\hat\rho$, using $\alpha_t \ge \alpha_T$, and in the end dividing both sides by $T\alpha_T$. ■

2.6 Bibliographic note

Lemma 2.1 and Lemma 2.10 are due to Yura Malitsky.

3 Smoothing and stochastic algorithms

In the last chapter, we focused on generic nonsmooth optimization template, for which $\mathcal{O}(1/\sqrt{K})$ is the optimal rate [NY83]. To enhance this result, we have to consider structured nonsmooth problems. In this chapter, we focus on convex problems with linear constraints.

To process the constraints randomly, our approach will be to use SGD and CD. A natural way to incorporate these algorithms for solving nonsmooth problems is via Nesterov's smoothing. Smoothing helps us formulate a sequence of smooth problems to solve the original nonsmooth problem. The idea in this chapter is to use SGD and accelerated proximal CD for solving the smoothed problems, along with a homotopy strategy to change the smoothness parameter to converge to the original problem.

Our SGD approach solves problems with infinitely many linear constraints with $\tilde{\mathcal{O}}(1/\sqrt{K})$ rate and our CD approach solves problems with finitely many linear constraints with $\mathcal{O}(1/K)$ rate, both of which are optimal. These results are among the first rate guarantees for the corresponding templates, with important advantages compared to contemporary developments.

This chapter is based on the joint works with Olivier Fercoq, Quoc Tran-Dinh, Ion Necoara and Volkan Cevher [ADFC17, FANC19].

3.1 Introduction

In this chapter, we focus on the following problem:

$$P_\star = \min_{x \in \mathbb{R}^d} \left\{ P(x) = f(Ax) + g(x) + h(x) \right\}, \tag{3.1}$$

where $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, and $h : \mathbb{R}^d \to \mathbb{R}$ are proper, l.s.c. and convex functions, $A \in \mathbb{R}^{n \times d}$ is a given matrix and $y_\star$ is a solution to the dual problem of (3.1). We particularly focus on linearly constrained problems when $f_i(A_i x) = \delta_{b_i}(A_i x)$ for $i \in [n]$:

$$\min_{x \in \mathbb{R}^d} g(x) + h(x), \quad \text{s.t. } A_i x \in b_i, \quad i \in [n], \tag{3.2}$$

where $A_i \in \mathbb{R}^{n_i \times d}$, $b_i \subseteq \mathbb{R}^{n_i}$. This is a key structure that we can exploit with SGD and CD methods to develop efficient algorithms. As we shall see, with SGD, we can even handle the case when n is not finite. In this chapter, we only focus on this case, however as we will remark later as well, it is straightforward to extend the developments in this chapter when f is not an indicator but a Lipschitz continuous function.

Particular instances of (3.2) involve primal support vector machines (SVM) classification and sparse regression which are central in machine learning [SSSSC11, GG09] Due to the huge volume of data that is used for these applications, storing or processing this data at once is often not preferred. Therefore, using these data points one by one or in mini batches in learning algorithms is becoming more important. One direction that the literature focused so far is solving unconstrained formulations of these problems, successes of which are amenable to regularization parameters that needs to be tuned. With stochastic methods capable of solving (3.4) directly, we present a parameter-free approach for solving these problems.

Approach. Our approach in this chapter is using Nesterov's smoothing to formulate

$$\min_{x \in \mathbb{R}^d} P_\beta(x) = g(x) + h(x) + f_\beta(Ax), \tag{3.3}$$

where $f_\beta(Ax) = \max_{y \in \mathbb{R}^n} \langle Ax, y \rangle - f^*(y) - \frac{\beta}{2}\|y - \dot{y}\|^2$, with center point $\dot{y}$. This problem is a smooth estimate of (3.2) and obviously its solution set is different. There are several approaches to pick β to ensure an approximate solution to (3.3) will give an approximate solution to (3.2). This choice may require the knowledge of the desired accuracy, number of maximum iterations or the diameters of the primal and/or dual domains as in [Nes05]. In order to make this choice flexible and our method applicable to constrained problems, we employ a homotopy strategy as in [TDFC18], to gradually decrease β to 0 and obtain approximate solutions to (3.2).

3.1.1 Contributions.

In Section 3.2, we consider the setting when n is not finite in (3.2), by using SGD.

▷ We provide a simple SGD-type algorithm without expensive projections.

▷ We prove $\tilde{\mathcal{O}}(1/\sqrt{k})$ convergence rate for general convex objectives and $\tilde{\mathcal{O}}(1/k)$ rate for restricted strongly convex objectives.

▷ We include generalizations of our framework for composite optimization with general nonsmooth Lispchitz continuous functions in addition to indicator functions.

▷ We provide numerical evidence and verify our theoretical results in practice.

In the second part, we show that by considering finite n in Equation (3.2) and using accelerated CD methods, we can improve the convergence rate.

▷ We propose a method combining accelerated proximal CD and smoothing to obtain $\mathcal{O}(1/K)$

rate which is optimal for the deterministic problem (3.1).

▷ We show an efficient implementation to take full advantage of CD-based method.

Brief explanation for alternatives in the literature. The only work we are aware for solving problems with infinitely many constraints was [PN17] that employed alternating projections on the sets $\{x : A_i x \in b_i\}$. Our algorithm is more general as these projections can be prohibitive depending on dimensions of A_i.

Prior to the developments in this chapter on CD, we were not aware of a CD method for this template with rate guarantees. After the publication of our results, we learned about [GXZ19] that studied randomized linearized ADMM. The advantage of our approach is using coordinate-wise Lipschitz constants of f and norms of A_i, whereas [GXZ19] requires global constants. These are among the most important properties to make CD based methods practical.

We provide more comparisons in Section 3.2.5.

3.2 Smoothing with SGD for infinitely many linear constraints

We formalize the problem with infinitely many constraints. There are several interpretations as considered in [NRP19]. Among these, we use *almost-sure constrained* approach. Let us denote ξ as the random variable with distribution $\mathbb{P}$ and with Assumption 3.1, define the problem

$$\min_{x\in\mathbb{R}^d}\{P(x) := g(x) + h(x)\} \quad \text{s.t.} \quad \mathbb{P}\left((A(\xi)x \in b(\xi)\right) = 1. \tag{3.4}$$

In words, we seek to satisfy the linear inclusion constraints in (3.4) *almost surely*. This change is what sets (3.4) apart from the standard stochastic convex optimization. We assume that $A(\xi)$ is a $n \times d$ matrix-valued random variable and $b(\xi) \subseteq \mathbb{R}^n$ is random nonempty, closed, convex. For the special case when $A(\xi)$ is an identity matrix, (3.4) is similar to the problem considered in [NRP19] with a constraint set defined as the the intersection of a infinitely many sets.

As mentioned before, the only method that we were aware for problems with infinitely many constraints was using possibly expensive projections on $\{x : A_i x \in b_i\}$ [PN17]. We take a different approach and use Nesterov's smoothing [Nes05]. In doing so, we avoid potentially expensive projections and only use much simpler projections to the set $b(\xi)$. We make use of the stochastic gradients of $h(\cdot)$, proximal operators of the nonsmooth component $g(\cdot)$.

3.2.1 Preliminaries

Space of random variables. In our formulation, we have infinite dimensional dual variables. We consider random variables of $\mathbb{R}^n$ belonging to the space

$$\mathcal{Y} = \{(y(\xi))_\xi \, : \, \mathbb{E}[\|y(\xi)\|^2] < +\infty\}$$

$\mathcal{Y}$ is a Hilbert space and its norm is $\|y\| = \sqrt{\mathbb{E}[\|y(\xi)\|^2]}$. We denote by μ the probability measure of the random variable ξ, endowed with the scalar product

$$\langle y, z\rangle = \mathbb{E}[y(\xi)^\top z(\xi)] = \int y(\xi)^\top z(\xi)\mu(d\xi).$$

Duality. We define the stochastic function

$$f_\xi(A(\xi)x) = \delta_{b(\xi)}(A(\xi)x). \tag{3.5}$$

As shown in [NRP19, Lemma 1] using simple arguments, (3.4) can be equivalently written as

$$\min_{x\in\mathbb{R}^d} \mathbb{E}[h_\xi(x)] + g(x) + \mathbb{E}[f_\xi(A(\xi)x)] =: P(x) + f(Ax), \tag{3.6}$$

where $A : \mathbb{R}^d \to \mathcal{Y}$ is defined as the linear operator such that $(Ax)(\xi) = A(\xi)x$ for all x and $f : \mathcal{Y} \to \mathbb{R}\cup\{+\infty\}$ is defined by

$$f(z) = \int \delta_{b(\xi)}(z(\xi))\mu(d\xi).$$

We will assume that

$$\|A\|_{2,\infty} = \sup_\xi \|A(\xi)\| < +\infty, \tag{3.7}$$

so that A is in fact continuous. Note that assuming a uniform bound on $\|A(\xi)\|$ is not restrictive since we can replace $A(\xi)x \in b(\xi)$ by

$$A'(\xi)x = \frac{A(\xi)x}{\|A(\xi)\|} \in b'(\xi) = \frac{b(\xi)}{\|A(\xi)\|},$$

without changing the set of vectors x satisfying the constraint, and projecting onto $b'(\xi)$ is as easy as projecting onto $b(\xi)$. We define the Lagrangian $\mathcal{L} : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}\cup\{+\infty\}$ as

$$\mathcal{L}(x, y) = P(x) + \int \langle A(\xi)x, y(\xi)\rangle - \mathrm{supp}_{b(\xi)}(y(\xi))\mu(d\xi).$$

We assume strong duality holds. For this Slater's condition is a sufficient condition [BC11]. We also assume existence of a primal solution. In the context of duality in Hilbert spaces, Slater's condition refers to $0 \in \mathrm{sri}(\mathrm{dom}\, G - A(\mathrm{dom}\, P))$, where $\mathrm{sri}(\cdot)$ refers to the strong relative interior.

Optimality conditions. We denote by $(x_\star, y_\star) \in \mathbb{R}^d \times \mathcal{Y}$ a saddle point of $\mathcal{L}(x, y)$. For the constrained problem, we say that x is an ε-solution if it satisfies

$$|P(x) - P(x_\star)| \le \varepsilon, \quad \sqrt{\mathbb{E}\left[\mathrm{dist}(A(\xi)x, b(\xi))^2\right]} \le \varepsilon. \tag{3.8}$$

We continue with the assumptions regarding (3.4).

Assumption 3.1.

▷ A primal solution exists and strong duality holds.

▷ h is L_h-smooth and g is proximable. h_ξ, g are proper, l.s.c. convex.

▷ We have $h = \mathbb{E}[h_\xi(x)]$ and $\sigma_h < +\infty$ such that $\mathbb{E}\|\nabla h_\xi(x) - \nabla h(x)\|^2 \le \sigma_h^2$.

▷ $A(\xi)$ is a $n \times d$ matrix-valued r.v., $b(\xi) \subseteq \mathbb{R}^n$ is a random nonempty, closed, convex set.

3.2.2 Algorithm

We derive the main step of our algorithm from smoothing framework. The problem in (3.4) is nonsmooth both due to $g(x)$ and the constraints encoded in $f_\xi(A(\xi)x)$ as in (3.5). We keep $g(x)$ intact since it is assumed to be proximable, and smooth f to get smoothed version of (3.6)

$$P_\beta(x) = \mathbb{E}\left[h_\xi(x)\right] + g(x) + \mathbb{E}\left[f_\beta(A(\xi)x, \xi)\right], \tag{3.9}$$

where $f_\beta(A(\xi)x, \xi) = \frac{1}{2\beta}\operatorname{dist}(A(\xi)x, b(\xi))^2$, in view of (3.3) and $\dot{y} = 0$. Note that $P_\beta(x)$ is $L_h + \frac{\|A\|_{2,2}^2}{\beta}$-smooth where

$$\|A\|_{2,2} = \sup_{x \neq 0} \frac{\sqrt{\mathbb{E}[\|A(\xi)x\|^2]}}{\|x\|} \le \|A\|_{2,\infty},$$

with $\|A\|_{2,\infty}$ being defined in (3.7). Note that (3.9) can also be viewed as a quadratic penalty (QP) formulation. We also define the smoothed gap function

$$S_\beta(x) = P_\beta(x) - P(x_\star) = P(x) - P(x_\star) + \frac{1}{2\beta}\int \operatorname{dist}(A(\xi)x, b(\xi))^2 \mu(d\xi). \tag{3.10}$$

The main idea of our method is to apply stochastic proximal gradient (SPG) [RVV20] iterations to (3.9) by using homotopy on the smoothness parameter β. Our algorithm has a double loop structure where for each value of β, we solve the problem (3.9) with SPG upto some accuracy. This strategy is similar to classical inexact quadratic penalty (QP) methods. In stark contrast to this approach, Algorithm 3.1 has explicit number of iterations for the inner loop which is determined by theoretical analysis, avoiding difficult-to-check stopping criteria for the inner loop in inexact QP methods. Decreasing β to 0 according to update rules from our analysis ensures the convergence to the original problem (3.4) rather than the smoothed problem (3.9).

In Algorithm 3.1, we present our method. Case 1 refers to parameters for general convex case and Case 2 refers to restricted strongly convex case.

It may look unusual at first glance that in the restricted strongly convex case, the step size α_s is decreasing faster. This is due to restricted strong convexity allowing us to decrease faster the smoothness parameter β_s, and the step size is driven by the smoothness of the approximation.

Algorithm 3.1 Stochastic approximation method for almost surely constrained problems (SASC, pronounced as "sassy")

```
$x_0^0 \in \mathbb{R}^d$
$\alpha_0 \le 3/(4L(\nabla f))$, and $\omega > 1$
Case 1: $m_0 \in \mathbb{N}_*$;    Case 2: $m_0 \ge \frac{\omega}{\mu\alpha_0}$.
for $s \in \mathbb{N}$ do
    $m_s = \lfloor m_0 \omega^s \rfloor$, and $\beta_s = 4\alpha_s \|A\|_{2,\infty}^2$
    Case 1: $\alpha_s = \alpha_0 \omega^{-s/2}$;    Case 2: $\alpha_s = \alpha_0 \omega^{-s}$.
    for $k \in \{0, \dots, m_s - 1\}$ do
        Draw $\xi = \xi_{k+1}^s$, and define $z = A(\xi) x_k^s$.
        $D(x_k^s, \xi) := \nabla h_\xi(x_k^s) + A(\xi)^\top \nabla_z f_{\beta_s}(A(\xi) x_k^s, \xi)$
        $x_{k+1}^s = \mathrm{prox}_{\alpha_s g}\left(x_k^s - \alpha_s D(x_k^s, \xi)\right)$
    end for
    $\bar{x}^s = \frac{1}{m_s} \sum_{k=1}^{m_s} x_k^s$
    Case 1: $x_0^{s+1} = x_{m_s}^s$;    Case 2: $x_0^{s+1} = \bar{x}^s$.
end for
```

3.2.3 Convergence

We present a key lemma which is instrumental in our analysis. It serves as a bridge between β_s with the smoothed gap in (3.10) and the optimality results in the sense of (3.8). This lemma can be seen as an extension of [TDFC18, Lemma 1] with infinite dimensional dual variables.

Lemma 3.1. *Let $(x_\star, y_\star)$ be a saddle point of*

$$\min_{x \in \mathbb{R}^d} \max_{y \in \mathcal{Y}} \mathcal{L}(x, y).$$

Then, the following hold:

$$\begin{aligned}
&S_\beta(x) \ge -\frac{\beta}{2}\|y_\star\|^2, \\
&P(x) - P(x_\star) \ge -\frac{1}{4\beta}\int \mathrm{dist}\,(A(\xi)x, b(\xi))^2 \mu(d\xi) - \beta\|y_\star\|^2, \\
&P(x) - P(x_\star) \le S_\beta(x), \\
&\int \mathrm{dist}\,(A(\xi)x, b(\xi))^2 \mu(d\xi) \le 4\beta^2\|y_\star\|^2 + 4\beta S_\beta(x).
\end{aligned}$$

The simple message of Lemma 3.1 is that if an algorithm decreases the smoothed gap function $S_\beta(x)$ and β simultaneously, then it obtains approximate solutions to (3.4) in the sense of (3.8), *i.e.* it decreases feasibility and objective suboptimality.

Recall from (3.9), $f_\beta(A(\xi)x, \xi) = \frac{1}{2\beta}\mathrm{dist}(A(\xi)x, b(\xi))^2$. The main technical challenge of applying SPG to (3.9) with homotopy stems from the stochastic term due to constraints:

$$\mathbb{E}[f_\beta(A(\xi)x, \xi)], \tag{3.11}$$

This term is in a suitable form to apply SPG, however its variance bound and Lipschitz constant of its gradient becomes worse and worse as $\beta_k \to 0$. A naive solution for this problem would be to decrease β_k slowly, so that these bounds will increase slowly and they can be dominated by the step size. Due to Lemma 3.1, the rate of decrease of β_k directly determines the convergence rate, so a slowly decaying β_k would result in slow convergence for the method. Our proof technique carefully balances the rate of β_k and the additional error terms due to using stochastic gradients of (3.11), so that the optimal rate of SPG is retained even with constraints.

We present the main theorems in the following two sections. The main proof strategy in Theorem 3.2 and Theorem 3.3 is to analyze the convergence of $S_\beta(x)$ and β_k and use Lemma 3.1 to translate the rates to objective residual and feasibility measures.

Convergence for General Convex Objectives.

Theorem 3.2. *Let Assumption 3.1 hold. Let* $M_s = \sum_{l=0}^{s} m_l$ *and pick* $\omega, \alpha_0, m_s, \beta_s$ *as Case 1 in Algorithm 3.1. Then, for all s,*

$$\mathbb{E}[P(\bar{x}^s) - P(x_\star)] \leq \frac{C_1}{\sqrt{M_s}}\left[C_2 + \frac{\log(M_s/m_0)}{\log(\omega)}C_3\right]$$

$$\mathbb{E}[P(\bar{x}^s) - P(x_\star)] \geq -\frac{2C_4}{\sqrt{M_s}}\|y_\star\|^2 - \frac{C_1}{\sqrt{M_s}}\left[C_2 + \frac{\log(M_s/m_0)}{\log(\omega)}C_3\right]$$

$$\sqrt{\mathbb{E}\left[\mathrm{dist}(A(\xi)\bar{x}^s, b(\xi))^2\right]} \leq \frac{1}{\sqrt{M_s}}\left[2C_4\|y_\star\| + 2\sqrt{C_1C_4}\sqrt{C_2 + \frac{\log(M_s/m_0)}{\log(\omega)}C_3}\right]$$

where $C_1 = \frac{\sqrt{m_0\omega}}{\alpha_0(m_0-1)\sqrt{\omega-1}}$, $C_2 = \frac{\|x_\star - x_0^0\|^2}{2} + 2\alpha_0 m_0 \sigma_h^2$, $C_3 = 2\alpha_0^2\|A\|_{2,\infty}^2 m_0\|y_\star\|^2 + 2\alpha_0 m_0\sigma_h^2$ *and* $C_4 = 4\alpha_0\sqrt{m_0}\|A\|_{2,\infty}^2\frac{\sqrt{\omega}}{\sqrt{\omega-1}}$.

Note that $\mathcal{O}(1/\sqrt{k})$ rate is optimal for solving (3.4) with SGD [PJ92, AWBR09]. In Theorem 3.2, we see that handling infinitely many constraints via SGD only brings overhead of a log.

Convergence for Restricted Strongly Convex Objectives. Now, we assume $P(x)$ in (3.4) to be restricted strongly convex in addition to Assumption 3.1: $P(x) \geq P(x_\star) + \frac{\mu}{2}\|x - x_\star\|^2$. Requiring restricted strong convexity of $P(x)$ is weaker than strong convexity of $h_\xi(x)$ or $g(x)$, see [NNG18] for more details.

Theorem 3.3. *Let Assumption 3.1 hold and P be* μ*-restricted strongly convex. Let* $M_s = \sum_{l=0}^{s} m_l$ *and pick* $\omega, \alpha_0, m_s, \beta_s$ *as Case 2 in Algorithm 3.1. Then, for all s,*

$$\mathbb{E}[P(\bar{x}^s) - P(x_\star)] \leq \frac{1}{M_s}\left[D_1 + \frac{\log(M_s/m_0)}{\log(\omega)}D_2\right]$$

$$\mathbb{E}[P(\bar{x}^s) - P(x_\star)] \geq -\frac{2D_3}{M_s}\|y_\star\|^2 - \frac{1}{M_s}\left[D_1 + \frac{\log(M_s/m_0)}{\log(\omega)}D_2\right]$$

$$\sqrt{\mathbb{E}\left[\mathrm{dist}(A(\xi)\bar{x}^s, b(\xi))^2\right]} \leq \frac{1}{M_s}\left[2D_3\|y_\star\| + 2\sqrt{D_3}\sqrt{D_1 + \frac{\log(M_s/m_0)}{\log(\omega)}D_2}\right]$$

where $D_1 = \frac{\omega}{\omega-1}\frac{m_0}{\alpha_0(m_0-1)}\frac{1}{2}\left\|x_0^0 - x_\star\right\|^2 + 2\alpha_0 m_0 \frac{\omega}{\omega-1}\sigma_h^2$, $D_2 = \frac{2m_0^2\alpha_0\omega}{(m_0-1)(\omega-1)}\left(\|A\|_{2,\infty}^2\|y_\star\|^2 + \sigma_h^2\right)$, $D_3 = 4\alpha_0 m_0\|A\|_{2,\infty}^2\frac{\omega}{\omega-1}$.

Similar comments to Theorem 3.2 can be made for Theorem 3.3. We have $\mathcal{O}(\log(k)/k)$ rate for both objective residual and feasibility under restricted strong convexity assumption. This rate is optimal up to a logarithmic factor for solving (3.4) even without constraints.

3.2.4 Extension

We can extend our method for solving problems considered in [OG12], which corresponds to (3.6) when $f_\xi(\cdot)$ is not an indicator function, but is L_f-Lipschitz continuous. This assumption is equivalent to $\mathrm{dom}(f^*)$ being bounded [BC11]. This special case with $g(x) = 0$ is studied in [OG12] with the specific assumptions in this section. Inspired by [Nes05], it has been shown in [OG12], that one has the following bound for the smooth approximation of $f_\xi(\cdot)$

$$\mathbb{E}[f_\xi(A(\xi)x)] \le \mathbb{E}[f_\beta(A(\xi)x,\xi)] + \frac{\beta}{2}L_f^2. \tag{3.12}$$

We can couple our main results with (3.12) to recover the guarantees of [OG12] with the addition of the nonsmooth proximable term $h(x)$. Essentially, after proving the bound for S_β, which is the first step in our proofs, we can directly use (3.12). We can also consider

$$\min_{x\in\mathbb{R}^d} \mathbb{E}\left[h_\xi(x) + f_{1,\xi}(A_1(\xi)x)\right] + g(x), \quad \text{s.t.} \quad \mathbb{P}\left(A_2(\xi)x \in b(\xi)\right) = 1,$$

where $f_{1,\xi}$ is Lipschitz continuous and we have the same assumptions as (3.4) for the constraints. As argued above, we can use our results from Section 3.2.3 to solve this template.

3.2.5 Related Works

Even though SGD is well studied, it applies to unconstrained problems [NJLS09, MB11, PJ92]. With simple constraints admitting efficient projection, and without almost sure constraints, projected SGD can be used [NJLS09]. In the case where $g(x)$ in (3.4) is a nonsmooth proximable function [RVV20] studied the convergence of stochastic proximal gradient (SPG) method which uses stochastic gradients of $h_\xi(x)$ with proximal operator of $g(x)$. This method generalize projected SGD, however, it still cannot process infinitely many constraints since it is not clear how to project onto the stochastic set in (3.4).

Methods based on alternating projections solve the following template

$$\min_{x\in\mathbb{R}^d} \mathbb{E}\left[h_\xi(x)\right] : \ x \in \mathcal{B}(:= \cap_{\xi\in\Omega}\mathcal{B}(\xi)). \tag{3.13}$$

Here, the feasible set $\mathcal{B}$ consists of the intersection of a possibly infinite number of convex sets. When $h_\xi(x) = 0$ this corresponds to the convex feasibility problem is studied in [NRP19]. For this setting, the authors combine the smoothing technique with minibatch SGD, leading to an alternating projections algorithm with linear convergence.

The most related to our work is [PN17] where the authors apply a proximal point type algorithm with alternating projections. Main idea behind [PN17] is to apply smoothing to $h_\xi(x)$ and apply stochastic gradient steps to the smoothed function, which corresponds to a stochastic proximal point type of update, combined with alternating projection steps. The authors show $\mathcal{O}(1/\sqrt{k})$ rate for general convex objectives and $\mathcal{O}(1/k)$ for smooth and strongly convex objectives. Smoothness requirement of [PN17] in the strongly convex case renders their results not applicable to our composite setting in (3.4). In addition, strong convexity is stronger than our assumption of restricted strong convexity. Lastly, [PN17] projects to $B(\xi)$ at each iteration, whereas we only project to $b(\xi)$. Unless $A(\xi)$ and $b(\xi)$ are of very small dimension, projection to $B(\xi)$ can be prohibitive due to solving a linear system at each iteration.

Stochastic forward-backward algorithms can also solve (3.4). However, the papers introducing those very general algorithms focused on proving convergence and did not present convergence rates [Bia15, BHS19, Sal18]. There are some other works [WCLG15, MYJ13, YNW17] that focus on (3.13) with finite number of constraints, which is more restricted than our setting.

When the number of constraints in (3.4) is finite and the objective is deterministic, Nesterov's smoothing framework is studied in [TDFC18, VNFC17, TDAFC19] with accelerated gradient methods. These methods obtain $\mathcal{O}(1/k)$ ($\mathcal{O}(1/k^2)$) rate when the number of constraints is finite and $h(x)$ is a (strongly) convex function whose gradient ∇h can be computed.

Another related work is [OG12] where the authors apply Nesterov's smoothing for Lipschitz f_ξ. Note that in our main template (3.4), $f_\xi(\cdot) = \delta_{b(\xi)}(\cdot)$, which is not Lipschitz continuous.

3.2.6 Numerical Experiments

Sparse regression with basis pursuit on synthetic data

We solve basis pursuit (BP) problem, widely used in ML and signal processing [Don06, AKSV18]:

$$\begin{aligned} &\min_{x\in\mathbb{R}^d} \|x\|_1 \\ &\text{st: } a^\top x = b, a.s. \end{aligned} \tag{3.14}$$

where $a \in \mathbb{R}^d$, $b \in \mathbb{R}$. We consider the setting where the measurements arrive in a streaming fashion, similar to [GG09]. For generating the data, we defined Σ as the matrix such that $\Sigma_{i,j} = \rho^{|i-j|}$ with $\rho = 0.9$. We generated random vector $x^* \in \mathbb{R}^d$, $d = 100$ with 10 nonzero coefficients and independent $\mathcal{N}(0,\Sigma)$ random variables a_i which are then centered and normalized. We define $b_i = a_i^\top x^*$. Because of the centering, there are multiple solutions to the infinite system $a^\top x = b$ a.s., and we wish to recover x^* as the solution of the basis pursuit problem (3.14).

We compare SASC (Algorithm 3.1), SGD [NJLS09] and SPP [PN17]. We manually tuned the step

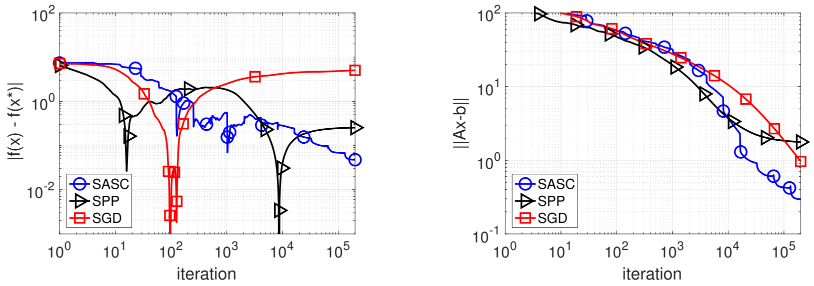

Figure 3.1 – Performance of SGD, SPP and SASC on synthetic basis pursuit problem.

sizes for the methods and included the best obtained results. Since the BP problem does not possess (restricted) strong convexity, we use the parameters from Case 1 in SASC and a fixed step size μ for SPP which is used for the analysis in [PN17, Corollary 6]. We used the parameters $\mu = 10^{-5}$ for SPP, $m_0 = 2$, $\omega = 2$, $\alpha_0 = 10^{-2}\|a_1 b_1\|_\infty$, where a_1 is the first measurement and b_1 is the corresponding result. We take $n = 10^5$ and make two passes over the data.

Figure 3.1 shows the results. We observe that SASC exhibits a $\tilde{O}(1/\sqrt{k})$ convergence in feasibility and objective suboptimality. The stair case shape of the curves comes from the double-loop nature of the method. SPP can also solve this problem since the projection onto a hyperplane is easy to do when the constraints are processed one by one. We see in Figure 3.1 that SPP is initially almost as fast as SASC, however, it stagnates once it reaches the pre-determined accuracy determined by the fixed step size μ. We also tried running SGD on $\min_x \frac{1}{2}\mathbb{E}\left\|a^\top x - b\right\|_2^2$ but this leads to non-sparse solutions, therefore SGD converges to another solution.

It is common in stochastic optimization to use mini-batches to parallelize and speed up computations. SPP projects onto linear constraints each iteration. When the data is processed in mini-batches, this requires solving linear systems with sizes depending on mini-batches. On the other hand, SASC handles mini-batches without any overhead.

Portfolio optimization

In this section, we consider Markowitz portfolio optimization with the task of maximizing the expected return given a maximum bound on the variance [AAEF07].

$$\begin{aligned} \min_{x \in \mathbb{R}^d} -\langle a_{avg}, x\rangle : \sum_{i=1}^{d} x_i = 1 & \\ |\langle a_i - a_{avg}, x\rangle| \leq \epsilon, \forall i \in [1, n], & \end{aligned} \tag{3.15}$$

where short positions are allowed and $a_{avg} = \mathbb{E}[a_i]$ is assumed to be known. This problem fits to our template (3.4), with a deterministic objective function, n linear constraints and one indicator function for enforcing $\sum_{i=1}^{d} x_i = 1$ constraint.

We implement SASC and SPP from [PN17]. Since the structure of (3.15) does not have restricted strong convexity, we apply the general convex version of SPP, which sets a smoothness

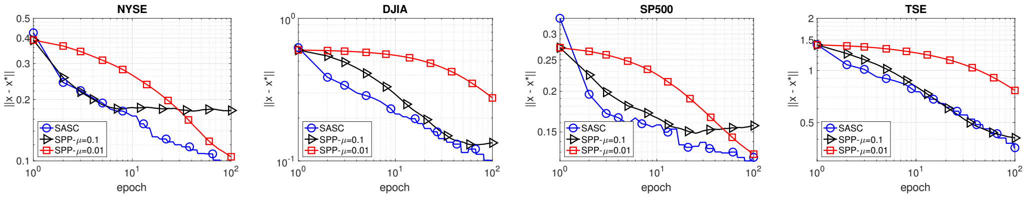

Figure 3.2 – Performance of SASC and SPP on portfolio optimization for four different datasets

parameter μ depending on the final accuracy. We run SPP with two different μ values 10^{-1} and 10^{-2}. We run SASC with the parameters $\alpha_0 = 1$, $\omega = 1.2$, $m_0 = 2$ and Case 1 in Algorithm 3.1. We use NYSE ($d = 36, n = 5651$), DJIA ($d = 30, n = 507$), SP500 ($d = 25, n = 1276$) and TSE ($d = 88, n = 1258$) where d corresponds to the number of stocks and n corresponds to the number of days for which the data is collected and we set ϵ in (3.15) to be 0.2. These datasets are also used in [BEYG04]. We compute the ground truth using cvx [GB14] and plotted the distance of the iterates of the algorithms to the solution $\|x - x^\star\|$ in Figure 3.2.

We can observe the behaviour of SPP from Figure 3.2 for different step size values μ. Larger μ causes a fast decrease in the beginning, however, it also affects the accuracy that the algorithm is going to reach. Therefore, large μ has the problem of stagnating at a low accuracy. Smaller μ causes SPP to reach to higher accuracies at the expense of slower initial behaviour. SASC has a steady behaviour and since it does not have a parameter depending on the final accuracy. It removes the necessity of tuning μ in SPP, as we can observe the steady decrease of SASC throughout, beginning from the initial stage of the algorithm.

Primal support vector machines without regularization parameter

In this section, we consider the classical setting of binary classification, with a small twist. For the standard setting, given a training set $\{a_1, a_2, \ldots, a_n\}$ and labels $\{b_1, b_2, \ldots, b_n\}$, where $a_i \in \mathbb{R}^p, \forall i$ and $b_i \in [-1, +1]$ the aim is to train a model that will classify the correct labels for the unseen examples. Primal hard margin SVM problem is

$$\min_{x \in \mathbb{R}^d} \frac{1}{2}\|x\|^2 : b_i \langle a_i, x \rangle \geq 1, \forall i. \tag{3.16}$$

Since this problem does not have a solution unless the data is linearly separable, the standard way is to relax the constraints, and solve the soft margin SVM problem with hinge loss instead:

$$\min_{x \in \mathbb{R}^d} \frac{1}{2}\|x\|^2 + C \sum_{i=1}^{n} \max\{0, 1 - b_i \langle a_i, x \rangle\}, \tag{3.17}$$

where C has the role of a regularization parameter to be tuned. The choice for C has a drastic effect on the performance of the classifier as also been studied in the literature [HRTZ04]. It is known that poor choices of C may lead to poor classification models.

We have a radically different approach for the SVM problem. Since the original formulation (3.16) fits to our template (3.4), we directly apply SASC. Even though the hard margin SVM problem does not necessarily have solution, applying SASC to (3.16) corresponds to solving a sequence of soft margin SVM problems with squared hinge loss, by changing regularization parameters. The advantage of such an approach will be that there will be no necessity for a regularization parameter C since this parameter will correspond to $\frac{1}{\beta}$ in our case where β is the smoothness parameter, for which we have theoretical guideline from our analysis.

We compare SASC with Pegasos algorithm [SSSSC11] which solves (3.17) by applying SGD. Since the selection of the regularization parameter C effects the performance of the model, we use 3 different values for the λ, namely $\{\lambda_1, \lambda_2, \lambda_3\} = \{10^{-3}/n, 1/n, 10^3/n\}$. We use datasets from libsvm database [CL11a]: `kdd2010 raw version (bridge to algebra)` with 19,264,997 training examples, 748,401 testing examples and 1,163,024 features, `rcv1.binary` with 20,242 training examples, 677,399 testing examples and 47,236 features. For the last dataset, `news20.binary`, since there was not a dedicated testing dataset, we randomly split examples for training and testing with 17.996 training examples, 2,000 testing examples and 1,355,191 features. For SASC, we use $\alpha_0 = 1/2$, $\omega = 2$ in all experiments and use the parameter choices in Case 2 in Algorithm 3.1 due to strong convexity in the objective. We computed the test errors for one pass over the data and compile the results in Figure 3.3.

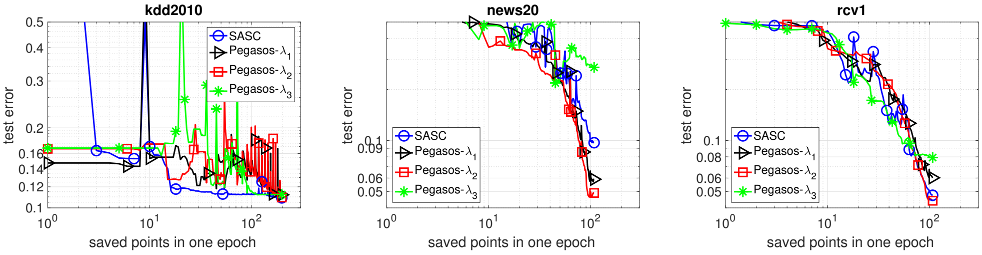

Figure 3.3 – Performance of SASC and Pegasos on SVM for three different datasets.

SASC seems to be comparable to Pegasos for different regularization parameters. As can be seen in Figure 3.3, Pegasos performs well for good selection of the regularization parameter. However, when the parameter is selected incorrectly, it might stagnate at a high test error which can be observed in the plots. On the other hand, SASC gets comparable, if not better, performance without the need to tune regularization parameter.

3.3 Smoothing with accelerated CD for linear constraints

In this section, we solve the problem in (3.1). As before, we particularize our results for the linearly constrained case, this time with finitely many constraints. When f is instead L_f-Lipschitz continuous, it is straightforward to extend our results as explained in Section 6.4.

Our algorithmic strategy in Algorithm 3.2 is to use Nesterov's smoothing along with accelerated proximal CD method of [FR15]. Since we no longer focus on stochastic optimization, we aim to get $\mathcal{O}(1/k)$ rate of convergence, which requires using acceleration. For CD setup, we introduce the following notation.

Notation. Let us decompose the variable vector x into m-blocks where each block is denoted by x_i and has the size $d_i \geq 1$ with $\sum_{i=1}^{m} d_i = d$. We decompose the identity matrix $\mathbb{I}_d$ of $\mathbb{R}^d$ into m block as $\mathbb{I}_d = [U_1, U_2, \cdots, U_m]$, where $U_i \in \mathbb{R}^{d \times d_i}$ has d_i unit vectors in its columns. In this case, any vector $x \in \mathbb{R}^d$ can be written as $x = \sum_{i=1}^{m} U_i x_i$, and each block becomes $x_i = U_i^\top x$ for $i \in [m]$. We define the partial gradients as $\nabla_i h(x) = U_i^\top \nabla h(x)$ for $i \in [m]$. We define the weighted norms: $\|x_i\|_{(i)}^2 = \langle H_i x_i, x_i \rangle, (\|y_i\|_{(i)}^*)^2 = \langle H_i^{-1} y_i, y_i \rangle$. Here, $H_i \in \mathbb{R}^{p_i \times p_i}$ is a symmetric positive definite matrix, and $L_i \in (0, \infty)$ for $i \in [n]$ and $\alpha > 0$.

At iteration k, we pick index i_k randomly w.p.. $q = (q_1, \ldots, q_m)$. As in [QR16] we use an arbitrary distribution, which may allow designing a good distribution that captures the underlying structure of specific problems. We define the σ algebra $\mathcal{F}_{k+1} = \sigma(i_0, \ldots, i_k)$ and $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot | \mathcal{F}_k]$.

Assumption 3.2. In (3.1), f, g and h are all proper, l.s.c. and convex. Moreover:

(a) The partial derivative $\nabla_i h(\cdot)$ Lipschitz continuous with $\hat{L}_i \in [0, +\infty)$, i.e., $\|\nabla_i h(x + U_i d_i) - \nabla_i h(x)\|_{(i)}^* \leq \hat{L}_i \|d_i\|_{(i)}$ for all $x \in \mathbb{R}^p, d_i \in \mathbb{R}^{p_i}$.

(b) The function g is separable, which has the following form $g(x) = \sum_{i=1}^{n} g_i(x^{(i)})$.

(c) For f, we have $f(Ax) := \delta_{\{b\}}(Ax)$, with $b \in \mathbb{R}^n$.

(d) A primal solution exists and strong duality holds.

In view of (3.3), as we focus on the case $f(\cdot) = \delta_{\{b\}}(\cdot)$, we have

$$\min_x P_\beta(x) = h(x) + g(x) + \frac{1}{2\beta}\|Ax - b\|^2 = g(x) + \psi_\beta(x).$$

It is easy to see that ψ_β is differentiable, and its block partial gradient

$$\nabla_i \psi_\beta(x) = \nabla_i h(x) + \frac{1}{\beta} A_i^\top (Ax - b) =: \nabla_i f(x) + A_i^\top y_\beta^*(Ax) \tag{3.18}$$

is Lipschitz continuous with the constant $L_i(\beta) := \hat{L}_i + \frac{\|A_i\|^2}{\beta}$, where $\hat{L}_i$ is given in Assumption 3.2, and $A_i \in \mathbb{R}^{n \times d_i}$ is the i-th column block of A.

3.3.1 Convergence

Theorem 3.4. *Let $\{\bar{x}^k\}$ be the sequence generated by Algorithm 3.2. In addition, let $\tau_0 := \min\{q_i \mid i \in [n]\} \in (0, 1]$ and $\beta_0 := (1 + \tau_0)\beta_1$ be given parameters. Then, we have*

$$\begin{cases} \mathbb{E}[P(\bar{x}_k) - P(x_\star)] & \leq \frac{C^*(x_0)}{\tau_0(k-1)+1} + \frac{\beta_1\|y_\star - \dot{y}\|^2}{2(\tau_0(k-1)+1)} + \|y_\star\|\mathbb{E}[\|A\bar{x}_k - b\|], \\ \mathbb{E}[\|A\bar{x}_k - b\|] & \leq \frac{\beta_1}{\tau_0(k-1)+1}\left[\|y_\star - \dot{y}\| + \left(\|y_\star - \dot{y}\|^2 + 2\beta_1^{-1}C^*(x_0)\right)^{1/2}\right], \end{cases} \tag{3.19}$$

where $C^*(x^0) := (1 - \tau_0)(P_{\beta_0}(x_0) - P(x_\star)) + \sum_{i=1}^{n} \frac{\tau_0 B_0^{(i)}}{2q_i}\|x_\star^{(i)} - x_0^{(i)}\|_{(i)}^2$. We note that the following lower bound always holds $-\|y_\star\|\mathbb{E}[\|A\bar{x}_k - b\|] \leq \mathbb{E}[P(\bar{x}_k) - P_\star]$, where $y_\star$ is any dual solution.

In particular, when we use uniform distribution, $\tau_0 = q_i = 1/n$, the convergence rate is $\mathcal{O}\left(\frac{n}{k}\right)$.

Algorithm 3.2 SMooth, Accelerate, Randomize The Coordinate Descent (SMART-CD)

```
Require: $\beta_1 > 0$ and $\alpha \in [0,1]$ as two input parameters. Choose $x_0 \in \mathbb{R}^d$.
Set $B_0^{(i)} := \hat{L}_i + \frac{\|A_i\|^2}{\beta_1}$ for $i \in [m]$. Compute $S_\alpha := \sum_{i=1}^m (B_0^{(i)})^\alpha$ and $q_i := \frac{(B_0^{(i)})^\alpha}{S_\alpha}$ for all $i \in [m]$.
Set $\tau_0 := \min\{q_i \mid 1 \le i \le m\} \in (0,1]$ for $i \in [m]$. Set $\bar{x}_0 = \tilde{x}_0 := x_0$.
for $k \leftarrow 0, 1, \cdots, k_{\max}$ do
    $\hat{x}_k := (1-\tau_k)\bar{x}_k + \tau_k\tilde{x}_k$
    $y_k^* := y_{\beta_{k+1}}^*(A\hat{x}_k) = \dot{y} + \beta_{k+1}^{-1}(A\hat{x}_k - b)$.
    Select a block coordinate $i_k \in [m]$ according to the probability distribution $q$.
    Set $\tilde{x}_{k+1} := \tilde{x}_k$, and compute the primal $i_k$-block coordinate:
        $$x_{k+1}^{(i_k)} := \arg\min_{x^{(i_k)} \in \mathbb{R}^{d_{i_k}}} \left\{ \langle \nabla_{i_k} f(\hat{x}_k) + A_{i_k}^\top y_k^*, x^{(i_k)} - \hat{x}_k^{(i_k)} \rangle + g_{i_k}(x^{(i_k)}) + \frac{\tau_k B_k^{(i_k)}}{2\tau_0} \|x^{(i_k)} - \tilde{x}_k^{(i_k)}\|_{(i_k)}^2 \right\}.$$
    $\bar{x}_{k+1} := \hat{x}_k + \frac{\tau_k}{\tau_0}(\tilde{x}_{k+1} - \tilde{x}_k)$.
    $\tau_{k+1} := \frac{\tau_k}{1+\tau_k}$, $\beta_{k+2} := (1-\tau_{k+1})\beta_{k+1}$, and $B_{k+1}^{(i)} := \hat{L}_i + \frac{\|A_i\|^2}{\beta_{k+2}}$ for $i \in [m]$.
end for
```

3.3.2 Efficient implementation

Since Algorithm 3.2 requires full vector updates at each iteration, we exploit the idea in [LS13, FR15] and show that these vectors can be updated in an efficient manner. We next show the equivalence between Algorithms 3.2 and 3.3, with its proof in Section 3.5.3.

Proposition 1. *Let $c_k = \prod_{l=0}^k (1-\tau_l)$, $\hat{z}_k = c_k u_k + \tilde{z}_k$ and $\bar{z}_k = c_{k-1} u_k + \tilde{z}_k$. Then, $\tilde{x}_k = \tilde{z}_k$, $\hat{x}_k = \hat{z}_k$ and $\bar{x}_k = \bar{z}_k$, for all $k \ge 0$, where $\tilde{x}_k$, $\hat{x}_k$, and $\bar{x}_k$ are defined in Algorithm 3.2.*

According to Algorithm 3.3, we never need to form or update full-dimensional vectors. Only times that we need $\hat{x}_k$ are when computing the gradient and the dual variable $y^*_{\beta_{k+1}}$. We present two special cases, which are common in ML, that admits an efficient implementation.

Remark 3.5. We characterize the per-iteration cost in an important case. Let $A, M \in \mathbb{R}^{n\times d}$ and

(a) h has the form $h(x) = \sum_{j=1}^m \varphi_j(e_j^\top Mx)$, where e_j is the j^{th} standard unit vector.

(b) f is separable since $f(Ax) = \delta_{\{b\}}(Ax)$.

We store and maintain the residuals $r_k^{u,h} = Mu_k$, $r_k^{\tilde{z},h} = M\tilde{z}_k$, $r_k^{u,f} = Au_k$, $r_k^{\tilde{z},f} = A\tilde{z}_k$, to have the per-iteration cost as $\mathcal{O}(\max\{|\{j \mid A_{ji} \neq 0\}|, |\{j \mid M_{ji} \neq 0\}|\})$ arithmetic operations. If f is partially separable as in [RT16], then the complexity of each iteration will remain moderate.

Algorithm 3.3 Efficient SMART-CD

```
Require: Choose parameters as Algorithm 3.2. Set $u_0 = \tilde{z}_0 := x_0$
for $k \leftarrow 0, 1, \cdots, k_{\max}$ do
    $y^*_{\beta_{k+1}}(c_k A u^k + A\tilde{z}^k) := \dot{y} + \beta^{-1}_{k+1}(c_k A u^k + A\tilde{z}^k - b)$.
    Select a block coordinate $i_k \in [n]$ according to the probability distribution $q$.
    Let $\nabla^k_i := \nabla_{i_k} h(c_k u^k + \tilde{z}^k) + A^\top_{i_k} y^*_{\beta_{k+1}}(c_k A u^k + A\tilde{z}^k)$. Compute
        $$t^{(i_k)}_{k+1} := \arg\min_{t \in \mathbb{R}^{d_{i_k}}} \left\{ \langle \nabla^{(i)}_k, t \rangle + g_{i_k}(t + \tilde{z}^{(i_k)}_k) + \frac{\tau_k B^{(i_k)}_k}{2\tau_0} \| t \|^2_{(i_k)} \right\}.$$
    $\tilde{z}^{(i_k)}_{k+1} := \tilde{z}^{(i_k)}_k + t^{(i_k)}_{k+1}$.
    $u^{(i_k)}_{k+1} := u^{(i_k)}_k - \frac{1-\tau_k/\tau_0}{c_k} t^{(i_k)}_{k+1}$.
    $\tau_{k+1} := \frac{\tau_k}{1+\tau_k}$, $\beta_{k+2} := (1-\tau_{k+1})\beta_{k+1}$, and $B^{(i)}_{k+1} := \hat{L}_i + \frac{\|A_i\|^2}{\beta_{k+2}}$ for $i \in [m]$.
end for
```

3.3.3 Restarting SMART-CD

Restarting accelerated methods significantly enhances practical performance when the underlying problem admits a (restricted) strong convexity condition. As a result, we describe below how to restart (i.e., the momentum term) in Efficient SMART-CD. If the restart is injected in the k-th iteration, then we restart the algorithm with the following steps:

$$u^{k+1} = 0, \quad r^{u,h}_{k+1} = 0, \quad r^{u,f}_{k+1} = 0, \quad \dot{y} = y^*_{\beta_{k+1}}(c_k r^{u,f}_k + r^{\tilde{z},f}_k), \quad \beta_{k+1} = \beta_1, \quad \tau_{k+1} = \tau_0, \quad c_k = 1.$$

First three steps of the restart procedure is for restarting the primal variable which is classical [OC15]. Restarting $\dot{y}$ is also suggested in [TDFC18]. The cost of this procedure is essentially equal to the cost of one iteration as described in Remark 3.5, therefore even restarting once every epoch will not cause a significant difference in terms of per-iteration cost.

3.3.4 Numerical experiments

A degenerate linear program. We consider the following degenerate LP from [TDFC18]:

$$\begin{cases} \min\limits_{x \in \mathbb{R}^d} & 2x^{(d)} \\ \text{s.t.} & \sum_{k=1}^{d-1} x^{(k)} = 1, \qquad x^{(d)} \geq 0, \\ & x^{(d)} - \sum_{k=1}^{d-1} x^{(k)} = 0, \qquad (2 \leq j \leq l). \end{cases} \tag{3.20}$$

Here, the constraint $x^{(d)} - \sum_{k=1}^{d-1} x^{(k)} = 0$ is repeated l times. This problem satisfies the linear constraint qualification condition, which guarantees the primal-dual optimality. We define $f(x) = 2x^{(d)}$, $\quad g(x) = \delta_{\{x^{(d)} \geq 0\}}(x^{(d)})$, $\quad h(Ax) = \delta_{\{b\}}(Ax)$, where

$$Ax = \left[\sum_{k=1}^{d-1} x^{(k)}, \; x^{(p)} - \sum_{k=1}^{d-1} x^{(k)}, \ldots, \; x^{(d)} - \sum_{k=1}^{d-1} x^{(k)} \right]^\top, \quad b = [1, 0, \ldots, 0]^\top,$$

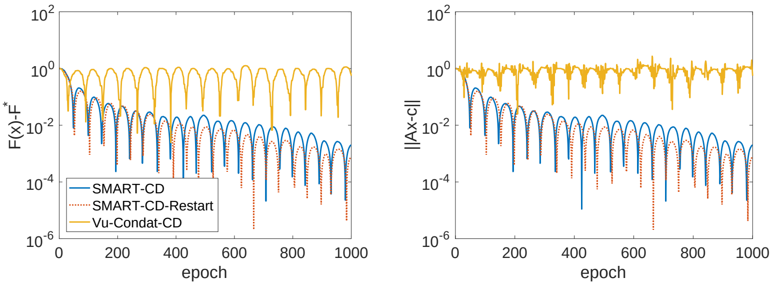

Figure 3.4 – The convergence behavior of 3 algorithms on a degenerate linear program.

we can fit this problem and its dual form into our template (3.1). We select $d = 10$ and $l = 200$. We implement our algorithm, its restarting variant and VC-CD. We use the same mapping to fit the problem into the template of VC-CD and show the results in Figure 3.4.

The explicit solution of the problem is used to generate the plot with suboptimality. We observe that degeneracy of the problem prevents VC-CD from making progress towards the solution, where SMART-CD preserves $\mathcal{O}(1/k)$ rate as predicted by theory. At the time of this experiment, [FB19] proved almost sure convergence for VC-CD without rates. Since the problem is non-strongly convex, restarting does not improve performance of SMART-CD.

Total Variation and ℓ_1-regularized least squares regression with fMRI data

In this experiment, we consider a computational neuroscience application where prediction is done based on a sequence of fMRI images. Since the images are high dimensional and the number of samples that can be taken is limited, TV-ℓ_1 regularization is used to get stable and predictive estimation results [DGTV14]. The problem we solve is

$$\min_{x\in\mathbb{R}^p} \frac{1}{2}\|Mx-b\|^2 + \lambda r\|x\|_1 + \lambda(1-r)\|x\|_{\mathrm{TV}}. \tag{3.21}$$

This problem fits to our template with $f(x) = \frac{1}{2}\|Mx-b\|^2,\ g(x) = \lambda r\|x\|_1,\ h(u) = \lambda(1-r)\|u\|_1$, where D is the 3D finite difference operator to define a total variation norm $\|\cdot\|_{\mathrm{TV}}$ and $u = Dx$. As mentioned before, our results can be easily extended to this case as h is Lipschitz.

We use an fMRI dataset where x is 3D image of the brain that contains 33177 voxels. Feature matrix M has 768 rows, each representing the brain activity for the corresponding example [DGTV14]. We compare our algorithm with Vu-Condat [Vũ13], FISTA [BT09], ASGARD [TDFC18], Chambolle-Pock [CP11], L-BFGS [BLNZ95] and VC-CD in Figure 3.5 with different values of λ and r. The simulation in Figure 3.5 is performed using benchmarking tool of [DGTV14]. The algorithms are tuned for the best parameters in practice. Per-iteration cost of SMART-CD and VC-CD is similar, therefore the behaviors of these two algorithms are similar in this experiment. Since Vu-Condat's, Chambolle-Pock's, FISTA and ASGARD methods work with full dimensional variables, they have slow convergence in time. L-BFGS has a close

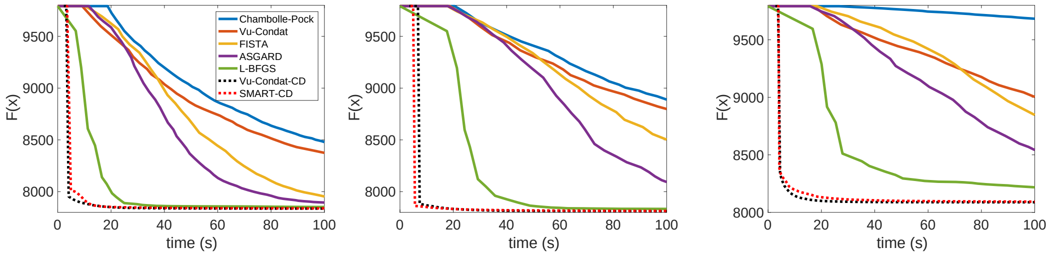

Figure 3.5 – The convergence of 7 algorithms for problem (3.21). The regularization parameters for the first plot are $\lambda = 0.001, r = 0.5$, for the second plot are $\lambda = 0.001, r = 0.9$, for the third plot are $\lambda = 0.01, r = 0.5$.

performance to CD methods.

Linear support vector machines with bias

In this section, we consider an application of our algorithm to support vector machines (SVM) problem for binary classification. Given a training set with n examples $\{a_1, a_2, \ldots, a_n\}$ such that $a_i \in \mathbb{R}^d$ and class labels $\{\bar{b}_1, \bar{b}_2, \ldots \bar{b}_n\}$ such that $\bar{b}_i \in \{-1, +1\}$, we define the soft margin primal support vector machines problem with bias as

$$\min_{w\in\mathbb{R}^d} \sum_{i=1}^{n} C_i \max\Big(0, 1 - \bar{b}_i(\langle a_i, w\rangle + w_0)\Big) + \tfrac{\lambda}{2}\|w\|^2. \tag{3.22}$$

As it is a common practice, we solve its dual formulation, which is a constrained problem:

$$\left\{\begin{array}{ll} \displaystyle\min_{x\in\mathbb{R}^n} & \left\{\tfrac{1}{2\lambda}\|MD(\bar{b})x\|^2 - \sum_{i=1}^n x^{(i)}\right\} \\ \text{s.t.} & 0 \le x^{(i)} \le C_i, \;\; i = 1, \cdots, m, \;\; \bar{b}^\top x = 0, \end{array}\right. \tag{3.23}$$

where $D(\bar{b})$ represents a diagonal matrix that has the class labels $\bar{b}_i$ in its diagonal and $M \in \mathbb{R}^{d\times n}$ is formed by the example vectors. We fit this problem into our template by

$$f(x) = \frac{1}{2\lambda}\|MD(\bar{b})x\|^2 - \sum_{i=1}^{n} x^{(i)}, \quad g_i(x^{(i)}) = \delta_{\{0\le x^{(i)}\le C_i\}}, \quad b = 0, \quad A = \bar{b}^\top.$$

We apply SMART-CD and compare with VC-CD and SDCA [SSZ13]. Even though SDCA is a popular for SVMs, it cannot handle the bias term. Hence, it only applies to (3.23) when $b^\top x = 0$ constraint is removed. This causes SDCA not to converge to the optimal solution with the bias term in (3.22). We summarize the properties of the classification datasets we used: `rcv1.binary` [CL11a, LYRL04] with $n = 20{,}242$, $d = 47{,}236$ in Figure 3.6, plot 1, a8a [CL11a, Lic13] $n = 22{,}696$, $d = 123$, Figure 3.6, plot 2; `gisette` [CL11a, GGBHD05] $n = 6{,}000$, $d = 5{,}000$, Figure 3.6, plot 3. We compile the results in Figure 3.6.

We compute the duality gap for each algorithm and present the results with epochs in the horizontal axis since per-iteration complexity of the algorithms is similar. As expected, SDCA gets stuck at a low accuracy since it ignores one of the constraints in the problem. We demonstrate this in the first experiment and then limit the comparison to SMART-CD and VC-CD.

Equipped with restart strategy, SMART-CD shows the fastest convergence behavior due to the restricted strong convexity of (3.23).

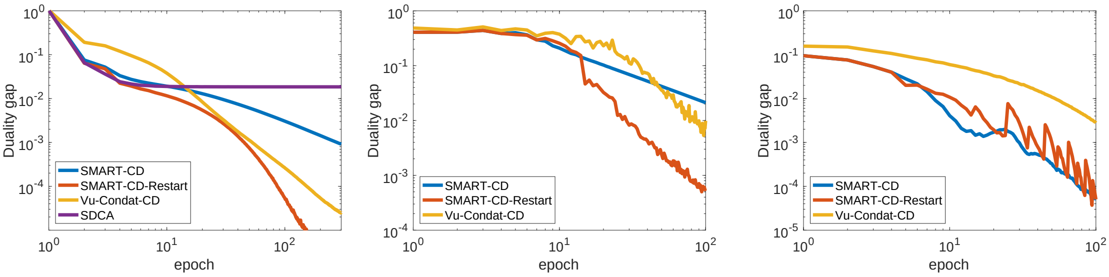

Figure 3.6 – Convergence of algorithms on the dual SVM (3.23) with bias. We only used SDCA in the first dataset since it stagnates at a very low accuracy.

3.4 Proofs for Section 3.2

We recall the definitions of P_β and S_β from (3.10), (3.9). We first prove Lemma 3.1.

Proof of Lemma 3.1. Optimal Lagrange multiplier $y_\star = (y_\star(\xi))_\xi$ is a random variable of $\mathcal{Y}$ with bounded variance due to the constraint qualification condition [BC11]. We start with:

$$-\int \langle A(\xi)x, y_\star(\xi)\rangle + \mathrm{supp}_{b(\xi)}(y_\star(\xi))\mu(d\xi) \le P(x) - P(x_\star) = S_\beta(x) - \frac{1}{2\beta}\int \mathrm{dist}(A(\xi)x, b(\xi))^2\mu(d\xi), \tag{3.24}$$

where the inequality is by saddle point definition, and the equality is by the definition of S_β. We continue by bounding the inner product $\langle A(\xi)x, y_\star(\xi)\rangle$. Let $z := A(\xi)x$, then

$$\begin{aligned}\langle z, y_\star(\xi)\rangle &= \langle z - \Pi_{b(\xi)}(z), y_\star(\xi)\rangle + \langle \Pi_{b(\xi)}(z), y_\star(\xi)\rangle \le \mathrm{dist}(z, b(\xi))\|y_\star(\xi)\| + \langle A(\xi)x_\star, y_\star(\xi)\rangle \\ &\le \frac{1}{4\beta}\mathrm{dist}(z, b(\xi))^2 + \beta\|y_\star(\xi)\|^2 + \mathrm{supp}_{b(\xi)}(y_\star(\xi)), \end{aligned} \tag{3.25}$$

where the first inequality is by Cauchy-Schwarz, optimality conditions, properties of Fenchel's transform: $A(\xi)x_\star \in \partial\,\mathrm{supp}_{b(\xi)}(y_\star(\xi)) \iff y_\star(\xi) \in \partial\delta_{b(\xi)}(A(\xi)x_\star) \iff \langle p - A(\xi)x_\star, y_\star(\xi)\rangle \le 0$, for all $p \in b(\xi)$, due to convexity of $\delta_{b(\xi)}(\cdot)$. The second inequality follows from Young's inequality and the definition $\mathrm{supp}_{b(\xi)}(y_\star(\xi)) = \sup_{u\in b(\xi)}\langle u, y_\star(\xi)\rangle$.

We use $\int \|y_\star(\xi)\|^2\mu(d\xi) = \|y_\star\|^2$, integrate (3.25) and plug in to (3.24) to obtain last inequality. Second and third inequalities directly follow from (3.24) and (3.25). For the first inequality:

$$\begin{aligned} S_\beta(x) &= P(x) + \frac{1}{2\beta}\int \mathrm{dist}(A(\xi)x, b(\xi))^2\mu(d\xi) - P(x_\star) \\ &= P(x) - P(x_\star) + \int \max_{y\in\mathbb{R}^d}\langle A(\xi)x, y\rangle - \mathrm{supp}_{b(\xi)}(y) - \frac{\beta}{2}\left\|y\right\|^2\mu(d\xi) \\ &\ge P(x) - P(x_\star) + \int \langle A(\xi)x, y_\star(\xi)\rangle - \mathrm{supp}_{b(\xi)}(y_\star(\xi)) - \frac{\beta}{2}\left\|y_\star(\xi)\right\|^2\mu(d\xi) \ge -\frac{\beta}{2}\|y_\star\|^2, \end{aligned}$$

where the second equality uses definition of smoothing and the last inequality is by (3.24). ■

3.4.1 General Convex Case

Lemma 3.6. *Let Assumption 3.1 hold and assume* $\forall s$, $\frac{L_h + \|A\|_{2,\infty}^2/\beta_s}{2\alpha_s} \le 0$, $2\alpha_s\|A\|_{2,\infty}^2 - \frac{\beta_s}{2} \le 0$.

$$\mathbb{E}\left[P_{\beta_S}(\bar{x}^S) - P_{\beta_S}(x_\star)\right] \le \frac{1}{2\alpha_S m_S}\|x_\star - x_0^0\|^2 + \frac{\sum_{s=0}^{S-1}\beta_s\alpha_s m_s}{2\alpha_S m_S}\|y_\star\|^2 + 2\frac{\sum_{s=0}^{S}\alpha_s^2 m_s}{\alpha_S m_S}\sigma_h^2. \tag{3.26}$$

Proof. Let us define $z = Ax \in \mathcal{Y}$ and $F_\beta(Ax) = \mathbb{E}[f_\beta(A(\xi)x, \xi)]$. We start by using smoothness of the function $h(x) + F_{\beta_s}(Ax)$

$$\begin{aligned} P_{\beta_s}(x_{k+1}^s) &\le h(x_k^s) + g(x_{k+1}^s) + F_{\beta_s}(Ax_k^s) + \langle \nabla h(x_k^s) + A^\top \nabla_z G_{\beta_s}(Ax_k^s), x_{k+1}^s - x_k^s\rangle \\ &+ \frac{L(\nabla h + \nabla_x F_{\beta_s})}{2}\|x_{k+1}^s - x_k^s\|^2 \\ &= h(x_k^s) + g(x_{k+1}^s) + F_{\beta_s}(x_k^s) + \langle \nabla h_\xi(x_k^s) + A(\xi)^\top \nabla_z f_{\beta_s}(A(\xi)x_k^s, \xi), x_{k+1}^s - x_k^s\rangle \\ &+ \langle \nabla h(x_k^s) - \nabla h_\xi(x_k^s) + A^\top \nabla_z F_{\beta_s}(Ax_k^s) - A(\xi)^\top \nabla_z f_{\beta_s}(A(\xi)x_k^s, \xi), x_{k+1}^s - x_k^s\rangle \\ &+ \frac{L(\nabla h + \nabla_x F_{\beta_s})}{2}\|x_{k+1}^s - x_k^s\|^2. \end{aligned} \tag{3.27}$$

We bound the linear terms in (3.27) separately. First, we use [Tse08, Property 1] with $x = x_\star$:

$$\begin{aligned} g(x_{k+1}^s) + \langle \nabla h_\xi(x_k^s) + A(\xi)^\top \nabla_z f_{\beta_s}(A(\xi)x_k^s, \xi), x_{k+1}^s - x_k^s\rangle &\le g(x_\star) - \frac{1}{2\alpha_s}\|x_{k+1}^s - x_k^s\|^2 \\ + \langle \nabla h_\xi(x_k^s) + A(\xi)^\top \nabla_z f_{\beta_s}(A(\xi)x_k^s, \xi), x_\star - x_k^s\rangle &+ \frac{1}{2\alpha_s}\|x_\star - x_k^s\|^2 - \frac{1}{2\alpha_s}\|x_\star - x_{k+1}^s\|^2 \end{aligned} \tag{3.28}$$

Further, by the fact that $f_{\beta_s}(\cdot, \xi)$ has $1/\beta_s$-Lipschitz gradient,

$$\begin{aligned} \langle A(\xi)^\top \nabla_z f_{\beta_s}(A(\xi)x_k^s, \xi), x_\star - x_k^s\rangle &\le f_{\beta_s}(A(\xi)x_\star, \xi) - f_{\beta_s}(A(\xi)x_k^s, \xi) \\ &- \frac{\beta_s}{2}\|\nabla_z f_{\beta_s}(A(\xi)x_k^s, \xi) - \nabla_z f_{\beta_s}(A(\xi)x_\star, \xi)\|^2 \\ &= f_{\beta_s}(A(\xi)x_\star, \xi) - f_{\beta_s}(A(\xi)x_k^s, \xi) - \frac{\beta_s}{2}\|\nabla_z f_{\beta_s}(A(\xi)x_k^s, \xi)\|^2, \end{aligned} \tag{3.29}$$

where the equality is by $\nabla_z f_{\beta_s}(A(\xi)x_\star, \xi) = 0$, due to the definition of $f_{\beta_s}(\cdot, \xi)$ and $A(\xi)x_\star \in b(\xi)$.

We now use the convexity, $\langle \nabla h_\xi(x_k^s), x_\star - x_k^s\rangle \le h_\xi(x_\star) - h_\xi(x_k^s)$ and (3.29) in (3.28) to get

$$\begin{aligned} g(x_{k+1}^s) + \langle \nabla h_\xi(x_k^s) + A(\xi)^\top \nabla_z f_{\beta_s}(A(\xi)x_k^s, \xi), x_{k+1}^s - x_k^s\rangle &\le g(x_\star) + h_\xi(x_\star) - h_\xi(x_k^s) \\ + f_{\beta_s}(A(\xi)x_\star, \xi) - f_{\beta_s}(A(\xi)x_k^s, \xi) - \frac{\beta_s}{2}\|\nabla_z f_{\beta_s}(A(\xi)x_k^s, \xi)\|^2 &+ \frac{1}{2\alpha_s}\|x_\star - x_k^s\|^2 - \frac{1}{2\alpha_s}\|x_\star - x_{k+1}^s\|^2 \\ &- \frac{1}{2\alpha_s}\|x_{k+1}^s - x_k^s\|^2 \end{aligned} \tag{3.30}$$

We define

$$T_{\alpha_s g}(x_k^s) = \operatorname{prox}_{\alpha_s g}(x_k^s - \alpha_s(\nabla h(x_k^s) + A^\top \nabla_z F_{\beta_s}(Ax_k^s))).$$

For the second linear term in (3.27), we apply conditional expectation knowing x_k^s

$$\begin{aligned}
&\mathbb{E}_k\left[\langle \nabla h(x_k^s) - \nabla h_\xi(x_k^s) + A^\top \nabla_z F_{\beta_s}(Ax_k^s) - A(\xi)^\top \nabla_z f_{\beta_s}(A(\xi)x_k^s,\xi), x_{k+1}^s - x_k^s\rangle\right] = \\
&\mathbb{E}_k\Big[\langle \nabla h(x_k^s) - \nabla h_\xi(x_k^s) + A^\top \nabla_z F_{\beta_s}(Ax_k^s) - A(\xi)^\top \nabla_z f_{\beta_s}(A(\xi)x_k^s,\xi), x_{k+1}^s - T_{\alpha_s g}(x_k^s)\rangle \\
&+ \langle \nabla h(x_k^s) - \nabla h_\xi(x_k^s) + A^\top \nabla_z F_{\beta_s}(Ax_k^s) - A(\xi)^\top \nabla_z f_{\beta_s}(A(\xi)x_k^s,\xi), T_{\alpha_s g}(x_k^s) - x_k^s\rangle\Big] \\
&= \mathbb{E}_k\left[\langle \nabla h(x_k^s) - \nabla h_\xi(x_k^s) + A^\top \nabla_z F_{\beta_s}(Ax_k^s) - A(\xi)^\top \nabla_z f_{\beta_s}(A(\xi)x_k^s,\xi), x_{k+1}^s - T_{\alpha_s g}(x_k^s)\rangle\right] \\
&\le \mathbb{E}_k\left[\|\nabla h(x_k^s) - \nabla h_\xi(x_k^s) + A^\top \nabla_z F_{\beta_s}(Ax_k^s) - A(\xi)^\top \nabla_z f_{\beta_s}(A(\xi)x_k^s,\xi)\| \|x_{k+1}^s - T_{\alpha_s g}(x_k^s)\|\right] \\
&\le \alpha_s \mathbb{E}_k\left[\|\nabla h(x_k^s) - \nabla h_\xi(x_k^s) + A^\top \nabla_z F_{\beta_s}(Ax_k^s) - A(\xi)^\top \nabla_z f_{\beta_s}(A(\xi)x_k^s,\xi)\|^2\right] \\
&\le 2\alpha_s \mathbb{E}_k\left[\|\nabla h(x_k^s) - \nabla h_\xi(x_k^s)\|^2\right] + 2\alpha_s \mathbb{E}_k\left[\|A^\top \nabla_z F_{\beta_s}(Ax_k^s) - A(\xi)^\top \nabla_z f_{\beta_s}(A(\xi)x_k^s,\xi)\|^2\right] \\
&\le 2\alpha_s \sigma_h^2 + 2\alpha_s \mathbb{E}_k\left[\|A(\xi)^\top \nabla_z f_{\beta_s}(A(\xi)x_k^s,\xi)\|^2\right] \\
&\le 2\alpha_s \sigma_h^2 + 2\alpha_s \sup_\xi \|A(\xi)\|^2 \mathbb{E}_k\left[\|\nabla_z f_{\beta_s}(A(\xi)x_k^s,\xi)\|^2\right],
\end{aligned} \tag{3.31}$$

where the second inequality is due to the definition of x_{k+1}^s, $T_{\alpha_s g}(x_k^s)$ and nonexpansiveness of proximal operator. Fourth inequality is due to the fact that $\mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] = \mathbb{E}\left[\|X\|^2\right] - (\mathbb{E}[X])^2$, for any random variable X and $\mathbb{E}_k\left[A(\xi)^\top \nabla_z f_{\beta_s}(A(\xi)x_k^s,\xi)\right] = A^\top \nabla_z F_{\beta_s}(Ax_k^s)$.

We take conditional expectation of (3.27), knowing x_k^s, and plug in (3.30), (3.31) to obtain

$$\begin{aligned}
\mathbb{E}_k[P_{\beta_s}(x_{k+1}^s)] &\le P_{\beta_s}(x_\star) + \frac{1}{2\alpha_s}\|x_\star - x_k^s\|^2 - \frac{1}{2\alpha_s}\mathbb{E}_k\left[\|x_\star - x_{k+1}^s\|^2\right] + 2\alpha_s \mathbb{E}_k\left[\|\nabla h_\xi(x_k^s)\|^2\right] \\
&+ \left(2\alpha_s\|A\|_{2,\infty}^2 - \frac{\beta_s}{2}\right)\mathbb{E}_k\left[\nabla_z f_{\beta_s}(A(\xi)x_k^s,\xi)\right] + \left(\frac{L(\nabla F) + \|A\|_{2,\infty}^2/\beta_s}{2} - \frac{1}{2\alpha_s}\right)\mathbb{E}_k\left[\|x_{k+1}^s - x_k^s\|^2\right].
\end{aligned}$$

We use the assumptions that $2\alpha_s\|A\|_{2,\infty}^2 - \frac{\beta_s}{2} \le 0$ and $\frac{L(\nabla h) + \|A\|_{2,\infty}^2/\beta_s}{2} - \frac{1}{2\alpha_s} \le 0$ to get

$$\mathbb{E}_k\left[P_{\beta_s}(x_{k+1}^s)\right] \le P_{\beta_s}(x_\star) + \frac{1}{2\alpha_s}\|x_\star - x_k^s\|^2 - \frac{1}{2\alpha_s}\mathbb{E}_k\left[\|x_\star - x_{k+1}^s\|^2\right] + 2\alpha_s\sigma_h^2.$$

We apply total expectation and sum for $k \in \{0, \dots, m_s - 1\}$ to obtain

$$\begin{aligned}
\mathbb{E}\left[P_{\beta_s}\left(\frac{1}{m_s}\sum_{k=1}^{m_s} x_k^s\right) - P_{\beta_s}(x_\star)\right] &\le \frac{1}{2\alpha_s m_s}\mathbb{E}\left[\|x_\star - x_0^s\|^2\right] - \frac{1}{2\alpha_s m_s}\mathbb{E}\left[\|x_\star - x_{m_s}^s\|^2\right] + \frac{2\alpha_s}{m_s}\sum_{k=0}^{m_s-1}\sigma_h^2 \\
&\le \frac{1}{2\alpha_s m_s}\mathbb{E}\left[\|x_\star - x_0^s\|^2\right] - \frac{1}{2\alpha_s m_s}\mathbb{E}\left[\|x_\star - x_{m_s}^s\|^2\right] + 2\alpha_s\sigma_h^2.
\end{aligned} \tag{3.32}$$

By Lemma 3.1, by using $P_{\beta_s}(x_\star) = P(x_\star)$, we have $P_{\beta_s}(x) - P_{\beta_s}(x_\star) \ge -\frac{\beta_s}{2}\|y_\star\|^2$. By the restarting rule of the inner loop, one has $x_{m_s}^s = x_0^{s+1}$. Using the previous inequality in (3.32):

$$\mathbb{E}\left[\|x_\star - x_0^{s+1}\|^2\right] \le \mathbb{E}\left[\|x_\star - x_0^s\|^2\right] + \beta_s\alpha_s m_s\|y_\star\|^2 + 4\alpha_s^2 m_s\sigma_h^2 \tag{3.33}$$

We now sum (3.33) for $s \in \{0,1,\dots,S-1\}$

$$\mathbb{E}\left[\|x_\star - x_0^S\|^2\right] \le \|x_\star - x_0^0\|^2 + \sum_{s=0}^{S-1} \beta_s \alpha_s m_s \|y_\star\|^2 + 4\sum_{s=0}^{S-1} \alpha_s^2 m_s \sigma_h^2 \tag{3.34}$$

We now use (3.34) in (3.32) to obtain the result. ■

Next, we estimate the rates of the parameters to determine the convergence rates:

Lemma 3.7. *Denote as $M_S = \sum_{s=0}^{S} m_s$ the total number of iterations to compute $\bar{x}^S$. Let $\omega, \alpha_0, m_0, m_s$ be chosen as Case 1 in Algorithm 3.1. Then, for all s, $\frac{L(\nabla h)+\|A\|_{2,\infty}^2/\beta_s}{2} - \frac{1}{2\alpha_s} \le 0$ and $2\alpha_s \|A\|_{2,\infty}^2 - \frac{\beta_s}{2} \le 0$. Moreover,*

$$\beta_s \le 4\alpha_0\sqrt{m_0}\|A\|_{2,\infty}^2 \frac{\sqrt{\omega}}{\sqrt{\omega-1}}\frac{1}{\sqrt{M_s}}$$
$$\alpha_s m_s \ge \alpha_0 \frac{(m_0-1)}{\sqrt{m_0}}\frac{\sqrt{\omega-1}}{\sqrt{\omega}}\sqrt{M_s}$$
$$\sum_{s=0}^{S-1} \beta_s\alpha_s m_s \le 4\alpha_0^2\|A\|_{2,\infty}^2 m_0 \frac{\log(M_s/m_0)}{\log(\omega)}$$
$$\sum_{s=0}^{S} \alpha_s^2 m_s \le \alpha_0 m_0 \left(\frac{\log(M_s/m_0)}{\log(\omega)} + 1\right)$$

Proof. By definition of β_s, we have $2\alpha_s\|A\|_{2,\infty}^2 - \frac{\beta_s}{2} = 0$. By using the definition of β_s, along with α_s being decreasing and the condition on α_0, we have $\frac{L_h+\|A\|_{2,\infty}^2/\beta_s}{2} - \frac{1}{2\alpha_s} \le 0$. Next,

$$M_S = \sum_{s=0}^{S} m_s = \sum_{s=0}^{S} \lfloor m_0\omega^s \rfloor \le \sum_{s=0}^{S} m_0\omega^s = m_0\frac{\omega^{S+1}-1}{\omega-1}, \tag{3.35}$$

which in turn gives

$$\omega^S \ge \frac{\omega-1}{\omega}\frac{M_S}{m_0} + \frac{1}{\omega} \ge \frac{\omega-1}{\omega}\frac{M_S}{m_0}. \tag{3.36}$$

We now use this bound to get

$$\beta_S = 4\alpha_S\|A\|_{2,\infty}^2 = 4\alpha_0\|A\|_{2,\infty}^2\omega^{-S/2} \le 4\alpha_0\|A\|_{2,\infty}^2\frac{\sqrt{\omega}}{\sqrt{\omega-1}}\frac{\sqrt{m_0}}{\sqrt{M_S}}$$
$$\alpha_S m_S = \alpha_0\omega^{-S/2}\lfloor m_0\omega^S\rfloor \ge \alpha_0 m_0\omega^{S/2} - \alpha_0\omega^{-S/2} \ge \alpha_0\frac{(m_0-1)}{\sqrt{m_0}}\frac{\sqrt{\omega-1}}{\sqrt{\omega}}\sqrt{M_S}.$$

We can also lower bound M_S as

$$M_S = \sum_{s=0}^{S} m_s = \sum_{s=0}^{S}\lfloor m_0\omega^s\rfloor = m_0 + \sum_{s=1}^{S}\lfloor m_0\omega^s\rfloor \ge m_0 + m_0\omega^S - 1 \ge m_0\omega^S,$$

since $m_0 \geq 1$. We thus get

$$S \leq \frac{\log(M_S/m_0)}{\log(\omega)} \tag{3.37}$$

Further,

$$\beta_s \alpha_s m_s = 4\alpha_0^2 \|A\|_{2,\infty}^2 \omega^{-s} \lfloor m_0 \omega^s \rfloor \leq 4\alpha_0^2 \|A\|_{2,\infty}^2 m_0.$$

Now we use (3.37) to show that

$$\sum_{s=0}^{S-1} \beta_s \alpha_s m_s \leq S \times 4\alpha_0^2 \|A\|_{2,\infty}^2 m_0 \leq 4\alpha_0^s \|A\|_{2,\infty}^2 m_0 \frac{\log(M_S/m_0)}{\log(\omega)}.$$

Lastly, we use the relation $\beta_s = 4\alpha_s \|A\|_{2,\infty}^2$ to conclude last bound. ■

Proof of Theorem 3.2. We first combine Lemma 3.6 and Lemma 3.7:

$$\begin{aligned}
\mathbb{E}[S_{\beta_S}(\bar{x}^S)] = \mathbb{E}\left[P_{\beta_S}(\bar{x}^S) - P_{\beta_S}(x_\star)\right] &\leq \frac{1}{2\alpha_S m_S}\|x_\star - x_0^0\|^2 + \frac{\sum_{s=0}^{S-1} \beta_s \alpha_s m_s}{2\alpha_S m_S}\|y_\star\|^2 + 2\frac{\sum_{s=0}^{S} \alpha_s^2 m_s}{\alpha_S m_S}\sigma_h^2 \\
&\leq \frac{\frac{\sqrt{m_0}}{(m_0-1)}\frac{\sqrt{\omega}}{\sqrt{\omega-1}}}{\alpha_0\sqrt{M_s}}\left[\frac{1}{2}\|x_\star - x_0^0\|^2 + \frac{4\alpha_0^2\|A\|_{2,\infty}^2 m_0 \frac{\log(M_s/m_0)}{\log(\omega)}}{2}\|y_\star\|^2 + 2\alpha_0 m_0\left(\frac{\log(M_s/m_0)}{\log(\omega)} + 1\right)\sigma_h^2\right] \\
&= \frac{C_1}{\sqrt{M_S}}\left[C_2 + \frac{\log(M_S/m_0)}{\log(\omega)}C_3\right]
\end{aligned}$$

We combine the inequality above with $\beta_S \leq 4\alpha_0\sqrt{m_0}\|A\|_{2,\infty}^2 \frac{\sqrt{\omega}}{\sqrt{\omega-1}}\frac{1}{\sqrt{M_s}} = \frac{C_4}{\sqrt{M_S}}$ and Lemma 3.1:

$$\begin{aligned}
\sqrt{\mathbb{E}\left[\operatorname{dist}(A(\xi)\bar{x}^s, b(\xi))^2\right]} &\leq \sqrt{4\beta_S^2\left\|y_\star\right\|^2 + 4\beta_S S_{\beta_S}(\bar{x}^S)} \\
&\leq \frac{2C_4\left\|y_\star\right\|}{\sqrt{M_S}} + \frac{2\sqrt{C_1 C_4}}{\sqrt{M_S}}\sqrt{C_2 + \frac{\log(M_S/m_0)}{\log(\omega)}C_3}
\end{aligned} \tag{3.38}$$

The other inequalities follow similarly using

$$S_{\beta_S}(\bar{x}^S) \geq P(\bar{x}^S) - P(x_\star) \geq -\frac{1}{4\beta_S}\int \operatorname{dist}(A(\xi)\bar{x}^S, b(\xi))^2 \mu(d\xi) - \beta_S\|y_\star\|^2 \geq -2\beta_S\|y_\star\|^2 - S_{\beta_S}(\bar{x}^S).$$

■

3.4.2 Restricted Strongly Convex Case

Lemma 3.8. *Let Assumption 3.1 hold. Assume that for all s,* $\frac{L(\nabla h)+\|A\|_{2,\infty}^2/\beta_s}{2\alpha_s} \leq 0$, $2\alpha_s\|A\|_{2,\infty}^2 - \frac{\beta_s}{2} \leq 0$ *and* $\mu\alpha_s m_s \geq \frac{1}{c}$, *for* $c < 1$. *Then,*

$$\mathbb{E}\left[P_{\beta_S}(x_k^S) - P_{\beta_S}(x_\star)\right] \leq \frac{c^S}{2\alpha_S m_S}\|x_\star - x_0^0\|^2 + \frac{\sum_{s=0}^{S-1} c^{S+1-s}\beta_s\alpha_s m_s}{2\alpha_S m_S}\|y_\star\|^2$$

$$+\frac{\sum_{s=0}^{S-1} 4c^{S+1-s}\alpha_s^2 m_s}{2\alpha_S m_S}\sigma_h^2 + 2\alpha_S\sigma_h^2. \quad (3.39)$$

Proof. We proceed same as the proof of Lemma 3.6, until (3.32). In the case where $F(x)+h(x)$ satisfies restricted strong convexity, we can derive

$$P_{\beta_s}(x) - P_{\beta_s}(x_\star) \geq -\frac{\beta_s}{2}\|y_\star\|^2 + \frac{\mu}{2}\|x - x_\star\|^2. \quad (3.40)$$

We use (3.40) in (3.32), along with the restarting rule $\bar{x}^s = x_0^{s+1}$ to get

$$\mu\alpha_s m_s \mathbb{E}\left[\|x_\star - x_0^{s+1}\|^2\right] \leq \mathbb{E}\left[\|x_\star - x_0^s\|^2\right] + \beta_s\alpha_s m_s\|y_\star\|^2 + 4\alpha_s^2 m_s\sigma_h^2. \quad (3.41)$$

Further, since $\mu\alpha_s m_s \geq \frac{1}{c}$, for $c < 1$:

$$\mathbb{E}\left[\|x_\star - x_0^{s+1}\|^2\right] \leq c\mathbb{E}\left[\|x_\star - x_0^s\|^2\right] + c\beta_s\alpha_s m_s\|y_\star\|^2 + 4c\alpha_s^2 m_s\sigma_h^2. \quad (3.42)$$

We now get, by recursively applying the inequality for $s \in \{0, 1, \ldots, S-1\}$

$$\mathbb{E}\left[\|x_\star - x_0^S\|^2\right] \leq c^S\|x_\star - x_0^0\|^2 + \sum_{s=0}^{S-1} c^{S-s}\beta_s\alpha_s m_s\|y_\star\|^2 + \sum_{s=0}^{S-1} 4c^{S-s}\alpha_s^2 m_s\sigma_h^2. \quad (3.43)$$

We plug (3.43) into (3.32) to obtain the result. ■

Next, we estimate the rates of the parameters in the restricted strongly convex case.

Lemma 3.9. *Denote as $M_S = \sum_{s=0}^S m_s$ the total number of iterations to compute $\bar{x}^S$. Let $\omega, \alpha_0, m_0, m_s$ be chosen as Case 2 in Algorithm 3.1 and $c = 1/\omega < 1$. Then, for all s, $\frac{L_h + \|A\|_{2,\infty}^2/\beta_s}{2} - \frac{1}{2\alpha_s} \leq 0$ and $2\alpha_s\|A\|_{2,\infty}^2 - \frac{\beta_s}{2} \leq 0$. Moreover,*

$$\beta_s \leq 4\alpha_0 m_0\|A\|_{2,\infty}^2 \frac{\omega}{\omega-1}\frac{1}{M_s}$$
$$\alpha_s m_s \geq \alpha_0(m_0 - 1)$$
$$\sum_{s=0}^{S-1} c^{S-s}\beta_s\alpha_s m_s \leq 4c^S\alpha_0^2\|A\|_{2,\infty}^2 m_0\left(\frac{\log(M_s/m_0)}{\log(\omega)}\right)$$
$$\sum_{s=0}^{S-1} c^{S-s}\alpha_s^2 m_s \leq c^S\alpha_0^2 m_0\left(\frac{\log(M_s/m_0)}{\log(\omega)}\right)$$
$$c^S \leq \frac{\omega}{\omega-1}\frac{m_0}{M_S}$$

Proof. We skip the proofs for the parts that are the same as Lemma 3.7. We have

$$\beta_s = 4\alpha_0\|A\|_{2,\infty}^2\omega^{-s} \leq 4\alpha_0 m_0\|A\|_{2,\infty}^2\frac{\omega}{\omega-1}\frac{1}{M_s}.$$

In addition,

$$\alpha_s m_s = \alpha_0 \omega^{-s} \lfloor m_0 \omega^s \rfloor \geq \alpha_0 \omega^{-s} (m_0 \omega^s - 1) \geq \alpha_0 (m_0 - 1),$$

where the last inequality follows since $\omega^s \geq 1$. We have

$$\sum_{s=0}^{S-1} c^{S-s} \beta_s \alpha_s m_s \leq \alpha_0 m_0 \sum_{s=0}^{S-1} c^{S-s} \beta_s \leq 4\alpha_0^2 \|A\|_{2,\infty}^2 m_0 c^S \sum_{s=0}^{S-1} (\omega c)^{-s} = S \times 4\alpha_0^2 \|A\|_{2,\infty}^2 m_0 c^S$$

Next, we have $c^S = \omega^{-S} \leq \frac{\omega}{\omega-1} \frac{m_0}{M_S}$. Fourth bound follows by combining the third bound with $\beta_s = 4\alpha_s \|A\|_{2,\infty}^2$. ■

Proof of Theorem 3.3. We first combine Lemma 3.8 and Lemma 3.9:

$$\begin{aligned}
\mathbb{E}\left[S_{\beta_S}(x_k^S)\right] &\leq \frac{c^S}{2\alpha_S m_S} \|x_\star - x_0^0\|^2 + \frac{\sum_{s=0}^{S-1} c^{S-s} \beta_s \alpha_s m_s}{2\alpha_S m_S} \|y_\star\|^2 + \frac{\sum_{s=0}^{S-1} 4c^{S-s} \alpha_s^2 m_s}{2\alpha_S m_S} \sigma_h^2 + 2\alpha_S \sigma_h^2 \\
&\leq \frac{\frac{\omega}{\omega-1} \frac{m_0}{M_S}}{\alpha_0(m_0 - 1)} \left[\frac{1}{2} \|x_\star - x_0^0\|^2 + \frac{4\alpha_0^2 \|A\|_{2,\infty}^2 m_0 \left(\frac{\log(M_s/m_0)}{\log(\omega)} \right)}{2} \|y_\star\|^2 + 2\alpha_0^2 m_0 \left(\frac{\log(M_s/m_0)}{\log(\omega)} \right) \sigma_h^2 \right] \\
&+ \frac{\beta_S}{2\|A\|_{2,\infty}^2} \sigma_h^2 \leq \frac{1}{M_S} \left[D_1 + \frac{\log(M_s/m_0)}{\log(\omega)} D_2 \right]
\end{aligned}$$

where $\beta_s \leq 4\alpha_0 m_0 \|A\|_{2,\infty}^2 \frac{\omega}{\omega-1} \frac{1}{M_s} = \frac{D_3}{M_S}$. We then use Lemma 3.1.

$$\begin{aligned}
\sqrt{\mathbb{E}\left[\mathrm{dist}(A(\xi)\bar{x}^s, b(\xi))^2\right]} &\leq \sqrt{4\beta_S^2 \left\|y_\star\right\|^2 + 4\beta_S S_{\beta_S}(\bar{x}^S)} \\
&\leq \frac{2D_3 \left\|y_\star\right\|}{M_S} + \frac{2\sqrt{D_3}}{M_S} \sqrt{D_1 + \frac{\log(M_S/m_0)}{\log(\omega)} D_2}. \qquad (3.44)
\end{aligned}$$

The other inequalities follow similarly. ■

3.5 Proofs for Section 3.3

3.5.1 Key lemmas

The following properties are key to design the algorithm, whose proofs are very similar to [TDFC18, Lemma 10] using a different norm, so we omit the proof here. The proof of the last property directly follows by using the explicit form of $h_\beta(u)$ in the case when $h^*(y) = \langle c, y \rangle$.

Lemma 3.10. *For any $u, \hat{u} \in \mathbb{R}^m$, let $f_\beta(u) = \max_y \langle u, y \rangle - f^*(y) - \frac{\beta}{2}\|y - \dot{y}\|^2$. Then,*

(a) $f_\beta(\cdot)$ is convex and smooth with $\nabla f_\beta(u) = y_\beta^*(u)$ being Lipschitz continuous with $L_{f_\beta} = \frac{1}{\beta}$.

(b) $f_\beta(u) + \langle \nabla f_\beta(u), \hat{u} - u \rangle + \frac{\beta}{2}\|y_\beta^*(u) - y_\beta^*(\hat{u})\|^2 \leq f_\beta(\hat{u})$.

(c) $f(\hat{u}) \geq f_\beta(u) + \langle \nabla f_\beta(u), \hat{u} - u \rangle + \frac{\beta}{2}\|y_\beta^*(u) - \dot{y}\|^2$.

(d) If $f^*(y) = \langle c, y\rangle$, then $f_\beta(u) = f_{\bar\beta}(u) + \frac{(\bar\beta-\beta)\beta}{2\bar\beta}\|y^*_\beta(u) - \dot y\|^2$.

The following lemma is motivated by [FR15].

Lemma 3.11. *Consider the iterates $\{\bar x_k, \tilde x_k\}_{k\geq 0}$ of Algorithm 3.2. Then, for $k \geq 0$ and $i \in [m]$, we can write $\{\bar x_k^{(i)}\}$ as a convex combination of $\{\tilde x_l^{(i)}\}_{l=0}^k$:*

$$\bar x_k^{(i)} = \sum_{l=0}^{k} \gamma_{k,l}^{(i)} \tilde x_l^{(i)}, \tag{3.45}$$

where $\gamma_{k,l}^{(i)} \geq 0$ and $\sum_{l=0}^k \gamma_{k,l}^{(i)} = 1$. Moreover, the coefficients $\gamma_{k,l}^{(i)}$ can explicitly be computed as

$$\gamma_{k+1,l}^{(i)} = \begin{cases} (1-\tau_k)\gamma_{k,l}^{(i)}, & \text{for } l = 0,\cdots,k-1,\\ (1-\tau_k)\gamma_{k,k}^{(i)} + \tau_k - \frac{\tau_k}{\tau_0}, & \text{for } l = k,\\ \frac{\tau_k}{\tau_0}, & \text{for } l = k+1. \end{cases} \tag{3.46}$$

Proof. Now, from the definition of $\bar x_{k+1}$ and $\hat x_k$, for $i \in [m]$, we can write

$$\bar x_{k+1}^{(i)} = (1-\tau_k)\bar x_k^{(i)} + \tau_k \tilde x_k^{(i)} + \frac{\tau_k}{\tau_0}(\tilde x_{k+1}^{(i)} - \tilde x_k^{(i)}) = (1-\tau_k)\bar x_k^{(i)} + (\tau_k - \frac{\tau_k}{\tau_0})\tilde x_k^{(i)} + \frac{\tau_k}{\tau_0}\tilde x_{k+1}^{(i)}. \tag{3.47}$$

We prove that $\bar x_k^{(i)} = \sum_{l=0}^k \gamma_{k,l}^{(i)} \tilde x_l^{(i)}$ for $i \in [m]$ such that $\gamma_{k,l}^{(i)} \geq 0$ and $\sum_{l=0}^k \gamma_{k,l}^{(i)} = 1$. For $k = 0$, we have $\bar x_0 = \tilde x_0$, which trivially holds by choosing $\gamma_{0,0}^{(i)} = 1$. Assume that this expression holds for $k \geq 1$, we prove it holds for $k+1$. From (3.47), using this induction assumption, we write

$$\bar x_{k+1}^{(i)} = (1-\tau_k)\sum_{l=0}^{k-1}\gamma_{k,l}^{(i)}\tilde x_l^{(i)} + \left[(1-\tau_k)\gamma_{k,k}^{(i)} + \tau_k - \frac{\tau_k}{\tau_0}\right]\tilde x_k^{(i)} + \frac{\tau_k}{\tau_0}\tilde x_{k+1}^{(i)} = \sum_{l=0}^{k+1}\gamma_{k+1,l}^{(i)}\tilde x_l^{(i)},$$

where constants $\gamma_{k+1,l}^{(i)}$ are as given in (3.46). It is easy to check $\sum_{l=0}^{k+1}\gamma_{k+1,l}^{(i)} = (1-\tau_k)\sum_{l=0}^k \gamma_{k,l}^{(i)} + \tau_k - \frac{\tau_k}{\tau_0} + \frac{\tau_k}{\tau_0} = (1-\tau_k) + \tau_k = 1$. Since $\{\tau_k\}_{k\geq 0}$ is non-increasing, $\gamma_{k,l}^{(i)} \geq 0$. ■

3.5.2 Convergence analysis of SMART-CD

Proof of Theorem 3.4. First, let us define the full primal proximal-gradient step as

$$\bar{\tilde x}_{k+1} := \operatorname*{argmin}_{x\in\mathbb{R}^d}\left\{\langle \nabla\psi_{\beta_{k+1}}(\hat x_k), x - \hat x_k\rangle + g(x) + \tau_k\sum_{i=1}^m \frac{B_k^{(i)}}{2\tau_0}\|x^{(i)} - \tilde x_k^{(i)}\|_{(i)}^2\right\}, \tag{3.48}$$

where $\nabla\psi_{\beta_{k+1}}(\hat x_k) = \nabla h(\hat x_k) + A^\top y^*_{\beta_{k+1}}(A\hat x_k)$. The primal coordinate step (Step 7) and Step 8 in Algorithm 3.2 can be written as

$$\tilde x_{k+1}^{(i)} = \begin{cases} \bar{\tilde x}_{k+1}^{(i)}, & \text{if } i = i_k,\\ \tilde x_k^{(i)}, & \text{otherwise.}\end{cases} \tag{3.49}$$

Moreover, using [Tse08, Property 2], we know that for all $x \in \mathbb{R}^d$ and for all $i \in [m]$,

$$
\begin{aligned}
g_i(\bar{\bar{x}}_{k+1}^{(i)}) \le g_i(x^{(i)}) + \langle \nabla_i \psi_{\beta_{k+1}}(\hat{x}_k), x^{(i)} - \bar{\bar{x}}_{k+1}^{(i)} \rangle + \frac{\tau_k B_k^{(i)}}{2\tau_0}\Big(\|x^{(i)} - \tilde{x}_k^{(i)}\|_{(i)}^2 - \|x^{(i)} - \bar{\bar{x}}_{k+1}^{(i)}\|_{(i)}^2\Big) \\
- \frac{\tau_k B_k^{(i)}}{2\tau_0}\|\bar{\bar{x}}_{k+1}^{(i)} - \tilde{x}_k^{(i)}\|_{(i)}^2. \qquad (3.50)
\end{aligned}
$$

Now, since the partial gradient $\nabla_{i_k} h$ is $\hat{L}_{i_k}$-Lipschitz continuous, using $\bar{x}_{k+1}^{(i_k)} = \hat{x}_k^{(i_k)} + \frac{\tau_k}{\tau_0}(\tilde{x}_{k+1}^{(i_k)} - \tilde{x}_k^{(i_k)})$ and $\bar{x}_{k+1}^{(i)} = \hat{x}_k^{(i)}$ for $i \neq i_k$, we have

$$
\begin{aligned}
h(\bar{x}_{k+1}) &\le h(\hat{x}_k) + \langle \nabla_{i_k} h(\hat{x}^k), \bar{x}_{k+1}^{(i_k)} - \hat{x}_k^{(i_k)} \rangle + \frac{\hat{L}_{i_k}}{2}\|\bar{x}_{k+1}^{(i_k)} - \hat{x}_k^{(i_k)}\|_{(i_k)}^2 \\
&= h(\hat{x}_k) + \frac{\tau_k}{\tau_0}\langle \nabla_{i_k} h(\hat{x}_k), \tilde{x}_{k+1}^{(i_k)} - \tilde{x}_k^{(i_k)} \rangle + \frac{\tau_k^2 \hat{L}_{i_k}}{2\tau_0^2}\|\tilde{x}_{k+1}^{(i_k)} - \tilde{x}_k^{(i_k)}\|_{(i_k)}^2. \qquad (3.51)
\end{aligned}
$$

Taking conditional expectation and noting (3.49), we obtain

$$
\mathbb{E}_k\big[h(\bar{x}_{k+1})\big] \le h(\hat{x}_k) + \frac{\tau_k}{\tau_0}\sum_{i=1}^{n} q_i \langle \nabla_i h(\hat{x}_k), \bar{\bar{x}}_{k+1}^{(i)} - \tilde{x}_k^{(i)} \rangle + \frac{\tau_k^2}{\tau_0^2}\sum_{i=1}^{m} q_i \frac{\hat{L}_i}{2}\|\bar{\bar{x}}_{k+1}^{(i)} - \tilde{x}_k^{(i)}\|_{(i)}^2. \qquad (3.52)
$$

We denote by $\varphi_\beta(x) := f_\beta(Ax)$. By Lemma 3.10, we see that $\varphi_{\beta_{k+1}}$ has block-coordinate Lipschitz gradient with $\frac{\|A_i\|^2}{\beta_{k+1}}$, where A_i is the i-th column block of A. Moreover, $\nabla_i \varphi_{\beta_{k+1}}(x) = A_i^\top y^*_{\beta_{k+1}}(Ax)$. Hence, using $\bar{x}_{k+1}^{(i_k)} = \hat{x}_k^{(i_k)} + \frac{\tau_k}{\tau_0}(\tilde{x}_{k+1}^{(i_k)} - \tilde{x}_k^{(i_k)})$ and $\bar{x}_{k+1}^{(i)} = \hat{x}_k^{(i)}$ for $i \neq i_k$, we write

$$
\begin{aligned}
\varphi_{\beta_{k+1}}(\bar{x}_{k+1}) &\le \varphi_{\beta_{k+1}}(\hat{x}_k) + \langle \nabla_{i_k} \varphi_{\beta_{k+1}}(\hat{x}_k), \bar{x}_{k+1}^{(i_k)} - \hat{x}_k^{(i_k)} \rangle + \frac{\|A_i\|^2}{2\beta_{k+1}}\|\bar{x}_{k+1}^{(i_k)} - \hat{x}_k^{(i_k)}\|_{(i_k)}^2 \\
&= \varphi_{\beta_{k+1}}(\hat{x}_k) + \frac{\tau_k}{\tau_0}\langle \nabla_{i_k} \varphi_{\beta_{k+1}}(\hat{x}_k), \tilde{x}_{k+1}^{(i_k)} - \tilde{x}_k^{(i_k)} \rangle + \frac{\tau_k^2\|A_i\|^2}{2\tau_0^2\beta_{k+1}}\|\tilde{x}_{k+1}^{(i_k)} - \tilde{x}_k^{(i_k)}\|_{(i_k)}^2.
\end{aligned}
$$

Taking the conditional expectation and noting (3.49), we get

$$
\begin{aligned}
\mathbb{E}_k\big[\varphi_{\beta_{k+1}}(\bar{x}_{k+1})\big] \le \varphi_{\beta_{k+1}}(\hat{x}_k) + \frac{\tau_k}{\tau_0}\sum_{i=1}^{m} q_i \langle \nabla_i \varphi_{\beta_{k+1}}(\hat{x}_k), \bar{\bar{x}}_{k+1}^{(i)} - \tilde{x}_k^{(i)} \rangle \\
+ \frac{\tau_k^2}{\tau_0^2}\sum_{i=1}^{m} q_i \frac{\|A_i\|^2}{2\beta_{k+1}}\|\bar{\bar{x}}_{k+1}^{(i)} - \tilde{x}_k^{(i)}\|_{(i)}^2. \qquad (3.53)
\end{aligned}
$$

Now, we define

$$
\hat{g}_k^{(i)} := \sum_{l=0}^{k} \gamma_{k,l}^{(i)} g_i(\tilde{x}_l^{(i)}) \quad \text{and} \quad \hat{g}_k := \sum_{i=1}^{m} \hat{g}_k^{(i)}. \qquad (3.54)
$$

Using Lemma 3.11, we can write

$$
\hat{g}_{k+1}^{(i)} = \sum_{l=0}^{k+1} \gamma_{k+1,l}^{(i)} g_i(\tilde{x}_l^{(i)}) = \sum_{l=0}^{k-1} (1-\tau_k)\gamma_{k,l}^{(i)} g_i(\tilde{x}_l^{(i)}) + \Big[(1-\tau_k)\gamma_{k,k}^{(i)} + \tau_k - \tfrac{\tau_k}{\tau_0}\Big] g_i(\tilde{x}_k^{(i)}) + \frac{\tau_k}{\tau_0} g_i(\tilde{x}_{k+1}^{(i)})
$$

$$
\begin{aligned}
&= (1-\tau_k)\sum_{l=0}^{k}\gamma_{k,l}^{(i)} g_i(\tilde{x}_l^{(i)}) + \tau_k g_i(\tilde{x}_k^{(i)}) + \frac{\tau_k}{\tau_0}\left(g_i(\tilde{x}_{k+1}^{(i)}) - g_i(\tilde{x}_k^{(i)})\right)\\
&= (1-\tau_k)\hat{g}_k^{(i)} + \tau_k g_i(\tilde{x}_k^{(i)}) + \frac{\tau_k}{\tau_0}\left(g_i(\tilde{x}_{k+1}^{(i)}) - g_i(\tilde{x}_k^{(i)})\right).
\end{aligned}
$$

Using the definition (3.54) of $\hat{g}_k$, this estimate implies

$$
\hat{g}_{k+1} = (1-\tau_k)\hat{g}_k + \sum_{i=1}^{m}\left[\tau_k g_i(\tilde{x}_k^{(i)}) + \frac{\tau_k}{\tau_0}\left(g_i(\tilde{x}_{k+1}^{(i)}) - g_i(\tilde{x}_k^{(i)})\right)\right].
$$

Now, by the expression (3.49), we can show that

$$
\mathbb{E}_k\left[g_i(\tilde{x}_{k+1}^{(i)})\right] = q_i g_i(\bar{\tilde{x}}_{k+1}^{(i)}) + (1-q_i) g_i(\tilde{x}_k^{(i)}).
$$

Combining the two last expressions, we derive

$$
\begin{aligned}
\mathbb{E}_k\left[\hat{g}_{k+1}\right] &= (1-\tau_k)\hat{g}_k + \sum_{i=1}^{m}\left[\tau_k g_i(\tilde{x}_k^{(i)}) + \frac{\tau_k}{\tau_0}\left(\mathbb{E}\left[g_i(\tilde{x}_{k+1}^{(i)})\right] - g_i(\tilde{x}_k^{(i)})\right)\right]\\
&= (1-\tau_k)\hat{g}_k + \tau_k\sum_{i=1}^{m} g_i(\tilde{x}_k^{(i)}) + \frac{\tau_k}{\tau_0}\sum_{i=1}^{n} q_i\left(g_i(\bar{\tilde{x}}_{k+1}^{(i)}) - g_i(\tilde{x}_k^{(i)})\right). \qquad (3.55)
\end{aligned}
$$

Let us define $\hat{F}_{\beta_k}^{k} := h(\bar{x}^k) + \hat{g}_k + f_{\beta_k}(A\bar{x}_k) \equiv h(\bar{x}_k) + \hat{g}_k + \varphi_{\beta_k}(\bar{x}_k)$. Then, from (3.52), (3.53) and (3.55), we have that

$$
\begin{aligned}
\mathbb{E}_k\left[\hat{F}_{\beta_{k+1}}^{k+1}\right] \le{}& \left[h(\hat{x}_k) + \frac{\tau_k}{\tau_0}\sum_{i=1}^{m} q_i\langle\nabla_i h(\hat{x}_k), \bar{\tilde{x}}_{k+1}^{(i)} - \tilde{x}_k^{(i)}\rangle\right] + \frac{\tau_k^2}{2\tau_0^2}\sum_{i=1}^{m} q_i\left(\hat{L}_i + \frac{\|A_i\|^2}{\beta_{k+1}}\right)\|\bar{\tilde{x}}_{k+1}^{(i)} - \tilde{x}_k^{(i)}\|_{(i)}^2\\
&+ \left[\varphi_{\beta_{k+1}}(\hat{x}_k) + \frac{\tau_k}{\tau_0}\sum_{i=1}^{m} q_i\langle\nabla_i\varphi_{\beta_{k+1}}(\hat{x}_k), \bar{\tilde{x}}_{k+1}^{(i)} - \tilde{x}_k^{(i)}\rangle\right]\\
&+ \left[(1-\tau_k)\hat{g}_k + \tau_k\sum_{i=1}^{m} g_i(\tilde{x}_k^{(i)}) + \frac{\tau_k}{\tau_0}\sum_{i=1}^{m} q_i\left(g_i(\bar{\tilde{x}}_{k+1}^{(i)}) - g_i(\tilde{x}_k^{(i)})\right)\right], \qquad (3.56)
\end{aligned}
$$

since $\nabla\psi_{\beta_{k+1}}(\hat{x}_k) = \nabla h(\hat{x}_k) + \nabla\varphi_{\beta_{k+1}}(\hat{x}_k)$. Now, using the estimate (3.50) into the last expression and noting that $B_k^{(i)} = \hat{L}_i + \frac{\|A_i\|^2}{\beta_{k+1}}$, we can further derive that for all x,

$$
\begin{aligned}
\mathbb{E}_k\left[\hat{F}_{\beta_{k+1}}^{k+1}\right] \le{}& \left[h(\hat{x}_k) + \frac{\tau_k}{\tau_0}\sum_{i=1}^{m} q_i\langle\nabla_i h(\hat{x}_k), x^{(i)} - \tilde{x}_k^{(i)}\rangle\right]\\
&+ \left[\varphi_{\beta_{k+1}}(\hat{x}_k) + \frac{\tau_k}{\tau_0}\sum_{i=1}^{m} q_i\langle\nabla_i\varphi_{\beta_{k+1}}(\hat{x}_k), x^{(i)} - \tilde{x}_k^{(i)}\rangle\right]\\
&+ \left[(1-\tau_k)\hat{g}_k + \tau_k\sum_{i=1}^{m} g_i(\tilde{x}_k^{(i)}) + \frac{\tau_k}{\tau_0}\sum_{i=1}^{m} q_i\left(g_i(x^{(i)}) - g_i(\tilde{x}_k^{(i)})\right)\right]\\
&+ \sum_{i=1}^{m} q_i\frac{\tau_k^2 B_k^{(i)}}{2\tau_0^2}\left(\|x^{(i)} - \tilde{x}_k^{(i)}\|_{(i)}^2 - \|x^{(i)} - \bar{\tilde{x}}_{k+1}^{(i)}\|_{(i)}^2\right). \qquad (3.57)
\end{aligned}
$$

Let us choose x such that for all $i \in [m]$, $x^{(i)} = \left(1 - \frac{\tau_0}{q_i}\right) \tilde{x}_k^{(i)} + \frac{\tau_0}{q_i} x_\star^{(i)}$. Note that as $\tau_0 \leq q_i$ for all i, $x^{(i)}$ is a convex combination of $\tilde{x}_k^{(i)}$ and $x_\star^{(i)}$. We obtain

$$\begin{aligned}\mathbb{E}_k\left[\hat{F}_{\beta_{k+1}}^{k+1}\right] \leq{}& [h(\hat{x}_k) + \tau_k\langle\nabla h(\hat{x}_k), x_\star - \tilde{x}_k\rangle] + \left[\varphi_{\beta_{k+1}}(\hat{x}_k) + \tau_k\langle\nabla\varphi_{\beta_{k+1}}(\hat{x}_k), x_\star - \tilde{x}_k\rangle\right] \\ &+ \left[(1-\tau_k)\hat{g}_k + \tau_k g(x_\star)\right] \\ &+ \sum_{i=1}^m q_i \frac{\tau_k^2 B_k^{(i)}}{2\tau_0^2}\left(\left\|\frac{\tau_0}{q_i}(x_\star^{(i)} - \tilde{x}_k^{(i)})\right\|_{(i)}^2 - \left\|\left(1-\frac{\tau_0}{q_i}\right)\tilde{x}_k^{(i)} + \frac{\tau_0}{q_i}x_\star^{(i)} - \bar{\bar{\tilde{x}}}_{k+1}^{(i)}\right\|_{(i)}^2\right). \qquad (3.58)\end{aligned}$$

We use $\|ax + (1-a)y - z\|^2 = a\|x-z\|^2 + (1-a)\|y-z\|^2 - a(1-a)\|x-y\|^2$ to get

$$\begin{aligned}&\left\|\left(1-\frac{\tau_0}{q_i}\right)\tilde{x}_k^{(i)} + \frac{\tau_0}{q_i}x_\star^{(i)} - \bar{\bar{\tilde{x}}}_{k+1}^{(i)}\right\|_{(i)}^2 \\ &\qquad = \left(1-\frac{\tau_0}{q_i}\right)\|\tilde{x}_k^{(i)} - \bar{\bar{\tilde{x}}}_{k+1}^{(i)}\|_{(i)}^2 + \frac{\tau_0}{q_i}\|x_\star^{(i)} - \bar{\bar{\tilde{x}}}_{k+1}^{(i)}\|_{(i)}^2 - \left(1-\frac{\tau_0}{q_i}\right)\frac{\tau_0}{q_i}\|\tilde{x}_k^{(i)} - x_\star^{(i)}\|_{(i)}^2 \\ &\qquad \geq \frac{\tau_0}{q_i}\|x_\star^{(i)} - \bar{\bar{\tilde{x}}}_{k+1}^{(i)}\|_{(i)}^2 - \left(1-\frac{\tau_0}{q_i}\right)\frac{\tau_0}{q_i}\|\tilde{x}_k^{(i)} - x_\star^{(i)}\|_{(i)}^2.\end{aligned}$$

Using this estimate gives

$$\begin{aligned}\mathbb{E}_k\left[\hat{F}_{\beta_{k+1}}^{k+1}\right] \leq{}& \left[h(\hat{x}_k) + \tau_k\langle\nabla h(\hat{x}_k), x_\star - \tilde{x}_k\rangle\right] + \left[\varphi_{\beta_{k+1}}(\hat{x}_k) + \tau_k\langle\nabla\varphi_{\beta_{k+1}}(\hat{x}_k), x_\star - \tilde{x}_k\rangle\right] \\ &+ \left[(1-\tau_k)\hat{g}_k + \tau_k g(x_\star)\right] + \sum_{i=1}^m \frac{\tau_k^2 B_k^{(i)}}{2\tau_0}\left(\|x_\star^{(i)} - \tilde{x}_k^{(i)}\|_{(i)}^2 - \|\bar{\bar{\tilde{x}}}_{k+1}^{(i)} - x_\star^{(i)}\|_{(i)}^2\right). \qquad (3.59)\end{aligned}$$

Using the convexity of h, we have $h(\hat{x}_k) + \langle\nabla h(\hat{x}_k), x_\star - \hat{x}_k\rangle \leq h(x_\star)$ and $h(\hat{x}_k) + \langle\nabla h(\hat{x}_k), \bar{x}_k - \hat{x}_k\rangle \leq h(\bar{x}_k)$. Moreover, since $\hat{x}_k = (1-\tau_k)\bar{x}_k + \tau_k\tilde{x}_k$, we have $\tau_k(x_\star - \tilde{x}_k) = (1-\tau_k)(\bar{x}_k - \hat{x}_k) + \tau_k(x_\star - \hat{x}_k)$. Combining these expressions, we obtain

$$h(\hat{x}^k) + \tau_k\langle\nabla h(\hat{x}_k), x_\star - \tilde{x}_k\rangle \leq (1-\tau_k)h(\bar{x}_k) + \tau_k h(x_\star). \qquad (3.60)$$

On the one hand, by the Lipschitz gradient and convexity of $\varphi_{\beta_{k+1}}$ in Lemma 3.10(b), we have

$$\varphi_{\beta_{k+1}}(\hat{x}_k) + \langle\nabla\varphi_{\beta_{k+1}}(\hat{x}_k), \bar{x}_k - \hat{x}_k\rangle \leq \varphi_{\beta_{k+1}}(\bar{x}_k) - \frac{\beta_{k+1}}{2}\|y_{\beta_{k+1}}^*(A\hat{x}_k) - y_{\beta_{k+1}}^*(A\bar{x}_k)\|^2.$$

On the other hand, by Lemma 3.10(c), we also have

$$\varphi_{\beta_{k+1}}(\hat{x}_k) + \langle\nabla\varphi_{\beta_{k+1}}(\hat{x}_k), x_\star - \hat{x}_k\rangle \leq h(Ax_\star) - \frac{\beta_{k+1}}{2}\|y_{\beta_{k+1}}^*(A\hat{x}_k) - \dot{y}\|^2$$

Combining these two inequalities and using $\tau_k(x_\star - \tilde{x}_k) = (1-\tau_k)(\bar{x}_k - \hat{x}_k) + \tau_k(x_\star - \hat{x}_k)$,

$$\begin{aligned}\varphi_{\beta_{k+1}}(\hat{x}_k) + \tau_k\langle\nabla\varphi_{\beta_{k+1}}(\hat{x}_k), x_\star - \tilde{x}_k\rangle \leq{}& (1-\tau_k)\varphi_{\beta_{k+1}}(\bar{x}_k) + \tau_k f(Ax_\star) \\ &- \frac{(1-\tau_k)\beta_{k+1}}{2}\|y_{\beta_{k+1}}^*(A\hat{x}_k) - y_{\beta_{k+1}}^*(A\bar{x}_k)\|^2 - \frac{\tau_k\beta_{k+1}}{2}\|y_{\beta_{k+1}}^*(A\hat{x}_k) - \dot{y}\|^2.\end{aligned}$$

Next, using Lemma 3.10(d), we can further estimate

$$\begin{aligned}
\varphi_{\beta_{k+1}}(\hat{x}_k) + \tau_k\langle\nabla\varphi_{\beta_{k+1}}(\hat{x}_k), x_\star - \tilde{x}_k\rangle &\leq (1-\tau_k)\varphi_{\beta_k}(\bar{x}_k) + \tau_k f(Ax_\star) - \frac{\tau_k\beta_{k+1}}{2}\|y^*_{\beta_{k+1}}(A\hat{x}_k) - \dot{y}\|^2 \\
&- \frac{(1-\tau_k)\beta_{k+1}}{2}\|y^*_{\beta_{k+1}}(A\hat{x}_k) - y^*_{\beta_{k+1}}(A\bar{x}_k)\|^2 + \frac{(1-\tau_k)(\beta_k - \beta_{k+1})\beta_{k+1}}{2\beta_k}\|y^*_{\beta_{k+1}}(A\bar{x}_k) - \dot{y}\|^2 \\
&\leq (1-\tau_k)\varphi_{\beta_k}(\bar{x}_k) + \tau_k f(Ax_\star) - \frac{(1-\tau_k)\beta_{k+1}}{2\beta_k}\left[\beta_{k+1} - (1-\tau_k)\beta_k\right]\|y^*_{\beta_{k+1}}(A\bar{x}_k) - \dot{y}\|^2. \quad (3.61)
\end{aligned}$$

Last inequality uses $(1-\tau)\|a-b\|^2 + \tau\|a\|^2 - \tau(1-\tau)\|b\|^2 = \|a-(1-\tau)b\|^2 \geq 0$ for any a, b, and $\tau \in [0,1]$. Substituting (3.60) and (3.61) into (3.59), and using $\beta_{k+1} = (1-\tau_k)\beta_k$, we obtain

$$\begin{aligned}
\mathbb{E}_k[\hat{F}^{k+1}_{\beta_{k+1}}] &\leq (1-\tau_k)\left[h(\bar{x}_k) + \hat{g}_k + \varphi_{\beta_k}(\bar{x}_k)\right] + \tau_k\left[h(x_\star) + g(x_\star) + f(Ax_\star)\right] \\
&+ \sum_{i=1}^{n}\frac{\tau_k^2 B_k^{(i)}}{2\tau_0}\left(\|x_\star^{(i)} - \tilde{x}_k^{(i)}\|^2_{(i)} - \|x_\star^{(i)} - \bar{\bar{\tilde{x}}}_{k+1}^{(i)}\|^2_{(i)}\right). \quad (3.62)
\end{aligned}$$

Next, let us denote by $Q_k := \sum_{i=1}^m \frac{\tau_k^2 B_k^{(i)}}{2\tau_0}\left[\|x_\star^{(i)} - \tilde{x}_k^{(i)}\|^2_{(i)} - \|x_\star^{(i)} - \bar{\bar{\tilde{x}}}_{k+1}^{(i)}\|^2_{(i)}\right]$. We can write Q_k as

$$\begin{aligned}
Q_k = \sum_{i=1}^m \frac{\tau_k^2 B_k^{(i)}}{2\tau_0}&\left[\|x_\star^{(i)} - \tilde{x}_k^{(i)}\|^2_{(i)} - \|x_\star^{(i)} - \bar{\bar{\tilde{x}}}_{k+1}^{(i)}\|^2_{(i)}\right] \\
&= \mathbb{E}_k\left[\frac{\tau_k^2 B_k^{(i_k)}}{2q_{i_k}\tau_0}\left(\|x_\star^{(i_k)} - \tilde{x}_k^{(i_k)}\|^2_{(i_k)} - \|x_\star^{(i_k)} - \tilde{x}_{k+1}^{(i_k)}\|^2_{(i_k)}\right)\right] \\
&= \mathbb{E}_k\left[\sum_{i=1}^m \frac{\tau_k^2 B_k^{(i)}}{2q_i\tau_0}\left(\|x_\star^{(i)} - \tilde{x}_k^{(i)}\|^2_{(i)} - \|x_\star^{(i)} - \tilde{x}_{k+1}^{(i)}\|^2_{(i)}\right)\right], \quad (3.63)
\end{aligned}$$

where the last equality follows from the fact that $\tilde{x}_{k+1}^{(i)} = \tilde{x}_k^{(i)}$ for $i \neq i_k$. Substituting this estimate to (3.62), using the definitions of $\hat{F}^k_{\beta_k}$, $P(x_\star) = P_\star$, and taking conditional expectation gives

$$\mathbb{E}\left[\hat{F}^{k+1}_{\beta_{k+1}} - P_\star + \sum_{i=1}^m \frac{\tau_k^2 B_k^{(i)}}{2q_i\tau_0}\|x_\star^{(i)} - \tilde{x}_{k+1}^{(i)}\|^2_{(i)}\right] \leq \mathbb{E}\left[(1-\tau_k)\left(\hat{F}^k_{\beta_k} - P_\star\right) + \sum_{i=1}^m \frac{\tau_k^2 B_k^{(i)}}{2q_i\tau_0}\|x_\star^{(i)} - \tilde{x}_k^{(i)}\|^2_{(i)}\right]. \quad (3.64)$$

To telescope this inequality we assume that $\tau_k^2 B_k^{(i)} \leq (1-\tau_k)\left(\tau_{k-1}^2 B_{k-1}^{(i)}\right)$, which is equivalent to

$$\tau_k^2\left(\hat{L}_i + \frac{\|A_i\|^2}{\beta_{k+1}}\right) \leq (1-\tau_k)\left[\tau_{k-1}^2\left(\hat{L}_i + \frac{\|A_i\|^2}{\beta_k}\right)\right]. \quad (3.65)$$

By $\beta_{k+1} = (1-\tau_k)\beta_k$, this condition becomes

$$\tau_k^2\left((1-\tau_k)\beta_k\hat{L}_i + \|A_i\|^2\right) \leq (1-\tau_k)^2\tau_{k-1}^2\left(\beta_k\hat{L}_i + \|A_i\|^2\right).$$

This condition holds if $\tau_k^2 = (1-\tau_k)^2\tau_{k-1}^2$, which leads to $\tau_k = \frac{\tau_{k-1}}{\tau_{k-1}+1}$, which is the update rule. It is easy to show that $\tau_k = \frac{1}{k+\tau_0^{-1}}$ and $\beta_k = \frac{\beta_1}{\tau_0(k-1)+1}$. We define $S_k = \sum_{i=1}^n \frac{\tau_k^2 B_i^k}{2q_i\tau_0}\|x_i^\star - \tilde{x}_i^{k+1}\|_{(i)}^2$. Then, we can show

$$\begin{aligned}\mathbb{E}\left[\hat{F}_{\beta_{k+1}}^{k+1} - P_\star + S_k\right] &\leq \prod_{i=1}^k (1-\tau_i)\mathbb{E}\left[\hat{F}_{\beta_1}^1 - P_\star + \sum_{i=1}^m \frac{\tau_0^2 B_0^{(i)}}{2q_i\tau_0}\|x_\star^{(i)} - \tilde{x}_1^{(i)}\|_{(i)}^2\right] \\ &\leq \prod_{i=1}^k (1-\tau_i)\left((1-\tau_0)(\hat{F}_{\beta_0}^0 - P_\star) + \sum_{i=1}^m \frac{\tau_0^2 B_0^{(i)}}{2q_i\tau_0}\|x_\star^{(i)} - \tilde{x}_0^{(i)}\|_{(i)}^2\right),\end{aligned}$$

where the second inequality is by (3.64). Since $\tau_k = \frac{1}{k+\tau_0^{-1}}$, it is easy to show that $\omega_{k+1} := \prod_{i=1}^k (1-\tau_i) \leq \prod_{i=1}^k \frac{i+\tau_0^{-1}-1}{i+\tau_0^{-1}} = \frac{1}{\tau_0 k+1}$. We have $P_{\beta_0}(x_0) = \hat{F}_{\beta_0}^0$, and $\tilde{x}_0 = x_0$; and by convexity of g with Lemma 3.11, we also have $g(\bar{x}_k) = g\left(\sum_{l=0}^k \gamma_{k,l}\tilde{x}_l\right) \leq \sum_{l=0}^k \gamma_{k,l} g(\tilde{x}_l) = \hat{g}_k$. Hence, we can write the above estimate as

$$\mathbb{E}\left[F_{\beta_k}(\bar{x}_k) - P^\star\right] \leq \frac{1}{\tau_0(k-1)+1}\left[(1-\tau_0)(P_{\beta_0}(x_0) - P_\star) + \sum_{i=1}^m \frac{\tau_0 B_0^{(i)}}{2q_i}\|x_\star^{(i)} - x_0^{(i)}\|_{(i)}^2\right]. \tag{3.66}$$

Recall that we denote as $y_\star$ a dual solution, existence of which is by Assumption 3.2. We define $D_{\beta_k}(x) := P(x) + f_{\beta_k}(Ax) - P_\star$ and apply [TDFC18, Lemma 1] to obtain the bounds

$$\left\{\begin{aligned} P(\bar{x}_k) - P_\star &\leq D_{\beta_k}(\bar{x}_k) + \left\|y_\star\right\| \|A\bar{x}_k - b\| + \tfrac{\beta_k}{2}\left\|y_\star - \dot{y}\right\|^2, \\ \|A\bar{x}_k - b\| &\leq \beta_k\left[\|y_\star - \dot{y}\| + \left(\|y_\star - \dot{y}\|^2 + 2\beta_k^{-1}D_{\beta_k}(\bar{x}_k)\right)^{1/2}\right]. \end{aligned}\right. \tag{3.67}$$

The result in (3.19) follows by taking expectation and using Jensen's inequality. ■

3.5.3 Equivalence of SMART-CD and Efficient SMART-CD

Proof of Proposition 1. We give a proof for the equivalence of Algorithm 3.2 and Algorithm 3.3 motivated by [FR15]. The claim trivially holds for $k = 0$ using the initialization of the parameters. Assume that the relations hold for some k. Using Step 5 of Algorithm 3.3, we have

$$\tilde{z}_{k+1}^{(i_k)} = \tilde{z}_k^{(i_k)} + t_{k+1}^{(i_k)}. \tag{3.68}$$

We can write from Step 4 of Algorithm 3.3 that

$$\begin{aligned} t_{k+1}^{(i_k)} &= \operatorname*{argmin}_{t\in\mathbb{R}^{d_{i_k}}}\left\{\langle\nabla_{i_k} h(c_k u_k + \tilde{z}_k) + A_{i_k}^\top y_{\beta_{k+1}}^*\left(c_k A u_k + A\tilde{z}_k\right), t\rangle + g_{i_k}(t + \tilde{z}_k^{(i_k)}) + \frac{\tau_k B_k^{(i_k)}}{2\tau_0}\|t\|_{(i_k)}^2\right\} \\ &= \operatorname*{argmin}_{t\in\mathbb{R}^{d_{i_k}}}\left\{\langle\nabla_{i_k} h(\hat{z}_k) + A_{i_k}^\top y_{\beta_{k+1}}^*\left(A\hat{z}_k\right), t\rangle + g_{i_k}(t + \tilde{z}_k^{(i_k)}) + \frac{\tau_k B_k^{(i_k)}}{2\tau_0}\|t\|_{(i_k)}^2\right\} \\ &= \operatorname*{argmin}_{t\in\mathbb{R}^{d_{i_k}}}\left\{\langle\nabla_{i_k} h(\hat{x}_k) + A_{i_k}^\top y_{\beta_{k+1}}^*\left(A\hat{x}_k\right), t\rangle + g_{i_k}(t + \tilde{x}_k^{(i_k)}) + \frac{\tau_k B_k^{(i_k)}}{2\tau_0}\|t\|_{(i_k)}^2\right\} \end{aligned}$$

$$
\begin{aligned}
&= -\tilde{x}_k^{(i_k)} + \arg\min_{x\in\mathbb{R}^{d_{i_k}}} \left\{ \langle \nabla_{i_k} h(\hat{x}_k) + A_{i_k}^\top y^*_{\beta_{k+1}}\big(A\hat{x}_k\big), x - \hat{x}_k^{(i_k)}\rangle + g_{i_k}(x) + \frac{\tau_k B_k^{(i_k)}}{2\tau_0}\|x - \tilde{x}_k^{(i_k)}\|^2_{(i_k)} \right\} \\
&= -\tilde{x}_k^{(i_k)} + \tilde{x}_{k+1}^{(i_k)}.
\end{aligned}
$$

By (3.68) and the inductive assumption on $\tilde{x}_k$, we obtain $\tilde{z}_{k+1} = \tilde{x}_{k+1}$. Next, using the definition of $\bar{z}_{k+1}$ and Step 6, we can derive

$$
\begin{aligned}
\bar{z}_{k+1} &= c_k u_{k+1} + \tilde{z}_{k+1} = c_k\left(u_k - \frac{1-\tau_k/\tau_0}{c_k}(\tilde{z}_{k+1} - \tilde{z}_k)\right) + \tilde{z}_{k+1} \\
&= c_k u_k + \tilde{z}_k + \frac{\tau_k}{\tau_0}(\tilde{z}_{k+1} - \tilde{z}_k) = \hat{z}_k + \frac{\tau_k}{\tau_0}(\tilde{z}_{k+1} - \tilde{z}_k) \\
&= \hat{x}_k + \frac{\tau_k}{\tau_0}(\tilde{x}_{k+1} - \tilde{x}_k) = \bar{x}_{k+1}.
\end{aligned}
$$

Finally, we use the definition of $\hat{z}_{k+1}$, c_k and Step 4 of Algorithm 3.2, we arrive at

$$
\begin{aligned}
\hat{z}_{k+1} &= c_{k+1} u_{k+1} + \tilde{z}_{k+1} = \frac{c_{k+1}}{c_k}(\bar{x}_{k+1} - \tilde{z}_{k+1}) + \tilde{z}_{k+1} \\
&= (1-\tau_{k+1})(\bar{z}_{k+1} - \tilde{z}_{k+1}) + \tilde{z}_{k+1} = (1-\tau_{k+1})(\bar{x}_{k+1} - \tilde{x}_{k+1}) + \tilde{x}_{k+1} \\
&= (1-\tau_{k+1})\bar{x}_{k+1} + \tau_{k+1}\tilde{x}_{k+1} = \hat{x}_{k+1}.
\end{aligned}
$$

Hence, we can conclude that Algorithm 3.2 and Algorithm 3.3 are equivalent. ■

3.6 Bibliographic note

Lemma 3.1 is due to Olivier Fercoq.

4 Convergence of stochastic primal-dual hybrid gradient algorithm

This chapter is motivated by the compelling practical performance of the algorithm Stochastic PDHG (SPDHG) [CERS18] in our experience. We focus on a similar problem template as the previous chapter. SPDHG belongs to the class of primal-dual splitting methods, which is an alternative approach to Nesterov's smoothing. We refer to coordinate descent based variants of such methods as PDCD methods.

In contrast to the favorable empirical performance we observed, theoretical guarantees of SPDHG in [CERS18] were surprisingly weak, especially in the general convex-concave case. This chapter provides a better analysis for this method with three new convergence results. Among these, we highlight the optimal rate $\mathcal{O}(1/k)$ for the standard optimality measure expected primal-dual gap. The difficulty in deriving this rate was already identified in one of the earliest papers on PDCD [DL14] and we introduce a generic technique for overcoming it.

This chapter is based on the joint work with Olivier Fercoq and Volkan Cevher [AFC21].

4.1 Introduction

In this chapter, we focus on the stochastic primal-dual hybrid gradient (SPDHG) algorithm proposed in [CERS18], for solving the optimization problem

$$\min_{x\in\mathcal{X}} \sum_{i=1}^{n} f_i(A_i x) + g(x), \tag{4.1}$$

where $f_i\colon \mathcal{Y}_i \to \mathbb{R}\cup\{+\infty\}$ and $g\colon \mathcal{X} \to \mathbb{R}\cup\{+\infty\}$ are proper l.s.c. convex functions and f is defined as the separable function such that $f(y) = \sum_{i=1}^{n} f_i(y^{(i)})$. $A_i\colon \mathcal{X} \to \mathcal{Y}_i$ is a linear mapping and A is defined such that $(Ax)_i = A_i x$. We recall that this structure covers the previously considered linealy constrained problem, in addition to empirical risk minimization and imaging problems [CERS18].

We refer to [CP16a] for a review of deterministic primal-dual methods for solving this template.

A common strategy for stochastic algorithms is to have coordinate-based updates for the separable dual variable [SSZ13, ZX17, CERS18]. These methods show competitive practical performance and are proven to converge linearly under the assumption that $f_i^*, \forall i$ and g are μ_i and μ_g-strongly convex functions, respectively. Step sizes of these methods in turn depend on μ_i, μ_g to obtain linear convergence. SPDHG belongs to this class, being the randomized version of PDHG [CP11, CP16b].

Chambolle et al. provide convergence analysis for SPDHG under various assumptions on the problem template [CERS18]. In the general convex case, [CERS18] proved that a particular Bregman distance between the iterates of SPDHG and any primal-dual solution converges almost surely to 0 and the ergodic sequence has a $\mathcal{O}(1/k)$ rate for this quantity. Note however in the general convex case, this result neither implies the almost sure convergence of the sequence to a solution nor convergence rate on the expected primal-dual gap. If f_i^* and g are strongly convex functions, SPDHG-μ, which is a variant of SPDHG with step sizes depending on strong convexity constants, is proven to converge linearly [CERS18, Theorem 6.1]. Estimation of strong convexity constants can be challenging in practice, restricting the use of SPDHG-μ.

In its most basic form, standard step sizes of SPDHG are determined using only $\|A_i\|$ [CERS18]. It is observed frequently in practice that the last iterate of PDHG or SPDHG with standard step sizes has competitive practical performance. Yet, existing results come short to prove even the most fundamental results about the algorithm such as iterate convergence or convergence rate for expected primal-dual gap [CERS18]. In this chapter, we focus on SPDHG with standard step sizes, and provide new theoretical results, paving the way for explaining its favorable convergence behavior in practice.

4.1.1 Contributions

We prove the following new results for SPDHG.

▷ We prove that the iterates of SPDHG converge almost surely to a solution. For this purpose, we introduce a representation of SPDHG as a fixed point operator in a duplicated space.

▷ For the ergodic sequence, we show that SPDHG has $\mathcal{O}(1/k)$ rate of convergence for the expected primal-dual gap. We also prove the same rate for objective residual and feasibility for linearly constrained problems. This the first time the optimal rate for the expected primal-dual gap is attained by PDCD methods. Our technique for obtaining this result is generic and can be of independent interest.

▷ When the problem is metrically subregular (see Section 4.6), we prove that SPDHG has linear convergence with standard step sizes. Our result shows that without any modification, basic SPDHG adapts to problem structure and attains linear rate when this assumption holds, which can help explain its favorable performance in practice.

▷ We show that SPDHG shows a competitive practical performance compared to SPDHG-μ of [CERS18] and other state-of-the-art methods such as variance reduction methods.

We summarize our results and compare with those of [CERS18] in Table 4.2 (Page 106).

4.2 Preliminaries

4.2.1 Notation

Recall that $\mathcal{X}, \mathcal{Y}$ are Euclidean spaces. We denote the partitioning of the dual space as $\mathcal{Y} = \prod_{i=1}^n \mathcal{Y}_i$. Given a vector $x \in \mathcal{X}$, we use bold symbol $\boldsymbol{x}$ to denote the duplicated version of this vector, which consists of n 'copies' of x, and the corresponding space is denoted by $\boldsymbol{\mathcal{X}} = \mathcal{X}^n$. Similarly, the duplicated dual space is $\boldsymbol{\mathcal{Y}} = \mathcal{Y}^n$ and $\boldsymbol{\mathcal{Z}} = \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{Y}}$. The 'copies' might be the same, or different, depending on how $\boldsymbol{x}$ is set. To access i^{th} copy, we use the notation $\boldsymbol{x}(i) \in \mathcal{X}$. For the operator $T \colon \boldsymbol{\mathcal{Z}} \to \boldsymbol{\mathcal{Z}}$, and a duplicated vector $\boldsymbol{q} \in \boldsymbol{\mathcal{Z}}$, we denote the output as $T(\boldsymbol{q}) = \begin{bmatrix} T_x(\boldsymbol{q}) \\ T_y(\boldsymbol{q}) \end{bmatrix}$ where, for example, i^{th} primal copy is denoted as $T_x(\boldsymbol{q})(i) \in \mathcal{X}$. Similarly, for the i^{th} primal copy in $\boldsymbol{q}$, we use $\boldsymbol{q}_x(i) \in \mathcal{X}$. To access i^{th} primal and dual copies, we use $\boldsymbol{q}(i) \in \mathcal{Z}$.

For example, when we pick one coordinate at a time, we can set $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}^n$, which would result in the duplicated spaces $\boldsymbol{\mathcal{X}} = \mathbb{R}^{dn}$, $\boldsymbol{\mathcal{Y}} = \mathbb{R}^{n^2}$, and $\boldsymbol{\mathcal{Z}} = \mathbb{R}^{dn+n^2}$.

Probability of selecting an index $i \in \{1, \dots, n\}$ is denoted as $p_i > 0$, with $\sum_{i=1}^n p_i = 1$. We define $P = \operatorname{diag}(p_1, \dots, p_n)$ and $\underline{p} = \min_i p_i$. Notation $\mathcal{F}_k$ defines the filtration generated by indices $\{i_1, \dots, i_{k-1}\}$, selected randomly every iteration. Let $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_k]$ denote the conditional expectation with respect to $\mathcal{F}_k$.

Using Fenchel conjugate, Problem (4.1) can be cast as the saddle point problem

$$\min_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \sum_{i=1}^n \langle A_i x, y^{(i)} \rangle - f_i^*(y^{(i)}) + g(x). \tag{4.2}$$

A primal-dual solution $(x^\star, y^\star) \in \mathcal{Z}^\star$ is characterized as

$$0 \in \begin{bmatrix} A^\top y_\star + \partial g(x_\star) \\ Ax_\star - \partial f^*(y_\star) \end{bmatrix} = F(x_\star, y_\star). \tag{4.3}$$

Given the functions g and f^* as in (4.2), we define

$$D_g(x, \bar{z}) = g(x) - g(\bar{x}) + \langle A^\top \bar{y}, x - \bar{x} \rangle, \tag{4.4}$$

$$D_{f^*}(y, \bar{z}) = f^*(y) - f^*(\bar{y}) - \langle A\bar{x}, y - \bar{y} \rangle. \tag{4.5}$$

If $\bar{z} = z_\star = (x_\star, y_\star)$, with $z_\star$ denoting a primal-dual solution as defined in (4.3), then (4.4) and (4.5) are Bregman distances generated by g and f^*. Respectively, these functions measure the distance between x and $x_\star$; and y and $y_\star$. Consequently, given z, $D_h(z, z^\star)$ is the Bregman distance generated by $h(z) = g(x)+f^*(y)$, to measure the distance between z and $z_\star$. When h is merely convex, a bound on this quantity does not imply a bound on the Euclidean distance. We also note that primal-dual gap function can be written as $\mathrm{Gap}(z) = \sup_{\bar{z}\in\mathcal{Z}} D_{f^*}(x,\bar{z}) + D_g(y,\bar{z})$.

4.2.2 Metric subregularity

For a set valued mapping $F\colon \mathcal{U} \rightrightarrows \mathcal{V}$, we denote the graph of F by $\operatorname{gra} F = \{(u,v) \in \mathcal{U}\times\mathcal{V}\colon\ v \in Fu\}$. We say that F is metrically subregular at $\bar{u}$ for $\bar{v}$, with $(\bar{u},\bar{v}) \in \operatorname{gra} F$, if there exists $\eta_0 > 0$ with a neighborhood of subregularity $\mathcal{N}(\bar{u})$ such that:

$$\operatorname{dist}(u, F^{-1}\bar{v}) \le \eta_0 \operatorname{dist}(\bar{v}, Fu), \quad \forall u \in \mathcal{N}(\bar{u}). \tag{4.6}$$

If $\mathcal{N}(\bar{u}) = \mathcal{U}$, then F is globally metrically subregular [DR09]. Absence of metric subregularity is signaled by $\eta_0 = +\infty$. This assumption is used in the context of deterministic and stochastic primal-dual algorithms in [LFP16, DL18, LFP19]. We study how the metric subregularity of the Karush-Kuhn-Tucker (KKT) operator F in (4.3) implies linear convergence of SPDHG.

We note that metric subregularity of F holds in following cases:

∘ f_i^* and g are strongly convex functions, since $\mathcal{N}(\bar{z}) = \mathcal{Z}$.

∘ The problem (4.1) is defined with piecewise linear quadratic (PLQ) functions and $\operatorname{dom} g$ and $\operatorname{dom} f^*$ are compact sets, in which case $\mathcal{N}(\bar{z}) = \operatorname{dom} g \times \operatorname{dom} f^*$. In particular the domain of a PLQ function can be represented as the union of finitely many polyhedral sets and in each set, the function is a quadratic (see [LFP19, Definition IV.3]). Problems with PLQ functions include Lasso, support vector machines, linear programs, etc.

Remark 4.1. In the first case above, compact domains are not needed since metric subregularity holds globally for these problems. One can also relax strong convexity in the first case, to weaker conditions as quadratic growth or restricted strong convexity, see [LP18, Lemma 4.3] for the details. Throughout the chapter, compact domain assumption is only needed in the second example above, for PLQs, as one sufficient condition for Assumption 4.2. The reason, as we will see in Theorem 4.8 is the lack of control on the low probability event that the trajectory may make an excursion far away. The same assumption for proving linear convergence of another primal-dual coordinate descent method is also needed in [LFP19].

Smoothed gap. In order to prove sublinear convergence rates for linearly constrained problems, we are going to utilize the smoothed gap framework introduced in [TDFC18]. For Problem (4.1), the smoothed gap function is defined as

$$\begin{aligned}\mathcal{G}_{\alpha,\beta}(x,y;\dot{x},\dot{y}) = \sup_{u,v}\, & g(x) + \langle Ax, v\rangle - f^*(v) \\ & - g(u) - \langle Au, y\rangle + f^*(y) - \frac{\alpha}{2}\|u-\dot{x}\|^2 - \frac{\beta}{2}\|v-\dot{y}\|^2. \end{aligned}\tag{4.7}$$

4.3 Algorithm

Chambolle et al. proposed SPDHG [CERS18], which fits into the PDCD class of algorithms, proposed before in [ZX17, FB19, DL14]. A comprehensive literature review is given in Section 4.5. Several variants of SPDHG are analyzed in [CERS18]. In this chapter, we focus on the standard SPDHG which we include as Algorithm 4.1.

Algorithm 4.1 Stochastic PDHG (SPDHG) [CERS18, Algorithm 1]

```
Input: Pick step sizes $\sigma_i, \tau$ by (4.8) and $x_0 \in \mathcal{X}$, $y_0 = y_1 = \bar{y}_1 \in \mathcal{Y}$. Given $P = \mathrm{diag}(p_1, \dots, p_n)$.
for $k = 1, 2, \dots$ do
    $x_k = \mathrm{prox}_{\tau, g}(x_{k-1} - \tau A^\top \bar{y}_k)$
    Draw $i_k \in \{1, \dots, n\}$ such that $\Pr(i_k = i) = p_i$.
    $y_{k+1}^{(i_k)} = \mathrm{prox}_{\sigma_{i_k}, f_{i_k}^*}(y_k^{(i_k)} + \sigma_{i_k} A_{i_k} x_k)$
    $y_{k+1}^{(i)} = y_k^{(i)}, \quad \forall i \neq i_k$
    $\bar{y}_{k+1} = y_{k+1} + P^{-1}(y_{k+1} - y_k),$
end for
```

Remark 4.2. We use serial sampling of blocks in our analysis for the ease of notation. We can extend our results with other samplings by using expected separable overapproximation (ESO) inequality as in [CERS18].

We focus on following standard step size rules for primal and dual step sizes τ, σ_i, which only depend on $\|A_i\|$ and not any other structural constants about the problem

$$p_i^{-1} \tau \sigma_i \|A_i\|^2 \leq \gamma^2 < 1. \tag{4.8}$$

Next section illustrates our novel theoretical results for SPDHG improving on [CERS18].

Assumption 4.1.

- f_i and g are proper, l.s.c., convex functions.
- The set of solutions to (4.1) is nonempty.
- Slater's condition holds [BC11].

Slater's condition is a standard sufficient assumption for strong duality, which is used in most works in the literature of primal-dual methods [BC11, CP11, CERS18, LFP19, TDFC18, FB19]. Strong duality ensures that a dual solution exists in (4.2) and the set of primal-dual solutions is characterized by (4.3).

4.4 Convergence

We start with a lemma analyzing one iteration behavior of the algorithm. This lemma is essentially the same as [CERS18, Lemma 4.4] up to minor modifications and is included for completeness, with its proof in Section 4.7.1.

For the lemma, we introduce the following notation

$$\begin{aligned} V(z) &= \frac{1}{2}\|x\|^2_{\tau^{-1}} + \frac{1}{2}\|y\|^2_{D(\sigma)^{-1}P^{-1}} + \langle Ax, P^{-1}y\rangle, \\ V_k(x,y) &= \frac{1}{2}\|x\|^2_{\tau^{-1}} - \langle Ax, P^{-1}(y_k - y_{k-1})\rangle + \frac{1}{2}\|y_k - y_{k-1}\|^2_{D(\sigma)^{-1}P^{-1}} + \frac{1}{2}\|y\|^2_{D(\sigma)^{-1}P^{-1}}. \end{aligned} \tag{4.9}$$

We also define the full dimensional dual update

$$\hat{y}^{(i)}_{k+1} = \operatorname{prox}_{\sigma_i, f_i^*}(y^{(i)}_k + \sigma_i A_i x_k), \quad \forall i \in \{1,\dots,n\}.$$

Lemma 4.3. *Let Assumption 4.1 hold. It holds for SPDHG that,* $\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}$,

$$\begin{aligned} D_g(x_k, z) + D_{f^*}(\hat{y}_{k+1}, z) \le V_k(x_{k-1} - x, y_k - y) - \mathbb{E}_k\left[V_{k+1}(x_k - x, y_{k+1} - y)\right] \\ - V(z_k - z_{k-1}). \end{aligned} \tag{4.10}$$

Moreover, under the step size rules in (4.8)*, we have with* $C_1 = 1 - \gamma$

$$V(z_k - z_{k-1}) \ge C_1\left(\frac{1}{2}\|x_k - x_{k-1}\|^2_{\tau^{-1}} + \frac{1}{2}\|y_k - y_{k-1}\|^2_{D(\sigma)^{-1}P^{-1}}\right), \tag{4.11}$$

$$V_k(x,y) \ge C_1\left(\frac{1}{2}\|x\|^2_{\tau^{-1}} + \frac{1}{2}\|y_k - y_{k-1}\|^2_{D(\sigma)^{-1}P^{-1}}\right) + \frac{1}{2}\|y\|^2_{D(\sigma)^{-1}P^{-1}}, \tag{4.12}$$

Lower bound in (4.11) specifically uses the structure of the vector $y_k - y_{k-1}$, therefore it would not be true for any y in the function $V(x,y)$. In all our proofs, we only need nonnegativity of $V(z_k - z_{k-1})$ which is proven at the end of proof of Lemma 4.3 in Section 4.7.1.

4.4.1 Almost sure convergence

In this section, we prove almost sure convergence of the iterates of SPDHG to a solution of (4.1). We first introduce an equivalent representation of SPDHG that is instrumental in our proofs. On a high level, this can be seen similar to the representation in [HY12] for PDHG. This representation shifts the update of the primal update so that the algorithm can be written as a fixed point operator. Since the definition of $\bar{y}_{k+1}$ depends on the selected index i_k at iteration k, the operator T is defined such that all the possible values of $\bar{y}_{k+1}$ and consequently, of x_{k+1} are captured.

Lemma 4.4. *Let us define $T\colon \mathcal{Z} \to \mathcal{Z}$ that to $(\boldsymbol{x}, \boldsymbol{y})$ associates $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ such that $\forall\, i \in \{1, \dots, n\}$,*

$$\begin{aligned}
\hat{\boldsymbol{y}}(i) &= \operatorname{prox}_{D(\sigma), f^*}(\boldsymbol{y}(i) + D(\sigma)A\boldsymbol{x}(i))\\
\bar{\boldsymbol{y}}(i) &= \boldsymbol{y}(i) + (1 + p_i^{-1})(\hat{\boldsymbol{y}}(i)_i - \boldsymbol{y}(i)_i)e(i)\\
\hat{\boldsymbol{x}}(i) &= \operatorname{prox}_{\tau, g}(\boldsymbol{x}(i) - \tau A^\top \bar{\boldsymbol{y}}(i))
\end{aligned}$$

where $\boldsymbol{x}(i) \in \mathcal{X}$, $\boldsymbol{y}(i) \in \mathcal{Y}$.

The fixed points of T are of the form $(\boldsymbol{x}(i), \boldsymbol{y}(i))$ such that $(\boldsymbol{x}(i), \boldsymbol{y}(i)) \in \mathcal{Z}_\star$, $\forall\, i \in \{1, \dots, n\}$. Moreover,

$$\big(x_{k+1}, \hat{y}_{k+1}\big) = \big(T_x(1 \otimes x_k, 1 \otimes y_k)(i_k), T_y(1 \otimes x_k, 1 \otimes y_k)(1)\big).$$

We also denote

$$\begin{aligned}
\bar{S} &= \operatorname{blkdiag}(\tau^{-1} I_{dn \times dn}, I_{n\times n} \otimes D(\sigma)^{-1}),\\
\bar{P} &= \operatorname{blkdiag}(p_1 I_{d\times d}, \dots, p_n I_{d\times d}, p_1 I_{n\times n}, \dots, p_n I_{n \times n}).
\end{aligned}$$

We then have,

$$\| T(1 \otimes x_k, 1 \otimes y_k) - (1 \otimes x_k, 1 \otimes y_k)\|^2_{\bar{S}\bar{P}} = \mathbb{E}_k \left[\|x_{k+1} - x_k\|^2_{\tau^{-1}} + \|y_{k+1} - y_k\|^2_{D(\sigma)^{-1}P^{-1}} \right].$$

Before the proof of the lemma, we use an example to illustrate the notation and the main idea.

Example 4.5. Let $d = 1$, $n = 2$, then $\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}(1) \\ \boldsymbol{x}(2) \end{bmatrix} \in \mathbb{R}^2$, $\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}(1) \\ \boldsymbol{y}(2) \end{bmatrix} \in \mathbb{R}^4$, and

$$\begin{aligned}
\bar{S} &= \operatorname{diag}(\tau^{-1}, \tau^{-1}, \sigma_1^{-1}, \sigma_2^{-1}, \sigma_1^{-1}, \sigma_2^{-1}) \in \mathbb{R}^{6\times 6},\\
\bar{P} &= \operatorname{diag}(p_1, p_2, p_1, p_1, p_2, p_2) \in \mathbb{R}^{6 \times 6}.
\end{aligned}$$

Then, we have by letting $\boldsymbol{x} = 1 \otimes x_k$, $\boldsymbol{y} = 1 \otimes y_k$,

$$\begin{aligned}
\hat{\boldsymbol{y}}(1) &= \operatorname{prox}_{D(\sigma), f^*}(y_k + D(\sigma)Ax_k), & \hat{\boldsymbol{y}}(2) &= \operatorname{prox}_{D(\sigma), f^*}(y_k + D(\sigma)Ax_k),\\
\bar{\boldsymbol{y}}(1) &= y_k + (1 + p_1^{-1}) \begin{bmatrix} \hat{\boldsymbol{y}}(1)^{(1)} - y_k^{(1)} \\ 0 \end{bmatrix}, & \bar{\boldsymbol{y}}(2) &= y_k + (1 + p_2^{-1}) \begin{bmatrix} 0 \\ \hat{\boldsymbol{y}}(2)^{(2)} - y_k^{(2)} \end{bmatrix},\\
\hat{\boldsymbol{x}}(1) &= \operatorname{prox}_{\tau, g}(x_k - \tau A^\top \bar{\boldsymbol{y}}(1)), & \hat{\boldsymbol{x}}(2) &= \operatorname{prox}_{\tau, g}(x_k - \tau A^\top \bar{\boldsymbol{y}}(2)).
\end{aligned}$$

We have $T(1 \otimes x_k, 1 \otimes y_k) = \left(\begin{bmatrix} \hat{\boldsymbol{x}}(1) \\ \hat{\boldsymbol{x}}(2) \end{bmatrix}, \begin{bmatrix} \hat{\boldsymbol{y}}(1) \\ \hat{\boldsymbol{y}}(2) \end{bmatrix} \right)$. By using the definition of $\hat{y}_{k+1}$ in Lemma 4.3, we see that $(x_{k+1}, \hat{y}_{k+1}) = (\hat{\boldsymbol{x}}(1), \hat{\boldsymbol{y}}(1))$ if $i_k = 1$ and $(x_{k+1}, \hat{y}_{k+1}) = (\hat{\boldsymbol{x}}(2), \hat{\boldsymbol{y}}(1))$ if $i_k = 2$. Note that we can take any copy of $\hat{\boldsymbol{y}}$ as $\hat{\boldsymbol{y}}(1) = \hat{\boldsymbol{y}}(2)$. Moreover, depending on i_k, one obtains y_{k+1} from $\hat{y}_{k+1}$ with a coordinate-wise update, as given in SPDHG (see Algorithm 4.1).

Proof of Lemma 4.4. Let $(\boldsymbol{x}, \boldsymbol{y})$ be a fixed point of T. Then it follows that

$\boldsymbol{y}(i) = \mathrm{prox}_{D(\sigma),f^*}(\boldsymbol{y}(i) + D(\sigma)A\boldsymbol{x}(i)), \forall i$, $\bar{\boldsymbol{y}}(i) = \boldsymbol{y}(i), \forall i$ and $\boldsymbol{x}(i) = \mathrm{prox}_{\tau,g}(\boldsymbol{x}(i) - \tau A^\top \boldsymbol{y}(i)), \forall i$. Hence, optimality conditions for each i are clearly the same as (4.3). Therefore fixed points of T are such that $(\boldsymbol{x}(i), \boldsymbol{y}(i)) \in \mathcal{Z}_\star, \forall i$.

The equality $(x_{k+1}, \hat{y}_{k+1}) = (T_x(1 \otimes x_k, 1 \otimes y_k)(i_k), T_y(1 \otimes x_k, 1 \otimes y_k)(1))$ is therefore another way to write the algorithm. Since when inputted $(1 \otimes x_k, 1 \otimes y_k)$, T outputs $(1 \otimes \hat{y}_{k+1})$ for the dual variable, we can simply take first copy for $\hat{y}_{k+1}$.

For the last result, we use $\|\hat{y}_{k+1} - y_k\|^2_{D(\sigma)^{-1}} = \mathbb{E}_k\big[\|y_{k+1} - y_k\|^2_{D(\sigma)^{-1}P^{-1}}\big]$ to show

$$\begin{aligned}
&\|T(1 \otimes x_k, 1 \otimes y_k) - (1 \otimes x_k, 1 \otimes y_k)\|^2_{\bar{S}\bar{P}} \\
&= \sum_{i=1}^n \Big(\|T_x(1 \otimes x_k, 1 \otimes y_k)(i) - x_k\|^2_{\tau^{-1}} p_i + \|T_y(1 \otimes x_k, 1 \otimes y_k)(i) - y_k\|^2_{D(\sigma)^{-1}} p_i\Big) \\
&= \sum_{i=1}^n \big(\|T_x(1 \otimes x_k, 1 \otimes y_k)(i) - x_k\|^2_{\tau^{-1}} p_i\big) + \|\hat{y}_{k+1} - y_k\|^2_{D(\sigma)^{-1}}\big(\sum_{i=1}^n p_i\big) \\
&= \mathbb{E}_k\Big[\|x_{k+1} - x_k\|^2_{\tau^{-1}} + \|y_{k+1} - y_k\|^2_{D(\sigma)^{-1}P^{-1}}\Big],
\end{aligned}$$

where we also used that $\sum_{i=1}^n p_i = 1$. ■

We proceed with the main theorem of this section. We will present the main ideas and the main ingredient that makes the proof possible in the following proof sketch. The details of the proof utilizing classical arguments from [CP15, Ber11, IBCH13] are deferred to Section 4.7.2.

Theorem 4.6. *Let Assumption 4.1 hold and define $\Delta_k = V_{k+1}(x_k - x_\star, y_{k+1} - y_\star)$. Then, it holds that $\mathbb{E}[V_k(x_{k-1} - x_\star, y_k - y_\star)] \le \Delta_0$, $\sum_{k=1}^\infty \mathbb{E}[V(z_k - z_{k-1})] \le \Delta_0$. Moreover, almost surely, there exists $(x_\star, y_\star) \in \mathcal{Z}_\star$, such that the iterates of SPDHG satisfy $(x_k, y_k) \to (x_\star, y_\star)$.*

Proof sketch. On (4.10), we pick $(x, y) = (x_\star, y_\star)$ and by convexity, $D_g(x_k, z_\star) \ge 0$, $D_{f^*}(\hat{y}_{k+1}, z_\star) \ge 0$. Next, by using the definition of Δ_k, we write (4.10) as

$$\mathbb{E}_k[\Delta_k] \le \Delta_{k-1} - V(z_k - z_{k-1}).$$

We apply Robbins-Siegmund lemma [RS71, Theorem 1] to get that almost surely, Δ_k converges to a finite valued random variable and $V(z_k - z_{k-1}) \to 0$. Consequently, by (4.11), $\|y_k - y_{k-1}\|$ converges to 0 almost surely. Since almost surely, Δ_k converges and $\|y_k - y_{k-1}\|$ converges to 0, by using the definition of V_k in (4.9), we have that $\|z_k - z_\star\|$ converges almost surely.

Next, we denote $\boldsymbol{q}_k = (1 \otimes x_k, 1 \otimes y_k)$ and use the arguments in [CP15, Proposition 2.3], [FB19, Theorem 1] to argue that there exists a set Ω with $\mathbb{P}(\Omega) = 1$ such that for every $z_\star \in \mathcal{Z}_\star$ and for every $\omega \in \Omega$, $\|z_k(\omega) - z_\star\|$ converges and $\|T(\boldsymbol{q}_k(\omega)) - \boldsymbol{q}_k(\omega)\| \to 0$. As for every $\omega \in \Omega$, $(z_k(\omega))_k$ is bounded, we denote by $\tilde{z} = (\tilde{x}, \tilde{y})$ one of its cluster points. Then, we denote $\tilde{\boldsymbol{q}} = (1 \otimes \tilde{x}, 1 \otimes \tilde{y})$ and say that $\tilde{\boldsymbol{q}}$ is a cluster point of $(\boldsymbol{q}_k(\omega))_k$.

The key step in our proof that enables the result is the fixed point characterization of T in Lemma 4.4. With this result, we derive $\tilde{z} \in \mathcal{Z}_\star$ as $\tilde{\boldsymbol{q}}$ is a fixed point of T.

To sum up, we have shown that at least on some subsequence, $z_k(\omega)$ converges to $\tilde{z} \in \mathcal{Z}_\star$. As for every $\omega \in \Omega$ and $z_\star \in \mathcal{Z}_\star$, $\|z_k(\omega) - z_\star\|$ converges, the result follows. ∎

4.4.2 Linear convergence

The standard approach for showing linear convergence with metric subregularity is to obtain a Fejer-type inequality of the form [LFP19]

$$\mathbb{E}_k\left[d(z_{k+1} - z_\star)\right] \le d(z_k - z_\star) - V(T(z_k) - z_k), \tag{4.13}$$

for suitably defined norms d and V and operator T. However, as evident from (4.10) and the definition of V_{k+1} in (4.9), one iteration result of SPDHG does not fit into this form. When $x = x_\star, y = y_\star$, $V_{k+1}(x_k - x_\star, y_{k+1} - y_\star)$ does not only measure distance to solution, but also the distance of subsequent iterates y_{k+1} and y_k. In addition, V_{k+1} includes $x_k - x_\star$ and $y_{k+1} - y_\star$ rather than $x_{k+1} - x_\star$ and $y_{k+1} - y_\star$, which further presents a challenge due to asymmetry, for using metric subregularity. Therefore, an intricate analysis is needed to control the additional terms and handle the asymmetry in V_{k+1}. In addition, Lemma 4.4 is necessary to identify T.

We need the following notation and lemma which builds on Lemma 4.4 for easier computations with metric subregularity. For the operators, we adopt the convention in [LFP19]. Operator C is the concatenation of subdifferentials, M is the skew symmetric matrix that is formed using matrix A. Operator F is the KKT operator and $\boldsymbol{H}$ is the “metric” that helps us write the algorithm in proximal point form (see Lemma 4.4). Due to duplication in Lemma 4.4, we need duplicated versions of C and M. Consistent with the notation of Lemma 4.4 , we use boldface to denote operators which operate in the duplicated space.

Lemma 4.7. *Under the notations of Lemma 4.4, to write compactly the operation of T, let us define the operators*

$$\begin{aligned}
&C\colon (x, y) \mapsto (\partial g(x), \partial f^*(y)),\\
&M\colon (x, y) \mapsto (A^\top y, -Ax),\\
&\boldsymbol{C}\colon (\boldsymbol{x}, \boldsymbol{y}) \mapsto (\partial g(\boldsymbol{x}(1)), \dots, \partial g(\boldsymbol{x}(n)), \partial f^*(\boldsymbol{y}(1)), \dots, \partial f^*(\boldsymbol{y}(n))),\\
&\boldsymbol{M}\colon (\boldsymbol{x}, \boldsymbol{y}) \mapsto (A^\top \boldsymbol{y}(1), \dots, A^\top \boldsymbol{y}(n), -A\boldsymbol{x}(1), \dots, -A\boldsymbol{x}(n)),\\
&F = C + M,
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{H}\colon (\boldsymbol{x}, \boldsymbol{y}) \mapsto \big(&\tau^{-1}\boldsymbol{x}(1) + A^\top(1 + p_1^{-1})E(1)\boldsymbol{y}(1), \dots,\\
&\tau^{-1}\boldsymbol{x}(n) + A^\top(1 + p_n^{-1})E(n)\boldsymbol{y}(n), D(\sigma)^{-1}\boldsymbol{y}(1), \dots, D(\sigma)^{-1}\boldsymbol{y}(n)\big).
\end{aligned}$$

Let $\boldsymbol{q}_k = (1 \otimes x_k, 1 \otimes y_k)$, $\hat{\boldsymbol{q}}_{k+1} = T(\boldsymbol{q}_k)$ *and* $\hat{z}_{k+1} = (x_{k+1}, \hat{y}_{k+1}) = ((\hat{\boldsymbol{q}}_{k+1})_x(i_k), (\hat{\boldsymbol{q}}_{k+1})_y(1))$. *Then, we have* $(\boldsymbol{H}-\boldsymbol{M})\boldsymbol{q}_k \in (\boldsymbol{C}+\boldsymbol{H})\hat{\boldsymbol{q}}_{k+1}$, $(\boldsymbol{M}-\boldsymbol{H})(\hat{\boldsymbol{q}}_{k+1}-\boldsymbol{q}_k) \in (\boldsymbol{C}+\boldsymbol{M})\hat{\boldsymbol{q}}_{k+1}$, *and*

$$\mathbb{E}_k\left[\operatorname{dist}^2(0, F\hat{z}_{k+1})\right] = \mathbb{E}_k\left[\operatorname{dist}^2(0, (C+M)\hat{z}_{k+1})\right] = \operatorname{dist}^2_{\bar{p}}(0, (\boldsymbol{C}+\boldsymbol{M})\hat{\boldsymbol{q}}_{k+1}).$$

Proof. We start by the representation in Lemma 4.4 by incorporating the update of $\bar{y}_{k+1}$, and recalling the definition of $E(i) = e_i e_i^\top$, $\forall i \in \{1, \dots, n\}$

$$\begin{aligned}\hat{y}(i) &= \operatorname{prox}_{D(\sigma), f^*}(y(i) + D(\sigma)Ax(i))\\ \hat{x}(i) &= \operatorname{prox}_{\tau,g}(x(i) - \tau A^\top\left[y(i) + (1+p_i^{-1})E(i)(\hat{y}(i) - y(i))\right])\\ &= \operatorname{prox}_{\tau,g}(x(i) - \tau A^\top(1+p_i^{-1})E(i)\hat{y}(i) + \tau A^\top(-I_{n\times n} + (1+p_i^{-1})E(i))y(i)).\end{aligned}$$

We now use the definition of proximal operator to obtain

$$\begin{aligned}&D(\sigma)^{-1}y(i) + Ax(i) \in \partial f^*(\hat{y}(i)) + D(\sigma)^{-1}\hat{y}(i)\\ &\tau^{-1}x(i) - A^\top y(i) + A^\top(1+p_i^{-1})E(i)y(i) \in \partial g(\hat{x}(i)) + \tau^{-1}\hat{x}(i) + A^\top(1+p_i^{-1})E(i)\hat{y}(i).\end{aligned}$$

We identify

$$\boldsymbol{H}\boldsymbol{q} = \begin{bmatrix}\tau^{-1}\boldsymbol{x}(1) + A^\top(1+p_1^{-1})E(1)\boldsymbol{y}(1)\\ \vdots\\ \tau^{-1}\boldsymbol{x}(n) + A^\top(1+p_n^{-1})E(n)\boldsymbol{y}(n)\\ D(\sigma)^{-1}\boldsymbol{y}(1)\\ \vdots\\ D(\sigma)^{-1}\boldsymbol{y}(n)\end{bmatrix}, \quad \boldsymbol{M}\boldsymbol{q} = \begin{bmatrix}A^\top\boldsymbol{y}(1)\\ \vdots\\ A^\top\boldsymbol{y}(n)\\ -A\boldsymbol{x}(1)\\ \vdots\\ -A\boldsymbol{x}(n)\end{bmatrix},$$

$$\boldsymbol{C}\hat{\boldsymbol{q}} = \begin{bmatrix}\partial g(\boldsymbol{x}(1))\\ \vdots\\ \partial g(\boldsymbol{x}(n))\\ \partial f^*(\boldsymbol{y}(1))\\ \vdots\\ \partial f^*(\boldsymbol{y}(n))\end{bmatrix}, \quad \boldsymbol{H}\hat{\boldsymbol{q}} = \begin{bmatrix}\tau^{-1}\hat{\boldsymbol{x}}(1) + A^\top(1+p_1^{-1})E(1)\hat{\boldsymbol{y}}(1)\\ \vdots\\ \tau^{-1}\hat{\boldsymbol{x}}(n) + A^\top(1+p_n^{-1})E(n)\hat{\boldsymbol{y}}(n)\\ D(\sigma)^{-1}\hat{\boldsymbol{y}}(1)\\ \vdots\\ D(\sigma)^{-1}\hat{\boldsymbol{y}}(n)\end{bmatrix},$$

and assign $\boldsymbol{q} = \boldsymbol{q}_k$ and $\hat{\boldsymbol{q}} = \hat{\boldsymbol{q}}_{k+1}$, by definition of T in Lemma 4.4 to obtain the first inclusion. The second inclusion follows by adding to both sides $\boldsymbol{M}\hat{\boldsymbol{q}}_{k+1}$ and rearranging.

For the equality, we write

$$\begin{aligned}\mathbb{E}_k\left[\operatorname{dist}^2(0, (C+M)\hat{z}_{k+1})\right] &= \sum_{i=1}^n \operatorname{dist}^2(0, (C+M)\hat{\boldsymbol{q}}_{k+1}(i))p_i\\ &= \operatorname{dist}^2_{\bar{p}}(0, (\boldsymbol{C}+\boldsymbol{M})\hat{\boldsymbol{q}}_{k+1}),\end{aligned}$$

where the first equality follows by $\hat{z}^{k+1} = (x^{k+1}, \hat{y}^{k+1}) = ((\hat{\boldsymbol{q}}_{k+1})_x(i_k), (\hat{\boldsymbol{q}}_{k+1})_y(1))$ and the second equality is by the definitions of C, M, $\boldsymbol{C}$, and $\boldsymbol{M}$ and $(\hat{\boldsymbol{q}}_{k+1})_y(i) = (\hat{\boldsymbol{q}}_{k+1})_y(1)$, $\forall i$. ■

We continue by presenting our assumption for linear convergence (see Section 4.2.2).

Assumption 4.2. Metric subregularity holds for KKT operator F in (4.3) at all $z_\star \in \mathcal{Z}_\star$ for 0 with constant $\eta > 0$ using the norm $\|\cdot\|_S$ with $S = \operatorname{diag}(\tau^{-1}1_p, \sigma_1^{-1}, \dots, \sigma_n^{-1})$, and the neighborhood of regularity $\mathcal{N}(z_\star)$ contains $\hat{z}_k, \forall k$.

In the next theorem, we show that SPDHG with step sizes in (4.8) attains linear convergence with Assumption 4.2. The proof idea is to utilize the negative term $-V(z_k - z_{k-1})$ in (4.10) to obtain contraction. For this, we use the results of Lemmas 4.4 and 4.7 to write this term with the fixed point characterization given in Lemma 4.4, which allows using metric subregularity. The full proof is deferred to Section 4.7.3.

For the proof, define the notations

$$(x_{\star,k-1}, y_{\star,k}) = \arg\min_{(x,y)\in\mathcal{Z}_\star} V_k(x_{k-1} - x, y_k - y),$$

which exists since V_k is a nonnegative quadratic function. We also define

$$\begin{aligned}\Delta_k &= V_{k+1}(x_k - x_{\star,k}, y_{k+1} - y_{\star,k+1}),\\ \Phi_k &= \Delta_k - \frac{C_1}{4\zeta}\|y_k - y_{\star,k}\|^2_{D(\sigma)^{-1}} \ge 0.\end{aligned}$$

Theorem 4.8. *Let Assumptions 4.1 and 4.2 hold. Then it holds that*

$$\mathbb{E}_k[\Delta_k] \le \Delta_{k-1} - V(z_k - z_{k-1}), \tag{4.14}$$

and

$$\mathbb{E}\left[\frac{C_1}{2}\|x_k - x_{\star,k}\|^2_{\tau^{-1}} + \frac{1}{2}\|y_{k+1} - y_{\star,k+1}\|^2_{D(\sigma)^{-1}P^{-1}}\right] \le (1-\rho)^k 2\Phi_0,$$

where, $\rho = \frac{C_1 p}{2\zeta}$, $\zeta = 2 + 2\eta^2\|\boldsymbol{H} - \boldsymbol{M}\|^2$, $C_1 = 1 - \gamma$.

One important remark about Theorem 4.8 is that the knowledge of the metric subregularity constant η is not needed for running the algorithm. Step sizes are chosen as (4.8) and linear convergence follows directly when Assumption 4.2 holds. Important examples where Assumption 4.2 holds are given in Section 4.2.2.

Even though Assumption 4.2 is more general than prior assumptions for linear convergence and our result is agnostic to the choice of the step size, we observe in practice that SPDHG can be much faster than the rate derived in Theorem 4.8. We reflect on this issue more in Chapter 8 and present open questions in this context.

Remark 4.9. Metric subregularity is used in Theorem 4.8 in the weighted norm

$$\operatorname{dist}_S(z, \mathcal{Z}_\star) \leq \eta \operatorname{dist}_S(0, Fz),$$

where $S = \operatorname{diag}(\tau^{-1} 1_p, \sigma_1^{-1}, \ldots, \sigma_n^{-1})$. In view of (4.6) if η_0 is the constant using the standard Euclidean norm, it is obvious that $\eta \leq \|S\| \|S^{-1}\| \eta_0$, but we use η in Theorem 4.8 since it can be smaller, resulting in a better rate.

4.4.3 Sublinear convergence

In this section, we prove optimal convergence rates for the ergodic sequence with different optimality measures. First, we study the expected primal dual gap and next, we study objective value and feasibility for linearly constrained problems.

Convergence of expected primal-dual gap

We recall the definition of the primal-dual gap function,

$$\operatorname{Gap}(\bar{x}, \bar{y}) = \sup_{z \in \mathcal{Z}} \mathcal{H}(\bar{x}, \bar{y}; x, y) := \sup_{z \in \mathcal{Z}} g(\bar{x}) + \langle A\bar{x}, y \rangle - f^*(y) - g(x) - \langle Ax, \bar{y} \rangle + f^*(\bar{y}). \tag{4.15}$$

It is also possible to consider the restricted primal-dual gap in the sense of [CERS18, CP11], which for any set $\mathcal{B} = \mathcal{B}_x \times \mathcal{B}_y \subseteq \mathcal{Z}$ would correspond to

$$\operatorname{Gap}_{\mathcal{B}}(\bar{x}, \bar{y}) = \sup_{z \in \mathcal{B}} \mathcal{H}(\bar{x}, \bar{y}; x, y). \tag{4.16}$$

The standard reference for validity of restricted primal-dual gap is [Nes07, Lemma 1].

The quantity of interest for stochastic algorithms is the expected (restricted) primal-dual gap $\mathbb{E}\left[\operatorname{Gap}_{\mathcal{B}}(\bar{x}, \bar{y})\right]$. As also mentioned in [DL14], showing convergence rate for this quantity is not straightforward, due to the coupling between supremum and expectation, In [DL14], convergence rate is shown in a relaxed quantity called "perturbed gap function". We are not aware of any results for a PDCD method with $\mathcal{O}(1/k)$ rate for expected primal-dual gap.

Even though this result was claimed in [CERS18], the proof has a technical issue, near the end of the proof in [CERS18, Theorem 4.3][1]. Since the supremum of expectation is upper bounded by the expectation of the supremum, which is in the definition of expected primal-dual gap, the order of expectation in the proof is incorrect. As we could not find a simple way of fixing the issue using the existing techniques, we introduce a new technique and provide a proof to show that the conclusions of [CERS18, Theorem 4.3], for the primal-dual gap, are still correct, with different constants in the bound.

Our technique in the following proof is inspired by the stochastic approximation literature of

[1] We communicated this with the authors who acknowledged the mistake

variational inequalities and saddle point problems (see [NJLS09, Lemma 3.1] for a reference). In this reference and followup works, such an analysis is used to obtain $\mathcal{O}(1/\sqrt{k})$ rates with decreasing step size and SGD-based methods. In the new proof, we adapt this idea by using the structure of PDCD to obtain the optimal $\mathcal{O}(1/k)$ rate of convergence with constant step size. Our technique uses Euclidean structure of the dual update of SPDHG, therefore might not be directly applicable to cases with Bregman distances being used for proximal operator.

We start with the lemma to decouple supremum and expectation in the proof.

Lemma 4.10. *Given a point $\tilde{y}_1 \in \mathcal{Y}$, for $k \geq 1$, we define the sequences*

$$v_{k+1} = y_k - \hat{y}_{k+1} - P^{-1}(y_k - y_{k+1}), \quad \textit{and,} \quad \tilde{y}_{k+1} = \tilde{y}_k - P v_{k+1}. \tag{4.17}$$

Then, we have for any $y \in \mathcal{Y}$,

$$\sum_{k=1}^{K} \langle \tilde{y}_k - y, v_{k+1} \rangle_{D(\sigma)^{-1}} \leq \frac{1}{2} \| \tilde{y}_1 - y \|^2_{D(\sigma)^{-1}P^{-1}} + \sum_{k=1}^{K} \frac{1}{2} \| v_{k+1} \|^2_{D(\sigma)^{-1}P}, \tag{4.18}$$

$$\mathbb{E}\left[\sum_{k=1}^{K} \frac{1}{2} \| v_{k+1} \|^2_{D(\sigma)^{-1}P} \right] \leq \frac{1}{C_1} \Delta_0. \tag{4.19}$$

Moreover, v_k and $\tilde{y}_k$ are $\mathcal{F}_k$-measurable and $\mathbb{E}_k \left[v_{k+1} \right] = 0$.

Proof. For brevity in this proof, we denote $\Upsilon = D(\sigma)^{-1}P^{-1}$. We have $\forall y \in \mathcal{Y}$,

$$\begin{aligned} \frac{1}{2} \| \tilde{y}_{k+1} - y \|^2_{\Upsilon} &= \frac{1}{2} \| \tilde{y}_k - y \|^2_{\Upsilon} - \langle P v_{k+1}, \tilde{y}_k - y \rangle_{\Upsilon} + \frac{1}{2} \| P v_{k+1} \|^2_{\Upsilon} \\ &= \frac{1}{2} \| \tilde{y}_k - y \|^2_{D(\sigma)^{-1}P^{-1}} - \langle v_{k+1}, \tilde{y}_k - y \rangle_{D(\sigma)^{-1}} + \frac{1}{2} \| v_{k+1} \|^2_{D(\sigma)^{-1}P}. \end{aligned}$$

Summing this equality gives the first result.

For the second result, we use $\mathbb{E}_k \left[P^{-1}(y_k - y_{k+1}) \right] = y_k - \hat{y}_{k+1}$, law of total expectation, and the definition of variance,

$$\begin{aligned} \mathbb{E}\left[\sum_{k=1}^{K} \frac{1}{2} \| v_{k+1} \|^2_{D(\sigma)^{-1}P} \right] &= \sum_{k=1}^{K} \frac{1}{2} \mathbb{E}\left[\mathbb{E}_k \left[\| v_{k+1} \|^2_{D(\sigma)^{-1}P} \right] \right] \\ &\leq \sum_{k=1}^{K} \frac{1}{2} \mathbb{E}\left[\mathbb{E}_k \left[\| P^{-1}(y_{k+1} - y_k) \|^2_{D(\sigma)^{-1}P} \right] \right] \\ &= \sum_{k=1}^{K} \frac{1}{2} \mathbb{E}\left[\| y_{k+1} - y_k \|^2_{D(\sigma)^{-1}P^{-1}} \right] \leq \frac{1}{C_1} \Delta_0, \end{aligned}$$

where the last inequality follows by $\sum_{k=1}^{\infty} \mathbb{E}\left[V(z_{k+1} - z_k) \right] \leq \Delta_0$ from Theorem 4.6 and $\frac{1}{2} \| y_{k+1} - y_k \|^2_{D(\sigma)^{-1}P^{-1}} \leq \frac{1}{1-\gamma} V(z_{k+1} - z_k)$ from Lemma 4.3.

Other results follow by the definition of the sequences and $\mathbb{E}_k \left[y_{k+1} - y_k \right] = P(\hat{y}_{k+1} - y_k)$. ■

We now describe our proof strategy for handling the abovementioned difficulty. Proof of Lemma 4.3, given in Section 4.7.1, proceeds by developing terms involving random quantities, by utilizing conditional expectations. In this case, however, our approach is to proceed without using conditional expectation since the quantity of interest requires us to take first supremum and then the expectation of the estimates. Our proof strategy will be to characterize the error term, and then utilize the results Lemma 4.10 to decouple and bound this term. First, we give the variant of Lemma 4.3 without taking expectations, with its proof given in Section 4.7.4.

Lemma 4.11. *We define* $f_P^*(y) = \sum_{i=1}^n p_i f_i^*(y^{(i)})$, *and similar to* (4.5) $D_{f^*}^P(\bar{y}, z) = \sum_{i=1}^n p_i f_i^*(\bar{y}^{(i)}) - p_i f_i^*(y^{(i)}) - \langle (Ax)_i, p_i(\bar{y}-y)^{(i)} \rangle$ *and recall the definitions of* V *and* V_k *from* (4.9) *and* $\mathcal{H}$ *from* (4.15).

Then, it holds that

$$\begin{aligned}\mathcal{H}(x_k, y_{k+1}; x, y) \le{}& V_k(x_{k-1} - x, y_k - y) - V_{k+1}(x_k - x, y_{k+1} - y) - V(z_k - z_{k-1}) \\ &+ \mathcal{E}_k + D_{f^*}^{P^{-1}-I}(y_k, z) - D_{f^*}^{P^{-1}-I}(y_{k+1}, z) - \langle y, v_{k+1} \rangle_{D(\sigma)^{-1}},\end{aligned} \tag{4.20}$$

where $v_{k+1} = y_k - \hat{y}_{k+1} - P^{-1}(y_k - y_{k+1})$ *and*

$$\begin{aligned}\mathcal{E}_k ={}& \frac{1}{2}\left[\|y_k\|^2_{D(\sigma)^{-1}} - \|\hat{y}_{k+1}\|^2_{D(\sigma)^{-1}} - \left(\|y_k\|^2_{D(\sigma)^{-1}P^{-1}} - \|y_{k+1}\|^2_{D(\sigma)^{-1}P^{-1}}\right)\right] \\ &+ \frac{1}{2}\|y_{k+1} - y_k\|^2_{D(\sigma)^{-1}P^{-1}} - \frac{1}{2}\|\hat{y}_{k+1} - y_k\|^2_{D(\sigma)^{-1}} + f^*(y_k) - f^*(\hat{y}_{k+1}) \\ &- (f^*_{P^{-1}}(y_k) - f^*_{P^{-1}}(y_{k+1})) - \langle Ax_k, y_k - \hat{y}_{k+1} - P^{-1}(y_k - y_{k+1}) \rangle,\end{aligned} \tag{4.21}$$

and also $\mathbb{E}_k[\mathcal{E}_k] = 0$.

With this lemma, we have identified the problematic inner product term for deriving the rate for expected gap, which is $\langle y, v_{k+1} \rangle$ in (4.20)). This is the only term coupling the free variable z and random term v_{k+1}. In the next theorem, we use Lemma 4.10 to manipulate this inner product. In particular, the idea in Lemma 4.10 was to bound the error term by $\|y_k - y_{k+1}\|^2$ which is proven to be small in Theorem 4.6, which is due to using PDCD updates. For the rest of the terms in (4.20), we observe that the terms with V_k will telescope and $\mathcal{E}_k$ has expectation 0 and it is independent of free variable z.

Theorem 4.12. *Let Assumption 4.1 hold. Define the sequences* $x_K^{\text{avg}} = \frac{1}{K}\sum_{k=1}^K x_k$ *and* $y_{K+1}^{\text{avg}} = \frac{1}{K}\sum_{k=1}^K y_{k+1}$, *where* x_k, y_k *are generated by SPDHG and recall the definition of* $\mathcal{H}$ *from* (4.15).

Then, for any bounded set $\mathcal{B} = \mathcal{B}_x \times \mathcal{B}_y \subseteq \mathcal{Z}$, *the following result holds for the expected primal dual gap defined in* (4.15)

$$\mathbb{E}\left[\sup_{z\in\mathcal{B}} \mathcal{H}(x_K^{\text{avg}}, y_{K+1}^{\text{avg}}; x, y)\right] = \mathbb{E}\left[\text{Gap}_{\mathcal{B}}(x_K^{\text{avg}}, y_{K+1}^{\text{avg}})\right] \le \frac{C_{\mathcal{B}}}{K}, \tag{4.22}$$

where

$$C_{\mathcal{B}} = \frac{3}{2} \sup_{x \in \mathcal{B}_x} \|x_0 - x\|^2_{\tau^{-1}} + \sup_{y \in \mathcal{B}_y} \|y_1 - y\|^2_{D(\sigma)^{-1}P^{-1}} + f^*_{P^{-1}-I}(y_1) + \left(\frac{1}{C_1} + \frac{2\gamma}{\underline{p}}\right)\Delta_0$$
$$+ \gamma\|x_0\|^2_{\tau^{-1}} + \frac{\gamma}{\underline{p}}\|y_1 - y_\star\|^2_{D(\sigma)^{-1}P^{-1}} + \sum_{i=1}^{n}\left(\frac{1}{p_i} - 1\right)\Big(-f_i^*(y_\star^{(i)}) + \|A_i x_\star\|_{D(\sigma)P}\sqrt{2\Delta_0}\Big).$$

Moreover, for the smoothed gap function (see (4.7)*), it holds that*

$$\mathbb{E}\left[\mathcal{G}_{\frac{1+2\gamma}{2K},\frac{1}{2K}}(x_K^{\text{avg}}, y_{K+1}^{\text{avg}}; x_0, y_1)\right] \le \frac{C_e}{K},$$

where

$$C_e = \gamma\|x_0\|^2_{\tau^{-1}} + \frac{\gamma}{\underline{p}}\|y_1 - y_\star\|^2_{D(\sigma)^{-1}P^{-1}} + \left(\frac{1}{C_1} + \frac{2\gamma}{\underline{p}}\right)\Delta_0$$
$$+ f^*_{P^{-1}-I}(y_1) + \sum_{i=1}^{n}\left(\frac{1}{p_i} - 1\right)\Big(-f_i^*(y_\star^{(i)}) + \|A_i x_\star\|_{D(\sigma)P}\sqrt{2\Delta_0}\Big).$$

We defer the proof to Section 4.7.5 due to its length. However, we remark that the main difficulty is solved by characterizing the error term in Lemma 4.11 and bounding it due to Lemma 4.10, as explained already. The rest of the proof estimates the exact constants.

Convergence of objective values

The guarantee for the expected primal-dual gap, which is recovered by setting $\mathcal{B} = \mathcal{X} \times \mathcal{Y}$ in (4.22) requires bounded primal and dual domains. In this section, we show that $\mathcal{O}(1/k)$ rate of convergence in terms of objective values and/or feasibility can be shown with possibly unbounded primal and dual domains.

Theorem 4.13. *Let Assumption 4.1 hold. We recall $x_K^{\text{avg}} = \frac{1}{K}\sum_{k=1}^K x_k$.*

- *If f is $L(f)$-Lipschitz continuous (whence the dual domain is bounded), and $y_1 \in \operatorname{dom} f^*$,*

$$\mathbb{E}\left[f(Ax_K^{\text{avg}}) + g(x_K^{\text{avg}}) - f(Ax_\star) - g(x_\star)\right] \le \frac{C_{e,1}}{K}.$$

- *If $f(\cdot) = \delta_{\{b\}}(\cdot)$ with $b \in \mathcal{Y}$,*

$$\mathbb{E}\left[g(x_K^{\text{avg}}) - g(x_\star)\right] \le \frac{C_{e,2}}{K}, \quad \mathbb{E}\left[\|Ax_K^{\text{avg}} - b\|_{D(\sigma)P}\right] \le \frac{C_{e,3}}{K},$$

where C_e is as defined in Theorem 4.12 and $C_{e,1} = C_e + \frac{2}{\underline{p}}L(f)^2 + \frac{1+2\gamma}{2}\|x_0 - x_\star\|^2_{\tau^{-1}}$,
$C_{e,3} = \frac{1}{2}\left\{\|y_\star - y_1\|_{D(\sigma)^{-1}P^{-1}} + \left(\|y_\star - y_1\|^2_{D(\sigma)^{-1}P^{-1}} + 4C_e + 6\|x_\star - x_0\|_{\tau^{-1}}\right)^{1/2}\right\}$,
$C_{e,2} = C_e + \frac{1}{2}\|y_\star - y_1\|^2_{D(\sigma)^{-1}P^{-1}} + \frac{1+2\gamma}{2}\|x_0 - x_\star\|^2_{\tau^{-1}} + \|y_\star\|_{D(\sigma)^{-1}P^{-1}}C_{e,3}$.

The proof of the theorem is a basic consequence of Theorem 4.12 and [TDFC18, Lemma 1]. We provide the proof in Section 4.7.6.

4.5 Related works

We summarize the comparison of the most related PDCD methods in Table 5.1 (Page 106).

Primal camp. Stochastic gradient based methods (SGD) can be applied to solve (4.1) [RM51, NJLS09]. However, this approach cannot get linear convergence except special cases [NRP19] due to properties of SGD. One alternative is variance reduction to obtain linear convergence under the assumption that functions f_i are smooth and g is strongly convex or f_i are smooth and strongly convex [JZ13, XZ14, AZ17]. Smoothness of f_i is equivalent to strong convexity of f_i^*. Therefore, the linear convergence results of these methods require the similar assumptions as [CERS18, Theorem 6.1]. Moreover, as in [CERS18], variance reduction based methods require knowing μ_i and μ_g to set the algorithm parameters to obtain linear convergence.

For the specific case of $f_i(\cdot) = \delta_{b_i}(\cdot)$, SGD-type methods are proposed in [PN17, Xu20, FANC19]. However, these methods only obtain $\mathcal{O}(1/k)$ rate with strong convexity of g, since they focus on the general problem where the objective can be given in expectation form. Even though this rate is optimal for their template, it is suboptimal for (4.1).

Primal-dual camp. A line of research utilizes coordinate descent type of schemes for solving (4.1). Coordinate descent with random sampling for unconstrained optimization is proposed in [Nes12] and later generalized and improved in [RT14, FR15]. These methods apply coordinate descent in the primal and obtain linear convergence rates with smooth and strongly convex f_i or smooth f_i and strongly convex g.

Another approach is to apply coordinate ascent in the dual to exploit separability of the dual in (4.1). Stochastic dual coordinate ascent (SDCA) and its accelerated variant are proposed in [SSZ13, SSZ14]. These methods require smoothness of f_i and strong convexity of g for linear convergence and the parameters depend on the smoothness and strong convexity constants.

SPDHG that we analyzed in this chapter is proposed in [CERS18]. The authors proved linear convergence of the modified method SPDHG-μ [CERS18, Theorem 6.1] by assuming strong convexity of f_i^*, g and special step sizes depending on strong convexity constants. Asymptotic convergence and the $\mathcal{O}(1/k)$ rate results in [CERS18, Theorem 4.3] are given in terms of Bregman distances which is not a valid and standard optimality measure. We prove linear convergence with standard step sizes in (4.8) and with weaker metric subregularity assumption, detailed in Section 4.2.2. Moreover, in the general convex case, we prove almost sure convergence of the iterates to a solution, which is the standard result and stronger than the corresponding result in [CERS18] with Bregman distances. Finally, we prove $\mathcal{O}(1/k)$ rate, with possibly unbounded domains, for the standard optimality measure expected primal-dual gap. The comparison of the results is also summarized in Table 4.2.

PDCD methods similar to SPDHG are proposed in [ZX17, DL14, FB19]. These variants assume strong convexity of f_i^*, g to guarantee linear convergence. Only [FB19] proved linear convergence with step sizes independent of strong convexity constants which provided a partial answer for adaptivity of SPDHG-type methods to strong convexity. However, as detailed in Table 5.1, with dense A matrix, and uniform sampling, this method requires step sizes n times smaller than (4.8) which is problematic in practice (see Section 4.6.1). For sublinear convergence, [FB19] proved $\mathcal{O}(1/\sqrt{k})$ rate on a randomly selected iterate, under similar assumption to ours whereas [ZX17] requires boundedness of dual domain, setting a horizon, and gives primal-only complexities (not anytime rates).

PDCD algorithms are also studied in [CP15, CP19, PR15]. As mentioned in [FB19, CERS18], operator theory-based proofs of these methods require using small step sizes depending on global information, which causes slow performance in practice.

CD methods for linearly constrained problems are studied in Chapter 3 and [DL14, LM18]. These methods obtain only sublinear convergence rates. In Theorem 4.13, the specific case of $f(\cdot) = \delta_{\{b\}}(\cdot)$ is studied in [LM18] and a similar result was derived. The rate in [LM18] has a different nature in the sense that it is an almost sure rate where the constant depends on trajectory, whereas our rate is in expectation.

Latafat et al. [LFP19] proposed a method called TriPD-BC and proved linear convergence for their method under metric subregularity. There exist two drawbacks of TriPD-BC for our setting. First, when A is not of special structure, such as block diagonal, one needs to use a complicated duplication strategy for an efficient implementation (see [FB19]). Second issue is that as in [FB19], this method needs to use n times smaller step sizes with dense A. For the details of duplication and small step sizes, we refer to [FB19]. The need to use small step sizes seriously affects the practical performance of the algorithm as illustrated in Section 4.6.1.

Some standard references for deterministic primal-dual algorithms are [CP11, CP16b, HY12, TDFC18, TDAFC19, EZC10]. As observed in [CERS18], coordinate descent-based variants significantly improve the practical performance of these deterministic methods.

Our results imply global linear convergence for deterministic PDHG when $n = 1$, answering the question posed in [CP11]: *"It would be interesting to understand whether the steps can be estimated in Algorithm 1 without the a priori knowledge of μ_i, μ_g."* We note that in the third part of Assumption 4.2, compact domains are not needed for this case. We highlight that such behaviour of deterministic primal-dual methods is investigated before in [LFP16, LFP19].

Linear programming. A related notion to metric subregularity for linear programming is Hoffman's lemma due to classical result in [Hof52], which is used by many researchers to show linear convergence of ADMM-type methods for LPs [YZH$^+$15, YH16, LYZZ18]. The drawback of these approaches is that one needs to know the constant η to run the algorithm which is difficult to estimate in general. Our analysis recovers these results specific to LPs with a much simpler algorithm that does not need the knowledge of η.

4.6 Numerical evidence

In this section, we support our theoretical findings by showing that SPDHG with step sizes in (4.8) obtains linear convergence for problems satisfying metric subregularity. The problems we solve, namely, basis pursuit, Lasso, and ridge regression all satisfy metric subregularity. Among them, only ridge regression is strongly convex-strongly concave, thus this is the only problem where linear convergence results from [CERS18] apply by using SPDHG-μ [CERS18, Theorem 6.1]. We show that even in this case, when strong convexity constants are small, applying SPDHG can be more beneficial for some datasets. SPDHG-μ is not applicable for other problems due to lack of strong convexity either in the primal or dual. We illustrate favorable behavior of SPDHG against popular methods SVRG [JZ13] and accelerated SVRG [ZSC18].

For space limitations, we include results with one or two datasets for each problem. For SPDHG, as in [CERS18], we use uniform sampling of coordinates and $\tau = \frac{\gamma}{n \max_i \|A_i\|}$ and $\sigma_i = \frac{\gamma}{\|A_i\|}$, with $\gamma = 0.99$ for all problems. For the other methods, we use the suggested theoretical step sizes in the respective papers and we do not fine tune any of the methods.

4.6.1 Sparse recovery with basis pursuit

We first solve the basis pursuit problem which is a fundamental problem in signal processing [CDS01] and also finds applications in machine learning [GS18, AKSV18]:

$$\min_{x \in \mathbb{R}^d} \|x\|_1 : Ax = b. \tag{4.23}$$

Since basis pursuit is PLQ, metric subregularity holds. The aim in this section is to illustrate the difference on the step sizes mentioned in Section 4.5, Table 5.1 and verify the empirical linear convergence of SPDHG. We compare SPDHG with coordinate descent version of Vu-Condat algorithm, developed in [FB19], which we refer to as FB-VC-CD. Note that [LFP19] requires duplication for an efficient implementation for this problem and it uses the same step sizes as [FB19]. For this reason, we only compare with FB-VC-CD and note that the practical performance of [LFP19] is expected to be similar to FB-VC-CD with same step sizes.

We generate the data matrix A synthetically where $n = 500$ and $d = 1000$ and entries of data matrix follow a normal distribution. We generate a covariance matrix $\Sigma_{i,j} = \rho^{|i-j|}$ with $\rho = 0.5$ and a sparse solution $x_\star$ with 100 nonzero entries. We then compute $b = Ax_\star$.

The analysis of SPDHG in [CERS18] shows $\mathcal{O}(1/k)$ rate on the Bregman distance to solution on the ergodic sequence whereas our analysis shows linear convergence on the last iterate. FB-VC-CD is proven to have $\mathcal{O}(1/\sqrt{k})$ rate for this problem [FB19]. FB-VC-CD is specially designed to exploit sparsity in the data. However, in our case the data is dense, which causes FB-VC-CD to use n times smaller step sizes. As shown in Figure 4.1, FB-VC-CD exhibits a slow rate whereas SPDHG converges faster, as predicted by our theoretical results.

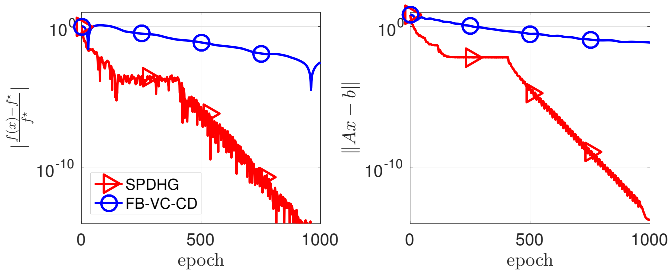

Figure 4.1 – Linear convergence of SPDHG for basis pursuit problem.

4.6.2 Lasso and ridge regression

In this section we solve ridge regression and Lasso problems, formulated as

$$\min_{x\in\mathbb{R}^d} \frac{1}{2}\|Ax-b\|^2 + \frac{\lambda}{2}\|x\|^2, \qquad \text{and,} \qquad \min_{x\in\mathbb{R}^d} \frac{1}{2}\|Ax-b\|^2 + \lambda\|x\|_1, \tag{4.24}$$

respectively. In terms of structure, the first problem is smooth and strongly convex, or equivalently, its Lagrangian is strongly convex-strongly concave. For this problem class, [CERS18] showed linear convergence for the method SPDHG-μ, which is a modified version of SPDHG using strong convexity and strong concavity constants for step sizes. In addition, SVRG and accelerated SVRG obtain linear convergence for this problem [XZ14, ZSC18, AZ17].

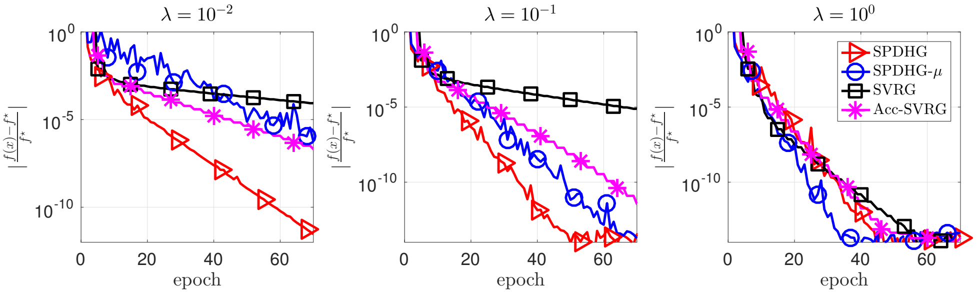

Figure 4.2 – Ridge regression, YearPredictionMSD, $n = 463,715, d = 90$.

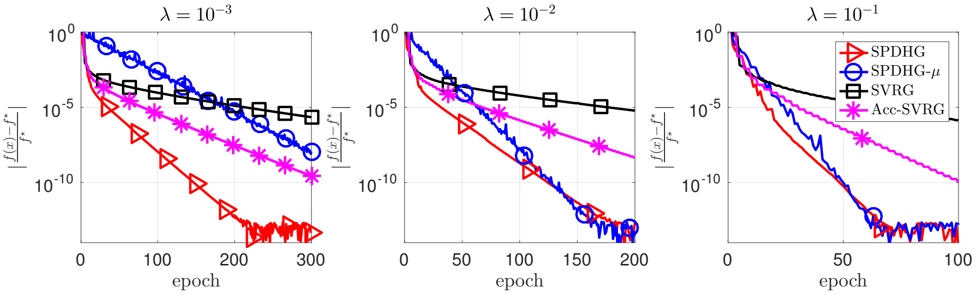

Figure 4.3 – Ridge regression, w8a, $n = 49,749, d = 300$.

We use regression datasets from libsvm [CL11b], perform row normalization, and use three different regularization parameters for each case and compile the results in Figures 4.2–4.4

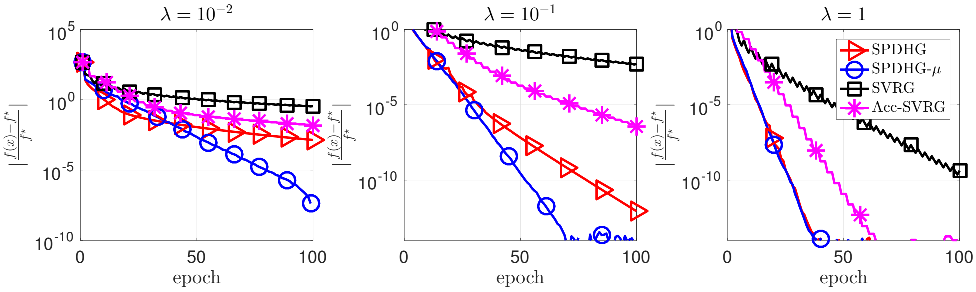

Figure 4.4 – Ridge regression, sector, $n = 6,412$, $d = 55,197$.

along with information on datasets and regularization parameters.

In the regime $n \geq d$, we observe in Figures 4.2 and 4.3 that, for large λ, or equivalently, large strong convexity constants, SPDHG-μ is faster than SPDHG, which is expected since SPDHG-μ is designed to use strong convexity, whereas our result holds generically without any modifications on the algorithm. However, we observe that, especially when λ is small, SPDHG gets a faster linear rate than SPDHG-μ, which may suggest robustness of SPDHG over SPDHG-μ. SPDHG exhibits competitive performance against SVRG and accelerated SVRG.

Goal of this experiment is not to argue that SPDHG gets the best performance in all cases since this is a very specific instance where most algorithms convergence linearly. Our goal is rather to illustrate that even though our linear convergence results apply to a broad class of problems, SPDHG is still competitive when compared to methods which are designed to exploit the structure of this specific setting.

In the regime $n \leq d$, we observe in Figure 4.4 that SPDHG-μ shows a faster behavior with small μ. This seems intuitive, since in this case the strong convexity *purely* comes from the regularization term and SPDHG-μ directly exploits this knowledge to get a better performance.

We then solve Lasso in (4.24), for which SPDHG-μ does not apply and accelerated SVRG cannot get linear rates in general. We compare with SVRG for varying regularization parameters, datasets with $n \leq d$ or $n \geq d$, and compile the results in Sections 4.6.2 and 4.6.2. We observe that SPDHG converges linearly and exhibits a better practical performance than SVRG.

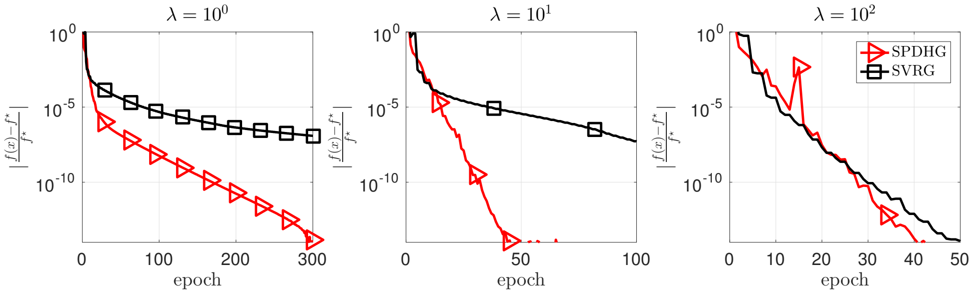

Figure 4.5 – Lasso, mnist scale, $n = 60,000, d = 780$.

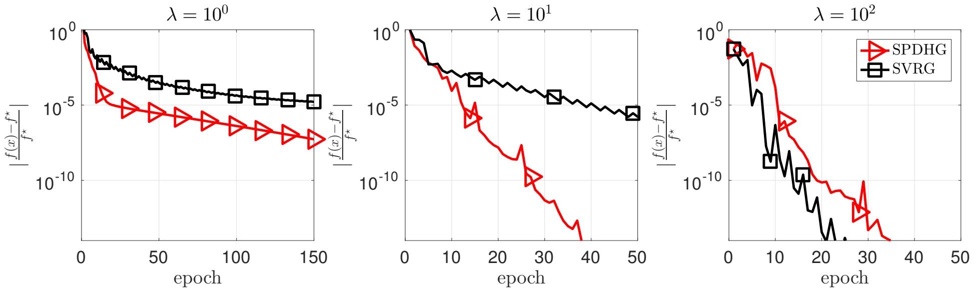

Figure 4.6 – Lasso, rcv1.binary, $n = 20,242$, $d = 47,236$.

4.7 Proofs

4.7.1 Proof of Lemma 4.3

Proof. As in [CERS18], we define the iterate $\hat{y}$ with full dimensional update

$$x_k = \mathrm{prox}_{\tau,g}(x_{k-1} - \tau A^\top \bar{y}_k), \tag{4.25}$$

$$\hat{y}_{k+1}^{(i)} = \mathrm{prox}_{\sigma_i, f_i^*}(y_k^{(i)} + \sigma_i A_i x_k). \tag{4.26}$$

By (1.7), we get, $\forall x \in \mathcal{X}$ and $\forall y \in \mathcal{Y}$ and $\forall i = \{1, \dots, n\}$

$$g(x) \geq g(x_k) + \langle x_k - x, A^\top \bar{y}_k\rangle + \frac{1}{2}\|x_k - x_{k-1}\|^2_{\tau^{-1}} + \frac{1}{2}\|x_k - x\|^2_{\tau^{-1}} - \frac{1}{2}\|x - x_{k-1}\|^2_{\tau^{-1}},$$

$$f_i^*(y^{(i)}) \geq f^*(\hat{y}_{k+1}^{(i)}) - \langle \hat{y}_{k+1}^{(i)} - y^{(i)}, A_i x_k\rangle + \frac{1}{2}\|\hat{y}_{k+1}^{(i)} - y_k^{(i)}\|^2_{\sigma_i^{-1}} + \frac{1}{2}\|\hat{y}_{k+1}^{(i)} - y^{(i)}\|^2_{\sigma_i^{-1}} - \frac{1}{2}\|y^{(i)} - y_k^{(i)}\|^2_{\sigma_i^{-1}}.$$

We sum the second inequality from $i = 1$ to n and add to the first inequality to obtain

$$\begin{aligned} 0 \geq\ & g(x_k) - g(x) + \langle x_k - x, A^\top \bar{y}_k\rangle + f^*(\hat{y}_{k+1}) - f^*(y) - \langle \hat{y}_{k+1} - y, Ax_k\rangle \\ & + \frac{1}{2}\left(-\|x_{k-1} - x\|^2_{\tau^{-1}} + \|x_k - x\|^2_{\tau^{-1}} + \|x_k - x_{k-1}\|^2_{\tau^{-1}}\right) \\ & + \frac{1}{2}\left(-\|y_k - y\|^2_{D(\sigma)^{-1}} + \|\hat{y}_{k+1} - y\|^2_{D(\sigma)^{-1}} + \|\hat{y}_{k+1} - y_k\|^2_{D(\sigma)^{-1}}\right). \end{aligned} \tag{4.27}$$

We recall

$$D_g(x_k, z) = g(x_k) - g(x) + \langle A^\top y, x_k - x\rangle, \tag{4.28}$$

$$D_{f^*}(\hat{y}_{k+1}, z) = f^*(\hat{y}_{k+1}) - f^*(y) - \langle Ax, \hat{y}_{k+1} - y\rangle. \tag{4.29}$$

We add and subtract $\langle A^\top y, x_k - x\rangle - \langle Ax, \hat{y}_{k+1} - y\rangle$ on the right hand side of (4.27) and use the definitions in (4.28) and (4.29) to get

$$0 \geq D_g(x_k, z) + D_{f^*}(\hat{y}_{k+1}, z) + \langle x_k - x, A^\top(\bar{y}_k - y)\rangle - \langle \hat{y}_{k+1} - y, A(x_k - x)\rangle$$

$$
\begin{aligned}
&+\frac{1}{2}\left(-\|x_{k-1}-x\|^2_{\tau^{-1}}+\|x_k-x\|^2_{\tau^{-1}}+\|x_k-x_{k-1}\|^2_{\tau^{-1}}\right)\\
&+\frac{1}{2}\left(-\|y_k-y\|^2_{D(\sigma)^{-1}}+\|\hat{y}_{k+1}-y\|^2_{D(\sigma)^{-1}}+\|\hat{y}_{k+1}-y_k\|^2_{D(\sigma)^{-1}}\right). \qquad (4.30)
\end{aligned}
$$

Note that at step k of SPDHG in Algorithm 4.1, we select an index $i_k \in \{1,\dots,n\}$ randomly with probability p_{i_k} and perform the following step on the dual variable

$$y_{k+1}^{(i_k)} = \hat{y}_{k+1}^{(i_k)}, \text{ and } y_{k+1}^{(i)} = y_k^{(i)}, \forall i \neq i_k. \qquad (4.31)$$

For any $Y \in \mathcal{Y}$ that is measurable with respect to $\mathcal{F}_k$, (4.31) immediately gives

$$
\begin{aligned}
&\mathbb{E}_k[y_{k+1}] = P\hat{y}_{k+1} + (I-P)\,y_k, \qquad (4.32)\\
&\mathbb{E}_k\left[\|y_{k+1}-Y\|^2_{D(\sigma)^{-1}}\right] = \|\hat{y}_{k+1}-Y\|^2_{D(\sigma)^{-1}P} + \|y_k-Y\|^2_{D(\sigma)^{-1}(I-P)}. \qquad (4.33)
\end{aligned}
$$

A simple manipulation of (4.32) and plugging in $Y=y$ and $Y=y_k$ in (4.33) respectively, gives

$$
\begin{aligned}
&\hat{y}_{k+1} = P^{-1}\mathbb{E}_k[y_{k+1}] - (P^{-1}-I)y_k \qquad (4.34)\\
&\|\hat{y}_{k+1}-y\|^2_{D(\sigma)^{-1}} = \mathbb{E}_k\left[\|y_{k+1}-y\|^2_{D(\sigma)^{-1}P^{-1}}\right] - \|y_k-y\|^2_{D(\sigma)^{-1}(P^{-1}-I)} \qquad (4.35)\\
&\|\hat{y}_{k+1}-y_k\|^2_{D(\sigma)^{-1}} = \mathbb{E}_k\left[\|y_{k+1}-y_k\|^2_{D(\sigma)^{-1}P^{-1}}\right]. \qquad (4.36)
\end{aligned}
$$

We apply (4.35) and (4.36) to the last line of (4.30) to get

$$
\begin{aligned}
&\frac{1}{2}\left(-\|y_k-y\|^2_{D(\sigma)^{-1}}+\|\hat{y}_{k+1}-y\|^2_{D(\sigma)^{-1}}+\|\hat{y}_{k+1}-y_k\|^2_{D(\sigma)^{-1}}\right)\\
&\quad=\frac{1}{2}\left(-\|y_k-y\|^2_{D(\sigma)^{-1}P^{-1}}+\mathbb{E}_k\left[\|y_{k+1}-y\|^2_{D(\sigma)^{-1}P^{-1}}+\|y_{k+1}-y_k\|^2_{D(\sigma)^{-1}P^{-1}}\right]\right). \qquad (4.37)
\end{aligned}
$$

In addition, we have for the bilinear term in (4.30) that

$$
\begin{aligned}
\langle x_k-x, A^\top(\bar{y}_k-y)\rangle - \langle \hat{y}_{k+1}-y, A(x_k-x)\rangle &= \langle A(x_k-x), \bar{y}_k-\hat{y}_{k+1}\rangle\\
&= \langle A(x_k-x), \bar{y}_k - P^{-1}\mathbb{E}_k[y_{k+1}] + (P^{-1}-I)y_k\rangle\\
&= -\mathbb{E}_k\left[\langle A(x_k-x), P^{-1}(y_{k+1}-y_k)\rangle\right] + \langle A(x_k-x), P^{-1}(y_k-y_{k-1})\rangle\\
&= -\mathbb{E}_k\left[\langle A(x_k-x), P^{-1}(y_{k+1}-y_k)\rangle\right] + \langle A(x_{k-1}-x), P^{-1}(y_k-y_{k-1})\rangle\\
&\quad + \langle A(x_k-x_{k-1}), P^{-1}(y_k-y_{k-1})\rangle. \qquad (4.38)
\end{aligned}
$$

where the second equality is by (4.34), and third equality is by the definition of $\bar{y}_k$.

We now insert (4.37) and (4.38) into (4.30) and also add and subtract $\frac{1}{2}\|y_k-y_{k-1}\|^2_{D(\sigma)^{-1}P^{-1}}$

$$
\begin{aligned}
0 \geq D_g(x_k,z) + D_{f^*}(\hat{y}_{k+1},z) &+ \frac{1}{2}\|y_k-y_{k-1}\|^2_{D(\sigma)^{-1}P^{-1}} - \frac{1}{2}\|y_k-y_{k-1}\|^2_{D(\sigma)^{-1}P^{-1}}\\
&- \mathbb{E}_k\left[\langle A(x_k-x), P^{-1}(y_{k+1}-y_k)\rangle\right] + \langle A(x_{k-1}-x), P^{-1}(y_k-y_{k-1})\rangle\\
&+ \langle A(x_k-x_{k-1}), P^{-1}(y_k-y_{k-1})\rangle
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{1}{2}\big(-\|x_{k-1}-x\|^2_{\tau^{-1}}+\|x_k-x\|^2_{\tau^{-1}}+\|x_k-x_{k-1}\|^2_{\tau^{-1}}\big)\\
&+\frac{1}{2}\Bigg(-\|y_k-y\|^2_{D(\sigma)^{-1}P^{-1}}+\mathbb{E}_k\left[\|y_{k+1}-y\|^2_{D(\sigma)^{-1}P^{-1}}\right]\\
&+\mathbb{E}_k\left[\|y_{k+1}-y_k\|^2_{D(\sigma)^{-1}P^{-1}}\right]\Bigg).
\end{aligned}
\tag{4.39}
$$

The first result follows by using the definitions of V and V_k.

It is straightforward to prove (4.11) and (4.12). Since $y_k^{(j)}=y_{k-1}^{(j)}, \forall j\neq i_{k-1}$,

$$
\begin{aligned}
|\langle Ax,P^{-1}(y_k-y_{k-1})\rangle| &= |\langle A_{i_{k-1}}x, p^{-1}_{i_{k-1}}(y_k^{(i_{k-1})}-y_{k-1}^{(i_{k-1})})\rangle|\\
&\leq \|A_{i_{k-1}}x\|p^{-1}_{i_{k-1}}\|y_k^{(i_{k-1})}-y_{k-1}^{(i_{k-1})}\|\\
&=\left(\tau^{1/2}\sigma^{1/2}_{i_{k-1}}p^{-1/2}_{i_{k-1}}\|A_{i_{k-1}}\|\right)\tau^{-1/2}\|x\|p^{-1/2}_{i_{k-1}}\sigma^{-1/2}_{i_{k-1}}\|y_k^{(i_{k-1})}-y_{k-1}^{(i_{k-1})}\|\\
&\leq\gamma\left(\tau^{-1/2}\|x\|p^{-1/2}_{i_{k-1}}\sigma^{-1/2}_{i_{k-1}}\|y_k^{(i_{k-1})}-y_{k-1}^{(i_{k-1})}\|\right)\\
&\leq\frac{\gamma}{2}\left(\|x\|^2_{\tau^{-1}}+\|y_k^{(i_{k-1})}-y_{k-1}^{(i_{k-1})}\|^2_{p^{-1}_{i_{k-1}}\sigma^{-1}_{i_{k-1}}}\right)=\frac{\gamma}{2}\left(\|x\|^2_{\tau^{-1}}+\|y_k-y_{k-1}\|^2_{D(\sigma)^{-1}P^{-1}}\right),
\end{aligned}
\tag{4.40}
$$

where the last step is due to $y_k^{(j)}=y_{k-1}^{(j)}, \forall j\neq i_{k-1}$. Inserting (4.40) into the definitions of $V(z_k-z_{k-1})$ and $V_k(z)$ in (4.9) is sufficient to prove (4.11) and (4.12). ■

4.7.2 Proof of Theorem 4.6

Proof. On (4.10), we pick $(x,y)=(x_\star,y_\star)$ and by convexity, $D_g(x^k,z_\star)\geq 0$, $D_{f^*}(\hat{y}^{k+1},z_\star)\geq 0$. Next, by using $\Delta^k=V_{k+1}(x^k-x_\star,y^{k+1}-y_\star)$, we write (4.10)

$$
\mathbb{E}_k[\Delta_k]\leq\Delta_{k-1}-V(z_k-z_{k-1}).
\tag{4.41}
$$

We denote $\boldsymbol{q}_k=(1\otimes x_k,1\otimes y_k)$. By taking total expectation and summing (4.41), and using Lemma 4.4, we have $\sum_{k=1}^\infty\mathbb{E}\left[\|T(\boldsymbol{q}_{k-1})-\boldsymbol{q}_{k-1}\|^2_{\bar{S}\bar{P}}\right]<+\infty$. We use Fubini-Tonelli theorem to exchange the infinite sum and the expectation to obtain $\mathbb{E}\left[\sum_{k=0}^\infty\|T(\boldsymbol{q}_{k-1})-\boldsymbol{q}_{k-1}\|^2_{\bar{S}\bar{P}}\right]<\infty$. Here, since $\sum_{k=0}^\infty\|T(\boldsymbol{q}_{k-1})-\boldsymbol{q}_{k-1}\|^2_{\bar{S}\bar{P}}$ is nonnegative, we conclude that $\sum_{k=0}^\infty\|T(\boldsymbol{q}_{k-1})-\boldsymbol{q}_{k-1}\|^2_{\bar{S}\bar{P}}$ is finite almost everywhere, which implies that $\|T(\boldsymbol{q}_{k-1})-\boldsymbol{q}_{k-1}\|^2_{\bar{S}\bar{P}}$ converges to 0 almost surely. Thus we established: there exists Ω_T with $\mathbb{P}(\Omega_T)=1$ such that $\forall\omega\in\Omega_T$, we have $T(\boldsymbol{q}_k(\omega))-\boldsymbol{q}_k(\omega)\to 0$.

We apply Robbins-Siegmund lemma [RS71, Theorem 1] on (4.41) to get that almost surely, Δ^k converges to a finite valued random variable and $V(z^k-z^{k-1})\to 0$. Consequently, by (4.11), $\|y^k-y^{k-1}\|$ converges to 0 almost surely. Then, since almost surely, Δ^k converges and $\|y^k-y^{k-1}\|$ converges to 0, we have that $\|z^k-z_\star\|$ converges almost surely.

In particular, we have shown that

$$\mathbb{P}\left\{\omega \in \Omega \colon \lim_{k\to\infty} \|z_k(\omega) - z^\star\| \text{ exists.}\right\} = 1. \tag{4.42}$$

The probability 1 set from which we select the trajectories is defined via $z_\star$. We define the set

$$\Omega_{z^\star} = \left\{\omega \in \Omega \colon \lim_{k\to\infty} \|z_k(\omega) - z_\star\| \text{ exists.}\right\} \tag{4.43}$$

Thus our statement is actually, for each $z_\star \in \mathcal{Z}_\star$, there exists a set $\Omega_{z_\star}$ with probability 1, such that $\forall \omega \in \Omega_{z^\star}$, $\lim_{k\to\infty} \|z_k(\omega) - z_\star\|$ exists.

We will now follow the arguments in [CP15, Proposition 2.3], [Ber11, Proposition 9], [IBCH13, Theorem 2], [FB19, Theorem 1] to strengthen this result.

Let us pick a set $\mathcal{C}$ which is a countable subset of $\mathrm{ri}(\mathcal{Z}_\star)$ that is dense in $\mathcal{Z}_\star$. Let us denote the elements of $\mathcal{C}$ as v_i for $i \in \mathbb{N}$.

We just proved that for all $v_i \in \mathcal{Z}_\star$, $\exists \Omega_{v_i}$ with $\mathbb{P}(\Omega_{v_i}) = 1$, such that $\forall \omega \in \Omega_{v_i}$, $\lim_{k\to\infty} \|z_k(\omega) - v_i\|$ exists. Let us denote $\Omega_{\mathcal{C}} = \cap_{i\in\mathbb{N}} \Omega_{v_i}$. As $\Omega_{\mathcal{C}}$ is the intersection of a countable number of sets of probability 1, $\mathbb{P}(\Omega_{\mathcal{C}}) = 1$.

Next, we set $\tilde{z} \in \mathcal{Z}_\star$. As $\mathcal{C}$ is dense in $\mathrm{ri}(\mathcal{Z}_\star)$, there exists a subsequence $v_{\varphi(i)}$, where $\varphi \colon \mathbb{N} \to \mathbb{N}$ is an increasing function, such that $v_{\varphi(i)} \to \tilde{z}$.

We pick $\omega \in \Omega_{\mathcal{C}}$ and study the existence of $\lim_{k\to\infty} \|z_k(\omega) - \tilde{z}\|$. By triangle inequality, $\forall i \in \mathbb{N}$,

$$\|z_k(\omega) - v_{\varphi(i)}\| - \|v_{\varphi(i)} - \tilde{z}\| \le \|z_k(\omega) - \tilde{z}\| \le \|z_k(\omega) - v_{\varphi(i)}\| + \|v_{\varphi(i)} - \tilde{z}\|.$$

Rearranging gives

$$-\|v_{\varphi(i)} - \tilde{z}\| \le \|z_k(\omega) - \tilde{z}\| - \|z_k(\omega) - v_{\varphi(i)}\| \le \|v_{\varphi(i)} - \tilde{z}\|.$$

As ω is chosen from $\Omega_{\mathcal{C}}$, and any element of $\Omega_{\mathcal{C}}$ is also an element of Ω_{v_i}, we know that $\lim_{k\to\infty} \|z_k(\omega) - v_{\varphi(i)}\|$ exists. Moreover, recall that $v_{\varphi(i)} \to \tilde{z}$.

We take limit as $k \to \infty$,

$$\begin{aligned} -\|v_{\varphi(i)} - \tilde{z}\| &\le \liminf_{k\to\infty} \|z_k(\omega) - \tilde{z}\| - \lim_{k\to\infty} \|z_k(\omega) - v_{\varphi(i)}\| \\ &\le \limsup_{k\to\infty} \|z_k(\omega) - \tilde{z}\| - \lim_{k\to\infty} \|z_k(\omega) - v_{\varphi(i)}\| \\ &\le \|v_{\varphi(i)} - \tilde{z}\|. \end{aligned}$$

As we take the limit along the subsequence defined by $\varphi(i)$, we have $\lim_{i\to\infty} \|v_{\varphi(i)} - \tilde{z}\| = 0$, which gives the equality of $\liminf$ and $\limsup$.

Thus, $\forall \omega \in \Omega_C$ with $\mathbb{P}(\Omega_C) = 1$ and $\forall \tilde{z} \in \mathcal{Z}_\star$, we have that $\lim_{k\to\infty} \|z_k(\omega) - \tilde{z}\|$ exists.

We pick $\omega \in \Omega_C \cap \Omega_T$ and as we have that $(z_k(\omega))_k$ is bounded, we denote by $\tilde{z} = (\tilde{x}, \tilde{y})$ one of its cluster points. Then, we denote $\tilde{\boldsymbol{q}} = (1 \otimes \tilde{x}, 1 \otimes \tilde{y})$ and say that $\tilde{\boldsymbol{q}}$ is a cluster point of $(\boldsymbol{q}_k(\omega))_k$.

As $T(\boldsymbol{q}_k(\omega)) - \boldsymbol{q}_k(\omega) \to 0$, by continuity of T we have $T(\tilde{\boldsymbol{q}}) - \tilde{\boldsymbol{q}} \to 0$, therefore $\tilde{\boldsymbol{q}}$ is a fixed point of T. We now use Lemma 4.4 to argue that fixed points of T which we denote as $(x_f(j), y_f(j))_{j=\{1,\ldots,n\}}$ are such that $(x_f(j), y_f(j)) \in \mathcal{Z}^\star, \forall j \in \{1,\ldots,n\}$. Since $\tilde{\boldsymbol{q}}$ is a fixed point of T, we conclude that $\tilde{z} \in \mathcal{Z}_\star$.

To sum up, we have shown that at least on some subsequence $z_k(\omega)$ converges to $\tilde{z} \in \mathcal{Z}_\star$. Then, the result follows due to existence of the limit, proven earlier. ■

4.7.3 Proof of Theorem 4.8

Proof. Starting from the result of Lemma 4.3, we have

$$D_g(x_k, z) + D_{f^*}(\hat{y}_{k+1}, z) \le -\mathbb{E}_k\left[V_{k+1}(x_k - x, y_{k+1} - y)\right] + V_k(x_{k-1} - x, y_k - y) \\ - V(z_k - z_{k-1}). \quad (4.44)$$

We pick $x = x_{\star,k-1}$, $y = y_{\star,k}$, with $z_{\star,k} = (x_{\star,k-1}, y_{\star,k})$ and use convexity to get $D_g(x_k, z_{\star,k}) \ge 0$ and $D_{f^*}(\hat{y}_{k+1}, z_{\star,k}) \ge 0$. In addition, we define

$$\Delta_{k-1} = V_k(x_{k-1} - x_{\star,k-1}, y_k - y_{\star,k})$$
$$\tilde{\Delta}_k = V_{k+1}(x_k - x_{\star,k-1}, y_{k+1} - y_{\star,k}).$$

We use these definitions in (4.44) to write

$$\mathbb{E}_k\left[\tilde{\Delta}_k\right] \le \Delta_{k-1} - V(z_k - z_{k-1}).$$

By definition of $(x_\star^k, y_\star^{k+1})$, we have $\Delta^k \le \tilde{\Delta}^k$, which implies that

$$\mathbb{E}_k\left[\Delta^k\right] \le \Delta^{k-1} - V(z^k - z^{k-1}).$$

Recursion of this inequality gives boundedness of the iterates x_k and y_k, in expectation. However, it is not possible to derive sure boundedness of the sequence. Without sure boundedness, the set that includes x_k, y_k depends on the specific trajectory of the algorithm, and it is not possible to find a set independent of these. As metric subregularity holds for PLQs with a bounded neighborhood (see Section 4.2.2), we cannot utilize this result and this is the main reason for the need for bounded domains in this case. This assumption ensures the existence of a uniform set bounding the sequence, to use metric subregularity assumption for PLQs.

We recall $S = \operatorname{diag}(\tau^{-1}1_p, \sigma_1^{-1}, \ldots, \sigma_n^{-1})$, $\bar{S}$ and $\bar{P}$ are as defined in Lemma 4.4, and $\operatorname{dist}_S^2(z_k, \mathcal{Z}_\star) = \|z_k - \mathcal{P}_{\mathcal{Z}_\star}^S(z_k)\|_S^2 = \|x_k - \tilde{x}_{\star,k}\|_{\tau^{-1}}^2 + \|y_k - y_{\star,k}\|_{D(\sigma)^{-1}}^2$ where $\tilde{x}_{\star,k}$ is the projection of x_k onto the set of solutions with respect to norm $\|\cdot\|_{\tau^{-1}}$. We now use Assumption 4.2 stating that $F = C + M$ is metrically subregular at $\mathcal{P}^S_{\mathcal{Z}_\star}(\hat{z}_{k+1})$ for 0. We recall, $\boldsymbol{q}_k = (1 \otimes x_k, 1 \otimes y_k)$ and $\hat{\boldsymbol{q}}_{k+1} = T(\boldsymbol{q}_k)$ and estimate as

$$\begin{aligned}\|x_k - \tilde{x}_{\star,k}\|^2_{\tau^{-1}} + \|y_k - y_{\star,k}\|^2_{D(\sigma)^{-1}} = \mathrm{dist}^2_S(z_k, \mathcal{Z}_\star) &\le \mathbb{E}_k\left[\|z_k - \mathcal{P}^S_{\mathcal{Z}_\star}(\hat{z}_{k+1})\|^2_S\right] \\ &\le 2\mathbb{E}_k\left[\|z_k - \hat{z}_{k+1}\|^2_S\right] + 2\mathbb{E}_k\left[\|\hat{z}_{k+1} - \mathcal{P}^S_{\mathcal{Z}_\star}(\hat{z}_{k+1})\|^2_S\right] \\ &\le 2\mathbb{E}_k\left[\|z_k - \hat{z}_{k+1}\|^2_S\right] + 2\eta^2\mathbb{E}_k\left[\mathrm{dist}^2_S(0, (C+M)\hat{z}_{k+1})\right] \\ &= 2\mathbb{E}_k\left[\|z_k - \hat{z}_{k+1}\|^2_S\right] + 2\eta^2\,\mathrm{dist}^2_{\bar{S}\bar{P}}(0, (\boldsymbol{C}+\boldsymbol{M})\hat{\boldsymbol{q}}_{k+1}) \\ &\le 2\mathbb{E}_k\left[\|z_k - \hat{z}_{k+1}\|^2_S\right] + 2\eta^2\|\boldsymbol{M} - \boldsymbol{H}\|^2\|\hat{\boldsymbol{q}}_{k+1} - \boldsymbol{q}_k\|^2_{\bar{S}\bar{P}}, \end{aligned} \tag{4.45}$$

where the first inequality is due to the definition of $\mathrm{dist}^2_S(z_k, \mathcal{Z}_\star)$, third inequality is due to metric subregularity of $C + M$ (see Remark 4.9) since $\mathrm{dist}^2_S(\hat{z}^{k+1}, \mathcal{Z}_\star) = \|\hat{z}^{k+1} - \mathcal{P}^S_{\mathcal{Z}_\star}(\hat{z}_{k+1})\|^2_S$. Second equality and fourth inequality are by Lemma 4.7 and Cauchy-Schwarz inequality.

First, we use $\|\hat{y}_{k+1} - y_k\|^2_{D(\sigma)^{-1}} = \mathbb{E}_k\left[\|y_{k+1} - y_k\|^2_{D(\sigma)^{-1}P^{-1}}\right]$ to estimate

$$\begin{aligned}\mathbb{E}_k\left[\|z_k - \hat{z}_{k+1}\|^2_S\right] &= \mathbb{E}_k\left[\|x_{k+1} - x_k\|^2_{\tau^{-1}}\right] + \|\hat{y}_{k+1} - y_k\|^2_{D(\sigma)^{-1}} \\ &= \mathbb{E}_k\left[\|x_{k+1} - x_k\|^2_{\tau^{-1}} + \|y_{k+1} - y_k\|^2_{D(\sigma)^{-1}P^{-1}}\right]. \end{aligned} \tag{4.46}$$

Second, we use Lemma 4.4 to obtain

$$\begin{aligned}\|\hat{\boldsymbol{q}}_{k+1} - \boldsymbol{q}_k\|^2_{\bar{S}\bar{P}} &= \|T(1 \otimes x_k, 1 \otimes y_k) - (1 \otimes x_k, 1 \otimes y_k)\|^2_{\bar{S}\bar{P}} \\ &= \mathbb{E}_k\left[\|x_{k+1} - x_k\|^2_{\tau^{-1}} + \|y_{k+1} - y_k\|^2_{D(\sigma)^{-1}P^{-1}}\right]. \end{aligned} \tag{4.47}$$

We combine (4.46) and (4.47) in (4.45) to get

$$\begin{aligned}&\frac{1}{2}\|x_k - \tilde{x}_{\star,k}\|^2_{\tau^{-1}} + \frac{1}{2}\|y_k - y_{\star,k}\|^2_{D(\sigma)^{-1}} \\ &\quad\le (2 + 2\eta^2\|N - H\|^2)\mathbb{E}_k\left[\frac{1}{2}\|x_{k+1} - x_k\|^2_{\tau^{-1}} + \frac{1}{2}\|y_{k+1} - y_k\|^2_{D(\sigma)^{-1}P^{-1}}\right]. \end{aligned} \tag{4.48}$$

Herein, we denote $\zeta = 2 + 2\eta^2\|\boldsymbol{H} - \boldsymbol{M}\|^2$.

By using (4.11), we have that, for all $\alpha \in [0, 1]$

$$\begin{aligned}\mathbb{E}_{k-1}\left[V(z_k - z_{k-1})\right] &\ge C_1\mathbb{E}_{k-1}\left[\frac{1}{2}\|x_k - x_{k-1}\|^2_{\tau^{-1}} + \frac{1}{2}\|y_k - y_{k-1}\|^2_{D(\sigma)^{-1}P^{-1}}\right] \\ &\ge \frac{C_1}{\zeta}\left(\frac{1}{2}\|x_{k-1} - \tilde{x}_{\star,k-1}\|^2_{\tau^{-1}} + \frac{1}{2}\|y_{k-1} - y_{\star,k-1}\|^2_{D(\sigma)^{-1}}\right) \\ &\ge \frac{C_1}{\zeta}\left(\frac{\alpha}{2}\|x_{k-1} - \tilde{x}_{\star,k-1}\|^2_{\tau^{-1}} + \frac{1}{2}\|y_{k-1} - y_{\star,k-1}\|^2_{D(\sigma)^{-1}}\right), \end{aligned} \tag{4.49}$$

where the second inequality is due to (4.48).

We have, by definition of $x_{\star,k-1}$ that

$$\begin{aligned}
\Delta_{k-1} &\leq V_k(x_{k-1}-\tilde{x}_{\star,k-1}, y_k - y_{\star,k}) \\
&= \frac{1}{2}\|x_{k-1}-\tilde{x}_{\star,k-1}\|^2_{\tau^{-1}} + \frac{1}{2}\|y_k - y_{\star,k}\|^2_{D(\sigma)^{-1}P^{-1}} + \frac{1}{2}\|y_k - y_{k-1}\|^2_{D(\sigma)^{-1}P^{-1}} \\
&\quad - \langle P^{-1}A(x_{k-1}-\tilde{x}_{\star,k-1}), y_k - y_{k-1}\rangle \\
&\leq \frac{1}{2}\|x_{k-1}-\tilde{x}_{\star,k-1}\|^2_{\tau^{-1}} + \frac{1}{2}\|y_k - y_{\star,k}\|^2_{D(\sigma)^{-1}P^{-1}} + \frac{1+\gamma}{2}\|y_k - y_{k-1}\|^2_{D(\sigma)^{-1}P^{-1}} + \frac{\gamma}{2}\|x_{k-1}-\tilde{x}_{\star,k-1}\|^2_{\tau^{-1}} \\
&= \frac{1+\gamma}{2}\|x_{k-1}-\tilde{x}_{\star,k-1}\|^2_{\tau^{-1}} + \frac{1+\gamma}{2\alpha}\|y_{k-1}-y_{\star,k-1}\|^2_{D(\sigma)^{-1}} + \frac{1+\gamma}{2}\|y_k - y_{k-1}\|^2_{D(\sigma)^{-1}P^{-1}} \\
&\quad - \frac{1+\gamma}{2\alpha}\|y_{k-1}-y_{\star,k-1}\|^2_{D(\sigma)^{-1}} + \frac{1}{2}\|y_k - y_{\star,k}\|^2_{D(\sigma)^{-1}P^{-1}},
\end{aligned}$$

where the second inequality is due to (4.8).

We now take conditional expectation of both sides and use (4.49) to get

$$\begin{aligned}
\mathbb{E}_{k-1}[\Delta_{k-1}] \leq \frac{(1+\gamma)\zeta}{C_1\alpha}\mathbb{E}_{k-1}[V(z_k - z_{k-1})] + \frac{1+\gamma}{2}\mathbb{E}_{k-1}\left[\|y_k - y_{k-1}\|^2_{D(\sigma)^{-1}P^{-1}}\right] \\
+ \frac{1}{2}\mathbb{E}_{k-1}\left[\|y_k - y_{\star,k}\|^2_{D(\sigma)^{-1}P^{-1}}\right] - \frac{1+\gamma}{2\alpha}\|y_{k-1}-y_{\star,k-1}\|^2_{D(\sigma)^{-1}}.
\end{aligned}$$

By using (4.11) and requiring that $\frac{(1+\gamma)}{C_1} \leq \frac{(1+\gamma)\zeta}{C_1\alpha}$, or equivalently $\zeta \geq \alpha$, which is not restrictive since α is finite, and one can increase η as in (4.6) to satisfy the requirement, we can combine the first two terms in the right hand side to get

$$\begin{aligned}
\mathbb{E}_{k-1}\big[\Delta_{k-1}\big] \leq \frac{2(1+\gamma)\zeta}{C_1\alpha}\mathbb{E}_{k-1}\big[V(z_k - z_{k-1})\big] + \frac{1}{2}\mathbb{E}_{k-1}\big[\|y_k - y_{\star,k}\|^2_{D(\sigma)^{-1}P^{-1}}\big] \\
- \frac{1+\gamma}{2\alpha}\|y_{k-1}-y_{\star,k-1}\|^2_{D(\sigma)^{-1}}.
\end{aligned}$$

We now insert this inequality into (4.14) and use that $\mathbb{E}_{k-1}[\mathbb{E}_k[\Delta_k]] = \mathbb{E}_{k-1}[\Delta_k]$

$$\begin{aligned}
\mathbb{E}_{k-1}[\Delta_k] \leq \mathbb{E}_{k-1}[\Delta_{k-1}] - \frac{C_1\alpha}{2(1+\gamma)\zeta}\mathbb{E}_{k-1}[\Delta_{k-1}] \\
+ \frac{C_1\alpha}{4(1+\gamma)\zeta}\mathbb{E}_{k-1}\left[\|y_k - y_{\star,k}\|^2_{D(\sigma)^{-1}P^{-1}}\right] - \frac{C_1}{4\zeta}\|y_{k-1}-y_{\star,k-1}\|^2_{D(\sigma)^{-1}}.
\end{aligned}$$

We take full expectation and rearrange to get

$$\begin{aligned}
&\mathbb{E}\left[\Delta_k - \frac{C_1\alpha}{4(1+\gamma)\zeta}\|y_k - y_{\star,k}\|^2_{D(\sigma)^{-1}P^{-1}}\right] \\
&\qquad \leq \left(1 - \frac{C_1\alpha}{2(1+\gamma)\zeta}\right)\mathbb{E}\left[\Delta_{k-1} - \frac{C_1}{4\zeta(1-\frac{C_1\alpha}{2(1+\gamma)\zeta})}\|y_{k-1}-y_{\star,k-1}\|^2_{D(\sigma)^{-1}}\right]. \qquad (4.50)
\end{aligned}$$

We require

$$C_2 = \frac{C_1 \alpha}{4\underline{p}(1+\gamma)\zeta} \leq \frac{C_1}{4\zeta} \leq \frac{C_1}{4\zeta(1-\frac{C_1\alpha}{2(1+\gamma)\zeta})} \iff \alpha \leq (1+\gamma)\underline{p}. \tag{4.51}$$

Let us pick $\alpha = (1+\gamma)\underline{p}$ so that $C_2 = \frac{C_1}{4\zeta}$ and define

$$\Phi_k = \Delta_k - C_2 \| y_k - y_{\star,k}\|^2_{D(\sigma)^{-1}}.$$

We note (4.49) and (4.14) to have

$$\| y_k - y_{\star,k}\|^2_{D(\sigma)^{-1}} \leq \frac{2\zeta}{C_1}\mathbb{E}_k\left[V(z_{k+1}-z_k)\right] \leq \frac{2\zeta}{C_1}\mathbb{E}_k\left[\Delta_k\right].$$

Then, we can lower bound Φ_k as

$$\mathbb{E}\big[\Phi_k\big] \geq \left(1 - C_2\frac{2\zeta}{C_1}\right)\mathbb{E}\big[\Delta_k\big] = \frac{1}{2}\mathbb{E}\big[\Delta_k\big]. \tag{4.52}$$

Therefore, it follows that $\mathbb{E}\left[\Phi_k\right]$ is nonnegative, by the definition of Δ_k and (4.12).

We can now rewrite (4.50) as

$$\mathbb{E}\big[\Phi_k\big] \leq (1-\rho)\mathbb{E}\big[\Phi_{k-1}\big],$$

where $\rho = \frac{C_1 \underline{p}}{2\zeta}$. We have shown that Φ_k converges linearly to 0 in expectation. By (6.35), it immediately follows that Δ_k converges linearly to 0.

To conclude, we note $\Delta_k = V_{k+1}(x_k - x_{\star,k}, y_{k+1} - y_{\star,k+1})$, and (4.12), from which we conclude linear convergence of $\|x_k - x_{\star,k}\|^2_{\tau^{-1}}$ and $\|y_{k+1} - y_{\star,k+1}\|^2_{D(\sigma)^{-1}P^{-1}}$.

It is obvious to see that $0 < \rho$ follows by the fact that η is finite by metric subregularity and $\rho < 1$ follows since $\gamma < 1$ and $\underline{p} \leq 1$. ■

4.7.4 Proof of Lemma 4.11

Proof. We note

$$\begin{aligned}
&\mathcal{H}(x_k, \hat{y}_{k+1}; x, y) = g(x_k) + \langle Ax_k, y\rangle - f^*(y) - g(x) - \langle Ax, \hat{y}_{k+1}\rangle + f^*(\hat{y}_{k+1}),\\
&\Delta_x = \frac{1}{2}\left[-\|x_{k-1} - x\|^2_{\tau^{-1}} + \|x_k - x\|^2_{\tau^{-1}} + \|x_k - x_{k-1}\|^2_{\tau^{-1}}\right],\\
&\Delta_y = \frac{1}{2}\left[-\|y_k - y\|^2_{D(\sigma)^{-1}} + \|\hat{y}_{k+1} - y\|^2_{D(\sigma)^{-1}} + \|\hat{y}_{k+1} - y_k\|^2_{D(\sigma)^{-1}}\right],
\end{aligned}$$

and, $\forall (x,y) \in \mathcal{X}\times\mathcal{Y}$. Then, we can write (4.27) as

$$0 \geq \mathcal{H}(x_k, \hat{y}_{k+1}; x, y) + \langle A(x - x_k), \hat{y}_{k+1} - \bar{y}_k\rangle + \Delta_x + \Delta_y. \tag{4.53}$$

We next note that

$$\begin{aligned}
\mathcal{H}(x_k,\hat{y}_{k+1};x,y) &= \mathcal{H}(x_k,y_{k+1};x,y) + \langle Ax, y_{k+1}-\hat{y}_{k+1}\rangle + f^*(\hat{y}_{k+1}) - f^*(y_{k+1}) \\
&- \left(f^*_{P^{-1}-I}(y_{k+1}) - f^*_{P^{-1}-I}(y_k)\right) + \left(f^*_{P^{-1}-I}(y_{k+1}) - f^*_{P^{-1}-I}(y_k)\right) \\
&+ \langle Ax, (P^{-1}-I)(y_{k+1}-y_k)\rangle - \langle Ax, (P^{-1}-I)(y_{k+1}-y_k)\rangle \\
&= \mathcal{H}(x_k,y_{k+1};x,y) + f^*(\hat{y}_{k+1}) - f^*(y_k) - (f^*_{P^{-1}}(y_{k+1}) - f^*_{P^{-1}}(y_k)) \\
&+ \langle Ax, y_k-\hat{y}_{k+1}-P^{-1}(y_k-y_{k+1})\rangle \\
&+ \left(f^*_{P^{-1}-I}(y_{k+1}) - f^*_{P^{-1}-I}(y_k)\right) - \langle Ax, (P^{-1}-I)(y_{k+1}-y_k)\rangle \\
&= \mathcal{H}(x_k,y_{k+1};x,y) + f^*(\hat{y}_{k+1}) - f^*(y_k) - (f^*_{P^{-1}}(y_{k+1}) - f^*_{P^{-1}}(y_k)) \\
&+ \langle Ax, y_k-\hat{y}_{k+1}-P^{-1}(y_k-y_{k+1})\rangle + D^{P^{-1}-I}_{f^*}(y_{k+1},y) - D^{P^{-1}-I}_{f^*}(y_k,y). \qquad (4.54)
\end{aligned}$$

By the definition of $\bar{y}_k$ in SPDHG, we have for the bilinear term in (4.53) that

$$\begin{aligned}
\langle A(x-x_k), \hat{y}_{k+1}-\bar{y}_k\rangle &= \langle A(x-x_k), \hat{y}_{k+1}-y_k-P^{-1}(y_k-y_{k-1})\rangle \\
&= \langle A(x-x_k), \hat{y}_{k+1}-y_k\rangle - \langle A(x-x_k), P^{-1}(y_k-y_{k-1})\rangle \\
&= \langle A(x-x_k), \hat{y}_{k+1}-y_k\rangle - \langle A(x-x_{k-1}), P^{-1}(y_k-y_{k-1})\rangle \\
&- \langle A(x_{k-1}-x_k), P^{-1}(y_k-y_{k-1})\rangle \\
&= \langle A(x-x_k), P^{-1}(y_{k+1}-y_k)\rangle - \langle A(x-x_{k-1}), P^{-1}(y_k-y_{k-1})\rangle \\
&- \langle A(x_{k-1}-x_k), P^{-1}(y_k-y_{k-1})\rangle \\
&+ \langle A(x-x_k), \hat{y}_{k+1}-y_k-P^{-1}(y_{k+1}-y_k)\rangle. \qquad (4.55)
\end{aligned}$$

We focus on Δ_y and get

$$\begin{aligned}
-\Delta_y = \frac{1}{2}\Bigg[&- \|\hat{y}_{k+1}-y_k\|^2_{D(\sigma)^{-1}} - \|\hat{y}_{k+1}-y\|^2_{D(\sigma)^{-1}} + \|y_k-y\|^2_{D(\sigma)^{-1}} \\
&+ \left(\|y_k-y\|^2_{D(\sigma)^{-1}P^{-1}} - \|y_{k+1}-y\|^2_{D(\sigma)^{-1}P^{-1}}\right) \\
&- \left(\|y_k-y\|^2_{D(\sigma)^{-1}P^{-1}} - \|y_{k+1}-y\|^2_{D(\sigma)^{-1}P^{-1}}\right)\Bigg] \\
= -\frac{1}{2}&\|y_{k+1}-y\|^2_{D(\sigma)^{-1}P^{-1}} + \frac{1}{2}\|y_k-y\|^2_{D(\sigma)^{-1}P^{-1}} - \frac{1}{2}\|\hat{y}_{k+1}-y_k\|^2_{D(\sigma)^{-1}} + \epsilon_k, \qquad (4.56)
\end{aligned}$$

where

$$\begin{aligned}
\epsilon_k &= \frac{1}{2}\left[\|y_k-y\|^2_{D(\sigma)^{-1}} - \|\hat{y}_{k+1}-y\|^2_{D(\sigma)^{-1}}\left(\|y_k-y\|^2_{D(\sigma)^{-1}P^{-1}} - \|y_{k+1}-y\|^2_{D(\sigma)^{-1}P^{-1}}\right)\right] \\
&= \frac{1}{2}\Bigg[\|y_k\|^2_{D(\sigma)^{-1}} - \|\hat{y}_{k+1}\|^2_{D(\sigma)^{-1}} - \left(\|y_k\|^2_{D(\sigma)^{-1}P^{-1}} - \|y_{k+1}\|^2_{D(\sigma)^{-1}P^{-1}}\right) \\
&- 2\langle y, y_k-\hat{y}_{k+1}-P^{-1}(y_k-y_{k+1})\rangle_{D(\sigma)^{-1}}\Bigg]. \qquad (4.57)
\end{aligned}$$

We use eqs. (4.54)–(4.56) in (4.53), add and subtract $\frac{1}{2}\|y_k - y_{k-1}\|^2_{D(\sigma)^{-1}P^{-1}}$ and use the definition $v_{k+1} = y_k - \hat{y}_{k+1} - P^{-1}(y_k - y_{k+1})$ from Lemma 4.10 to obtain

$$\begin{aligned}
\mathcal{H}(x_k, y_{k+1}; x, y) \le{}& -\frac{1}{2}\|x_k - x\|^2_{\tau^{-1}} + \frac{1}{2}\|x_{k-1} - x\|^2_{\tau^{-1}} \\
&- \langle A(x - x_k), P^{-1}(y_{k+1} - y_k)\rangle + \langle A(x - x_{k-1}), P^{-1}(y_k - y_{k-1})\rangle \\
&- \frac{1}{2}\|x_k - x_{k-1}\|^2_{\tau^{-1}} - \frac{1}{2}\|y_k - y_{k-1}\|^2_{D(\sigma)^{-1}P^{-1}} \\
&- \langle A(x_k - x_{k-1}), P^{-1}(y_k - y_{k-1})\rangle - \frac{1}{2}\|y_{k+1} - y\|^2_{D(\sigma)^{-1}P^{-1}} \\
&+ \frac{1}{2}\|y_k - y\|^2_{D(\sigma)^{-1}P^{-1}} - \frac{1}{2}\|\hat{y}_{k+1} - y_k\|^2_{D(\sigma)^{-1}} + \frac{1}{2}\|y_k - y_{k-1}\|^2_{D(\sigma)^{-1}P^{-1}} \\
&+ \frac{1}{2}\left[\|y_k\|^2_{D(\sigma)^{-1}} - \|\hat{y}_{k+1}\|^2_{D(\sigma)^{-1}} - \left(\|y_k\|^2_{D(\sigma)^{-1}P^{-1}} - \|y_{k+1}\|^2_{D(\sigma)^{-1}P^{-1}}\right)\right] \\
&+ f^*(y_k) - f^*(\hat{y}_{k+1}) - (f^*_{P^{-1}}(y_k) - f^*_{P^{-1}}(y_{k+1})) - \langle y, v_{k+1}\rangle_{D(\sigma)^{-1}} \\
&- \langle Ax_k, y_k - \hat{y}_{k+1} - P^{-1}(y_k - y_{k+1})\rangle + D^{P^{-1}-I}_{f^*}(y_k, z) - D^{P^{-1}-I}_{f^*}(y_{k+1}, z).
\end{aligned} \tag{4.58}$$

We obtain the first result of the lemma by using the definitions of V_k and V from Lemma 4.3, and definition of $\mathcal{E}_k$ from (4.21).

Second, on $\mathcal{E}_k$ (see (4.21)), we use the following conditional expectation estimations $\mathbb{E}_k\left[P^{-1}(y_k - y_{k+1})\right] = y_k - \hat{y}_{k+1}$, $\mathbb{E}_k\left[f^*_{P^{-1}}(y_k) - f^*_{P^{-1}}(y_{k+1})\right] = f^*(y_k) - f^*(\hat{y}_{k+1})$, $\mathbb{E}_k\left[\|y_{k+1} - y_k\|^2_{D(\sigma)^{-1}P^{-1}}\right] = \|\hat{y}_{k+1} - y_k\|^2_{D(\sigma)^{-1}}$

$$\begin{aligned}
&\mathbb{E}_k\left[\mathcal{E}_k\right] = -\frac{1}{2}\|\hat{y}_{k+1} - y_k\|^2_{D(\sigma)^{-1}} + \frac{1}{2}\mathbb{E}_k\left[\|y_{k+1} - y_k\|^2_{D(\sigma)^{-1}P^{-1}}\right] \\
&+ \frac{1}{2}\left(\|y_k\|^2_{D(\sigma)^{-1}} - \|\hat{y}_{k+1}\|^2_{D(\sigma)^{-1}}\right) - \frac{1}{2}\mathbb{E}_k\left[\|y_k\|^2_{D(\sigma)^{-1}P^{-1}} - \|y_{k+1}\|^2_{D(\sigma)^{-1}P^{-1}}\right] \\
&+ f^*(y_k) - f^*(\hat{y}_{k+1}) - \mathbb{E}_k\left[f^*_{P^{-1}}(y_k) - f^*_{P^{-1}}(y_{k+1})\right] - \langle Ax_k, y_k - \hat{y}_{k+1} - \mathbb{E}_k\left[P^{-1}\left(y_k - y_{k+1}\right)\right]\rangle \\
&= 0.
\end{aligned}$$

■

4.7.5 Proof of Theorem 4.12

Proof. We will continue from the result of Lemma 4.11. We have for the last error term in (4.20)

$$-\langle y, v_{k+1}\rangle_{D(\sigma)^{-1}} = \langle \tilde{y}_k - y, v_{k+1}\rangle_{D(\sigma)^{-1}} - \langle \tilde{y}_k, v_{k+1}\rangle_{D(\sigma)^{-1}}, \tag{4.59}$$

where $\tilde{y}_k$ is the random sequence defined in Lemma 4.10.

We sum (4.20) after using (4.59) and Lemma 4.3

$$\sum_{k=1}^{K} \mathcal{H}(x_k, y_{k+1}; x, y) \le -V_{K+1}(x_K - x, y_{K+1} - y) + V_1(x_0 - x, y_1 - y) + D^{P^{-1}-I}_{f^*}(y_1; z)$$

$$-D_{f^*}^{P^{-1}-I}(y_{K+1};z)+\sum_{k=1}^{K}\left(\langle \tilde{y}_k-y,v_{k+1}\rangle_{D(\sigma)^{-1}}-\langle \tilde{y}_k,v_{k+1}\rangle_{D(\sigma)^{-1}}+\mathcal{E}^k\right), \quad (4.60)$$

Next, by Young's inequality

$$-\langle A(x-x_K),P^{-1}(y_{K+1}-y_K)\rangle \le \frac{\gamma}{2}\|x-x_K\|^2_{\tau^{-1}}+\frac{\gamma}{2}\|y_{K+1}-y_K\|^2_{D(\sigma)^{-1}P^{-1}}. \quad (4.61)$$

On (4.60), we can use (4.18) from Lemma 4.10 with $\tilde{y}_1=y_1=y_0$ and (4.61) with the definition of $V_{K+1}(x_K-x,y_{K+1}-y)$ from Lemma 4.3, and by $\gamma<1$ from the step size rules in (4.8) to get

$$\begin{aligned}\sum_{k=1}^{K}\mathcal{H}(x_k,y_{k+1};x,y)\le{}&\frac{1}{2}\|x_0-x\|^2_{\tau^{-1}}+\|y_1-y\|^2_{D(\sigma)^{-1}P^{-1}}+f^*_{P^{-1}-I}(y_1)-f^*_{P^{-1}-I}(y_{K+1})\\&+\langle Ax,(P^{-1}-I)(y_{K+1}-y_1)\rangle+\sum_{k=1}^{K}\left(\frac{1}{2}\|v_{k+1}\|^2_{D(\sigma)^{-1}P}-\langle\tilde{y}_k,v_{k+1}\rangle_{D(\sigma)^{-1}}+\mathcal{E}_k\right). \quad (4.62)\end{aligned}$$

We have $\langle Ax,(P^{-1}-I)(y_{K+1}-y_1)\rangle \le \frac{\gamma}{2}\|x\|^2_{\tau^{-1}}+\frac{\gamma}{2\underline{p}}\|y_{K+1}-y_1\|^2_{D(\sigma)^{-1}P^{-1}}$ and $\frac{\gamma}{2}\|x\|^2_{\tau^{-1}}\le\gamma\|x-x_0\|^2_{\tau^{-1}}+\gamma\|x_0\|^2_{\tau^{-1}}$ by Young's inequality and (4.8).

We use these inequalities, arrange (4.62), and divide both sides by K

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathcal{H}(x_k,y_{k+1};x,y)\le\frac{1}{K}\Bigg\{&\frac{1+2\gamma}{2}\|x_0-x\|^2_{\tau^{-1}}+\|y_1-y\|^2_{D(\sigma)^{-1}P^{-1}}+\gamma\|x_0\|^2_{\tau^{-1}}\\&+\frac{\gamma}{2\underline{p}}\|y_{K+1}-y_1\|^2_{D(\sigma)^{-1}P^{-1}}+f^*_{P^{-1}-I}(y_1)-f^*_{P^{-1}-I}(y_{K+1})\\&+\sum_{k=1}^{K}\left(\frac{1}{2}\|v_{k+1}\|^2_{D(\sigma)^{-1}P}-\langle\tilde{y}_k,v_{k+1}\rangle_{D(\sigma)^{-1}}+\mathcal{E}_k\right)\Bigg\}. \quad (4.63)\end{aligned}$$

We now take supremum of (4.63) with respect to z, note that only the first two terms on the right hand side depend on $z=(x,y)$, and x_0, y_1 are deterministic. Then we take expectation of both sides of (4.63) and use $\gamma<1$

$$\begin{aligned}\mathbb{E}\left[\sup_{z\in\mathcal{B}}\frac{1}{K}\sum_{k=1}^{K}\mathcal{H}(x_k,y_{k+1};x,y)\right]\le\frac{1}{K}\Bigg\{&\sup_{z\in\mathcal{B}}\left\{\frac{3}{2}\|x_0-x\|^2_{\tau^{-1}}+\|y_1-y\|^2_{D(\sigma)^{-1}P^{-1}}\right\}\\&+\mathbb{E}\left[\frac{\gamma}{2\underline{p}}\|y_{K+1}-y_1\|^2_{D(\sigma)^{-1}P^{-1}}+f^*_{P^{-1}-I}(y_1)-f^*_{P^{-1}-I}(y_{K+1})\right]+\gamma\|x_0\|^2_{\tau^{-1}}\\&+\sum_{k=1}^{K}\frac{1}{2}\mathbb{E}\left[\|v_{k+1}\|^2_{D(\sigma)^{-1}P}\right]-\sum_{k=1}^{K}\mathbb{E}\left[\langle\tilde{y}_k,v_{k+1}\rangle_{D(\sigma)^{-1}}\right]+\sum_{k=1}^{K}\mathbb{E}\left[\mathcal{E}_k\right]\Bigg\}. \quad (4.64)\end{aligned}$$

First, as $\tilde{y}_k$ is $\mathcal{F}_k$-measurable, $\mathbb{E}_k\left[v_{k+1}\right]=0$, by Lemma 4.10, and by the law of total expectation,

$$\mathbb{E}\left[\sum_{k=1}^{K}\langle\tilde{y}_k,v_{k+1}\rangle_{D(\sigma)^{-1}}\right]=\sum_{k=1}^{K}\mathbb{E}\left[\mathbb{E}_k\left[\langle\tilde{y}_k,v_{k+1}\rangle_{D(\sigma)^{-1}}\right]\right]$$

$$= \sum_{k=1}^{K} \mathbb{E}\left[\langle \tilde{y}_k, \mathbb{E}_k[v_{k+1}]\rangle_{D(\sigma)^{-1}}\right] = 0. \tag{4.65}$$

On (4.64), we use (4.19) from Lemma 4.10, (4.65), and $\sum_{k=1}^{K}\mathbb{E}[\mathcal{E}_k] = \sum_{k=1}^{K}\mathbb{E}\left[\mathbb{E}_k\left[\mathcal{E}_k\right]\right] = 0$, which follows from Lemma 4.11 along with the law of total expectation, to obtain

$$\begin{aligned}
\mathbb{E}\left[\sup_{z\in\mathcal{B}} \frac{1}{K}\sum_{k=1}^{K} \mathcal{H}(x_k, y_{k+1}; x, y)\right] \le{}& \sup_{z\in\mathcal{B}}\left\{\frac{3}{2K}\|x_0 - x\|^2_{\tau^{-1}} + \frac{1}{K}\|y_1 - y\|^2_{D(\sigma)^{-1}P^{-1}}\right\}\\
&+ \frac{\gamma}{2K\underline{p}}\mathbb{E}\left[\|y_{K+1} - y_1\|^2_{D(\sigma)^{-1}P^{-1}}\right] + \frac{\gamma}{K}\|x_0\|^2_{\tau^{-1}}\\
&+ \frac{1}{K}\mathbb{E}\left[f^*_{P^{-1}-I}(y_1) - f^*_{P^{-1}-I}(y_{K+1})\right] + \frac{1}{C_1 K}\Delta_0.
\end{aligned} \tag{4.66}$$

By Theorem 4.6 and Lemma 4.3, $\mathbb{E}\left[\|y_{K+1} - y_\star\|^2_{D(\sigma)^{-1}P^{-1}}\right] \le 2\Delta_0$, and by Jensen's inequality and concavity of square root, $\mathbb{E}\left[\|y_{K+1} - y_\star\|_{D(\sigma)^{-1}P^{-1}}\right] \le \sqrt{2\Delta_0}$. With these estimations we have

$$\mathbb{E}\left[\|y_{K+1} - y_1\|^2_{D(\sigma)^{-1}P^{-1}}\right] \le 2\|y_1 - y_\star\|^2_{D(\sigma)^{-1}P^{-1}} + 4\Delta_0. \tag{4.67}$$

As f_i is proper, lower semicontinuous, convex, and $A_i x^\star \in \partial f_i^*(y_i^\star)$, we additionally note that

$$f_i^*(y_i^{K+1}) \ge f_i^*(y_i^\star) + \langle A_i x^\star, y_{K+1}^{(i)} - y_\star^{(i)}\rangle \ge f_i^*(y_\star^{(i)}) - \|A_i x^\star\|_{D(\sigma)P}\|y_{K+1}^{(i)} - y_\star^{(i)}\|_{D(\sigma)^{-1}P^{-1}},$$

and

$$\begin{aligned}
\mathbb{E}[f^*_{P^{-1}-I}(y_{K+1})] &= \sum_{i=1}^{n}\left(\frac{1}{p_i} - 1\right)\mathbb{E}\left[f_i^*(y_{K+1}^{(i)})\right]\\
&\ge \sum_{i=1}^{n}\left(\frac{1}{p_i} - 1\right)\left(f_i^*(y_\star^{(i)}) - \|A_i x_\star\|_{D(\sigma)P}\mathbb{E}\left[\|y_{K+1}^{(i)} - y_\star^{(i)}\|_{D(\sigma)^{-1}P^{-1}}\right]\right)\\
&\ge \sum_{i=1}^{n}\left(\frac{1}{p_i} - 1\right)\left(f_i^*(y_\star^{(i)}) - \|A_i x^\star\|_{D(\sigma)P}\sqrt{2\Delta_0}\right).
\end{aligned} \tag{4.68}$$

We now use (4.67) and (4.68) in (4.66) to obtain

$$\begin{aligned}
\mathbb{E}\left[\sup_{z\in\mathcal{B}} \frac{1}{K}\sum_{k=1}^{K} \mathcal{H}(x_k, y_{k+1}; x, y)\right] \le{}& \frac{3}{2K}\sup_{x\in\mathcal{B}_x}\|x_0 - x\|^2_{\tau^{-1}} + \frac{1}{K}\sup_{y\in\mathcal{B}_y}\|y_1 - y\|^2_{D(\sigma)^{-1}P^{-1}}\\
&+ \frac{\gamma}{K\underline{p}}\|y_1 - y_\star\|^2_{D(\sigma)^{-1}P^{-1}} + \frac{2\gamma}{K\underline{p}}\Delta_0 + \frac{\gamma}{K}\|x_0\|^2_{\tau^{-1}} + \frac{1}{K}f^*_{P^{-1}-I}(y_1)\\
&+ \frac{1}{K}\sum_{i=1}^{n}\left(\frac{1}{p_i} - 1\right)\left(-f_i^*(y_\star^{(i)}) + \|A_i x_\star\|_{D(\sigma)P}\sqrt{2\Delta_0}\right) + \frac{1}{C_1 K}\Delta_0 =: \frac{C_{\mathcal{B}}}{K}.
\end{aligned}$$

We define as $C_{\mathcal{B}}$ the constant of right hand side and use Jensen's inequality on the left hand side with definitions of x_K^{avg} and y_{K+1}^{avg} to get the first result.

For the second part of the theorem, we proceed the same until (4.63). Then, we move the terms $\frac{1+2\gamma}{2K}\|x_0 - x\|^2_{\tau^{-1}}$ and $\frac{1}{K}\|y_1 - y\|^2_{D(\sigma)^{-1}P^{-1}}$ to the left hand side, take supremum, use the definition of smoothed gap, then take expectations of both sides and use the same estimations as in the first part to conclude. ■

4.7.6 Proof of Theorem 4.13

Proof. We have, from Theorem 4.12, the following bound for the smoothed gap (see also (4.7))

$$\mathbb{E}\left[\mathcal{G}_{\frac{1+2\gamma}{2K},\frac{1}{2K}}(x_K^{\text{avg}}, y_{K+1}^{\text{avg}}; x_0, y_1)\right] \le \frac{C_e}{K}.$$

• When f is Lipschitz continuous in the norm $\|\cdot\|_{D(\sigma)}$, we will argue as in [FB19, Theorem 11]. On (4.7), with the parameters used in this theorem, we make the following observations. By [BC11, Corollary 17.19], when f is $L(f)$-Lipschitz continuous in the norm $\|\cdot\|_{D(\sigma)}$, it follows that $\|y_1 - y\|^2_{D(\sigma)^{-1}} \le 4L(f)^2$. By Lipschitzness and the definition of conjugate function, we can pick $y \in \partial f(Ax_K^{\text{avg}}) \neq \emptyset$ such that $\langle Ax_K^{\text{avg}}, y\rangle - f^*(y) = f(Ax_K^{\text{avg}})$. Next by Fenchel-Young inequality, $f^*(y_{K+1}^{\text{avg}}) - \langle A^\top y_{K+1}^{\text{avg}}, x_\star\rangle \ge -f(Ax_\star)$. We also use $\underline{p} = \min_i p_i$ to obtain (see (4.7))

$$\begin{aligned}\mathbb{E}\left[\mathcal{G}_{\frac{1+2\gamma}{K},\frac{1}{K}}(x_K^{\text{avg}}, y_{K+1}^{\text{avg}}; x_0, y_1)\right] &\ge \mathbb{E}\left[f(Ax_K^{\text{avg}}) + g(x_K^{\text{avg}}) - f(Ax_\star) - g(x_\star)\right] \\ &\quad - \frac{2}{\underline{p}K}L(f)^2 - \frac{1+2\gamma}{2K}\|x_0 - x_\star\|^2_{\tau^{-1}},\end{aligned}$$

where the result directly follows.

• When $f(\cdot) = \delta_{\{b\}}(\cdot)$, we use [TDFC18, Lemma 1], to obtain the bounds

$$\begin{aligned}&\mathbb{E}\left[g(x_K^{\text{avg}}) - g(x_\star)\right] \le \mathbb{E}\left[\mathcal{G}_{\frac{1+2\gamma}{2K},\frac{1}{2K}}(x_K^{\text{avg}}, y_{K+1}^{\text{avg}}; x_0, y_1)\right] \\ &+ \frac{1+2\gamma}{2K}\|x_0 - x_\star\|^2_{\tau^{-1}} - \mathbb{E}\left[\langle y_\star, Ax_K^{\text{avg}} - b\rangle\right] + \frac{1}{2K}\|y_\star - y_1\|^2_{D(\sigma)^{-1}P^{-1}}, \\ &\mathbb{E}\left[\|Ax_K^{\text{avg}} - b\|_{D(\sigma)P}\right] \le \frac{1}{2K}\Bigg\{\|y_\star - y_1\|_{D(\sigma)^{-1}P^{-1}} + \Bigg(\|y_\star - y_1\|^2_{D(\sigma)^{-1}P^{-1}} \\ &+ 4K\mathbb{E}\left[\mathcal{G}_{\frac{1+2\gamma}{2K},\frac{1}{2K}}(x_K^{\text{avg}}, y_{K+1}^{\text{avg}}; x_0, y_1)\right] + 2(1+2\gamma)\|x_0 - x_\star\|^2_{\tau^{-1}}\Bigg)^{1/2}\Bigg\}.\end{aligned}$$

We use Cauchy-Schwarz inequality on $\langle y_\star, Ax_K^{\text{avg}} - b\rangle$ and use the bound of $\mathbb{E}\left[\|Ax_K^{\text{avg}} - b\|_{D(\sigma)P}\right]$ to conclude. ■

4.8 Bibliographic Note

Lemma 4.4 and Theorem 4.8 are mainly due to Olivier Fercoq. The results in Section 4.4.3 are mainly due to the author of this dissertation. The remaining results are joint between coauthors.

	Linear convergence	Rates with only convexity	Step sizes for linear convergence*
[CERS18]	$f_i^*: \mu_i$-s.c. $g: \mu_g$-s.c.	Ergodic $\mathcal{O}\left(\frac{1}{k}\right)$for Bregman distance to solution	$\|A_i\|, \mu_i, \mu_g$
[ZX17]	$f_i^*: \mu_i$-s.c. $g: \mu_g$-s.c.	Nonergodic $\mathcal{O}\left(\frac{1}{k}\right)$with bounded dual domain and fixed horizon	$\|A_i\|, \mu_i, \mu_g$
[FB19]	$f_i^*: \mu_i$-s.c. $g: \mu_g$-s.c.	Randomly selected iterate $\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$	$n^2\tau\sigma_i\|A_i\|^2 < 1$
[LFP19]	F is MS (see (4.3))	×	$n^2\tau\sigma_i\|A_i\|^2 < 1$
This chapter	F is MS (see (4.3))	Ergodic $\mathcal{O}\left(\frac{1}{k}\right)$ for primal-dual gap, objective values and feasibility	$n\tau\sigma_i\|A_i\|^2 < 1$

Table 4.1 – Comparison of primal dual coordinate descent methods. s.c. denotes strongly convex, MS denotes metrically subregular. Please see Section 4.5 for a thorough comparison. Please see Section 7.2 for comparison of MS and s.c. assumptions. *Step sizes are for optimization with a potentially dense A matrix and uniform sampling: $p_i = 1/n$.

	a.s. convergence	Linear convergence	Ergodic rates
[CERS18]	$D_h(z_k, z_\star) \to 0$, for any $z_\star$ where D_h is Bregman distance generated by $h(z) = f^*(y) + g(x)$	Assumption: f_i^*, g s.c. step sizes depending on μ_i, μ_g	$D_h(z_k^{\text{avg}}, z_\star) = \mathcal{O}(1/k)$
This chapter	$z_k \to z_\star$, for some $z_\star$.	Assumption: F in (4.3) is MS Step sizes: $n\tau\sigma_i\|A_i\|^2 < 1^*$	• Restricted primal-dual gap $\mathbb{E}\left[\text{Gap}_{\mathcal{B}}(x_k^{\text{avg}}, y_k^{\text{avg}})\right] = \mathcal{O}(1/k)$ • f is Lipschitz† $\mathbb{E}\left[\lvert P(x_k^{\text{avg}}) - P(x_\star)\rvert\right] = \mathcal{O}(1/k)$ • $f(\cdot) = \delta_b(\cdot)$ $\mathbb{E}\left[\lvert g(x_k^{\text{avg}}) - g(x_\star)\rvert\right] = \mathcal{O}(1/k)$ $\mathbb{E}\left[\|Ax_k^{\text{avg}} - b\|\right] = \mathcal{O}(1/k)$

Table 4.2 – Comparison of our results and previous results on SPDHG. *Step size condition is for uniform sampling: $p_i = 1/n$. †In this case $P(x) := f(Ax) + g(x)$.

5 A sparsity aware primal-dual coordinate descent algorithm

In this chapter, we continue studying PDCD methods and propose a new algorithm to improve SPDHG analyzed in the previous chapter. The new method bridges the benefits of SPDHG with other PDCD methods designed for sparse data [FB19, LFP19]. In addition to rigorous convergence guarantees, we show that the method has per-iteration cost depending on the number of nonzeros of the data matrix, which was not the case for SPDHG. As predicted by our theory, the new method attains a compelling empirical performance with both dense and sparse datasets.

This chapter is based on the joint work with Olivier Fercoq and Volkan Cevher [AFC20].

5.1 Introduction

In this chapter, we consider the problem

$$\min_{x\in\mathcal{X}} f(x)+g(x)+h(Ax), \tag{5.1}$$

where $f,g\colon \mathcal{X}\to\mathbb{R}\cup\{+\infty\}$ and $h\colon \mathcal{Y}\to\mathbb{R}\cup\{+\infty\}$ are proper, lower semicontinuous, convex functions, $A\colon \mathcal{X}\to\mathcal{Y}$ is a linear operator. $\mathcal{X}$ and $\mathcal{Y}$ are Euclidean spaces such that $\mathcal{X}=\prod_{i=1}^n \mathcal{X}_i$, and $\mathcal{Y}=\prod_{j=1}^m \mathcal{Y}_j$. Moreover, f is assumed to have coordinate-wise Lipschitz continuous gradients and g,h admit easily computable proximal operators.

We recall that the advantage of coordinate-based methods is that they access to blocks of A and update a subset of variables, resulting in cheap per iteration costs. Moreover, they utilize larger step sizes depending on the properties of the problem in selected blocks. Existing PDCD methods fail to retain both these advantages, as sparsity of A varies. In particular, methods that have cheap per-iteration costs with sparse A [FB19, LFP19], are restricted to use small step sizes with dense A. On the other hand, methods that can use large step sizes with dense A such as SPDHG [CERS18], have high per-iteration costs with sparse A.

Contributions. In this chapter, we identify random extrapolation as the key to design a

	Step sizes with dense data	iteration cost	block-wise Lipschitz	probability law	Efficient implementation
[CERS18]	$n\tau_i\sigma\|A_i\|^2 < 1$	m	N/A	arbitrary	direct[†]
[FB19]	$n^2\tau_i\sigma\|A_i\|^2 < 1$	$\lvert J(i)\rvert^*$	Yes	uniform	direct or dupl.
[LFP19]	$n^2\tau_i\sigma\|A_i\|^2 < 1$	$\lvert J(i)\rvert^*$	No	arbitrary	duplication[‡]
PURE-CD	$n\tau_i\sigma\|A_i\|^2 < 1$	$\lvert J(i)\rvert^*$	Yes	arbitrary	direct

Table 5.1 – Comparison of PDCD methods. We only compare here the most related methods to ours and include a comprehensive literature review in Section 5.5. In the last column, we refer to the way one needs to implement the algorithm, for it to be efficient in both sparse and dense settings. [*] $J(i)$ is defined in (5.2). [†]SPDHG only has implementation for dense setting and not for sparse. [‡]The concept of duplication for PDCD is described in [FB19].

method that combines the benefits of the methods in two camps and propose the primal-dual method with random extrapolation and coordinate descent (PURE-CD).

▷ PURE-CD exhibits the advantages of [FB19, LFP19] in the sparse setting and the advantages of [CERS18] in the dense setting simultaneously, achieving the best of both worlds.

▷ As PURE-CD has the favorable properties in both ends of the spectrum, it has the best performance in the regime in between: moderately sparse data (see also Section 5.6.1). Table 5.1 compiles a summary for the comparison of PURE-CD and previous methods.

▷ In addition to adapting to the sparsity of A, we prove that PURE-CD also adapts to unknown structures in the problem, and obtains linear rate of convergence, without any modifications in the step sizes, with metric subregularity.

▷ In the general convex case, we prove that the iterates of PURE-CD converges almost surely to a solution of problem (5.1).

▷ We show that in this case, the ergodic sequence obtains the optimal $\mathcal{O}(1/k)$ rate of convergence on the expected primal-dual gap.

5.2 Preliminaries

5.2.1 Notation

For $u \in \mathcal{X}_i$, $U_i(u) \in \mathcal{X}$ is such that each element of $U_i(u)$ is 0, except the block i which contains u. We use block size of 1 for simplicity. We use the following notation for the sparse setting,

$$\begin{aligned} J(i) &= \{j \in \{1,\dots,m\} \colon\ A_{j,i} \neq 0\} \\ I(j) &= \{i \in \{1,\dots,n\} \colon\ A_{j,i} \neq 0\}. \end{aligned} \tag{5.2}$$

In words, given a matrix A and $i \in \{1,\dots,n\}$, $J(i)$ denotes the row indices that correspond to nonzero values in the column indexed by i. Similarly, with $j \in \{1,\dots,m\}$, $I(j)$ gives the column indices corresponding to nonzero values in the row indexed by j.

Moreover, given positive probabilities $(p_i)_{1\le i\le n}$, we define

$$\pi_j = \sum_{i\in I(j)} p_i. \tag{5.3}$$

In the simple case of $p_i = 1/n$, it is easy to see that $n\pi_j$ corresponds to number of nonzeros in the row indexed by j.

At iteration k, the algorithm randomly picks an index $i_{k+1} \in \{1,\dots,n\}$. To govern the selection rule, we define the probability matrix $P = \operatorname{diag}(p_1,\dots,p_n)$, where $p_i = \Pr(i_{k+1} = i)$, and $\underline{p} = \min_i p_i$. We define as $\mathcal{F}_k$ the filtration generated by the random indices $\{i_1,\dots,i_k\}$.

Denoting $z = (x, y)$, we define the functions

$$\begin{aligned}
D_p(x_{k+1}, z) &= f(x_{k+1}) + g(x_{k+1}) - f(x) - g(x) + \langle A^\top y, x_{k+1} - x\rangle,\\
D_d(\bar{y}_{k+1}, z) &= h^*(\bar{y}_{k+1}) - h^*(y) - \langle Ax, \bar{y}_{k+1} - y\rangle.
\end{aligned}$$

Optimality. Problem (5.1) has the following saddle point formulation

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} f(x) + g(x) + \langle Ax, y\rangle - h^*(y).$$

KKT conditions state that the vector $z_\star = (x_\star, y_\star)$ is a primal-dual solution when

$$0 \in \begin{bmatrix} \nabla f(x_\star) + \partial g(x_\star) + A^\top y_\star \\ Ax_\star - \partial h^*(y_\star) \end{bmatrix} =: F(z_\star). \tag{5.4}$$

We call $\mathcal{Z}_\star$ the set of such solutions.

Metric subregularity. We use the metric subregularity assumption for proving linear convergence. We refer to Section 4.2.2 for the details. We are interested in the metric subregularity of KKT operator F (see (5.4)) for 0. Intuitively speaking, as $0 \in F(z_\star), \forall z^\star \in \mathcal{Z}_\star$, metric subregularity of F for 0 essentially gives us a way to characterize the behavior of the iterates around the solution set.

We state our main assumptions which are standard in the literature [FB19, CERS18, LFP19]:

Assumption 5.1. *f, g and h are proper, lower semicontinuous, convex.*

▷ g is separable, i.e., $g(x) = \sum_{i=1}^n g_i(x^{(i)})$, and f has coordinatewise Lipschitz gradients such that $\forall x \in \mathcal{X}, \forall u \in \mathcal{X}_i$,

$$f(x + U_i(u)) \le f(x) + \langle \nabla_i f(x), u\rangle + \frac{\beta_i}{2}\|u\|^2. \tag{5.5}$$

▷ Set of solutions to problem (5.1) is nonempty.

▷ Slater's condition holds.

5.3 Algorithm

In this section, we sketch the main ideas behind our algorithm. PDHG[1], due to [CP11, Con13, Vũ13] reads as

$$\begin{aligned} \bar{x}_{k+1} &= \mathrm{prox}_{\tau,g}\left(\bar{x}_k - \tau\left(\nabla f(\bar{x}_k) + A^\top \bar{y}_k\right)\right) \\ \bar{y}_{k+1} &= \mathrm{prox}_{\sigma,h^*}\left(\bar{y}_k + \sigma A(2\bar{x}_{k+1} - \bar{x}_k)\right). \end{aligned} \tag{5.6}$$

The main intuition behind PDCD methods proposed by [ZX17, FB19, CERS18] is to incorporate coordinate based updates. Among these methods, [ZX17] specializes in strongly convex-strongly concave problems, whereas the other other ones focus on convex-concave problems.

A closely related approach concentrated on the following interpretation of primal-dual method (5.6) which is named as TriPD in [LFP19, Algorithm 1]

$$\begin{aligned} \bar{y}_{k+1} &= \mathrm{prox}_{\sigma,h^*}\left(\hat{y}_k + \sigma A\bar{x}_k\right) \\ \bar{x}_{k+1} &= \mathrm{prox}_{\tau,g}\left(\bar{x}_k - \tau\left(\nabla f(\bar{x}_k) + A^\top \bar{y}_{k+1}\right)\right) \\ \hat{y}_{k+1} &= \bar{y}_{k+1} + \sigma A(\bar{x}_{k+1} - \bar{x}_k). \end{aligned} \tag{5.7}$$

By moving the $\bar{y}_{k+1}$ update in TriPD to take place after $\hat{y}_{k+1}$ update, one obtains (5.6).

As observed in [LFP19], this particular interpretation of primal-dual method is useful for randomization. TriPD-BC as proposed in [LFP19] iterates as

$$\begin{aligned} &\bar{y}_{k+1} = \mathrm{prox}_{\sigma,h^*}\left(y_k + \sigma A x_k\right) \\ &\bar{x}_{k+1} = \mathrm{prox}_{\tau,g}\left(x_k - \tau\left(\nabla f(x_k) + A^\top \bar{y}_{k+1}\right)\right) \\ &\hat{y}_{k+1} = \bar{y}_{k+1} + \sigma A(\bar{x}_{k+1} - x_k) \\ &\text{Draw an index } i_{k+1} \in \{1,\dots,n\} \text{ randomly.} \\ &x_{k+1}^{(i_{k+1})} = \bar{x}_{k+1}^{(i_{k+1})}, \quad x_{k+1}^{(j)} = x_k^{(j)}, \forall j \neq i_{k+1} \\ &y_{k+1}^{(j)} = \hat{y}_{k+1}^{(j)}, \forall j \in J(i_{k+1}), \quad y_{k+1}^{(j)} = y_k^{(j)}, \forall j \not\in J(i_{k+1}). \end{aligned}$$

One immediate limitation of TriPD-BC is that to update y_{k+1}, it needs $\bar{x}_{k+1}$, whereas only $\bar{x}_{k+1}^{(i_{k+1})}$ is needed to update x_{k+1}. As also discussed in [LFP19], this scheme is suitable when A has special structure such as sparsity. When A is dense, the method updates all elements of y_{k+1} and $\hat{y}_{k+1}$, in which case both $\bar{y}_{k+1}$ and $\bar{x}_{k+1}$ are computed, which has the same cost as a deterministic algorithm. For an efficient implementation in the dense setting, one can use duplication of dual variables as described in [FB19]. However, in this case the method is restricted to use small step sizes as discussed in [FB19]. Compared to SPDHG in [CERS18], the step sizes can be n times worse, deteriorating the performance of the method considerably in the dense setting.

On the other hand, the drawback of SPDHG is that it needs to update all dual variables (or all primal variables for the formulation in Chapter 4) at every iteration, whereas the methods

[1] This method is also known as Vũ-Condat algorithm.

in [FB19, LFP19] update only a subset of dual variables depending on the sparsity of A. When the dual dimension is high, per iteration cost of [CERS18] can become prohibitive.

Our idea, inspired by [CERS18], to make TriPD-BC efficient for dense setting is to use x_{k+1} rather than $\bar{x}_{k+1}$ in the update of $\hat{y}_{k+1}$. Although simple to state, this modification makes $\hat{y}_{k+1}$ random, rendering the analysis of [LFP19] and other analyses working with monotone operators not applicable.

This leads to our algorithm, primal-dual method with random extrapolation and coordinate descent (PURE-CD). Our method uses large step sizes as in [CERS18] in the dense setting, while staying efficient in terms of per iteration costs in the sparse setting as in [FB19, LFP19]. These make PURE-CD the first PDCD algorithm that provably obtains favorable properties in both sparse and dense settings.

Algorithm 5.1 Primal-dual method with random extrapolation and coordinate descent (PURE-CD)

```
Input: Diagonal matrices $\theta, \tau, \sigma > 0$, chosen according to (5.8), (5.9).
for $k = 0, 1 \dots$ do
    $\bar{y}_{k+1} = \mathrm{prox}_{\sigma, h^*}\left(y_k + \sigma A x_k\right)$
    $\bar{x}_{k+1} = \mathrm{prox}_{\tau, g}\left(x_k - \tau\left(\nabla f(x_k) + A^\top \bar{y}_{k+1}\right)\right)$
    Draw $i_{k+1} \in \{1, \dots, n\}$ with $\Pr(i_{k+1} = i) = p_i$
    $x_{k+1}^{(i_{k+1})} = \bar{x}_{k+1}^{(i_{k+1})}$
    $x_{k+1}^{(j)} = x_k^{(j)}, \forall j \neq i_{k+1}$
    $y_{k+1}^{(j)} = \bar{y}_{k+1}^{(j)} + \sigma_j \theta_j (A(x_{k+1} - x_k))^{(j)}, \forall j \in J(i_{k+1}),\ y_{k+1}^{(j)} = y_k^{(j)}, \forall j \notin J(i_{k+1})$
end for
```

5.4 Convergence Analysis

In this section, we analyze the convergence behavior of Algorithm 5.1 under various assumptions. We first start with a lemma analyzing one iteration of the algorithm.

Lemma 5.1. *Let Assumption 5.1 hold. Recall the definitions of D_p and D_d from Section 5.2.1 and let $\theta = \mathrm{diag}(\theta_1, \dots, \theta_m)$ and $\pi = \mathrm{diag}(\pi_1, \dots, \pi_m)$ be chosen as*

$$\theta_j = \frac{\pi_j}{\underline{p}}, \text{ where } \pi_j = \sum_{i \in I(j)} p_i, \text{ and } \underline{p} = \min_i p_i. \tag{5.8}$$

We define the functions, given z,

$$V(z) = \frac{\underline{p}}{2}\|x\|^2_{\tau^{-1}P^{-1}} + \frac{\underline{p}}{2}\|y\|^2_{\sigma^{-1}\pi^{-1}},$$
$$\tilde{V}(z) = \frac{\underline{p}}{2}\|x\|^2_{C(\tau)} + \frac{\underline{p}}{2}\|y\|^2_{\sigma^{-1}},$$

where $C(\tau)_i = \frac{2p_i}{\underline{p}\tau_i} - \frac{1}{\tau_i} - p_i \sum_{j=1}^m \pi_j^{-1} \sigma_j \theta_j^2 A_{j,i}^2 - \frac{\beta_i p_i}{\underline{p}}$.

Then, for the iterates of Algorithm 5.1, $\forall z \in \mathcal{Z}$, it holds that:

$$\mathbb{E}_k\left[D_p(x_{k+1},z)\right] + \underline{p}D_d(\bar{y}_{k+1},z) + \mathbb{E}_k\left[V(z_{k+1}-z)\right] \\ \leq (1-\underline{p})D_p(x_k,z) + V(z_k - z) - \tilde{V}(\bar{z}_{k+1}-z_k).$$

The main technical challenge in the proof of the lemma, compared to the corresponding results in [LFP19] and [CERS18] is handling stochasticity in both variables x_{k+1}, y_{k+1} (and also $\hat{y}_{k+1}$ for [LFP19]). Using coordinatewise Lipschitz constants of f with arbitrary sampling also requires an intricate analysis.

The result of Lemma 5.1 is promising for deriving convergence results for Algorithm 5.1. When $z = z_\star$ in Lemma 5.1, as $D_p(x_{k+1}, z_\star) \geq 0$, $D_d(\bar{y}_{k+1}, z_\star) \geq 0$ and when step sizes are chosen such that $\tilde{V}$ is a squared norm, Lemma 5.1 describes a stochastic monotonicity property similar to [FB19]. In particular, it shows that $D_p(x_{k+1}, z_\star) + V(z_{k+1} - z_\star)$ which measures the distance to solution in a Bregman distance sense, is monotonically nonincreasing in expectation.

5.4.1 Almost sure convergence

Almost sure convergence is a fundamental property for randomized methods describing the limiting behavior of the iterates in different realization of the algorithm. The following theorem states that the iterates of Algorithm 5.1 converge almost surely to a point in the solution set.

Theorem 5.2. *Let Assumption 5.1 hold and let θ, π be as in Lemma 5.1. Choose step sizes τ, σ such that*

$$\tau_i < \frac{2p_i - \underline{p}}{\beta_i p_i + \underline{p}^{-1} p_i \sum_{j=1}^m \pi_j \sigma_j A_{j,i}^2}. \tag{5.9}$$

For the iterates z_k of Algorithm 5.1, almost surely there exist $z_\star \in \mathcal{Z}_\star$ such that $z_k \to z_\star$.

We analyze the step size rule (5.9) in Theorem 5.2 and compare with existing efficient methods in dense and sparse settings.

Remark 5.3.
▷ Let A be dense, with all its elements being nonzero, $p_i = 1/n$ and $f(\cdot) = 0$, then the step size rule reduces to

$$\tau_i < \frac{1}{n\sigma \|A_i\|^2},$$

which is the step size rule of SPDHG [CERS18, AFC21], making it favorable in the dense setting (see Section 4.6.1). In contrast, step size rules of [FB19, LFP19] are n times worse due to duplication, in this case.

⊳ Let A be diagonal, and we use $p_i = \frac{1}{n}$, which results in $\pi_j = \frac{1}{n}$. Then,

$$\tau_i < \frac{1}{\beta_i + \sum_{j=1}^{m} \sigma_j A_{j,i}^2},$$

which is the step size rule of Vu-Condat-CD [FB19], upon using the definition of $J(i)$ from (5.2). Similarly, Algorithm 5.1 updates 1 dual coordinate and 1 primal coordinate, in this case. In contrast, SPDHG [CERS18] updates m dual coordinates, resulting in m times higher per iteration cost.

We note that the step sizes of TriPD-BC [LFP19] depend on global Lipschitz constant of f rather than the coordinatewise ones. Using coordinatewise Lipschitz constants in practice is important for the success of coordinate descent, as they give larger step sizes [Nes12, RT14, FR15].

The takeaway from Remark 5.3 is that Algorithm 5.1 recovers the characteristics of the best performing methods in fully dense and fully sparse settings. Moreover, as it is the only method with the desirable dependencies in both cases, it has the best properties in the moderate sparse cases. We validate this observation with numerical experiments in Section 5.6.

5.4.2 Linear convergence

Linear convergence of primal-dual methods in practice is a widely observed phenomenon [CP11, LFP16]. We show that Algorithm 5.1 also shares this property and obtains linear convergence under metric subregularity, without any modification on the algorithm.

We define the Bregman-type projection onto the solution set

$$z_{\star,k} = \arg\min_{u \in \mathcal{Z}_\star} D_p(x_k, u) + V(z_k - u). \tag{5.10}$$

We now show that $z_{\star,k}$ is well-defined under our assumptions. First, the solution set is convex and closed. Second, $D_p(x_k, u) \geq 0$ for all $u \in \mathcal{Z}_\star$ and it is also lower semicontinuous. Third, we remark that $V(z_k - u)$ is a squared norm (see Lemma 5.1), thus coercive, therefore the sum is coercive and lower semicontinuous over $\mathcal{Z}_\star$. Hence, $z_{\star,k}$ exists.

The definition of $z_{\star,k}$ in (5.10) is more involved compared to the corresponding quantity in [LFP19]. This is in fact due to us using coordinatewise Lipschitz constants in our step sizes, rather than the global Lipschitz constant in [LFP19].

Assumption 5.2.
KKT operator F is metrically subregular at all $z_\star \in \mathcal{Z}_\star$ for 0, and $\bar{z}_k \in \mathcal{N}(z_\star), \forall z_\star, \forall k$.

Theorem 5.4. *Let Assumptions 5.1 and 5.2 hold. Let θ and the step sizes τ, σ be chosen according to* (5.8) *and* (5.9)*, respectively. Moreover, $z_{\star,k} = (x_{\star,k}, y_{\star,k})$ is as defined in* (5.10)*. Then, for z_k*

generated by Algorithm 5.1, it follows that

$$\mathbb{E}\left[\frac{p}{2}\|x_k - x_{\star,k}\|^2_{\tau^{-1}P^{-1}} + \frac{p}{2}\|y_k - y_{\star,k}\|^2_{\sigma^{-1}\pi^{-1}}\right] \leq (1-\rho)^k \Delta_0,$$

where $\rho = \min\left(\underline{p}, \frac{C_{2,\tilde{V}}}{C_{V,2}((2+2c)+(1+c)(\eta\|H-M\|+\bar{\beta}))^2}\right)$, $\Delta_0 = D_p(x_0, z_{\star,0}) + V(z_0 - z_0^\star)$, $\bar{\beta}$ *is the global Lipschitz constant of* f*,*
$C_{2,\tilde{V}} = \frac{p}{2}\min\left\{\min_i C(\tau)_i, \min_j \sigma_j^{-1}\right\}$, $C_{V,2} = \frac{1}{2}\max\left\{\max_i \frac{1}{\tau_i}, \max_j \frac{1}{\sigma_j}\right\}$, $c = C_{2,V}\sqrt{\|A\|/2}$*,*
$C_{2,V} = \sqrt{\frac{2}{\underline{p}\min\left\{\min_i \tau_i^{-1}p_i^{-1}, \min_j \sigma_j^{-1}\pi_j^{-1}\right\}}}$*, and*

$$H = \begin{bmatrix} \tau^{-1} & A^\top \\ 0 & \sigma^{-1} \end{bmatrix}, \quad M = \begin{bmatrix} 0 & A^\top \\ -A & 0 \end{bmatrix}.$$

The proof is given in Section 5.8.3. The first remark about Theorem 5.4 is similar to Chapter 4. Since metric subregularity constant η is not required in the algorithm, the step sizes to achieve linear convergence are the same step sizes as (5.9). Therefore, PURE-CD adapts to structures on the problem, without any need to modify the algorithm, and attains linear rate of convergence.

We refer to Section 4.2.2 for example problems when metric subregularity holds and as a result PURE-CD obtains linear convergence. Compared with the linear convergence rate in [LFP19] for TriPD-BC, our result have a similar contraction factor, however, due to larger step sizes (see Remark 5.3), the rate comes with a better constant.

5.4.3 Ergodic rates

In this section, we study Algorithm 5.1 in the general case, under Assumption 5.1, and show the optimal $\mathcal{O}(1/k)$ convergence rate on the ergodic sequence. The quantity of interest is the primal-dual gap function [CP11]

$$\mathrm{Gap}(\bar{x}, \bar{y}) = \sup_{z \in \mathcal{Z}} f(\bar{x}) + g(\bar{x}) + \langle A\bar{x}, y\rangle - h^*(y) - f(x) - g(x) - \langle Ax, \bar{y}\rangle + h^*(\bar{y}). \tag{5.11}$$

A related quantity is the restricted gap function [CP11], which, for any set $\mathcal{C} \subset \mathcal{Z}$ is defined as

$$\mathrm{Gap}_{\mathcal{C}}(\bar{x}, \bar{y}) = \sup_{z \in \mathcal{C}} f(\bar{x}) + g(\bar{x}) + \langle A\bar{x}, y\rangle - h^*(y) - f(x) - g(x) - \langle Ax, \bar{y}\rangle + h^*(\bar{y}). \tag{5.12}$$

See [Nes07, Lemma 1] for validity of restricted gap function as an optimality measure.

Due to randomization in PDCD, we are interested in the expected primal-dual gap, denoted as $\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(\bar{x}, \bar{y})\right]$. As described in Chapter 4, it is technically challenging to prove rates for this quantity as it is the expectation of supremum. We use the technique we introduced in Chapter 4 to show convergence of expected primal-dual gap for SPDHG of [CERS18]. This rate is for ergodic sequence averaging x_k and the full dual variable $\bar{y}_k$. We can use this technique for our analysis. However, there remains another challenge as full dual variable is not computed in PURE-CD. Thus, averaging $\bar{y}_k$ is not feasible in our case.

In addition to Assumption 5.1, in this section we will assume separability of h, to be able to do an efficient averaging with the dual iterate.

Due to the asymmetric nature of Algorithm 5.1, there are fundamental difficulties for proving a rate with averaging y_{k+1}. On this front, we propose a new type of analysis for the dual variable. To start with, we define the following iterate which has the same cost to compute as y_{k+1} each iteration. Let $\breve{y}_1 = y_1 = \bar{y}_1$,

$$\begin{aligned} \breve{y}^{(j)}_{k+1} &= \bar{y}^{(j)}_{k+1}, \quad \forall j \in J(i_{k+1}), \\ \breve{y}^{(j)}_{k+1} &= \breve{y}^{(j)}_{k}, \quad \forall j \notin J(i_{k+1}). \end{aligned} \tag{5.13}$$

We note that $\breve{y}_k$ is $\mathcal{F}_k$-measurable and more useful properties of $\breve{y}_k$ for analysis are given in Lemma 5.11 in Section 5.8.4.

Due to the definition of $\breve{y}_k$, it is now feasible to compute and average this iterate. We can show the convergence of expected primal-dual gap by averaging $\breve{y}_k$ and x_k. We remark that we use some coarse inequalities to give simple constants for Theorem 5.5 and Theorem 5.7. Therefore, the bounds are not optimized with respect to dimension dependence. In Section 5.8.4, we give these theorems with their original, tighter bounds and we show how we transform the tighter bounds into the constants we give in this section.

Theorem 5.5. *Let Assumption 5.1 hold and θ, τ, σ are chosen as in* (5.8), (5.9)*. Moreover, let h be separable. Let $x^{av}_K = \frac{1}{K}\sum_{k=1}^K x_k$ and $y^{av}_K = \frac{1}{K}\sum_{k=1}^K \breve{y}_k$, where $\breve{y}_k$ is defined in* (5.13)*, then for any bounded set $\mathcal{C} = \mathcal{C}_x \times \mathcal{C}_y \subset \mathcal{Z}$, with iterates of Algorithm 5.1, we have*

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(x^{av}_K, y^{av}_K)\right] \le \frac{C_g}{\underline{p}K},$$

where $C_g = \sum_{i=1}^4 C_{g,i}$, $C_{\tau,\tilde{V}} = \min_i C(\tau)_i \tau_i$,
$C_{g,1} = \sup_{z\in\mathcal{C}} \left\{2\underline{p}\|x_0 - x\|^2_{\tau^{-1}P^{-1}} + 2\underline{p}\|y_0 - y\|^2_{\sigma^{-1}\pi^{-1}}\right\} + 4\sqrt{\Delta_0 \underline{p}^{-1}}\|A\| \sup_{y\in\mathcal{C}_y} \|y\|_{\tau P}$
$+ 2\sqrt{\Delta_0(\underline{p}^{-1} + 2\underline{p}^{-3}C^{-1}_{\tau,\tilde{V}})}\|A\| \sup_{x\in\mathcal{C}_x} \|x\|_{\sigma\pi}$,
$\sum_{i=2}^4 C_{g,i} = \Delta_0\left(5 + 9\underline{p}^{-1} + C^{-1}_{\tau,\tilde{V}}\left(1 + 10\underline{p}^{-1} + 14\underline{p}^{-2}\right)\right) + (1 - \underline{p})(f(x_0) + g(x_0) - f(x_\star) - g(x_\star)) +$
$h^*(y_0) - h^*(y_\star) + \underline{p}\|Ax_\star\|^2_{\sigma\pi^{-1}} + \|A^\top y_\star\|^2_{\tau P}$.

Remark 5.6. When implementing averaging of x_k, and $\breve{y}_k$, one should use a technique similar to [DL15b]. The main idea is to only update the averaged vector at the coordinates where an update occurred. For this, we remember for each coordinate, the last time it is updated, wait until a coordinate is selected again and update the averaged vector using this information.

The result in Theorem 5.5 would give a rate for primal-dual gap when $\mathcal{C} = \mathcal{Z}$. However, in general such a rate is not desirable as taking a supremum over $\mathcal{Z}$ might result in an unbounded constant. This rate would be meaningful when both primal and dual domains are bounded in which case one would take the supremum in $C_{g,1}$ over the bounded domains.

Alternatively, in the following theorem, we show that for two important special cases, we can extend this result to show guarantees without bounded domains. Namely, we show the same rate for the case when $h(\cdot) = \delta_{\{b\}}(\cdot)$, $b \in \mathbb{R}^m$ to cover linearly constrained problems. Moreover, we show the result for the case when h is Lipschitz continuous.

Theorem 5.7. *Let Assumption 5.1 hold. We use the same parameters θ, τ, σ and the definitions for x_K^{av} and y_K^{av} as Theorem 5.5. We consider two cases separately:*
▷ *If $h(\cdot) = \delta_{\{b\}}(\cdot)$, we obtain*

$$\mathbb{E}[f(x_K^{av}) + g(x_K^{av}) - f(x_\star) - g(x_\star)] \le \frac{C_o}{\underline{p}K}.$$

$$\mathbb{E}[\|Ax_K^{av} - b\|] \le \frac{C_f}{\underline{p}K}.$$

▷ *If h is L_h-Lipschitz continuous, we obtain*

$$\mathbb{E}[f(x_K^{av}) + g(x_K^{av}) + h(Ax_K^{av}) - f(x_\star) - g(x_\star) - h(Ax_\star)] \le \frac{C_l}{\underline{p}K},$$

where $C_f = 3c_1\|x_\star - x_0\|_{\tau^{-1}P^{-1}} + 2\sqrt{c_1 C_s} + 4c_1\|y_\star - y_0\|_{\sigma^{-1}\pi^{-1}}$,
$C_o = C_s + \|y_\star\|_{\sigma^{-1}\pi^{-1}} C_f + 2c_1 \underline{p}^{-1} V(z_0 - z_\star)$,
$C_l = C_s + c_1\|x_\star - x_0\|^2_{\tau^{-1}P^{-1}} + 4c_1 L_h^2$,
$C_s = C_{g,2} + C_{g,5} + C_{g,6}$, with $c_1 = 2\underline{p} + 2$, $C_{g,2}$ as defined in the statement of Theorem 5.5 in Section 5.8.4 and $C_{g,5}, C_{g,6}$ are defined in the proof in (5.81), (5.82).

5.5 Related works

One of the first PDCD methods is SPDC, which is proposed in [ZX17], that solves a special case of problem (5.1) with $f = 0$. SPDC has linear convergence when g, h^* are strongly convex and the step sizes are selected according to strong convexity constants. In the general convex case, SPDC has perturbation-based analysis, which needs to set an ϵ, requires knowing $\|x_\star\|^2$, and shows ϵ-based iteration complexity results, and not anytime convergence rates. Almost sure convergence of the iterates of SPDC is not proven in the general convex case. Moreover, the step sizes of SPDC are scalar and they depend on the maximum block norm of A. It is shown in [ZX17] that in the specific cases when $g(x) = \|x\|^2$ or $g(x) = \|x\|_1 + \|x\|^2$, one can use a special implementation for efficiency with sparse data.

[TQMZ20] proposed a new method similar to SPDC with the same type of guarantees as [ZX17]. Due to similar analysis techniques, this method inherits the abovementioned drawbacks of SPDC. For this method, [TQMZ20] showed a new implementation technique for sparse data, that can be used with any separable $g(x)$.

An early PDCD method is proposed by [DL14] where the authors focused on showing sublinear convergence rates. The authors showed guarantees for a weaker version of expected primal-dual gap function in (5.11).

Building on [DL14], block-coordinate variants of alternating direction method of multipliers (ADMM) are proposed in [GXZ19, XZ18]. These papers focus on linearly constrained problems and show ergodic sublinear convergence rates. Moreover, [XZ18] showed that under strong convexity assumption and special decomposition of the blocks, the method achieves linear convergence. This linear convergence result, similar to [ZX17] requires knowing the strong convexity constants to set the algorithmic parameters. Moreover, these results generally set step sizes depending on global Lipschitz constants and norm of whole matrix A.

Another early PDCD variant to solve problem (5.1) in its full generality, where f, g, h are all non-separable, is by [FB19]. This method uses coordinatewise Lipschitz constants of the smooth part and it is designed to exploit sparsity of A. This method has almost sure convergence guarantees as well as linear convergence when g, h^* are strongly convex. As opposed to most results in this nature, it is not required to know strong convexity constants to set the step sizes. In the general convex case, the method has $\mathcal{O}(1/\sqrt{k})$ rate for a randomly selected iterate. As argued in Section 5.4.1, main limitation of [FB19] is that small step sizes are required when matrix A is dense. Moreover, the results in this paper are restricted to uniform probability law for selecting coordinates.

One of the most related works to ours, and a building block of PURE-CD is TriPD-BC [LFP19]. The authors showed almost sure convergence of the iterates and linear convergence under metric subregularity, by using global Lipschitz constants of f for the step sizes. This work did not have any sublinear convergence rates in the general convex case. Similar to [FB19], TriPD-BC is designed for sparse setting and an efficient implementation is by duplication of dual variables, which as explained in [FB19] results in small step sizes (see Sections 5.4.1 and 6.2).

Another building block of PURE-CD is SPDHG by [CERS18], to solve (5.1) when $f = 0$. Linear convergence result of SPDHG by [CERS18] is similar to [ZX17] and requires setting step sizes with strong convexity constants. In the partially strongly convex case, [CERS18] proved $\mathcal{O}(1/k^2)$ rates. Moreover, our analysis in Chapter 4 gave stronger results for the method. As explained before, even though SPDHG is fast with dense data, it needs to update all the dual coordinates, resulting in high per iterations costs with sparse data (see Sections 5.4.1 and 6.2).

5.6 Numerical experiments

5.6.1 Effect of sparsity

As explained in Section 5.4.1, and Remark 5.3, PURE-CD brings together the benefits of different methods that are designed for dense and sparse cases. We now compare the empirical

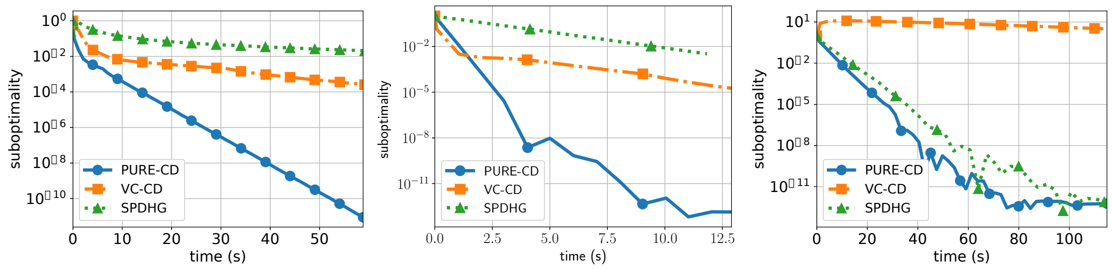

Figure 5.1 – Lasso: Left: rcv1, $n = 20,242$, $m = 47,236$, density = 0.16%, $\lambda = 10$; Middle: w8a, $n = 49,749$, $m = 300$, density = 3.9%, $\lambda = 10^{-1}$; Right: covtype, $n = 581,012$, $m = 54$, density = 22.1%, $\lambda = 10$.

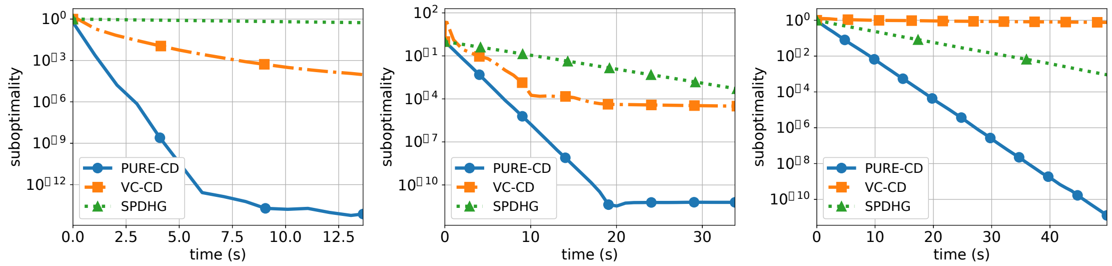

Figure 5.2 – Ridge regression: Left: sector, $n = 6,412$, $m = 55,197$, density = 0.3%, $\lambda = 0.1$; Middle: a9a, $n = 32,561$, $m = 123$, density = 11.3%, $\lambda = 0.1$; Right: mnist, $n = 60,000$, $m = 780$, density = 19.2%, $\lambda = 1$.

performance of PURE-CD with Vu-Condat-CD from [FB19] which has desirable properties with sparse data and SPDHG from [CERS18] which has desirable properties with dense data.

We select uniform sampling, $p_i = 1/n$, so (5.9) simplifies to

$$\tau_i < \frac{1}{\sum_{j=1}^{m} \theta_j \sigma_j A_{j,i}^2}. \tag{5.14}$$

We provide a step size policy inspired by the step size rules chosen in [CERS18] and [FB19]. We use the following step sizes, for $\gamma < 1$

$$\sigma_j = \frac{1}{\theta_j \max_{i'} \|A_{i'}\|}, \quad \tau_i = \frac{\gamma \max_{i'} \|A_{i'}\|}{\|A_i\|^2}.$$

We note that in contrast to [CERS18], step sizes are both diagonal. In our case, it is important to utilize diagonal step sizes for both primal and dual variables since we perform coordinate-wise updates for both primal and dual variables and the step sizes need to be set appropriately to obtain good practical performance. For SPDHG and Vu-Condat-CD, we use step sizes suggested in the papers.

In the edge cases (one nonzero element per row or fully dense), it is easy to see that our step size policy reduces to the suggested step sizes of [CERS18] and [FB19].

For experiments, we used the generic coordinate descent solver, implemented in Cython, by [Fer19], which includes an implementation of Vu-Condat-CD with duplication and we implemented SPDHG and PURE-CD. We solve Lasso and ridge regression, where we let $g(x) = \lambda\|x\|_1$, $h(Ax) = \frac{1}{2}\|Ax - b\|^2$, $f = 0$, and $g(x) = \frac{\lambda}{2}\|x\|^2$, $h(Ax) = \frac{1}{2}\|Ax - b\|^2$, $f = 0$, respectively, in our template (5.1). Then, we apply all the methods to the dual problems of these, to access data by rows.

We use datasets from LIBSVM with different sparsity levels [CL11b]. The properties of each data matrix are given in the caption of the corresponding figures. For preprocessing, we removed all-zero rows and all-zero columns of A and we performed row normalization. The results are compiled in Figure 5.2.

We observe the behavior predicted by theory. With sparse data such as rcv1, where density level is 0.16%, SPDHG makes very little progress in the given time window. The reason is that the per iteration cost of SPDHG in this case is updating 47,236 dual variables, whereas for PURE-CD and Vu-Condat-CD, the cost is updating 75 dual variables. We note that PURE-CD is faster than Vu-Condat-CD due to better step sizes. On the other hand, with moderate sparsity, SPDHG and Vu-Condat-CD is comparable, whereas PURE-CD exhibits the best performance. For denser data, SPDHG and PURE-CD exhibit similar behavior where Vu-Condat-CD is slower than both due to smaller step sizes.

5.6.2 Comparison with specialized methods

In this section, we compare the practical performance of PURE-CD with state-of-the-art algorithms that are designed for strongly convex-strongly concave problems. We defer some of the plots and more details about experiments to Section 5.7. We focus on the problem $\min_x \frac{1}{n}\sum_{i=1}^n h_i(A_i x) + \frac{\lambda}{2}\|x\|^2$, where $h_i(x) = (x - b_i)^2$. Each h_i is smooth with Lipschitz constants $L_i = 2$ and the second component is strongly convex, which results in strong convexity in both primal and dual problems.

In this case, the algorithms SDCA [SSZ13], ProxSVRG [XZ14], Accelerated SVRG [ZSC18], SPDC [ZX17] are all designed to use the strong convexity to obtain linear convergence. These algorithms use the strong convexity constant λ for setting the algorithmic parameters. Moreover, as all the abovementioned algorithms have special implementations to exploit sparsity in this specific case, we make the comparison with respect to number of passes of the data, rather than time. The results are compiled for two datasets in Figure 5.3 and more datasets are included in Section 5.7. We use theoretical step sizes for all the algorithms, given in the respective papers.

• PURE-CD-λ: This variant uses the non-agnostic step sizes, using λ, which still satisfy the theoretical requirement (5.14).

$$\sigma_j = \frac{n}{\theta_j \sqrt{n\lambda}\max_{i'}\|A_{i'}\|}, \quad \tau_i = \frac{\gamma\sqrt{n\lambda}\max_{i'}\|A_{i'}\|}{n\|A_i\|^2}.$$

- PURE-CD: This variant is with the standard agnostic step sizes.

$$\sigma_j = \frac{n}{\theta_j \max_{i'} \|A_{i'}\|}, \quad \tau_i = \frac{\gamma \max_{i'} \|A_{i'}\|}{n\|A_i\|^2}.$$

We observe that PURE-CD has a consistent linear convergence behavior as predicted by theory. In most of the datasets (see Section 5.7), it has the fastest convergence behavior. However, in some datasets, as λ gets smaller, we observed that the linear rate of PURE-CD slowed down, which motivated us to try PURE-CD-λ, which incorporates the knowledge of λ as the other methods. It seems to show favorable behavior when PURE-CD slows down.

The takeaway message is that PURE-CD, which is designed for a general problem, adapts to strong convexity well with agnostic step sizes in most cases. However, in some cases, it does not perform as good as the algorithms which are designed to exploit strong convexity. In those cases however, one can choose separating step sizes of PURE-CD accordingly, and use PURE-CD-λ to get better performance.

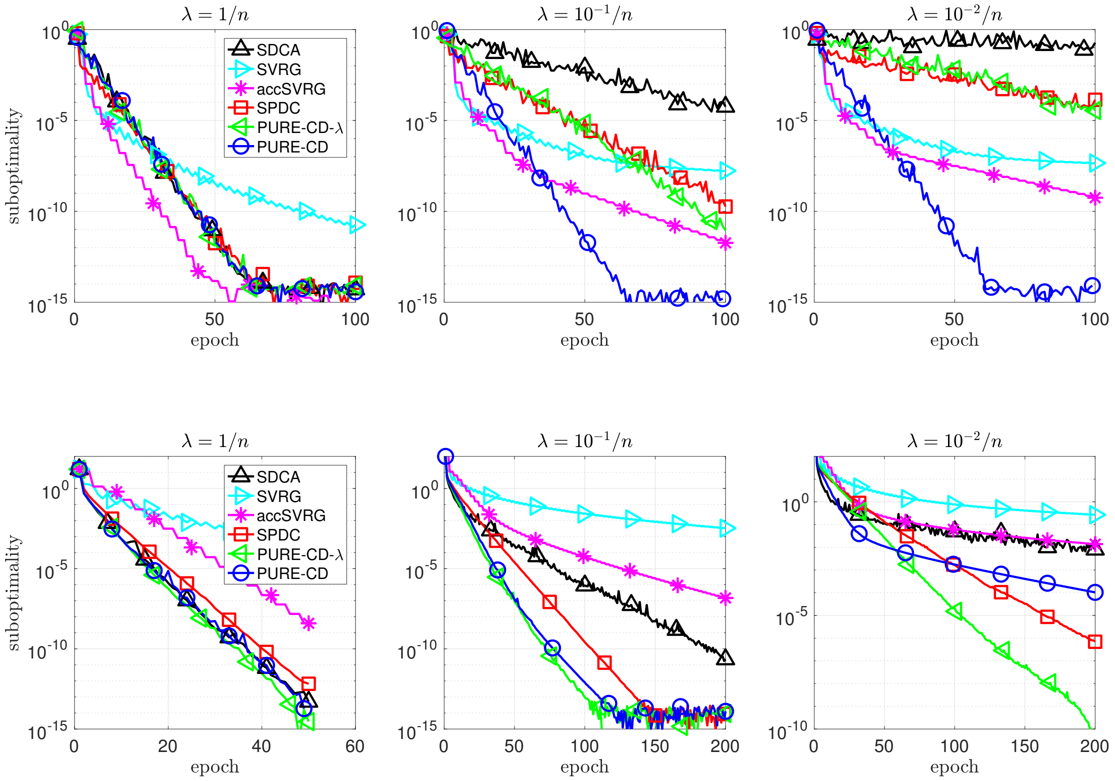

Figure 5.3 – top: a9a, $n = 32,561$, $m = 123$, bottom: sector, $n = 6,412$, $m = 55,197$.

5.7 More experimental results

In this section, we compare the practical performance of PURE-CD with state-of-the-art algorithms that are designed to exploit problem structures. In particular, we focus on the problem

$$\min_x \frac{1}{n}\sum_{i=1}^n h_i(A_i x) + \frac{\lambda}{2}\|x\|^2, \tag{5.15}$$

where $h_i(x) = (x - b_i)^2$. Each h_i is smooth with Lipschitz constants $L_i = 2$ and the second component is strongly convex. This is equivalent to strong convexity in both primal and dual problems.

In this case, the algorithms SDCA, SVRG, Accelerated SVRG/Katyusha, SPDC are all designed to use the strong convexity to obtain linear convergence. These algorithms use the strong convexity constant λ for setting the algorithmic parameters (with the exception of SVRG which theoretically needs it to set number of inner loop iterations). Moreover, as all the abovementioned algorithms have special implementations to exploit sparsity, for fairness, as all algorithms have different structures, we did not try to implement them in the most efficient manner, therefore we make the comparison with respect to number of passes of the data, rather than time.

We use PURE-CD with the agnostic step size and also with a non-agnostic step size that uses λ. Both step size rules are supported by theory. Moreover, similar to Section 5.6, we apply PURE-CD to the dual problem of (5.15) to access the data row-wise as other methods.

The details of parameters for each algorithm:

• SVRG: We use the theoretical step size given in [XZ14, Theorem 3.1]

• Accelerated SVRG/Katyusha: We use the theoretically suggested step size parameter and acceleration parameter [ZSC18, Theorem 1, Table 2]

• SDCA: We use directly the specialization of SDCA for ridge regression, as decribed in [SSZ13, Section 6.2]

• SPDC: We use the step sizes from [ZX17, Theorem 1]

• PURE-CD-λ: This variant uses the non-agnostic step sizes, using λ. We note that the step sizes satisfy the theoretical requirement (5.14).

$$\sigma_j = \frac{n}{\theta_j \sqrt{n\lambda}\max_{i'}\|A_{i'}\|}, \quad \tau_i = \frac{\gamma\sqrt{n\lambda}\max_{i'}\|A_{i'}\|}{n\|A_i\|^2}.$$

• PURE-CD: This variant is with the standard agnostic step sizes, as in Section 5.6. We note that the step sizes are scaled by n since the problem is scaled by $1/n$, compared to Section 5.6.

$$\sigma_j = \frac{n}{\theta_j \max_{i'}\|A_{i'}\|}, \quad \tau_i = \frac{\gamma\max_{i'}\|A_{i'}\|}{n\|A_i\|^2}.$$

We use datasets from LIBSVM, and try three different regularization parameters $\frac{1}{n}$, $\frac{10^{-1}}{n}$, and $\frac{10^{-2}}{n}$. We performed preprocessing by removing the all-zero rows and columns from the data matrix, then we normalized row norms of A. We chose the parameters as described above and did not perform any tuning for any algorithm.

We observe that PURE-CD has a consistent linear convergence behavior as predicted by theory. In most of the datasets, it has the fastest convergence behavior. However, in some datasets, as λ gets smaller, we observed that the linear rate of PURE-CD slowed down, which motivated us to try PURE-CD-λ, which incorporates the knowledge of λ as the other methods. It seems to show favorable behavior when PURE-CD slows down.

The takeaway message is that PURE-CD adapts very well with agnostic step sizes in most cases. However, in some cases, it does not perform as good as the algorithms which are designed to exploit structure. In those cases however, one can choose separating step sizes of PURE-CD accordingly, and use PURE-CD-λ to get better performance.

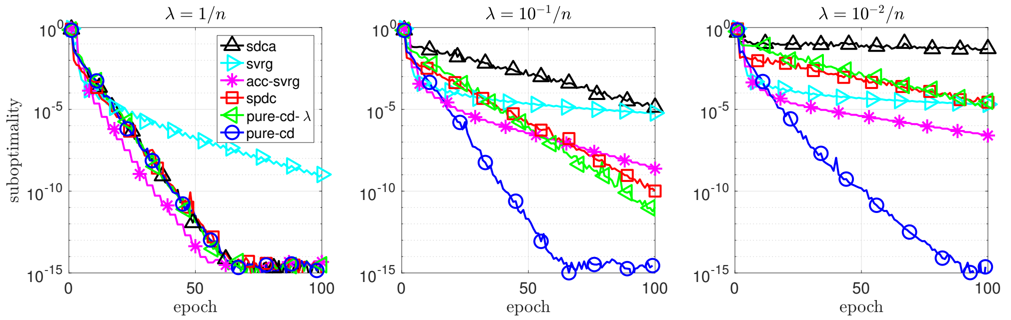

Figure 5.4 – w8a, $n = 49,749$, $m = 300$.

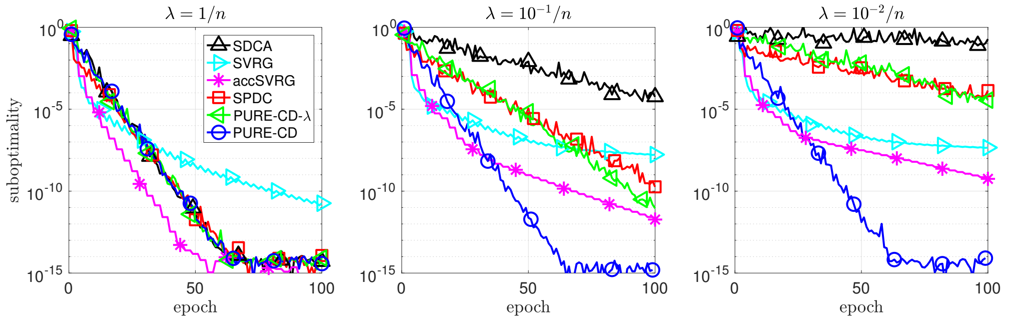

Figure 5.5 – a9a, $n = 32,561$, $m = 123$.

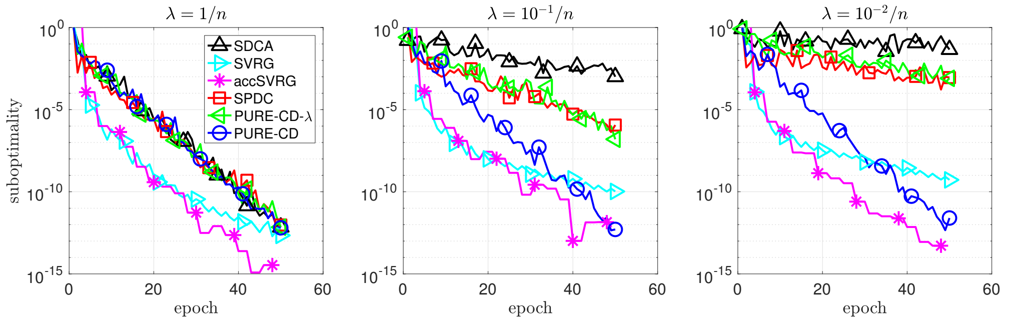

Figure 5.6 – covtype, $n = 581,012$, $m = 54$.

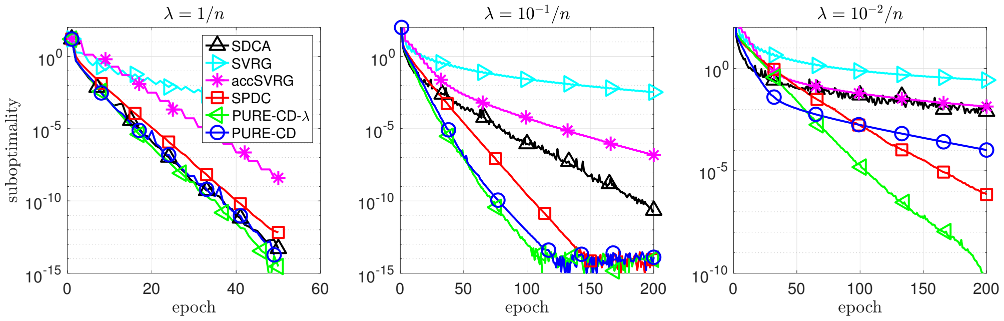

Figure 5.7 – sector, $n = 6,412$, $m = 55,197$.

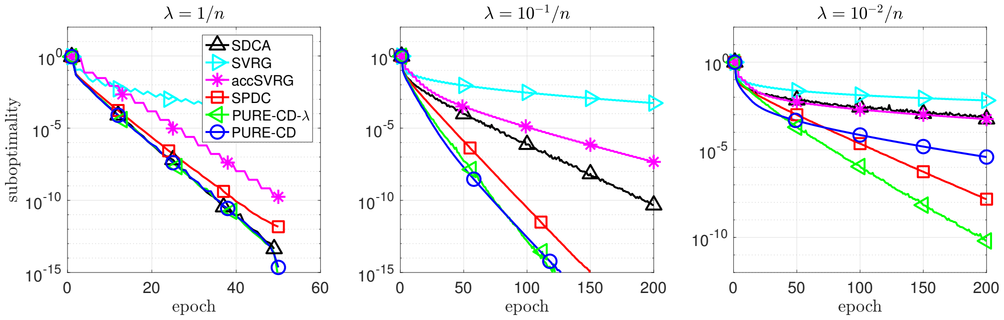

Figure 5.8 – rcv1.binary, $n = 20,242$, $m = 47,236$.

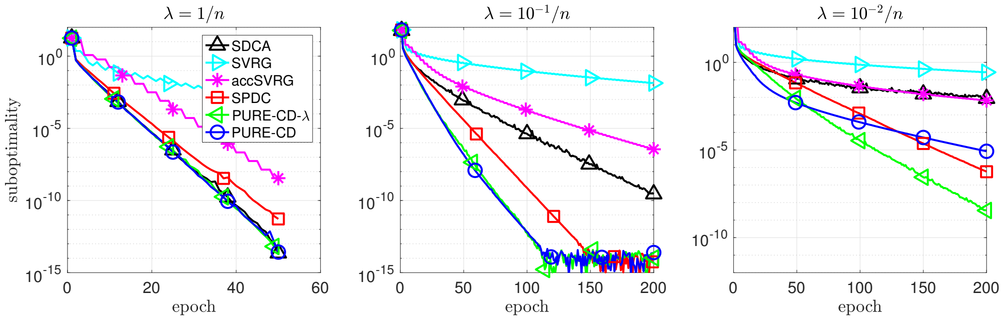

Figure 5.9 – news20, $n = 15,935$, $m = 62,061$.

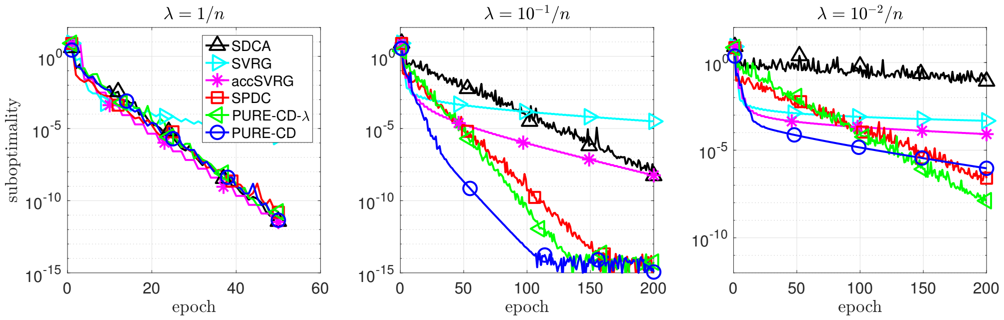

Figure 5.10 – mnist, $n = 60,000$, $m = 780$.

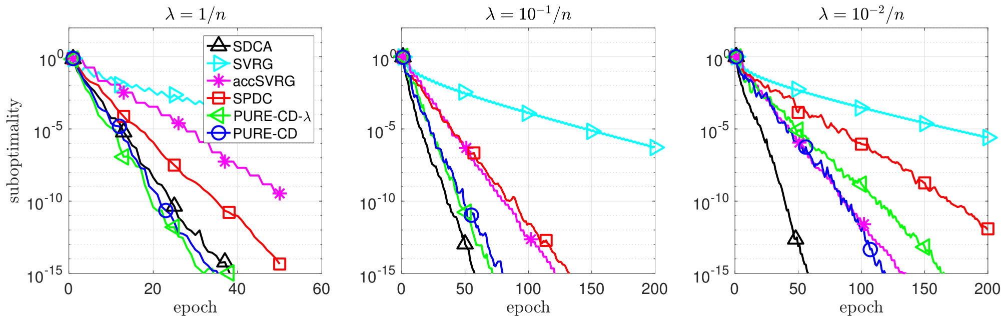

Figure 5.11 – leukemia, $n = 38$, $m = 7,129$.

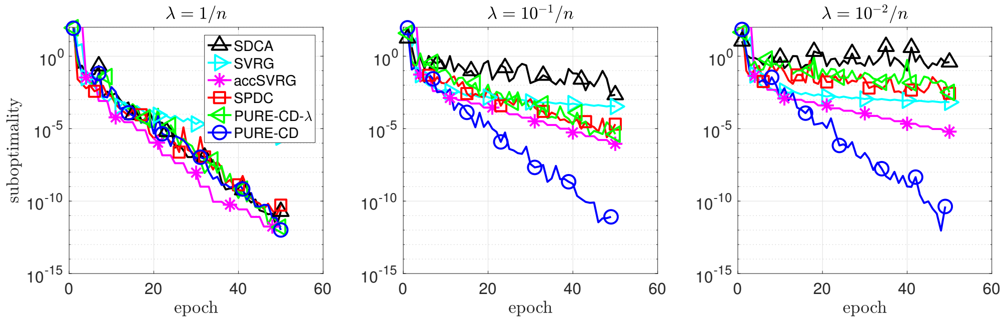

Figure 5.12 – YearPredictionMSD, $n = 463,715$, $m = 90$.

5.7.1 Further details about experiments

The experiments are done on a computer with Intel Core i7 CPUs at 2.9 GHz.

5.8 Proofs

5.8.1 Proofs for one iteration result

We start with technical lemmas. Our first result computes the conditional expectation of y_{k+1}.

Lemma 5.8. *Let y_{k+1} as defined in Algorithm 5.1, and recall the definitions of π and P from Section 5.2.1. Then it holds that for any $\mathcal{F}_k$-measurable Y and $\forall \gamma = \{\gamma_1, \dots, \gamma_m\}$, with $\gamma_i > 0$,*

$$\mathbb{E}_k\left[\|y_{k+1} - Y\|_\gamma^2\right] = \|\bar{y}_{k+1} - Y\|_{\gamma\pi}^2 - \|y_k - Y\|_{\gamma\pi}^2 + \|y_k - Y\|_\gamma^2 + 2\langle \bar{y}_{k+1} - Y, \gamma\sigma\theta AP(\bar{x}_{k+1} - x_k)\rangle$$
$$+ \sum_{i=1}^n \sum_{j=1}^m p_i \gamma_j \sigma_j^2 \theta_j^2 A_{j,i}^2 (\bar{x}_{k+1}^{(i)} - x_k^{(i)})^2.$$

Proof. First, we note that for $\mathcal{F}_k$-measurable Y, it follows that

$$\mathbb{E}_k\left[\|y_{k+1} - Y\|_\gamma^2\right] = \mathbb{E}_k\left[\sum_{j=1}^m \gamma_j \left(y_{k+1}^{(j)} - Y^{(j)}\right)^2\right]$$

$$
\begin{aligned}
&= \mathbb{E}_k \left[\sum_{j\in J(i_{k+1})} \gamma_j (\bar{y}_{k+1}^{(j)} + \sigma_j \theta_j (A(x_{k+1}-x_k))^{(j)} - Y^{(j)})^2 + \sum_{j\notin J(i_{k+1})} \gamma_j (y_k^{(j)} - Y^{(j)})^2 \right] \\
&= \sum_{i=1}^n p_i \left[\sum_{j\in J(i)} \gamma_j \left(\bar{y}_{k+1}^{(j)} + \sigma_j\theta_j A_{j,i}(\bar{x}_{k+1}^{(i)} - x_k^{(i)}) - Y^{(j)}\right)^2 + \sum_{j\notin J(i)} \gamma_j \left(y_k^{(j)} - Y^{(j)}\right)^2 \right] \\
&= \sum_{i=1}^n \sum_{j\in J(i)} p_i\gamma_j (\bar{y}_{k+1}^{(j)} - Y^{(j)})^2 + 2\sum_{i=1}^n \sum_{j\in J(i)} p_i\gamma_j\sigma_j\theta_j A_{j,i}(\bar{x}_{k+1}^{(i)} - x_k^{(i)})(\bar{y}_{k+1}^{(j)} - Y^{(j)}) \\
&+ \sum_{i=1}^n \sum_{j\in J(i)} p_i\gamma_j \left(\sigma_j\theta_j A_{j,i}(\bar{x}_{k+1}^{(i)} - x_k^{(i)})\right)^2 + \sum_{i=1}^n \sum_{j\notin J(i)} p_i\gamma_j (y_k^{(j)} - Y^{(j)})^2 \qquad (5.16)
\end{aligned}
$$

where for the third equality, we used the fact that x_{k+1} is different from x_k only on the coordinate i_{k+1} and $\bar{x}_{k+1}^{(i_{k+1})} = x_{k+1}^{(i_{k+1})}$, which gives

$$(A(x_{k+1}-x_k))^{(j)} = (A((x_{k+1}^{(i_{k+1})} - x_k^{(i_{k+1})})e_{i_{k+1}}))^{(j)} = A_{j,i_{k+1}}(\bar{x}_{k+1}^{(i_{k+1})} - x_k^{(i_{k+1})}).$$

We focus on the last term on the right hand side of (5.16)

$$
\begin{aligned}
\sum_{i=1}^n \sum_{j\notin J(i)} p_i\gamma_j (y_k^{(j)} - Y^{(j)})^2 &= \sum_{i=1}^n \sum_{j=1}^n p_i\gamma_j (y_k^{(j)} - Y^{(j)})^2 - \sum_{i=1}^n \sum_{j\in J(i)} p_i\gamma_j (y_k^{(j)} - Y^{(j)})^2 \\
&= \|y_k - Y\|_\gamma^2 - \sum_{j=1}^m \sum_{i\in I(j)} p_i\gamma_j (y_k^{(j)} - Y^{(j)})^2 = \|y_k - Y\|_\gamma^2 - \sum_{j=1}^m \pi_j\gamma_j (y_k^{(j)} - Y^{(j)})^2 \\
&= \|y_k - Y\|_\gamma^2 - \|y_k - Y\|_{\gamma\pi}^2 \qquad (5.17)
\end{aligned}
$$

where we use the fact that $\sum_{i=1}^n \sum_{j\in J(i)} \gamma'_{j,i} = \sum_{j=1}^m \sum_{i\in I(j)} \gamma'_{j,i}$, for any γ', due to the definition of $J(i)$, $I(j)$ and $\pi_j = \sum_{i\in I(j)} p_i$.

We estimate the first term of (5.16), similar to (5.17)

$$\sum_{i=1}^n \sum_{j\in J(i)} p_i\gamma_j (\bar{y}_{k+1}^{(j)} - Y^{j})^2 = \sum_{j=1}^m \sum_{i\in I(j)} p_i\gamma_j (\bar{y}_{k+1}^{(j)} - Y^{(j)})^2 = \|\bar{y}_{k+1} - Y\|_{\gamma\pi}^2. \qquad (5.18)$$

We lastly estimate the second and third terms of (5.16). We use the fact that $A_{j,i} = 0$, if $j \notin J(i)$ to obtain

$$
\begin{aligned}
&2\sum_{i=1}^n \sum_{j\in J(i)} p_i\gamma_j\sigma_j\theta_j A_{j,i}(\bar{x}_{k+1}^{(i)} - x_k^{(i)})(\bar{y}_{k+1}^{(j)} - Y^{(j)}) + \sum_{i=1}^n \sum_{j\in J(i)} p_i\gamma_j \left(\sigma_j\theta_j A_{j,i}(\bar{x}_{k+1}^{(i)} - x_k^{(i)})\right)^2 \\
&= 2\sum_{i=1}^n \sum_{j=1}^m p_i\gamma_j\sigma_j\theta_j A_{j,i}(\bar{x}_{k+1}^{(i)} - x_k^{(i)})(\bar{y}_{k+1}^{(j)} - Y^{(j)}) + \sum_{i=1}^n \sum_{j=1}^m p_i\gamma_j \left(\sigma_j\theta_j A_{j,i}(\bar{x}_{k+1}^{(i)} - x_k^{(i)})\right)^2 \\
&= 2\langle \bar{y}_{k+1} - Y, \gamma\sigma\theta AP(\bar{x}_{k+1} - x_k)\rangle + \sum_{i=1}^n \sum_{j=1}^m p_i\gamma_j \left(\sigma_j\theta_j A_{j,i}(\bar{x}_{k+1}^{(i)} - x_k^{(i)})\right)^2. \qquad (5.19)
\end{aligned}
$$

We use (5.17), (5.18), and (5.19) in (5.16) to obtain the final result. ■

We continue with the following lemma which handles necessary manipulations for the terms involving the primal variable, to handle arbitrary probabilities.

Lemma 5.9. *We recall that $P = \operatorname{diag}(p_1, \dots, p_n)$, $\underline{p} = \min_i p_i$, and define*

$$x' = x_k + P^{-1}\underline{p}(x - x_k) = P^{-1}\underline{p}x + (1 - P^{-1}\underline{p})x_k,$$

and

$$g_P(x) = \sum_{i=1}^n p_i g_i(x^{(i)}).$$

Then, for a function $g(x) = \sum_{i=1}^n g_i(x^{(i)})$, the following conclusions hold:

$$\begin{aligned} g_P(x') &\le \underline{p}g(x) - \underline{p}g(x_k) + g_P(x_k), \\ \|x' - x_{k+1}\|^2_{\tau^{-1}} - \|x' - x_k\|^2_{\tau^{-1}} &= \underline{p}\|x - x_{k+1}\|^2_{\tau^{-1}P^{-1}} + \|x_{k+1} - x_k\|^2_{\tau^{-1}} - \underline{p}\|x_{k+1} - x_k\|^2_{\tau^{-1}P^{-1}} \\ &\quad - \underline{p}\|x - x_k\|^2_{\tau^{-1}P^{-1}}. \end{aligned}$$

Proof. We have that $x'^{(i)} = p_i^{-1}\underline{p}x^{(i)} + (1 - p_i^{-1}\underline{p})x_k^{(i)}$. It follows by convexity of g_i that

$$g_P(x') = \sum_{i=1}^n p_i g_i(x'^{(i)}) \le \sum_{i=1}^n \underline{p}g_i(x^{(i)}) + (p_i - \underline{p})g_i(x_k^{(i)}) = \underline{p}g(x) - \underline{p}g(x_k) + g_P(x_k).$$

Moreover, since for any $0 \le c \le 1$ and any u, v, it is true that $\|cu + (1-c)v\|^2 = c\|u\|^2 + (1-c)\|v\|^2 - c(1-c)\|u - v\|^2$, we obtain

$$\begin{aligned} \|x' - x_{k+1}\|^2_{\tau^{-1}} = \underline{p}\|x - x_{k+1}\|^2_{\tau^{-1}P^{-1}} &+ \|x_{k+1} - x_k\|^2_{\tau^{-1}} - \underline{p}\|x_{k+1} - x_k\|^2_{\tau^{-1}P^{-1}} \\ &- \underline{p}\|x - x_k\|^2_{\tau^{-1}P^{-1}} + \underline{p}^2\|x - x_k\|^2_{\tau^{-1}P^{-2}}. \end{aligned}$$

Lastly, plugging in $x' = x_k + P^{-1}\underline{p}(x - x_k)$ to $\|x' - x_k\|^2$ gives

$$-\|x' - x_k\|^2_{\tau^{-1}} = -\underline{p}^2\|x - x_k\|^2_{\tau^{-1}P^{-2}}.$$ ■

We are now ready to prove Lemma 5.1 which describes the one iteration behavior of the algorithm.

Proof of Lemma 5.1. By the definition of proximal operator and convexity, $\forall x' \in \mathcal{X}$, $\forall y \in \mathcal{Y}$,

$$\begin{aligned} p_i g_i(x'^{(i)}) \ge p_i g_i(\bar{x}^{(i)}_{k+1}) &- p_i \langle \nabla_i f(x_k) + (A^\top \bar{y}_{k+1})^{(i)}, x'^{(i)} - \bar{x}^{(i)}_{k+1} \rangle \\ &+ \frac{1}{2}\left(\|x_k^{(i)} - \bar{x}^{(i)}_{k+1}\|^2_{\tau_i^{-1}p_i} + \|x'^{(i)} - \bar{x}^{(i)}_{k+1}\|^2_{\tau_i^{-1}p_i} - \|x'^{(i)} - x_k^{(i)}\|^2_{\tau_i^{-1}p_i}\right), \\ \underline{p}h^*(y) \ge \underline{p}h^*(\bar{y}_{k+1}) + \underline{p}\langle Ax_k, y - \bar{y}_{k+1}\rangle &+ \frac{\underline{p}}{2}\left(\|y_k - \bar{y}_{k+1}\|^2_{\sigma^{-1}} + \|y - \bar{y}_{k+1}\|^2_{\sigma^{-1}} - \|y - y_k\|^2_{\sigma^{-1}}\right). \end{aligned}$$

We sum the first inequality for $i = 1$ to n, then add it to the second inequality and use the definition $g_P(x) = \sum_{i=1}^{n} p_i g_i(x^{(i)})$ to derive

$$\begin{aligned}
g_P(x') + \underline{p}h^*(y) \geq g_P(\bar{x}_{k+1}) + \underline{p}h^*(\bar{y}_{k+1}) - \langle \nabla f(x_k), P(x' - \bar{x}_{k+1})\rangle - \langle A^\top \bar{y}_{k+1}, P(x' - \bar{x}_{k+1})\rangle \\
+ \underline{p}\langle Ax_k, y - \bar{y}_{k+1}\rangle + \frac{1}{2}\Big(\|x_k - \bar{x}_{k+1}\|^2_{\tau^{-1}P} + \|x' - \bar{x}_{k+1}\|^2_{\tau^{-1}P} - \|x' - x_k\|^2_{\tau^{-1}P}\Big) \\
+ \frac{\underline{p}}{2}\left(\|y_k - \bar{y}_{k+1}\|^2_{\sigma^{-1}} + \|y - \bar{y}_{k+1}\|^2_{\sigma^{-1}} - \|y - y_k\|^2_{\sigma^{-1}}\right). \qquad (5.20)
\end{aligned}$$

First, we note that for $\mathcal{F}_k$-measurable X and any $\gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_n)$, such that $\gamma_i > 0$, the following hold

$$\begin{aligned}
&\mathbb{E}_k[g(x_{k+1})] = g_P(\bar{x}_{k+1}) - g_P(x_k) + g(x_k), && (5.21) \\
&\mathbb{E}_k[x_{k+1}] = P\bar{x}_{k+1} - Px_k + x_k, \\
&\mathbb{E}_k\left[\|x_{k+1} - X\|^2_\gamma\right] = \|\bar{x}_{k+1} - X\|^2_{\gamma P} - \|x_k - X\|^2_{\gamma P} + \|x_k - X\|^2_\gamma. && (5.22)
\end{aligned}$$

We use (5.22) with $\gamma = \tau^{-1}$ and $X = x'$ to obtain

$$\begin{aligned}
&\frac{1}{2}\Big(\|x_k - \bar{x}_{k+1}\|^2_{\tau^{-1}P} + \|x' - \bar{x}_{k+1}\|^2_{\tau^{-1}P} - \|x' - x_k\|^2_{\tau^{-1}P}\Big) \\
&\qquad = \frac{1}{2}\Big(\|x_k - \bar{x}_{k+1}\|^2_{\tau^{-1}P} + \mathbb{E}_k\left[\|x' - x_{k+1}\|^2_{\tau^{-1}}\right] - \|x' - x_k\|^2_{\tau^{-1}}\Big). \qquad (5.23)
\end{aligned}$$

We use $\gamma = \pi^{-1}\sigma^{-1}$ and $Y = y$ in Lemma 5.8, then

$$\begin{aligned}
\|\bar{y}_{k+1} - y\|^2_{\sigma^{-1}} = \mathbb{E}_k\left[\|y_{k+1} - y\|^2_{\sigma^{-1}\pi^{-1}}\right] + \|y_k - y\|^2_{\sigma^{-1}} - \|y_k - y\|^2_{\sigma^{-1}\pi^{-1}} \\
- 2\langle \bar{y}_{k+1} - y, \pi^{-1}\theta AP(\bar{x}_{k+1} - x_k)\rangle - \sum_{i=1}^{n}\sum_{j=1}^{m} p_i \pi_j^{-1}\sigma_j \theta_j^2 A_{j,i}^2 (\bar{x}^i_{k+1} - x^i_k)^2. \qquad (5.24)
\end{aligned}$$

We let $x'^i = p_i^{-1}\underline{p}x^i + (1 - p_i^{-1}\underline{p})x^i_k$, and use Lemma 5.9 to get

$$g_P(x') \leq \underline{p}g(x) - \underline{p}g(x_k) + g_P(x_k). \qquad (5.25)$$

Combined with (5.21), the last inequality gives

$$g_P(\bar{x}_{k+1}) - g_P(x') \geq \mathbb{E}_k\left[g(x_{k+1})\right] - g(x_k) + \underline{p}g(x_k) - \underline{p}g(x). \qquad (5.26)$$

We further use $x' = P^{-1}\underline{p}x + (1 - P^{-1}\underline{p})x_k = x_k + P^{-1}\underline{p}(x - x_k)$ in (5.20) to obtain

$$-\langle \nabla f(x_k), P(x' - \bar{x}_{k+1}) = -\underline{p}\langle \nabla f(x_k), x - x_k\rangle - \langle \nabla f(x_k), P(x_k - \bar{x}_{k+1})\rangle, \qquad (5.27)$$

and

$$-\langle A^\top \bar{y}_{k+1}, P(x' - \bar{x}_{k+1})\rangle = -\underline{p}\langle A^\top \bar{y}_{k+1}, x - x_k\rangle - \langle A^\top \bar{y}_{k+1}, P(x_k - \bar{x}_{k+1})\rangle. \qquad (5.28)$$

Moreover, by using Lemma 5.9 in (5.23), we get

$$\begin{aligned}\frac{1}{2}\mathbb{E}_k\left[\|x'-x_{k+1}\|^2_{\tau^{-1}}\right]-\frac{1}{2}\|x'-x_k\|^2_{\tau^{-1}} &= \frac{\underline{p}}{2}\mathbb{E}_k\left[\|x-x_{k+1}\|^2_{\tau^{-1}P^{-1}}\right]+\frac{1}{2}\mathbb{E}_k\left[\|x_{k+1}-x_k\|^2_{\tau^{-1}}\right]\\ &\quad-\frac{\underline{p}}{2}\mathbb{E}_k\left[\|x_{k+1}-x_k\|^2_{\tau^{-1}P^{-1}}\right]-\frac{\underline{p}}{2}\|x-x_k\|^2_{\tau^{-1}P^{-1}}. \end{aligned}\tag{5.29}$$

We also note that $\mathbb{E}_k\left[\|x_{k+1}-x_k\|^2_{\tau^{-1}P^{-1}}\right]=\|\bar{x}_{k+1}-x_k\|^2_{\tau^{-1}}$.

In (5.20), we collect eqs. (5.23), (5.24) and (5.26)–(5.29) to obtain

$$\begin{aligned}0 &\geq \mathbb{E}_k[g(x_{k+1})]-g(x_k)+\underline{p}g(x_k)-\underline{p}g(x)+\underline{p}h^*(\bar{y}_{k+1})-\underline{p}h^*(y)+\underline{p}\langle Ax_k,y-\bar{y}_{k+1}\rangle\\ &\quad-\underline{p}\langle\nabla f(x_k),x-x_k\rangle-\langle\nabla f(x_k),P(x_k-\bar{x}_{k+1})\rangle-\underline{p}\langle A^\top\bar{y}_{k+1},x-x_k\rangle-\langle A^\top\bar{y}_{k+1},P(x_k-\bar{x}_{k+1})\rangle\\ &\quad+\frac{\underline{p}}{2}\mathbb{E}_k\left[\|x_{k+1}-x\|^2_{\tau^{-1}P^{-1}}\right]-\frac{\underline{p}}{2}\|x_k-x\|^2_{\tau^{-1}P^{-1}}+\|\bar{x}_{k+1}-x_k\|^2_{\tau^{-1}P}-\frac{\underline{p}}{2}\|\bar{x}_{k+1}-x_k\|^2_{\tau^{-1}}\\ &\quad+\frac{\underline{p}}{2}\|\bar{y}_{k+1}-y_k\|^2_{\sigma^{-1}}+\frac{\underline{p}}{2}\mathbb{E}_k\left[\|y_{k+1}-y\|^2_{\sigma^{-1}\pi^{-1}}\right]-\frac{\underline{p}}{2}\|y_k-y\|^2_{\sigma^{-1}\pi^{-1}}\\ &\quad-\langle\bar{y}_{k+1}-y,\underline{p}\pi^{-1}\theta AP(\bar{x}_{k+1}-x_k)\rangle-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\underline{p}p_i\pi_j^{-1}\sigma_j\theta_j^2A_{j,i}^2(\bar{x}^i_{k+1}-x^i_k)^2. \end{aligned}\tag{5.30}$$

We use coordinatewise smoothness of f to obtain

$$\begin{aligned}-\langle\nabla f(x_k),P(x_k-\bar{x}_{k+1})\rangle=-\langle\nabla f(x_k),\mathbb{E}_k[x_k-x_{k+1}]\rangle &\geq \mathbb{E}_k\left[f(x_{k+1})-f(x_k)-\frac{1}{2}\|x_{k+1}-x_k\|^2_\beta\right]\\ &\geq \mathbb{E}_k[f(x_{k+1})]-f(x_k)-\frac{1}{2}\|\bar{x}_{k+1}-x_k\|^2_{\beta P}. \end{aligned}\tag{5.31}$$

Next, by using the definition of $\tilde{V}$ and $C(\tau)$, we note

$$\begin{aligned}\tilde{V}(\bar{z}_{k+1}-z_k) &= \frac{\underline{p}}{2}\|\bar{y}_{k+1}-y_k\|^2_{\sigma^{-1}}+\|\bar{x}_{k+1}-x_k\|^2_{\tau^{-1}P}-\frac{\underline{p}}{2}\|\bar{x}_{k+1}-x_k\|^2_{\tau^{-1}}\\ &\quad-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\underline{p}p_i\pi_j^{-1}\sigma_j\theta_j^2A_{j,i}^2(\bar{x}^{(i)}_{k+1}-x^{(i)}_k)^2-\frac{1}{2}\|\bar{x}_{k+1}-x_k\|^2_{\beta P}. \end{aligned}\tag{5.32}$$

Using definition of V, D_p, D_d (see Section 5.2.1) in (5.30) along with $-\underline{p}\langle\nabla f(x_k),x-x_k\rangle\geq-\underline{p}f(x)+\underline{p}f(x_k)$ and the last inequality yields

$$\begin{aligned}0 &\geq \mathbb{E}_k\left[D_p(x_{k+1};z)\right]-(1-\underline{p})D_p(x_k;z)+\underline{p}D_d(\bar{y}_{k+1};z)+\tilde{V}(\bar{z}_{k+1}-z_k)+\mathbb{E}_k\left[V(z_{k+1}-z)\right]\\ &\quad-V(z_k-z)+\underline{p}\langle Ax_k,y-\bar{y}_{k+1}\rangle-\underline{p}\langle A^\top\bar{y}_{k+1},x-x_k\rangle-\langle A^\top\bar{y}_{k+1},P(x_k-\bar{x}_{k+1})\rangle\\ &\quad-\langle\bar{y}_{k+1}-y,\underline{p}\pi^{-1}\theta AP(\bar{x}_{k+1}-x_k)\rangle-\langle A^\top y,x_{k+1}-x\rangle+\underline{p}\langle Ax,\bar{y}_{k+1}-y\rangle\\ &\quad+\langle A^\top y,x_k-x\rangle-\underline{p}\langle A^\top y,x_k-x\rangle. \end{aligned}\tag{5.33}$$

We work on the bilinear terms to get

$$\langle A^\top y,x_k-x\rangle-\langle A^\top y,x_{k+1}-x\rangle-\langle A^\top\bar{y}_{k+1},P(x_k-\bar{x}_{k+1})\rangle-\langle y-\bar{y}_{k+1},\underline{p}\pi^{-1}\theta AP(x_k-\bar{x}_{k+1})\rangle$$

$$
\begin{aligned}
&= \langle A^\top y, x_k - x_{k+1}\rangle - \langle A^\top \bar{y}_{k+1}, \mathbb{E}_k[x_k - x_{k+1}]\rangle - \langle y - \bar{y}_{k+1}, \underline{p}\pi^{-1}\theta A \mathbb{E}_k[x_k - x_{k+1}]\rangle \\
&= \mathbb{E}_k\left[\langle A^\top(y - \bar{y}_{k+1}), x_k - x_{k+1}\rangle - \langle y - \bar{y}_{k+1}, \underline{p}\pi^{-1}\theta A(x_k - x_{k+1})\rangle\right] = 0, \qquad (5.34)
\end{aligned}
$$

where the last equality is due to the requirement in (5.8) as

$$
\underline{p}\pi^{-1}\theta = I \iff \theta_j = \frac{\pi_j}{\underline{p}}, \quad \forall j \in \{1, \dots, m\}.
$$

We continue to estimate the remaining bilinear terms as

$$
\begin{aligned}
&\underline{p}\left[\langle Ax_k, y - \bar{y}_{k+1}\rangle - \langle A^\top \bar{y}_{k+1}, x - x_k\rangle + \langle Ax, \bar{y}_{k+1} - y\rangle - \langle A^\top y, x_k - x\rangle\right] = \\
&\qquad \underline{p}\left[\langle A^\top(y - \bar{y}_{k+1}), x - x_k\rangle + \langle y - \bar{y}_{k+1}, A(x_k - x)\rangle\right] = 0. \qquad (5.35)
\end{aligned}
$$

We use the last estimation, eqs. (5.34) and (5.35) in (5.33) to finish the proof. ■

5.8.2 Proof for almost sure convergence

We include proof of Theorem 5.2.

Proof of Theorem 5.2. Equipped with Lemma 5.1, we will follow the standard arguments as in [FB19]. We refer to [FB19, Theorem 1] for the finer details of the arguments.

We first invoke the main result of Lemma 5.1 with $z = z_\star = (x_\star, y_\star)$ where $z_\star \in \mathcal{Z}_\star$:

$$
\begin{aligned}
\mathbb{E}_k\left[S_p(x_{k+1})\right] + \underline{p}S_d(\bar{y}_{k+1}) + \mathbb{E}_k\left[V(z_{k+1} - z_\star)\right] \le (1 - \underline{p})S_p(x_k) + V(z_k - z_\star) \\
- \tilde{V}(\bar{z}_{k+1} - z_k), \qquad (5.36)
\end{aligned}
$$

where we have used the definitions

$$
S_p(x_{k+1}) = D_p(x_{k+1}; z_\star) \ge 0, \quad S_d(\bar{y}_{k+1}) = D_d(\bar{y}_{k+1}; z_\star) \ge 0.
$$

Nonnegativity of these quantities follow from the definition of $z_\star$ as in (5.4) and convexity.

Moreover, we recall the definitions of V, $\tilde{V}$ from Lemma 5.1. Then, we see that $\tilde{V}(z)$ is nonnegative by the choice of step sizes τ, σ and $\theta_j = \frac{\pi_j}{\underline{p}}$. It is also immediate that $V(z)$ is nonnegative.

Then, we can write (5.36) as

$$
\mathbb{E}_k\left[S_p(x_{k+1}) + V(z_{k+1} - z_\star)\right] \le S_p(x_k) + V(z_k - z_\star) - \underline{p}\left(S_p(x_k) + S_d(\bar{y}_{k+1})\right). \qquad (5.37)
$$

We use Robbins-Siegmund lemma on this inequality and nonnegativity of $S_p(x_k), S_d(\bar{y}_k), V(z), \tilde{V}(z)$ to conclude that $V(z_k - z_\star)$ converges almost surely, $\sum_k S_p(x_k) + S_d(\bar{y}_{k+1}) < \infty$, therefore $S_p(x_k)$ and $S_d(\bar{y}_k)$ converges to 0 almost surely. Then, we strengthen these conclusion by arguing as [FB19, Theorem 1], [IBCH13], [CP15, Proposition 2.3], to get the result: there exist Ω_1 with $\mathbb{P}(\Omega_1) = 1$, such that $\forall w \in \Omega_1$ and $\forall z_\star \in \mathcal{Z}_\star$, $V(z_k(\omega) - z_\star)$ converges.

We now take full expectation of (5.36), use the nonnegativity of $S_p(x_k)$, $S_d(\bar{y}_k)$, and sum the inequality to obtain,

$$\sum_{k=0}^{\infty} \mathbb{E}\left[\tilde{V}(\bar{z}_{k+1} - z_k)\right] \leq S_p(x_0) + V(z_0 - z_\star) := \Delta_0 < \infty. \tag{5.38}$$

By Fubini-Tonelli theorem, $\mathbb{E}\left[\sum_{k=0}^{\infty} \tilde{V}(\bar{z}_{k+1} - z_k)\right] < \infty$. Since $\tilde{V}(\bar{z}_{k+1} - z_k)$ is nonnegative we have that $\tilde{V}(\bar{z}_{k+1} - z_k)$ converges almost surely to 0. Then, by the definition of $\tilde{V}(z)$ and the choice of step sizes, it follows that $\bar{z}_{k+1} - z_k$ converges to 0 almost surely. We say that there exist Ω_2 with $\mathbb{P}(\Omega_2) = 1$ such that $\forall \omega \in \Omega_2$, $\bar{z}_{k+1}(\omega) - z_k(\omega) \to 0$.

We define $T \colon \mathcal{Z} \to \mathcal{Z}$ such that

$$\begin{aligned}
\bar{y} &= \operatorname{prox}_{\sigma,h^*}\left(y + \sigma A x\right), \\
\bar{x} &= \operatorname{prox}_{\tau,g}\left(x - \tau(\nabla f(x) + A^\top \bar{y})\right).
\end{aligned}$$

It is easy to see that $\bar{z}_{k+1} = T(z_k)$.

We use the definition of proximal operator in the definition of T and compare with the definition of a saddle point in (5.4) to conclude that the fixed points of T correspond to the set of saddle points $\mathcal{Z}_\star$.

We now fix $\omega \in \Omega_1 \cap \Omega_2$ and then it follows that $z_k(\omega)$ is a bounded sequence, from what we have proved beforehand. As $z_k(\omega)$ is bounded, it converges on at least one subsequence. We denote by $\check{z}$ the cluster point of this subsequence. By the fact that $\bar{z}_{k+1}(\omega) = T(z_k(\omega))$ and $\bar{z}_{k+1}(\omega) - z_k(\omega) \to 0$, we conclude that $T(z_k(\omega)) - z_k(\omega) \to 0$. As T is continuous, by the nonexpansiveness of proximal operator, we get $\check{z}$ is a fixed point of T and that $\check{z} \in \mathcal{Z}_\star$.

Since we know that for any $\omega \in \Omega_1 \cap \Omega_2$ and for any $z_\star \in \mathcal{Z}_\star$, $V(z_k(\omega) - z_\star)$ converges, and we have proved V($z_k(\omega) - \check{z}$) converges to 0 at least on one subsequence with $\check{z} \in \mathcal{Z}_\star$, we conclude that the sequence z_k converges to a point $\check{z} \in \mathcal{Z}_\star$, almost surely. ■

5.8.3 Proof for linear convergence

In this section, we include the proof of Theorem 5.4. First, we need a lemma to characterize the specific choice of projection onto the solution set, given in (5.10).

Lemma 5.10. *Let us denote*

$$\begin{aligned}
z_{\star,k} &= \arg\min_{u \in \mathcal{Z}_\star} D_p(x_k; u) + V(z_k - u). \\
z^e_{\star,k} &= \arg\min_{u \in \mathcal{Z}_\star} V(z_k - u).
\end{aligned}$$

We have that $V^{1/2}(z_{\star,k} - z^e_{\star,k}) \leq cV^{1/2}(z_k - z^e_{\star,k})$, where $c = C_{2,V}\sqrt{\frac{\|A\|}{2}}$ and $C_{2,V}$ is such that for any z, $\|z\| \leq C_{2,V} V(z)^{1/2}$.

Proof. Let us first remark that since $u = (u_x, u_y) \in \mathcal{Z}_*$ is a saddle point of the Lagrangian $L(x, y) = f(x) + g(x) + \langle Ax, y\rangle - h^*(y)$, we have that $D_p(x_k, u) = L(x_k, u_y) - L(u_x, u_y)$ is independent of u_x and affine in u_y. In particular, for any primal solution $x_\star$, there exists a constant, $C(x_k, x_\star) = f(x_k) + g(x_k) - f(x_\star) - g(x_\star)$ such that $D_p(x_k, u) = C(x_k, x_\star) + \langle A(x_k - x_\star), u_y\rangle$, for all $u \in \mathcal{Z}_\star$. We have also used here the fact that for two different primal solutions $x_{\star,1}, x_{\star,2}$, it follows that $L(x_{\star,1}, u_y) = L(x_{\star,2}, u_y)$, where u_y is a dual solution.

By prox inequality, we have for any $z \in \mathcal{Z}$,

$$
\begin{aligned}
C(x_k, x_\star) + \langle A(x_k - x_\star), y_{\star,k}\rangle + V(z - z^\star_k) &\leq C(x_k, x_\star) + \langle A(x_k - x_\star), y\rangle + V(z_k - z) \\
&\qquad - V(z_{\star,k} - z_k), \\
V(z^e_{\star,k} - z) &\leq V(z_k - z) - V(z^e_{\star,k} - z_k),
\end{aligned}
$$

where $x_\star$ is any primal solution. For the first inequality, we plug in $z = z^e_{\star,k}$ and for the second inequality, we plug in $z = z_{\star,k}$. Summing both equalities and rearranging yields

$$
2V(z^e_{\star,k} - z_{\star,k}) \leq \langle A(x_k - x_\star), y^e_{\star,k} - y_{\star,k}\rangle
$$

Since the inequality holds for any $x_\star \in \mathcal{X}_\star$, we can plug in $x^e_{\star,k}$ and use Cauchy-Schwarz inequality to get

$$
2V(z^e_{\star,k} - z_{\star,k}) \leq C^2_{2,V}\|A\| V(z_k - z^e_{\star,k})^{1/2} V(z_{\star,k} - z^e_{\star,k})^{1/2}.
$$

Lastly $C_{2,V} = \sqrt{\frac{2}{\underline{p}\min\left\{\min_i \tau_i^{-1} p_i^{-1}, \min_j \sigma_j^{-1}\pi_j^{-1}\right\}}}$. ■

Proof of Theorem 5.4. We note the definitions, as in [LFP19],

$$
\begin{aligned}
&A\colon (x, y) \mapsto (\partial g(x) + \partial h^*(y)) \\
&M\colon (x, y) \mapsto (A^\top y, -Ax) \\
&C\colon (x, y) \mapsto (\nabla f(x), 0) \\
&H\colon (x, y) \mapsto (\tau^{-1}x + A^\top y, \sigma^{-1}y).
\end{aligned}
$$

Under this notations, KKT operator defined in (5.4) can be written as

$$
F = A + M + C. \tag{5.39}
$$

Moreover, $\bar{z}_{k+1} = (H + A)^{-1}(H - M - C)z_k$, which in fact (without the term $\nabla f(x_k)$) is the well-known Arrow-Hurwicz operator [AAHU58].

Moreover, we will use the following inequalities regarding V and $\tilde{V}$ (see Lemma 5.1 for the definitions)

$$\tilde{V}(z) \geq C_{2,\tilde{V}}\left(\|x\|^2+\|y\|^2\right) := \frac{\underline{p}}{2}\min\left\{\min_i C(\tau)_i, \min_j \sigma_j^{-1}\right\}\left(\|x\|^2+\|y\|^2\right), \tag{5.40}$$

$$V(z) \leq C_{V,2}\left(\|x\|^2+\|y\|^2\right) := \frac{1}{2}\max\left\{\max_i \frac{1}{\tau_i}, \max_j \frac{1}{\sigma_j}\right\}\left(\|x\|^2+\|y\|^2\right). \tag{5.41}$$

We recall the definition of $z_{\star,k}$

$$z_{\star,k} = \arg\min_{u\in\mathcal{Z}_\star} D_p(x_k, u) + V(z_k - u).$$

We now argue that $z_{\star,k}$ is well-defined under our assumptions. Under Assumption 5.1, we know that the solution set is convex and closed; $D_p(x_k, u) \geq 0$ for all $u \in \mathcal{Z}_\star$ and it is also lower semicontinuous. Next, we remark that $V(z_k - u)$ is a squared norm, thus coercive, therefore the sum is coercive and lower semicontinuous over $\mathcal{Z}_\star$. Hence, $z_{\star,k}$ exists.

We use the result of Lemma 5.1 with $z = z_{\star,k}$ and $D_d(\bar{y}_{k+1}, z_{\star,k}) \geq 0$

$$\mathbb{E}_k\Big[D_p(x_{k+1}; z_k^\star) + V(z_{k+1} - z_k^\star)] \leq (1-\underline{p})D_p(x_k; z_k^\star) + V(z_k - z_k^\star) - \tilde{V}(\bar{z}_{k+1} - z_k).$$

We use the definition of $z_{\star,k+1}$ to deduce

$$\begin{aligned}\mathbb{E}_k\Big[D_p(x_{k+1}, z_{\star,k+1}) + V(z_{k+1} - z_{\star,k+1})] \leq (1-\underline{p})D_p(x_k, z_{\star,k}) + V(z_k - z_{\star,k})& \\ - \tilde{V}(\bar{z}_{k+1} - z_k).&\end{aligned} \tag{5.42}$$

In addition to Bregman projections $z_{\star,k}$ and $\bar{z}_{\star,k+1}$, we introduce the definitions for Euclidean projections (see Lemma 5.10)

$$\begin{aligned} z^e_{\star,k} &= \arg\min_{u\in\mathcal{Z}_\star} V(z_k - u), \\ \bar{z}^e_{\star,k+1} &= \arg\min_{u\in\mathcal{Z}_\star} V(\bar{z}_{k+1} - u). \end{aligned}$$

Now, we use triangle inequalities and Lemma 5.10 to get

$$\begin{aligned} V(z_k - z_{\star,k})^{1/2} &\leq V(z_k - z^e_{\star,k})^{1/2} + V(z^e_{\star,k} - z_{\star,k})^{1/2} \leq (1+c)V(z_k - z^e_{\star,k})^{1/2} \\ &\leq (1+c)(V(z_k - \bar{z}_{k+1})^{1/2} + V(\bar{z}_{k+1} - \bar{z}^e_{\star,k+1})^{1/2} + V(z^e_{\star,k} - \bar{z}^e_{\star,k+1})^{1/2}) \end{aligned}$$

We use nonexpansiveness with metric V on the last term, to obtain

$$V(z_k - z_{\star,k})^{1/2} \leq (2+2c)V(z_k - \bar{z}_{k+1})^{1/2} + (1+c)V(\bar{z}^e_{\star,k+1} - \bar{z}_{k+1})^{1/2}.$$

We use the definition of $\bar{z}^e_{\star,k+1}$ to get $V(\bar{z}_{k+1} - \bar{z}^e_{\star,k+1}) \leq V(\bar{z}_{k+1} - P_{\mathcal{Z}_\star}(\bar{z}_{k+1}))$, where $P_{\mathcal{Z}_\star}$ is defined to be standard Euclidean projection, and then we use the relation between V and

Euclidean norm from (5.41)

$$\begin{aligned} V(z_k - z_{\star,k})^{1/2} &\leq (2+2c)\sqrt{C_{V,2}}\|z_k - \bar{z}_{k+1}\| + (1+c)\sqrt{C_{V,2}}\|P_{\mathcal{Z}_\star}(\bar{z}_{k+1}) - \bar{z}_{k+1}\| \\ &= (2+2c)\sqrt{C_{V,2}}\|z_k - \bar{z}_{k+1}\| + (1+c)\sqrt{C_{V,2}}\operatorname{dist}(\bar{z}_{k+1}, \mathcal{Z}_\star). \qquad (5.43) \end{aligned}$$

We now use metric subregularity of F for 0 (see Assumption 5.2), and the assumption that $\bar{z}_{k+1} \in \mathcal{N}(z_\star), \forall z_\star$,

$$\operatorname{dist}(\bar{z}_{k+1}, \mathcal{Z}_\star) \leq \eta \operatorname{dist}(0, F(\bar{z}_{k+1})).$$

We now use that $(H - M - C)(z_k - \bar{z}_{k+1}) \in F(\bar{z}_{k+1})$, which can be obtained by using (5.39) and $\bar{z}_{k+1} = (H+A)^{-1}(H - M - C)z_k$. Therefore

$$\operatorname{dist}(\bar{z}_{k+1}, \mathcal{Z}_\star) \leq \eta\|(H - M - C)(z_k - \bar{z}_{k+1})\| \leq \eta(\|H - M\| + \bar{\beta})\|z_k - \bar{z}_{k+1}\|,$$

where $\bar{\beta}$ is the global Lipschitz constant of f.

We plug this inequality into (5.43) to obtain

$$V(z_k - z_{\star,k}) \leq C_{V,2}((2+2c) + (1+c)(\eta\|H - M\| + \bar{\beta}))^2\|\bar{z}_{k+1} - z_k\|^2.$$

Moreover, since $\tilde{V}(\bar{z}_{k+1} - z_k)$ is a squared norm, under the step size condition, it follows that $\tilde{V}(\bar{z}_{k+1} - z_k) \geq C_{2,\tilde{V}}\|\bar{z}_{k+1} - z_k\|^2$, as in (5.40), therefore,

$$V(z_k - z_{\star,k}) \leq \frac{C_{V,2}((2+2c) + (1+c)(\eta\|H - M\| + \bar{\beta}))^2}{C_{2,\tilde{V}}}\tilde{V}(\bar{z}_{k+1} - z_k).$$

We use this inequality in (5.42) to obtain

$$\begin{aligned} \mathbb{E}_k\Big[D_p(x_{k+1}, z^\star_{k+1}) + V(z_{k+1} - z_{\star,k+1})] &\leq (1 - \underline{p})D_p(x_k; z_{\star,k}) \\ &\quad + \left(1 - \frac{C_{2,\tilde{V}}}{C_{V,2}((2+2c) + (1+c)(\eta\|H - M\| + \bar{\beta}))^2}\right)V(z_k - z_{\star,k}), \end{aligned}$$

where the constants $C_{2,\tilde{V}}$, $C_{V,2}$ are as defined in (5.40), (5.41).

We take full expectation and define $\rho = \min\left(\underline{p}, \frac{C_{2,\tilde{V}}}{C_{V,2}((2+2c)+(1+c)(\eta\|H-M\|+\bar{\beta}))^2}\right)$. Then, we have that

$$\mathbb{E}\left[D_p(x_{k+1}, z_{\star,k+1}) + V(z_{k+1} - z_{\star,k+1})\right] \leq (1-\rho)\mathbb{E}\left[D_p(x_k, z_{\star,k}) + V(z_k - z_{\star,k})\right].$$

We have that $0 < \rho < 1$, as metric subregularity constant $\eta > 0$. Hence, linear convergence of $D_p(x_k, z_{\star,k})$ and $V(z_k - z_{\star,k})$ follows. We obtain the final result after using the definition of V, and the fact that $D_p(x_{k+1}, z_{\star,k+1}) \geq 0$. ■

5.8.4 Ergodic convergence rates

We introduce the following lemma, which establishes the properties of the sequence $\check{y}_k$, defined in (5.13).

Lemma 5.11. *We recall the definition of $\check{y}_k$ from* (5.13)*: let $\check{y}_1 = y_1 = \bar{y}_1$ and*

$$\begin{aligned} \check{y}^{(j)}_{k+1} &= \bar{y}^{(j)}_{k+1}, \quad \forall j \in J(i_{k+1}) \\ \check{y}^{(j)}_{k+1} &= \check{y}^{(j)}_{k}, \quad \forall j \not\in J(i_{k+1}), \end{aligned}$$

where computing $\check{y}_{k+1}$ requires the same number of operations as computing y_{k+1} every iteration, and $\check{y}_k$ is $\mathcal{F}_k$-measurable.

Moreover, if it holds for a function l that $l(y) = \sum_{j=1}^m l_j(y^{(j)})$ and

$$l_\gamma(y) = \sum_{j=1}^m \gamma_j l_j(y^{(j)}),$$

we have the following for the sequence $\check{y}_k$, and $\mathcal{F}_k$-measurable Y:

$$\begin{aligned} &\mathbb{E}_k\left[\check{y}^{(j)}_{k+1} - \check{y}^{(j)}_k\right] = \pi_j\left(\bar{y}^{(j)}_{k+1} - \check{y}^{(j)}_k\right), \forall j \\ &\mathbb{E}_k\left[\|\check{y}_{k+1} - Y\|^2_\gamma\right] = \|\bar{y}_{k+1} - Y\|^2_{\gamma\pi} - \|\check{y}_k - Y\|^2_{\gamma\pi} + \|\check{y}_k - Y\|^2_\gamma \\ &\mathbb{E}_k\left[l(\check{y}_{k+1})\right] = l_\pi(\bar{y}_{k+1}) - l_\pi(\check{y}_k) + l(\check{y}_k) \\ &\mathbb{E}_k\left[\|\check{y}_{k+1} - y_{k+1}\|^2_\gamma\right] = \|\bar{x}_{k+1} - x_k\|^2_{B(\gamma)} + \|\check{y}_k - y_k\|^2_\gamma - \|\check{y}_k - y_k\|^2_{\gamma\pi} \\ &\sum_{k=1}^K \mathbb{E}\left[\|\check{y}_{k+1} - \check{y}_k\|^2_\gamma\right] \le 2\sum_{k=1}^K \mathbb{E}\left[\|\bar{y}_{k+1} - y_k\|^2_{\gamma\pi}\right] + 2\sum_{k=1}^K \mathbb{E}\left[\|\bar{x}_{k+1} - x_k\|^2_{B(\gamma)}\right] \\ &\mathbb{E}\left[\|\check{y}_k - Y\|^2_{\gamma\pi}\right] \le 2\mathbb{E}\left[\|y_k - Y\|^2_{\gamma\pi}\right] + 2\sum_{k=1}^K \mathbb{E}\left[\|\bar{x}_{k+1} - x_k\|^2_{B(\gamma)}\right], \end{aligned}$$

where $B(\gamma)_i = p_i \sum_{j=1}^m \theta_j^2 \gamma_j \sigma_j^2 A_{j,i}^2$, and $\pi_j = \sum_{i\in I(j)} p_i$.

Proof. We first use the definition of $\check{y}_k$ to get the first result. For any j,

$$\begin{aligned} \mathbb{E}_k[\check{y}^{(j)}_{k+1}] &= \mathbb{E}_k\left[\mathbb{1}_{j\in J(i_{k+1})}\bar{y}^{(j)}_{k+1} + \mathbb{1}_{j\not\in J(i_{k+1})}\check{y}^{(j)}_k\right] = \sum_{i=1}^n p_i\left[\mathbb{1}_{j\in J(i)}(\bar{y}^{(j)}_{k+1}) + \mathbb{1}_{j\not\in J(i)}\check{y}^{(j)}_k\right] \\ &= \sum_{i\in I(j)} p_i \bar{y}^{(j)}_{k+1} + \sum_{i\not\in I(j)} p_i \check{y}^{(j)}_k = \sum_{i\in I(j)} p_i \bar{y}^{(j)}_{k+1} + \sum_{i=1}^n p_i \check{y}^j_k - \sum_{i\in I(j)} p_i \check{y}^{(j)}_k \\ &= \check{y}^{(j)}_k + \pi_j\left(\bar{y}^{(j)}_{k+1} - \check{y}^{(j)}_k\right). \end{aligned}$$

For the second result, we estimate similar to Lemma 5.8

$$\mathbb{E}_k\left[\|\check{y}_{k+1} - Y\|^2_\gamma\right] = \mathbb{E}_k\left[\sum_{j\in J(i_{k+1})} \gamma_j(\bar{y}^{(j)}_{k+1} - Y^{(j)})^2 + \sum_{j\not\in J(i_{k+1})} \gamma_j(\check{y}^{(j)}_k - Y^{(j)})^2\right]$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} p_i \left[\sum_{j \in J(i)} \gamma_j (\bar{y}_{k+1}^{(j)} - Y^{(j)})^2 + \sum_{j \notin J(i)} \gamma_j (\breve{y}_k^{(j)} - Y^{(j)})^2 \right] \\
&= \sum_{j=1}^{m} \sum_{i \in I(j)} p_i \gamma_j (\bar{y}_{k+1}^{(j)} - Y^{(j)})^2 + \sum_{j=1}^{m} \sum_{i=1}^{n} p_i \gamma_j (\breve{y}_k^{(j)} - Y^{(j)})^2 - \sum_{j=1}^{m} \sum_{i \in I(j)} p_i \gamma_j (\breve{y}_k^{(j)} - Y^{(j)})^2 \\
&= \| \bar{y}_{k+1} - Y \|_{\gamma\pi}^2 + \| \breve{y}_k - Y \|_{\gamma}^2 - \| \breve{y}_k - Y \|_{\gamma\pi}^2 .
\end{aligned}
$$

We derive the third result using similar estimations

$$
\begin{aligned}
\mathbb{E}_k \left[l(\breve{y}_{k+1}) \right] &= \mathbb{E}_k \left[\sum_{j \in J(i_{k+1})} l_j(\bar{y}_{k+1}^{(j)}) + \sum_{j \notin J(i_{k+1})} l_j(\breve{y}_k^{(j)}) \right] \\
&= \sum_{i=1}^{n} p_i \left[\sum_{j \in J(i)} l_j(\bar{y}_{k+1}^{(j)}) + \sum_{j \notin J(i)} l_j(\breve{y}_k^{(j)}) \right] \\
&= \sum_{j=1}^{m} \pi_j l_j(\bar{y}_{k+1}^{(j)}) + \sum_{j=1}^{m} \sum_{i=1}^{n} p_i l_j(\breve{y}_k^{(j)}) - \sum_{j=1}^{m} \pi_j l_j(\breve{y}_k^{(j)}) \\
&= l_\pi(\bar{y}_{k+1}) - l_\pi(\breve{y}_k) + l(\breve{y}_k).
\end{aligned}
$$

For the fourth result, we use the definitions of both $\breve{y}_{k+1}$ and y_{k+1} (see Algorithm 5.1 and (5.16)),

$$
\begin{aligned}
\mathbb{E}_k \Big[\| \breve{y}_{k+1} - y_{k+1} \|_{\gamma}^2 \Big] = \mathbb{E}_k \Big[&\sum_{j \in J(i_{k+1})} \gamma_j \left(\bar{y}_{k+1}^{(j)} - (\bar{y}_{k+1}^{(j)} + \sigma_j \theta_j A_{j,i_{k+1}} (x_{k+1}^{(i_{k+1})} - x_k^{(i_{k+1})})) \right)^2 \\
&+ \sum_{j \notin J(i_{k+1})} \gamma_j \left(\breve{y}_k^{(j)} - y_k^{(j)} \right)^2 \Big] \\
= \sum_{i=1}^{n} p_i \left[\sum_{j \in J(i)} \gamma_j \left(\sigma_j \theta_j A_{j,i} (\bar{x}_{k+1}^{(i)} - x_k^{(i)}) \right)^2 + \sum_{j \notin J(i)} \gamma_j \left(\breve{y}_k^{(j)} - y_k^{(j)} \right)^2 \right] \\
= \sum_{i=1}^{n} p_i \sum_{j \in J(i)} \gamma_j \sigma_j^2 A_{j,i}^2 \theta_i^2 (\bar{x}_{k+1}^{(i)} - x_k^{(i)})^2 + \sum_{i=1}^{n} p_i \sum_{j=1}^{m} \gamma_j (\breve{y}_k^{(j)} - y_k^{j})^2 - \sum_{i=1}^{n} \sum_{j \in J(i)} p_i \gamma_j (\breve{y}_k^{(j)} - y_k^{(j)})^2 \\
= \| \bar{x}_{k+1} - x_k \|_{B(\gamma)}^2 + \| \breve{y}_k - y_k \|_{\gamma}^2 - \| \breve{y}_k - y_k \|_{\gamma\pi}^2 , \qquad (5.44)
\end{aligned}
$$

where for the second equality, we used the fact that x_{k+1} is different from x_k only on the coordinate i_{k+1}, which gives

$$
(A(x_{k+1} - x_k))^{(j)} = (A((x_{k+1}^{(i_{k+1})} - x_k^{(i_{k+1})}) e_{i_{k+1}}))^{(j)} = A_{j,i_{k+1}} (\bar{x}_{k+1}^{(i_{k+1})} - x_k^{(i_{k+1})}),
$$

and for the last equality, we noted $A_{j,i} = 0, \forall\, j \notin J(i)$ and defined

$$
B(\gamma)_i = p_i \sum_{j=1}^{m} \gamma_j \sigma_j^2 \theta_j^2 A_{j,i}^2 .
$$

For the fifth result, we first take full expectation and then sum the inequality (5.44)

$$
\sum_{k=1}^{K} \mathbb{E} \Big[\| \breve{y}_k - y_k \|_{\gamma\pi}^2 \Big] \le \sum_{k=1}^{K} \mathbb{E} \Big[\| \bar{x}_{k+1} - x_k \|_{B(\gamma)}^2 \Big] + \| y_1 - \breve{y}_1 \|_{\gamma}^2 . \qquad (5.45)
$$

Then, we write from the second result that

$$\mathbb{E}_k\left[\|\breve{y}_{k+1}-\breve{y}_k\|_\gamma^2\right]=\|\bar{y}_{k+1}-\breve{y}_k\|_{\gamma\pi}^2\le 2\|\bar{y}_{k+1}-y_k\|_{\gamma\pi}^2+2\|\breve{y}_k-y_k\|_{\gamma\pi}^2.$$

We take full expectation and sum to get

$$\begin{aligned}\sum_{k=1}^K\mathbb{E}\left[\|\breve{y}_{k+1}-\breve{y}_k\|_\gamma^2\right]&\le 2\sum_{k=1}^K\mathbb{E}\left[\|\bar{y}_{k+1}-y_k\|_{\gamma\pi}^2\right]+2\sum_{k=1}^K\mathbb{E}\left[\|\breve{y}_k-y_k\|_{\gamma\pi}^2\right]\\&\le 2\sum_{k=1}^K\mathbb{E}\left[\|\bar{y}_{k+1}-y_k\|_{\gamma\pi}^2\right]+2\sum_{k=1}^K\mathbb{E}\left[\|\bar{x}_{k+1}-x_k\|_{B(\gamma)}^2\right],\end{aligned}$$

where we have used (5.45) and the fact that $\breve{y}_1=y_1$.

For the last result, we note that

$$\mathbb{E}\|\breve{y}_k-Y\|_{\gamma\pi}^2\le 2\mathbb{E}\|y_k-Y\|_{\gamma\pi}^2+2\mathbb{E}\|\breve{y}_k-y_k\|_{\gamma\pi}^2,$$

and we use (5.45) with $\breve{y}_1=y_1$, for the second term. ■

We continue with the restatement and the proof of Theorem 5.5. This length of the proof is due to the complications discussed earlier. First, as discussed in [AFC21], the order of expectation and supremum requires a special proof which delays taking expectations of the estimates (which prohibits simplifications and results in long expressions). Lemma 5.12 thus can be seen as a version of Lemma 5.1 with expectation not taken.

However, this is not enough due to the special structure of our new method suited for sparse settings. In particular, we have to manipulate the terms with dual variable carefully, as we cannot average $\bar{y}_k$ (see Lemma 5.1). Therefore, the treatment with $\breve{y}_k$, which is characterized in Lemma 5.11 and Lemma 5.13, is an intricate part of our proof.

Lemma 5.12. *Let Assumption 5.1 hold. Given the definitions of D_p and D_d given from Lemma 5.1, it follows that*

$$\begin{aligned}0\ge\ &D_p(x_{k+1},z)+\underline{p}D_d(\bar{y}_{k+1},z)-(1-\underline{p})D_p(x_k;z)+\tilde{V}(\bar{z}_{k+1}-z_k)+\frac{1}{2}\|\bar{x}_{k+1}-x_k\|_{\beta P}^2+S_1+S_2\\&+V(z_{k+1}-z)-V(z_k-z),\end{aligned}$$

where

$$\begin{aligned}S_1=\ &g_P(\bar{x}_{k+1})-g_P(x_k)-\big(g(x_{k+1})-g(x_k)\big)\\&-f(x_{k+1})+f(x_k)-\underline{p}f(x_k)+\underline{p}f(x)-\underline{p}\langle\nabla f(x_k),x-x_k\rangle-\langle\nabla f(x_k),P(x_k-\bar{x}_{k+1})\rangle\\&+\frac{\underline{p}}{2}\|\bar{x}_{k+1}\|_{\tau^{-1}}^2-\frac{\underline{p}}{2}\|x_k\|_{\tau^{-1}}^2-\left(\frac{\underline{p}}{2}\|x_{k+1}\|_{\tau^{-1}P^{-1}}^2-\frac{\underline{p}}{2}\|x_k\|_{\tau^{-1}P^{-1}}^2\right)\\&+\frac{\underline{p}}{2}\|\bar{y}_{k+1}\|_{\sigma^{-1}}^2-\frac{\underline{p}}{2}\|y_k\|_{\sigma^{-1}}^2+\underline{p}\langle\bar{y}_{k+1},\pi^{-1}\theta AP(\bar{x}_{k+1}-x_k)\rangle\end{aligned}$$

$$
\begin{aligned}
&+\frac{p}{2}\sum_{i=1}^{n} p_i \sum_{j=1}^{m} \pi_j^{-1}\sigma_j\theta_j^2 A_{j,i}^2(\bar{x}_{k+1}^{(i)}-x_k^{(i)})^2-\frac{p}{2}\Big(\|y_{k+1}\|^2_{\sigma^{-1}\pi^{-1}}-\|y_k\|^2_{\sigma^{-1}\pi^{-1}}\Big),\\
S_2 &= \langle y, A(x_k-x_{k+1})-AP(x_k-\bar{x}_{k+1})\rangle+\underline{p}\langle x, x_k-\bar{x}_{k+1}-P^{-1}(x_k-x_{k+1})\rangle_{\tau^{-1}}\\
&-\underline{p}\langle y, \pi^{-1}\sigma^{-1}(y_k-y_{k+1})-\sigma^{-1}(y_k-\bar{y}_{k+1})+\pi^{-1}\theta AP(\bar{x}_{k+1}-x_k)\rangle.
\end{aligned}
$$

Proof. We now follow the proof of Lemma 1 without taking conditional expectations, similar to ergodic convergence rate proof of [AFC21].

First, we have, from (5.20)

$$
\begin{aligned}
g_P(x')+\underline{p}h^*(y) \geq{}& \underbrace{g_P(\bar{x}_{k+1})+\underline{p}h^*(\bar{y}_{k+1})-\langle A^\top\bar{y}_{k+1}, P(x'-\bar{x}_{k+1})\rangle+\underline{p}\langle Ax_k, y-\bar{y}_{k+1}\rangle}_{T_1}\\
&\underbrace{-\langle\nabla f(x_k), P(x'-\bar{x}_{k+1})\rangle}_{T_2}+\underbrace{\frac{1}{2}\Big(\|x_k-\bar{x}_{k+1}\|^2_{\tau^{-1}P}+\|x'-\bar{x}_{k+1}\|^2_{\tau^{-1}P}-\|x'-x_k\|^2_{\tau^{-1}P}\Big)}_{T_3}\\
&\underbrace{+\frac{p}{2}\big(\|y_k-\bar{y}_{k+1}\|^2_{\sigma^{-1}}+\|y-\bar{y}_{k+1}\|^2_{\sigma^{-1}}-\|y-y_k\|^2_{\sigma^{-1}}\big)}_{T_4}. \qquad (5.46)
\end{aligned}
$$

We start with T_1 and add and subtract $\langle A^\top y, x_{k+1}-x\rangle-\underline{p}\langle Ax, \bar{y}_{k+1}-y\rangle-\langle A^\top y, x_k-x\rangle+\underline{p}\langle A^\top y, x_k-x\rangle+g(x_{k+1})+g_P(x_k)-g(x_k)-\underline{p}\langle y-\bar{y}_{k+1}, \pi^{-1}\theta AP(x_k-\bar{x}_{k+1})\rangle$

$$
\begin{aligned}
T_1 ={}& g(x_{k+1})-g(x_k)+g_P(x_k)+g_P(\bar{x}_{k+1})-g(x_{k+1})+g(x_k)-g_P(x_k)+\underline{p}h^*(\bar{y}_{k+1})\\
&-\langle A^\top\bar{y}_{k+1}, P(x'-\bar{x}_{k+1})\rangle+\underline{p}\langle Ax_k, y-\bar{y}_{k+1}\rangle+\underline{p}\langle y-\bar{y}_{k+1}, \pi^{-1}\theta AP(x_k-\bar{x}_{k+1})\rangle\\
&+\langle A^\top y, x_{k+1}-x\rangle-\underline{p}\langle Ax, \bar{y}_{k+1}-y\rangle+\langle A^\top y, x_k-x\rangle-\underline{p}\langle A^\top y, x_k-x\rangle\\
&-\langle A^\top y, x_{k+1}-x\rangle+\underline{p}\langle Ax, \bar{y}_{k+1}+y\rangle-\langle A^\top y, x_k-x\rangle+\underline{p}\langle A^\top y, x_k-x\rangle\\
&-\underline{p}\langle y-\bar{y}_{k+1}, \pi^{-1}\theta AP(x_k-\bar{x}_{k+1})\rangle. \qquad (5.47)
\end{aligned}
$$

We first use $x'^{(i)}=p_i^{-1}\underline{p}x^{(i)}+(1-p_i^{-1}\underline{p})x_k^{(i)}$ as in Lemma 5.9 to get

$$
-\langle A^\top\bar{y}_{k+1}, P(x'-\bar{x}_{k+1})\rangle=-\underline{p}\langle A^\top\bar{y}_{k+1}, x-x_k\rangle-\langle A^\top\bar{y}_{k+1}, P(x_k-\bar{x}_{k+1})\rangle.
$$

Next, we use that

$$
\begin{aligned}
&\underline{p}\big[-\langle A^\top y, x_k-x\rangle+\langle Ax, \bar{y}_{k+1}-y\rangle+\langle Ax_k, y-\bar{y}_{k+1}\rangle-\langle A^\top\bar{y}_{k+1}, x-x_k\rangle\big]=\\
&\qquad\qquad\underline{p}\big[\langle A^\top(y-\bar{y}_{k+1}), x-x_k\rangle+\langle y-\bar{y}_{k+1}, A(x_k-x)\rangle\big]=0,
\end{aligned}
$$

to obtain

$$
\begin{aligned}
T_1 ={}& g(x_{k+1})-g(x_k)+g_P(x_k)+g_P(\bar{x}_{k+1})-g(x_{k+1})+g(x_k)-g_P(x_k)+\underline{p}h^*(\bar{y}_{k+1})\\
&-\langle A^\top\bar{y}_{k+1}, P(x_k-\bar{x}_{k+1})\rangle+\underline{p}\langle y-\bar{y}_{k+1}, \pi^{-1}\theta AP(x_k-\bar{x}_{k+1})\rangle
\end{aligned}
$$

$$
\begin{aligned}
&+ \langle A^\top y, x_{k+1} - x\rangle - \underline{p}\langle Ax, \bar{y}_{k+1} - y\rangle + \langle A^\top y, x_k - x\rangle \\
&- \langle A^\top y, x_{k+1} - x\rangle - \langle A^\top y, x_k - x\rangle + \underline{p}\langle A^\top y, x_k - x\rangle \\
&- \underline{p}\langle y - \bar{y}_{k+1}, \pi^{-1}\theta AP(x_k - \bar{x}_{k+1})\rangle. \qquad (5.48)
\end{aligned}
$$

We use $\theta_j = \frac{\pi_j}{\underline{p}}$ to deduce

$$
\begin{aligned}
\langle A^\top y, x_k - x_{k+1}\rangle - \langle A^\top \bar{y}_{k+1}, P(x_k - \bar{x}_{k+1})\rangle - \langle y - \bar{y}_{k+1}, \underline{p}\pi^{-1}\theta AP(x_k - \bar{x}_{k+1})\rangle& \\
= -\langle y, AP(x_k - \bar{x}_{k+1})\rangle.& \qquad (5.49)
\end{aligned}
$$

We use the last identity and recall the definitions of D_p and D_d to write T_1 as

$$
\begin{aligned}
T_1 &= D_p(x_{k+1}, z) - f(x_{k+1}) + f(x) + \underline{p}D_d(\bar{y}_{k+1}, z) + \underline{p}h^*(y) - (1-\underline{p})D_p(x_k; z) \qquad (5.50)\\
&- \underline{p}\big(g(x_k) - g(x)\big) + g_P(x_k) + g_P(\bar{x}_{k+1}) - g(x_{k+1}) + g(x_k) - g_P(x_k) \\
&+ \langle y, A(x_k - x_{k+1}) - AP(x_k - \bar{x}_{k+1})\rangle + \underline{p}\langle y - \bar{y}_{k+1}, \pi^{-1}\theta AP(x_k - \bar{x}_{k+1})\rangle + (1-\underline{p})(f(x_k) - f(x)).
\end{aligned}
$$

Second, for T_2, we use $x' = P^{-1}\underline{p}x + (1 - P^{-1}\underline{p})x_k = x_k + P^{-1}\underline{p}(x - x_k)$ to obtain

$$
T_2 = -\langle \nabla f(x_k), P(x' - \bar{x}_{k+1}) = -\underline{p}\langle \nabla f(x_k), x - x_k\rangle - \langle \nabla f(x_k), P(x_k - \bar{x}_{k+1})\rangle.
$$

We now combine these two estimates

$$
\begin{aligned}
T_1 + T_2 = D_p(x_{k+1}, z) &+ \underline{p}D_d(\bar{y}_{k+1}, z) - (1-\underline{p})D_p(x_k, z) + g_P(x_k) + \underline{p}g(x) - \underline{p}g(x_k) + \underline{p}h^*(y) \\
&- f(x_{k+1}) + f(x_k) - \underline{p}f(x_k) + \underline{p}f(x) - \underline{p}\langle \nabla f(x_k), x - x_k\rangle - \langle \nabla f(x_k), P(x_k - \bar{x}_{k+1})\rangle \\
&+ g_P(\bar{x}_{k+1}) - g(x_{k+1}) + g(x_k) - g_P(x_k) + \langle y, A(x_k - x_{k+1}) - AP(x_k - \bar{x}_{k+1})\rangle \\
&+ \underline{p}\langle y - \bar{y}_{k+1}, \pi^{-1}\theta AP(x_k - \bar{x}_{k+1})\rangle. \qquad (5.51)
\end{aligned}
$$

We now work on T_3 in (5.46), in order to make terms depending on x telescope. First, we note that by Lemma 5.9, with the slight change of using $\bar{x}_{k+1}$ instead of x_{k+1} and $\tau^{-1}P$ instead of τ^{-1} in the metric, we get

$$
\begin{aligned}
\frac{1}{2}\|x' - \bar{x}_{k+1}\|^2_{\tau^{-1}P} - \frac{1}{2}\|x' - x_k\|^2_{\tau^{-1}P} = \frac{p}{2}\|x - \bar{x}_{k+1}\|^2_{\tau^{-1}} - \frac{p}{2}\|x - x_k\|^2_{\tau^{-1}}& \\
+ \frac{1}{2}\|\bar{x}_{k+1} - x_k\|^2_{\tau^{-1}P} - \frac{p}{2}\|\bar{x}_{k+1} - x_k\|^2_{\tau^{-1}}.&
\end{aligned}
$$

Thus, on T_3, we add and subtract $\frac{p}{2}\|x - x_{k+1}\|^2_{\tau^{-1}P^{-1}} - \frac{p}{2}\|x - x_k\|^2_{\tau^{-1}P^{-1}}$

$$
\begin{aligned}
T_3 = \|\bar{x}_{k+1} - x_k\|^2_{\tau^{-1}P} &- \frac{p}{2}\|\bar{x}_{k+1} - x_k\|^2_{\tau^{-1}} + \frac{p}{2}\|x - x_{k+1}\|^2_{\tau^{-1}P^{-1}} - \frac{p}{2}\|x - x_k\|^2_{\tau^{-1}P^{-1}} \\
+ \frac{p}{2}\|x - \bar{x}_{k+1}\|^2_{\tau^{-1}} &- \frac{p}{2}\|x - x_k\|^2_{\tau^{-1}} - \left(\frac{p}{2}\|x - x_{k+1}\|^2_{\tau^{-1}P^{-1}} - \frac{p}{2}\|x - x_k\|^2_{\tau^{-1}P^{-1}}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \|\bar{x}_{k+1} - x_k\|^2_{\tau^{-1}P} - \frac{\underline{p}}{2}\|\bar{x}_{k+1} - x_k\|^2_{\tau^{-1}} + \frac{\underline{p}}{2}\|x - x_{k+1}\|^2_{\tau^{-1}P^{-1}} - \frac{\underline{p}}{2}\|x - x_k\|^2_{\tau^{-1}P^{-1}} + \frac{\underline{p}}{2}\|\bar{x}_{k+1}\|^2_{\tau^{-1}} \\
&- \frac{\underline{p}}{2}\|x_k\|^2_{\tau^{-1}} + \underline{p}\langle x, x_k - \bar{x}_{k+1}\rangle_{\tau^{-1}} - \left(\frac{\underline{p}}{2}\|x_{k+1}\|^2_{\tau^{-1}P^{-1}} - \frac{\underline{p}}{2}\|x_k\|^2_{\tau^{-1}P^{-1}} + \underline{p}\langle x, (x_k - x_{k+1})\rangle_{\tau^{-1}P^{-1}}\right) \\
&+ \frac{\underline{p}}{2}\|\bar{x}_{k+1}\|^2_{\tau^{-1}} - \frac{\underline{p}}{2}\|x_k\|^2_{\tau^{-1}} - \left(\frac{\underline{p}}{2}\|x_{k+1}\|^2_{\tau^{-1}P^{-1}} - \frac{\underline{p}}{2}\|x_k\|^2_{\tau^{-1}P^{-1}}\right) \\
&+ \underline{p}\langle x, x_k - \bar{x}_{k+1} - P^{-1}(x_k - x_{k+1})\rangle_{\tau^{-1}}. \qquad (5.52)
\end{aligned}
$$

We estimate T_4 in (5.46) similarly. First note that $\theta_j = \frac{\pi_j}{\underline{p}}$ and on T_4, we add and subtract

$$
\begin{aligned}
\frac{\underline{p}}{2}\Bigg(&\|y - y_{k+1}\|^2_{\sigma^{-1}\pi^{-1}} - \|y - y_k\|^2_{\sigma^{-1}\pi^{-1}} + 2\langle y - \bar{y}_{k+1}, \pi^{-1}\theta AP(x_k - \bar{x}_{k+1})\rangle \\
&+ \sum_{i=1}^{n} p_i \sum_{j=1}^{m} \pi_j^{-1}\sigma_j\theta_j^2 A_{j,i}\left(\bar{x}^{(i)}_{k+1} - x^{(i)}_k\right)^2\Bigg). \qquad (5.53)
\end{aligned}
$$

In particular, we have

$$
\begin{aligned}
T_4 &= \frac{\underline{p}}{2}\|y_k - \bar{y}_{k+1}\|^2_{\sigma^{-1}} + \frac{\underline{p}}{2}\|y - y_{k+1}\|^2_{\sigma^{-1}\pi^{-1}} - \frac{\underline{p}}{2}\|y - y_k\|^2_{\sigma^{-1}\pi^{-1}} \\
&+ \frac{\underline{p}}{2}\|y - \bar{y}_{k+1}\|^2_{\sigma^{-1}} - \frac{\underline{p}}{2}\|y - y_k\|^2_{\sigma^{-1}} + \underline{p}\langle \bar{y}_{k+1} - y, \pi^{-1}\theta AP(\bar{x}_{k+1} - x_k)\rangle \\
&+ \frac{\underline{p}}{2}\sum_{i=1}^{n} p_i \sum_{j=1}^{m} \pi_j^{-1}\theta_j^2\sigma_j A^2_{j,i}(\bar{x}^{(i)}_{k+1} - x^{(i)}_k)^2 - \left(\frac{\underline{p}}{2}\|y - y_{k+1}\|^2_{\sigma^{-1}\pi^{-1}} - \frac{\underline{p}}{2}\|y - y_k\|^2_{\sigma^{-1}\pi^{-1}}\right) \\
&- \underline{p}\langle \bar{y}_{k+1} - y, \pi^{-1}\theta AP(\bar{x}_{k+1} - x_k)\rangle - \frac{\underline{p}}{2}\sum_{i=1}^{n} p_i \sum_{j=1}^{m} \pi_j^{-1}\sigma_j\theta_j^2 A^2_{j,i}(\bar{x}^{(i)}_{k+1} - x^{(i)}_k)^2 \\
&= \frac{\underline{p}}{2}\|y_k - \bar{y}_{k+1}\|^2_{\sigma^{-1}} + \frac{\underline{p}}{2}\|y - y_{k+1}\|^2_{\sigma^{-1}\pi^{-1}} - \frac{\underline{p}}{2}\|y - y_k\|^2_{\sigma^{-1}\pi^{-1}} \\
&+ \frac{\underline{p}}{2}\|\bar{y}_{k+1}\|^2_{\sigma^{-1}} - \frac{\underline{p}}{2}\|y_k\|^2_{\sigma^{-1}} + \underline{p}\langle \bar{y}_{k+1}, \pi^{-1}\theta AP(\bar{x}_{k+1} - x_k)\rangle \\
&+ \frac{\underline{p}}{2}\sum_{i=1}^{n} p_i \sum_{j=1}^{m} \pi_j^{-1}\sigma_j\theta_j^2 A^2_{j,i}(\bar{x}^{(i)}_{k+1} - x^{(i)}_k)^2 + \underline{p}\langle y, y_k - \bar{y}_{k+1}\rangle_{\sigma^{-1}} - \underline{p}\langle y, \pi^{-1}\theta AP(\bar{x}_{k+1} - x_k)\rangle \\
&- \frac{\underline{p}}{2}\left(\|y_{k+1}\|^2_{\sigma^{-1}\pi^{-1}} - \|y_k\|^2_{\sigma^{-1}\pi^{-1}}\right) - \underline{p}\langle y, y_k - y_{k+1}\rangle_{\sigma^{-1}\pi^{-1}} \\
&- \underline{p}\langle \bar{y}_{k+1} - y, \pi^{-1}\theta AP(\bar{x}_{k+1} - x_k)\rangle - \frac{\underline{p}}{2}\sum_{i=1}^{n} p_i \sum_{j=1}^{m} \pi_j^{-1}\sigma_j\theta_j^2 A^2_{j,i}(\bar{x}^{(i)}_{k+1} - x^{(i)}_k)^2. \qquad (5.54)
\end{aligned}
$$

To simplify, let us introduce some more definitions to have simpler expression when we combine T_1, T_2, T_3, T_4 from eqs. (5.51), (5.52) and (5.54). On a high level, S_1 will collect zero mean terms independent of (x, y) and S_2 will collect zero mean terms dependent on (x, y).

$$
\begin{aligned}
S_1 = g_P(\bar{x}_{k+1}) - g_P(x_k) - \big(g(x_{k+1}) - g(x_k)\big) \\
- f(x_{k+1}) + f(x_k) - \underline{p}f(x_k) + \underline{p}f(x) - \underline{p}\langle \nabla f(x_k), x - x_k\rangle - \langle \nabla f(x_k), P(x_k - \bar{x}_{k+1})\rangle \\
+ \frac{\underline{p}}{2}\|\bar{x}_{k+1}\|^2_{\tau^{-1}} - \frac{\underline{p}}{2}\|x_k\|^2_{\tau^{-1}} - \left(\frac{\underline{p}}{2}\|x_{k+1}\|^2_{\tau^{-1}P^{-1}} - \frac{\underline{p}}{2}\|x_k\|^2_{\tau^{-1}P^{-1}}\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{\underline{p}}{2}\|\bar{y}_{k+1}\|^2_{\sigma^{-1}} - \frac{\underline{p}}{2}\|y_k\|^2_{\sigma^{-1}} + \underline{p}\langle \bar{y}_{k+1}, \pi^{-1}\theta AP(\bar{x}_{k+1} - x_k)\rangle \\
&+\frac{\underline{p}}{2}\sum_{i=1}^{n} p_i \sum_{j=1}^{m} \pi_j^{-1}\sigma_j\theta_j^2 A_{j,i}^2(\bar{x}^{(i)}_{k+1} - x^{(i)}_k)^2 - \frac{\underline{p}}{2}\left(\|y_{k+1}\|^2_{\sigma^{-1}\pi^{-1}} - \|y_k\|^2_{\sigma^{-1}\pi^{-1}}\right) \qquad (5.55)\\
S_2 &= \langle y, A(x_k - x_{k+1}) - AP(x_k - \bar{x}_{k+1})\rangle + \underline{p}\langle x, x_k - \bar{x}_{k+1} - P^{-1}(x_k - x_{k+1})\rangle_{\tau^{-1}} \\
&- \underline{p}\langle y, \pi^{-1}\sigma^{-1}(y_k - y_{k+1}) - \sigma^{-1}(y_k - \bar{y}_{k+1}) + \pi^{-1}\theta AP(\bar{x}_{k+1} - x_k)\rangle
\end{aligned}
$$

We can now collect T_1, T_2, T_3, T_4, and use the definitions of S_1, S_2, in (5.46)

$$
\begin{aligned}
g_P(x') + \underline{p}h^*(y) &\geq D_p(x_{k+1}; z) + \underline{p}D_d(\bar{y}_{k+1}; z) - (1-\underline{p})D_p(x_k; z) + g_P(x_k) + \underline{p}g(x) - \underline{p}g(x_k) \\
&+ \underline{p}h^*(y) + \frac{\underline{p}}{2}\|x - x_{k+1}\|^2_{\tau^{-1}P^{-1}} - \frac{\underline{p}}{2}\|x - x_k\|^2_{\tau^{-1}P^{-1}} + \frac{\underline{p}}{2}\|y - y_{k+1}\|^2_{\sigma^{-1}\pi^{-1}} - \frac{\underline{p}}{2}\|y - y_k\|^2_{\sigma^{-1}\pi^{-1}} \\
&+ \|\bar{x}_{k+1} - x_k\|^2_{\tau^{-1}P} - \frac{\underline{p}}{2}\|\bar{x}_{k+1} - x_k\|^2_{\tau^{-1}} - \frac{\underline{p}}{2}\sum_{i=1}^{n} p_i \sum_{j=1}^{m} \pi_j^{-1}\sigma_j\theta_j^2 A_{j,i}^2(\bar{x}^i_{k+1} - x^i_k)^2 \\
&- \frac{1}{2}\|\bar{x}_{k+1} - x_k\|^2_{\beta P} + \frac{1}{2}\|\bar{x}_{k+1} - x_k\|^2_{\beta P} + \frac{\underline{p}}{2}\|y_k - \bar{y}_{k+1}\|^2_{\sigma^{-1}} + S_1 + S_2. \qquad (5.56)
\end{aligned}
$$

We make few observations on this inequality. First, by Lemma 5.9, as in (5.25)

$$
g_P(x_k) + \underline{p}g(x) - \underline{p}g(x_k) - g_P(x') \geq 0.
$$

Second, we have, as in (5.32)

$$
\begin{aligned}
\tilde{V}(\bar{z}_{k+1} - z_k) &= \frac{\underline{p}}{2}\|\bar{y}_{k+1} - y_k\|^2_{\sigma^{-1}} + \|\bar{x}_{k+1} - x_k\|^2_{\tau^{-1}P} - \frac{\underline{p}}{2}\|\bar{x}_{k+1} - x_k\|^2_{\tau^{-1}} \\
&- \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} \underline{p}p_i\pi_j^{-1}\sigma_j\theta_j^2 A_{j,i}^2(\bar{x}^{(i)}_{k+1} - x^{(i)}_k)^2 - \frac{1}{2}\|\bar{x}_{k+1} - x_k\|^2_{\beta P} \\
&= \frac{\underline{p}}{2}\|\bar{y}_{k+1} - y_k\|^2_{\sigma^{-1}} + \frac{\underline{p}}{2}\|\bar{x}_{k+1} - x_k\|^2_{C(\tau)},
\end{aligned}
$$

where

$$
C(\tau)_i = \frac{2p_i}{\underline{p}\tau_i} - \frac{1}{\tau_i} - p_i\sum_{j=1}^{m}\pi_j^{-1}\sigma_j\theta_j^2 A_{j,i}^2 - \frac{\beta_i p_i}{\underline{p}}.
$$

We use these estimates in (5.56) and the definition of V to conclude. ■

Lemma 5.13. *Let Assumption 5.1 hold and let h be separable. Given the definitions of D_p and D_d from Lemma 5.1, we define*

$$
D_d^{\gamma}(\bar{y}_{k+1}, z) = \sum_{j=1}^{m}\gamma_j\left(h_j^*(\bar{y}^{(j)}_{k+1}) - h_j^*(y^{(j)}) - \langle (Ax)^{(j)}, \bar{y}^{(j)}_{k+1} - y^{(j)}\rangle\right).
$$

Moreover let S_1, S_2 be as Lemma 5.12, and $\breve{y}_k$ as Lemma 5.11, it follows that

$$
0 \geq \underline{p}D_p(x_{k+1}, z) + \underline{p}D_d(\breve{y}_{k+1}, z) + \tilde{V}(\bar{z}_{k+1} - z_k) + \frac{1}{2}\|\bar{x}_{k+1} - x_k\|^2_{\beta P} + V(z_{k+1} - z) - V(z_k - z)
$$

$$+ S_1 + S_2 + (1-\underline{p})D_p(x_{k+1},z) - (1-\underline{p})D_p(x_k,z) + \underline{p}D_d^{\pi^{-1}-I}(\breve{y}_{k+1},z) - \underline{p}D_d^{\pi^{-1}-I}(\breve{y}_k,z)$$
$$+ \underline{p}h^*(\bar{y}_{k+1}) - \underline{p}h^*(\breve{y}_k) - \underline{p}\big(h^*_{\pi^{-1}}(\breve{y}_{k+1}) - h^*_{\pi^{-1}}(\breve{y}_k)\big) + \underline{p}\langle Ax, \breve{y}_k - \bar{y}_{k+1} - \pi^{-1}\left(\breve{y}_k - \breve{y}_{k+1}\right)\rangle.$$

Proof. This lemma is an intricate part of the proof of Theorem 5.5. If we finish the estimations as in Chapter 4, then we will end up needing to average $\bar{y}_k$. However, this is not feasible in our algorithm, since we do not update full dual vector, thus we do not compute $\bar{y}_k$ unless the data is fully dense. We will use Lemma 5.11 to go from $\bar{y}_k$ to $\breve{y}_k$. Let us repeat the definition of $\breve{y}_k$ from Lemma 5.11: Let $\breve{y}_1 = y_1 = \bar{y}_1$, and

$$\breve{y}^{(j)}_{k+1} = \bar{y}^{(j)}_{k+1}, \quad \forall j \in J(i_{k+1})$$
$$\breve{y}^{(j)}_{k+1} = \breve{y}^{(j)}_{k}, \quad \forall j \not\in J(i_{k+1}).$$

We now work on $D_d(\bar{y}_{k+1};z)$ and note that h^*_γ is defined as in Lemma 5.11.

$$\begin{aligned}
D_d(\bar{y}_{k+1};z) &= D_d(\breve{y}_{k+1};z) - \langle Ax, \bar{y}_{k+1}\rangle + h^*(\bar{y}_{k+1}) + \langle Ax, \breve{y}_{k+1}\rangle - h^*(\breve{y}_{k+1}) \\
&= D_d(\breve{y}_{k+1};z) - \langle Ax, \bar{y}_{k+1}\rangle + h^*(\bar{y}_{k+1}) + \langle Ax, \breve{y}_{k+1}\rangle - h^*(\breve{y}_{k+1}) \\
&+ h^*_{I-\pi^{-1}}(\breve{y}_{k+1}) - h^*_{I-\pi^{-1}}(\breve{y}_{k+1}) + h^*_{I-\pi^{-1}}(\breve{y}_{k}) - h^*_{I-\pi^{-1}}(\breve{y}_{k}) \\
&+ \langle Ax, (I-\pi^{-1})(\breve{y}_{k+1} - \breve{y}_k)\rangle - \langle Ax, (I-\pi^{-1})(\breve{y}_{k+1} - \breve{y}_k) \\
&= D_d(\breve{y}_{k+1};z) + h^*(\bar{y}_{k+1}) - h^*(\breve{y}_k) - \big(h^*_{\pi^{-1}}(\breve{y}_{k+1}) - h^*_{\pi^{-1}}(\breve{y}_k)\big) \\
&+ h^*_{\pi^{-1}-1}(\breve{y}_{k+1}) - h^*_{\pi^{-1}-1}(\breve{y}_k) + \langle Ax, \breve{y}_k - \bar{y}_{k+1} - \pi^{-1}\left(\breve{y}_k - \breve{y}_{k+1}\right)\rangle \\
&+ \langle Ax, (I-\pi^{-1})(\breve{y}_{k+1} - \breve{y}_k)\rangle \\
&= D_d(\breve{y}_{k+1};z) + h^*(\bar{y}_{k+1}) - h^*(\breve{y}_k) - \big(h^*_{\pi^{-1}}(\breve{y}_{k+1}) - h^*_{\pi^{-1}}(\breve{y}_k)\big) \\
&+ \langle Ax, \breve{y}_k - \bar{y}_{k+1} - \pi^{-1}\left(\breve{y}_k - \breve{y}_{k+1}\right)\rangle + D_d^{\pi^{-1}-I}(\breve{y}_{k+1};z) - D_d^{\pi^{-1}-I}(\breve{y}_k;z)
\end{aligned}$$

We insert this estimate into the result of Lemma 5.12 to finish the proof. ■

The following lemma is similar to Lemma 4.10.

Lemma 5.14. *Given a Euclidean space $\mathcal{S}$, a fixed diagonal matrix $\gamma \geq 0$, let the random sequences $u_k, v_k \in \mathcal{S}$ be $\mathcal{F}_k$-measurable with*

$$u_{k+1} = v_{k+1} - \mathbb{E}_k\left[v_{k+1}\right].$$

Let $\tilde{x}_1$ be arbitrary and set for $k \geq 1$,

$$\tilde{x}_{k+1} = \tilde{x}_k + u_{k+1}.$$

Then, $\tilde{x}_k$ is $\mathcal{F}_k$-measurable and we have

$$\sum_{k=1}^K \langle x, u_{k+1}\rangle_\gamma \leq \frac{1}{2}\|\tilde{x}_1 - x\|^2_\gamma + \sum_{k=1}^K \langle \tilde{x}_k, u_{k+1}\rangle_\gamma + \frac{1}{2}\sum_{k=1}^K \|v_{k+1}\|^2_\gamma,$$

with $\mathbb{E}\left[\sum_{k=1}^K \langle \tilde{x}_k, u_{k+1}\rangle_\gamma\right] = 0$ *and for any* $S \subset \mathcal{S}$

$$\mathbb{E} \sup_{x\in S} \sum_{k=1}^K \langle x, u_{k+1}\rangle_\gamma \le \sup_{x\in S} \frac{1}{2}\|\tilde{x}_1 - x\|_\gamma^2 + \frac{1}{2}\sum_{k=1}^K \mathbb{E}\|v_{k+1}\|_\gamma^2.$$

Proof. First, by the definition of $\tilde{x}_{k+1}$, for all $x \in \mathcal{S}$

$$\frac{1}{2}\|\tilde{x}_{k+1} - x\|_\gamma^2 = \frac{1}{2}\|\tilde{x}_k - x\|_\gamma^2 + \langle \tilde{x}_k - x, u_{k+1}\rangle_\gamma + \frac{1}{2}\|u_{k+1}\|_\gamma^2.$$

Summing this inequality gives

$$\sum_{k=1}^K \langle x, u_{k+1}\rangle_\gamma \le \frac{1}{2}\|\tilde{x}_1 - x\|_\gamma^2 + \sum_{k=1}^K \langle \tilde{x}_k, u_{k+1}\rangle_\gamma + \sum_{k=1}^K \frac{1}{2}\|u_{k+1}\|_\gamma^2.$$

We take supremum and expectation of both sides to get

$$\mathbb{E}\left[\sup_{x\in S}\sum_{k=1}^K \langle x, u_{k+1}\rangle_\gamma\right] \le \sup_{x\in S}\frac{1}{2}\|\tilde{x}_1 - x\|_\gamma^2 + \sum_{k=1}^K \mathbb{E}\left[\langle \tilde{x}_k, u_{k+1}\rangle_\gamma\right] + \sum_{k=1}^K \frac{1}{2}\mathbb{E}\left[\|u_{k+1}\|_\gamma^2\right].$$

By the law of total expectation, $\mathcal{F}_k$-measurability of $\tilde{x}_k$ and $\mathbb{E}_k\left[u_{k+1}\right] = 0$, we have

$$\sum_{k=1}^K \mathbb{E}\left[\langle \tilde{x}_k, u_{k+1}\rangle_\gamma\right] = \sum_{k=1}^K \mathbb{E}\left[\mathbb{E}_k\left[\langle \tilde{x}_k, u_{k+1}\rangle_\gamma\right]\right] = \sum_{k=1}^K \mathbb{E}\left[\langle \tilde{x}_k, \mathbb{E}_k\left[u_{k+1}\right]\rangle_\gamma\right] = 0.$$

Finally, we use the definition of u_k and the inequality $\mathbb{E}\|X - \mathbb{E}X\|^2 \le \mathbb{E}\|X\|^2$ which holds for any random variable X. ■

As mentioned in the main text, we will give the theorems in the appendix with tighter, but more complicated constants. After Theorem 5.7, we show how we obtained the simplified bounds in our main text.

Theorem 5.5. Let Assumption 5.1 hold and θ, τ, σ are chosen as in (5.8), (5.9). Moreover, let h be separable.

We define $x_K^{av} = \frac{1}{K}\sum_{k=1}^K x_k$ and $y_K^{av} = \frac{1}{K}\sum_{k=1}^K \breve{y}_k$, where $\breve{y}_k$ is defined in (5.13), then for any bounded set $\mathcal{C} = \mathcal{C}_x \times \mathcal{C}_y \subset \mathcal{Z}$ with iterates of Algorithm 5.1, it holds that

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(x_K^{av}, y_K^{av})\right] \le \frac{C_g}{\underline{p}K},$$

where $C_g = C_{g,1} + C_{g,2} + C_{g,3} + C_{g,4}$, $C_{\tau,\tilde{V}} = \min_i C(\tau)_i\tau_i$,
$C_{g,1} = \sup_{z\in\mathcal{C}}\left\{2\underline{p}\|x_0 - x\|_{\tau^{-1}P^{-1}}^2 + 2\underline{p}\|y_0 - y\|_{\sigma^{-1}\pi^{-1}}^2\right\} + (1-\underline{p})4\sqrt{\Delta\underline{p}^{-1}}\|A\|\sup_{y\in\mathcal{C}_y}\|y\|_{\tau P}$
$+2\underline{p}\sqrt{\Delta_0\underline{p}^{-1} + \|2P - \underline{p}\|\Delta_0\underline{p}^{-3}C_{\tau,\tilde{V}}^{-1}}\|A\|\|\pi^{-1} - I\|\sup_{x\in\mathcal{C}_x}\|x\|_{\sigma\pi}$,
$C_{g,2} = \|2P - \underline{p}\|\left(1 + \|P/\underline{p}\| + \|\tau^{1/2}P^{1/2}A^\top\pi^{-1/2}\sigma^{1/2}\|^2\right)\frac{2\Delta_0}{\underline{p}C_{\tau,\tilde{V}}} + \Delta_0 C_{\tau,\tilde{V}}^{-1}$

$$
\begin{aligned}
&+ (4+4\|\tau^{1/2}P^{1/2}A^\top \pi^{-1/2}\sigma^{1/2}\|^2)\Delta_0,\\
&C_{g,3} = (1-\underline{p})\Big(f(x_0)+g(x_0)-f(x_\star)-g(x_\star)+\|A^\top y_\star\|_{\tau P}\sqrt{2\Delta_0\underline{p}^{-1}}\Big),\\
&C_{g,4} = \underline{p}h^*_{\pi^{-1}-I}(\breve{y}_0)+\underline{p}\textstyle\sum_{j=1}^m(\pi_j^{-1}-1)h_j^*(y_\star^j)+\frac{\underline{p}}{2}\|Ax_\star\|^2_{\sigma\pi^{-1}}+\Delta_0+\frac{\|2P-\underline{p}\|\Delta_0}{\underline{p}^2C_{\tau,\tilde{V}}}.
\end{aligned}
$$

Proof. We start with the result of Lemma 5.13. First, we will manipulate the terms arising in $S_2 + \langle Ax, \breve{y}_k - \bar{y}_{k+1} - \pi^{-1}(\breve{y}_k - \breve{y}_{k+1})$ (see definition of S_2 in Lemma 5.12).

$$
\begin{aligned}
-\Big(S_2 + \underline{p}\langle Ax, \breve{y}_k - \bar{y}_{k+1} - \pi^{-1}(\breve{y}_k - \breve{y}_{k+1})\rangle\Big) &= -\langle y, A(x_k - x_{k+1}) - AP(x_k - \bar{x}_{k+1})\rangle\\
-\underline{p}\langle x, x_k - \bar{x}_{k+1} - P^{-1}(x_k - x_{k+1})\rangle_{\tau^{-1}} &+ \underline{p}\langle y, \pi^{-1}\sigma^{-1}(y_k - y_{k+1}) - \sigma^{-1}(y_k - \bar{y}_{k+1})\\
&+ \pi^{-1}\theta AP(\bar{x}_{k+1} - x_k)\rangle - \underline{p}\langle Ax, \breve{y}_k - \bar{y}_{k+1} - \pi^{-1}(\breve{y}_k - \breve{y}_{k+1})\rangle \quad (5.57)
\end{aligned}
$$

For the four terms on the right hand side, we will apply Lemma 5.14. First, $\mathbb{E}_k\left[\pi^{-1}\left(\breve{y}_k - \breve{y}_{k+1}\right)\right] = \breve{y}_k - \bar{y}_{k+1}$ from Lemma 5.11, $\mathbb{E}_k\left[P^{-1}(x_k - x_{k+1}) = x_k - \bar{x}_{k+1}\right]$, by coordinate wise updates. Finally, as in the proof of Lemma 5.11, we can derive, as $A_{j,i} = 0, \forall i \not\in I(j)$,

$$
\begin{aligned}
\mathbb{E}_k[y_{k+1}^{(j)}] &= \sum_{i=1}^n p_i\Big[\mathbb{1}_{j\in J(i)}\big(\bar{y}_{k+1}^{(j)} + \sigma_j\theta_j A_{j,i}(\bar{x}_{k+1}^{(i)} - x_k^{(i)})\big) + \mathbb{1}_{j\not\in J(i)}y_k^{(j)}\Big]\\
&= y_k^{(j)} + \sum_{i\in I(j)} p_i\big(\bar{y}_{k+1}^{(j)} - y_k^{(j)}\big) + \sum_{i=1}^n p_i\sigma_j\theta_j A_{j,i}(\bar{x}_{k+1}^{(i)} - x_k^{(i)})\\
&= y_k^{(j)} + \pi_j(\bar{y}_{k+1}^{(j)} - y_k^{(j)}) + \sigma_j\theta_j(AP(\bar{x}_{k+1} - x_k))^{(j)}\\
\mathbb{E}_k[y_{k+1}] &= y_k + \pi(\bar{y}_{k+1} - y_k) + \sigma\theta AP(\bar{x}_{k+1} - x_k)\\
&\iff \mathbb{E}_k\left[y_k - y_{k+1}\right] = \pi(y_k - \bar{y}_{k+1}) - \sigma\theta AP(\bar{x}_{k+1} - x_k).
\end{aligned}
$$

In particular, for (5.57), we set in Lemma 5.14

$$
\begin{aligned}
&u_{k+1} = -\underline{p}^{-1}\sigma\pi A(x_k - x_{k+1}) + \underline{p}^{-1}\sigma\pi AP(x_k - \bar{x}_{k+1}), && \gamma = \sigma^{-1}\pi^{-1}\underline{p}, \mathcal{S} = \mathcal{Y}, && \tilde{x}_1 = y_1.\\
&u_{k+1} = (x_k - x_{k+1}) - P(x_k - \bar{x}_{k+1}), && \gamma = \tau^{-1}P^{-1}, \quad \mathcal{S} = \mathcal{X}, && \tilde{x}_1 = x_1,\\
&u_{k+1} = \big(y_k - y_{k+1}\big) - \pi\left(y_k - \bar{y}_{k+1}\right) + \sigma\theta AP(\bar{x}_{k+1} - x_k), && \gamma = \sigma^{-1}\pi^{-1}, \quad \mathcal{S} = \mathcal{Y}, && \tilde{x}_1 = y_1,\\
&u_{k+1} = \tau PA^\top\left(\pi^{-1}\left(\breve{y}_k - \breve{y}_{k+1}\right) - \left(\breve{y}_k - \bar{y}_{k+1}\right)\right), && \gamma = \tau^{-1}P^{-1}, \quad \mathcal{S} = \mathcal{X}, && \tilde{x}_1 = x_1,
\end{aligned}
$$

Then, we can apply Lemma 5.14 for these cases to bound (5.57) as

$$
\begin{aligned}
\mathbb{E}\sup_{z\in\mathcal{C}}[(5.57)] \le \sup_{z\in\mathcal{C}}\Big\{\underline{p}\|x - x_1\|^2_{\tau^{-1}P^{-1}} + \underline{p}\|y - y_1\|^2_{\sigma^{-1}\pi^{-1}}\Big\} &+ \sum_{k=1}^K\frac{\underline{p}}{2}\mathbb{E}\big[\|x_k - x_{k+1}\|^2_{\tau^{-1}P^{-1}}\big]\\
+\mathbb{E}\big[\|y_k - y_{k+1}\|^2_{\sigma^{-1}\pi^{-1}}\big] + \sum_{k=1}^K\frac{1}{2\underline{p}}\mathbb{E}\|\sigma\pi A(x_k - x_{k+1})\|^2_{\sigma^{-1}\pi^{-1}}&\\
+\sum_{k=1}^K\frac{\underline{p}}{2}\mathbb{E}\|\tau PA^\top\pi^{-1}(\breve{y}_k - \breve{y}_{k+1})\|^2_{\tau^{-1}P^{-1}}. \quad (5.58)&
\end{aligned}
$$

We now recall the definition of S_1 from (5.55), and use the identities (5.21), (5.22), Lemma 5.8, along with the law of total expectation to estimate

$$\begin{aligned}
\mathbb{E}\left[S_1 + \frac{1}{2}\|\bar{x}_{k+1} - x_k\|^2_{\beta P}\right] &= \mathbb{E}\left[-f(x_{k+1}) + f(x_k) - \langle \nabla f(x_k), P(x_k - \bar{x}_{k+1})\rangle + \frac{1}{2}\|\bar{x}_{k+1} - x_k\|^2_{\beta P}\right] \\
&\quad + \underline{p}\mathbb{E}\left[f(x) - f(x_k) - \langle \nabla f(x_k), x - x_k\rangle\right] \\
&\geq \mathbb{E}\left[-f(x_{k+1}) + f(x_k) - \mathbb{E}_k\left[\langle \nabla f(x_k), x_k - x_{k+1}\rangle\right] + \frac{1}{2}\|\bar{x}_{k+1} - x_k\|^2_{\beta P}\right] \\
&\geq \mathbb{E}\left[-\frac{1}{2}\mathbb{E}_k\left[\|x_k - x_{k+1}\|^2_{\beta}\right] + \frac{1}{2}\|\bar{x}_{k+1} - x_k\|^2_{\beta P}\right] \\
&= \mathbb{E}\left[-\frac{1}{2}\|x_k - \bar{x}_{k+1}\|^2_{\beta P} + \frac{1}{2}\|\bar{x}_{k+1} - x_k\|^2_{\beta P}\right] \\
&= 0, \qquad (5.59)
\end{aligned}$$

where the first inequality is by convexity, second inequality is by coordinatewise smoothness of f. Furthermore, for the result of Lemma 5.13, by Lemma 5.11 and the law of total expectation

$$\mathbb{E}\left[h^*(\check{y}_{k+1}) - h^*(\check{y}_k) - \left(h^*_{\pi^{-1}}(\check{y}_{k+1}) - h^*_{\pi^{-1}}(\check{y}_k)\right)\right] = 0. \qquad (5.60)$$

We rearrange and sum the result of Lemma 5.13, take supremum and expectation, plug in eqs. (5.58)–(5.60), and use $\tilde{V}$ is a squared norm

$$\begin{aligned}
\mathbb{E}\left[\sup_{z\in\mathcal{C}} \sum_{k=1}^{K} \underline{p}\left(D_p(x_k, z) + D_d(\check{y}_k, z)\right)\right] &\leq \sup_{z\in\mathcal{C}} \frac{3\underline{p}}{2}\left(\|x - x_0\|^2_{\tau^{-1}P^{-1}} + \|y - y_0\|^2_{\sigma^{-1}\pi^{-1}}\right) \\
&\quad + \mathbb{E}\sup_{z\in\mathcal{C}}(1 - \underline{p})\left(D_p(x_0, z) - D_p(x_K, z)\right) + \mathbb{E}\sup_{z\in\mathcal{C}} \underline{p}D_d^{\pi^{-1}-I}(\check{y}_0, z) - \underline{p}D_d^{\pi^{-1}-I}(\check{y}_K, z) \\
&\quad + \sum_{k=1}^{K} \frac{\underline{p}}{2}\mathbb{E}\|x_k - x_{k+1}\|^2_{\tau^{-1}P^{-1}} + \|y_k - y_{k+1}\|^2_{\sigma^{-1}\pi^{-1}} + \sum_{k=1}^{K} \frac{1}{2\underline{p}}\mathbb{E}\|\sigma\pi A(x_k - x_{k+1})\|^2_{\sigma^{-1}\pi^{-1}} \\
&\quad + \sum_{k=1}^{K} \frac{\underline{p}}{2}\mathbb{E}\|\tau P A^\top \pi^{-1}(\check{y}_k - \check{y}_{k+1})\|^2_{\tau^{-1}P^{-1}}. \qquad (5.61)
\end{aligned}$$

We first note by (5.38)

$$\sum_{k=1}^{\infty} \mathbb{E}\left[\tilde{V}(\bar{z}_{k+1} - z_k)\right] \leq \Delta_0. \qquad (5.62)$$

$$\tilde{V}(\bar{z}_{k+1} - z_k) \geq \frac{\underline{p}C_{\tau,\tilde{V}}}{2}\|\bar{x}_{k+1} - x_k\|^2_{\tau^{-1}} + \frac{\underline{p}}{2}\|\bar{y}_{k+1} - y_k\|^2_{\sigma^{-1}}, \qquad (5.63)$$

with $C_{\tau,\tilde{V}} = \min_i C(\tau)_i \tau_i$, where we used the definition of $\tilde{V}$ from Lemma 5.1.

We have

$$\mathbb{E}\sup_{z\in\mathcal{C}} D_p(x_0, z) - D_p(x_K, z) = \mathbb{E}\sup_{z\in\mathcal{C}_y} f(x_0) + g(x_0) - f(x_K) - g(x_K) + \langle A^\top y, x_0 - x_K\rangle$$

$$\leq \mathbb{E} \sup_{z \in \mathcal{C}_y} f(x_0) + g(x_0) - f(x_K) - g(x_K) + \|A\| \|y\|_{\tau P} \|x_0 - x_K\|_{\tau^{-1} P^{-1}}. \tag{5.64}$$

Then, we use the optimality conditions, convexity, and (5.36)

$$\begin{aligned} \mathbb{E}\left[f(x_K) + g(x_K)\right] &\geq \mathbb{E}\left[f(x_\star) + g(x_\star) - \langle A^\top y_\star, x_K - x_\star \rangle\right] \\ &\geq \mathbb{E}\left[f(x_\star) + g(x_\star) - \|A^\top y_\star\|_{\tau P} \|x_K - x_\star\|_{\tau^{-1} P^{-1}}\right] \\ &\geq f(x_\star) + g(x_\star) - \|A^\top y_\star\|_{\tau P} \sqrt{\frac{2\Delta_0}{\underline{p}}}. \end{aligned} \tag{5.65}$$

to obtain for this estimation

$$\begin{aligned} \mathbb{E} \sup_{z \in C} D_p(x_0, z) - D_p(x_K, z) \leq f(x_0) + g(x_0) - f(x_\star) - g(x_\star) + \|A^\top y_\star\|_{\tau P} \sqrt{2\Delta_0 \underline{p}^{-1}} \\ + 4\sqrt{2\Delta_0 \underline{p}^{-1}} \|A\| \sup_{y \in C_y} \|y\|_{\tau P}. \end{aligned} \tag{5.66}$$

We estimate similarly to obtain

$$\begin{aligned} \mathbb{E} \sup_{z \in \mathcal{C}} \underline{p} D_d^{\pi^{-1} - I}(\breve{y}_0; z) - \underline{p} D_d^{\pi^{-1} - I}(\breve{y}_K; z) &= \underline{p} \mathbb{E} h^*_{\pi^{-1} - I}(\breve{y}_0) - h^*_{\pi^{-1} - I}(\breve{y}_K) - \langle Ax, (\pi^{-1} - I)(\breve{y}_0 - \breve{y}_K) \rangle. \\ &\leq \underline{p} \mathbb{E} h^*_{\pi^{-1} - I}(\breve{y}_0) - h^*_{\pi^{-1} - I}(\breve{y}_K) + \|A\| \|\pi^{-1} - I\| \|x\|_{\sigma\pi} \|\breve{y}_0 - \breve{y}_K\|_{\sigma^{-1}\pi^{-1}}. \end{aligned} \tag{5.67}$$

By convexity and Lemma 5.11

$$\begin{aligned} \mathbb{E}\Big[h^*_{\pi^{-1} - I}(\breve{y}_K)\Big] &= \mathbb{E}\left[\sum_{j=1}^m (\pi_j^{-1} - 1) h_j^*(\breve{y}_K^{(j)}) \geq \sum_{j=1}^m (\pi_j^{-1} - 1)\left(h_j^*(y_\star^{(j)}) + \langle (Ax_\star)_j, \breve{y}_K^{(j)} - y_\star^{(j)} \rangle\right)\right] \\ &\geq \mathbb{E}\left[\sum_{j=1}^m (\pi_j^{-1} - 1) h_j^*(y_\star^{(j)}) - \frac{1}{2}\|Ax_\star\|^2_{\sigma\pi^{-1}} - \frac{1}{2}\|\breve{y}_K - y_\star\|^2_{\sigma^{-1}\pi^{-1}}\right] \\ &\geq \mathbb{E}\left[\sum_{j=1}^m (\pi_j^{-1} - 1) h_j^*(y_\star^{(j)}) - \frac{1}{2}\|Ax_\star\|^2_{\sigma\pi^{-1}} - \|y_K - y_\star\|^2_{\sigma^{-1}\pi^{-1}} - \sum_{k=1}^K \|\bar{x}_{k+1} - x_k\|^2_{B(\pi^{-2}\sigma^{-1})}\right] \\ &\geq \sum_{j=1}^m (\pi_j^{-1} - 1) h_j^*(y_\star^{(j)}) - \frac{1}{2}\|Ax_\star\|^2_{\sigma\pi^{-1}} - \frac{\Delta_0}{\underline{p}} - \frac{\|2P - \underline{p}\|\Delta_0}{\underline{p}^3 C_{\tau,\tilde{V}}}, \end{aligned} \tag{5.68}$$

where we used

$$\|x\|^2_{B(\pi^{-2}\sigma^{-1})} \leq \frac{1}{\underline{p}^2}\|2P - \underline{p}\| \|x\|^2_{\tau^{-1}}, \tag{5.69}$$

which follows by using the step size rule from (5.9) and definition of $B(\gamma)$ from Lemma 5.11.

Thus, the final bound for (5.67), after using Lemma 5.11

$$\mathbb{E} \sup_{z \in \mathcal{C}} \underline{p} D_d^{\pi^{-1} - I}(\breve{y}_0; z) - \underline{p} D_d^{\pi^{-1} - I}(\breve{y}_K; z) \leq \underline{p} h^*_{\pi^{-1} - I}(\breve{y}_0) + \underline{p} \sum_{j=1}^m (\pi_j^{-1} - 1) h_j^*(\breve{y}_\star^{(j)}) + \frac{\underline{p}}{2}\|Ax_\star\|^2_{\sigma\pi^{-1}}$$

$$+\Delta_0 + \frac{\|2P - \underline{p}\|\Delta_0}{\underline{p}^2 C_{\tau,\tilde{V}}} + 2\underline{p}\sqrt{\Delta_0 \underline{p}^{-1} + \|2P - \underline{p}\|\Delta_0 \underline{p}^{-3} C_{\tau,\tilde{V}}^{-1}} \|A\| \|\pi^{-1} - I\| \sup_{x \in \mathcal{C}_x} \|x\|_{\sigma\pi}. \tag{5.70}$$

We continue to estimate, by Lemma 5.8, the definition of $B(\gamma)$ from Lemma 5.11, and the definition of π_j

$$\begin{aligned}
\mathbb{E}_k\big[\|y_{k+1} - y_k\|^2_{\sigma^{-1}\pi^{-1}}\big] &= \|\bar{y}_{k+1} - y_k\|^2_{\sigma^{-1}} + 2\langle \bar{y}_{k+1} - y_k, \pi^{-1}\theta AP(\bar{x}_{k+1} - x_k)\rangle \\
&\qquad + \sum_{i=1}^{n}\sum_{j=1}^{m} p_i \pi_j^{-1} \sigma_j \theta_j^2 A_{j,i}^2 (\bar{x}_{k+1}^{(i)} - x_k^{(i)})^2 \\
&\le \|\bar{y}_{k+1} - y_k\|^2_{\sigma^{-1}} + \|\bar{y}_{k+1} - y_k\|^2_{\sigma^{-1}} + \sum_{i=1}^{n}\sum_{j=1}^{m} p_i^2 \pi_j^{-2} \sigma_j \theta_j^2 A_{j,i}^2 (\bar{x}_{k+1}^{(i)} - x_k^{(i)})^2 \\
&\qquad + \sum_{i=1}^{n}\sum_{j=1}^{m} p_i \pi_j^{-1} \sigma_j \theta_j^2 A_{j,i}^2 (\bar{x}_{k+1}^{(i)} - x_k^{(i)})^2 \\
&\le 2\|\bar{y}_{k+1} - y_k\|^2_{\sigma^{-1}} + \Big(1 + \|P/\underline{p}\|\Big) \|\bar{x}_{k+1} - x_k\|^2_{B(\pi^{-1}\sigma^{-1})}, &\text{(5.71)}\\
\sum_{k=1}^{K} \frac{p}{2}\mathbb{E}_k\left[\|x_{k+1} - x_k\|^2_{\tau^{-1}P^{-1}}\right] &= \frac{p}{2}\sum_{k=1}^{K} \|\bar{x}_{k+1} - x_k\|^2_{\tau^{-1}} \le \frac{\Delta_0}{C_{\tau,\tilde{V}}}. &\text{(5.72)}
\end{aligned}$$

We will continue with estimating the last two terms of (5.61):

$$\begin{aligned}
\frac{1}{2\underline{p}}\mathbb{E}_k\left[\|\sigma\pi A(x_k - x_{k+1})\|^2_{\sigma^{-1}\pi^{-1}}\right] &= \mathbb{E}_k\left[\frac{1}{2\underline{p}}\sum_{j=1}^{m} \sigma_j \pi_j ((A(x_k - x_{k+1}))^{(j)})^2\right] \\
&= \mathbb{E}_k\left[\frac{1}{2\underline{p}}\sum_{j=1}^{m} \sigma_j \pi_j A_{j,i_{k+1}}^2 (x_k^{(i_{k+1})} - \bar{x}_{k+1}^{(i_{k+1})})^2\right] \\
&= \frac{1}{2\underline{p}}\sum_{j=1}^{m}\sum_{i=1}^{n} p_i \sigma_j \pi_j A_{j,i}^2 (x_k^{(i)} - \bar{x}_{k+1}^{(i)})^2 \\
&= \frac{p}{2}\|\bar{x}_{k+1} - x_k\|^2_{B(\pi^{-1}\sigma^{-1})}. &\text{(5.73)}
\end{aligned}$$

Finally,

$$\begin{aligned}
\sum_{k=1}^{K} \mathbb{E}\Big[\|\tau P A^\top \pi^{-1}(\check{y}_{k+1} - \check{y}_k)\|^2_{\tau^{-1}P^{-1}}\Big] &\le \|\tau^{1/2}P^{1/2}A^\top \pi^{-1/2}\sigma^{1/2}\|^2 \sum_{k=1}^{K} \mathbb{E}\|\check{y}_{k+1} - \check{y}_k\|^2_{\sigma^{-1}\pi^{-1}} \\
&\le \|\tau^{1/2}P^{1/2}A^\top \pi^{-1/2}\sigma^{1/2}\|^2 \sum_{k=1}^{K} \mathbb{E}\Big[2\|\bar{y}_{k+1} - y_k\|^2_{\sigma^{-1}} + 2\|\bar{x}_{k+1} - x_k\|^2_{B(\pi^{-1}\sigma^{-1})}\Big], &\text{(5.74)}
\end{aligned}$$

where we use Lemma 5.11 for the last inequality.

By, eqs. (5.71)–(5.74), we finalized the bound for the last three terms in (5.61)

$$\sum_{k=1}^{K} \frac{p}{2}\mathbb{E}\|x_k - x_{k+1}\|^2_{\tau^{-1}P^{-1}} + \|y_k - y_{k+1}\|^2_{\sigma^{-1}\pi^{-1}} + \sum_{k=1}^{K} \frac{1}{2\underline{p}}\mathbb{E}\|\sigma\pi A(x_k - x_{k+1})\|^2_{\sigma^{-1}\pi^{-1}}$$

$$+\sum_{k=1}^{K}\frac{p}{2}\mathbb{E}\,\|\tau PA^\top\pi^{-1}(\check{y}_k-\check{y}_{k+1})\|^2_{\tau^{-1}P^{-1}}\le C_{g,2}, \tag{5.75}$$

where

$$C_{g,2}=\|2P-\underline{p}\|\Big(1+\|P/\underline{p}\|+\|\tau^{1/2}P^{1/2}A^\top\pi^{-1/2}\sigma^{1/2}\|^2\Big)\frac{2\Delta_0}{\underline{p}C_{\tau,\tilde{V}}}+\frac{\Delta_0}{C_{\tau,\tilde{V}}}$$
$$+(4+4\|\tau^{1/2}P^{1/2}A^\top\pi^{-1/2}\sigma^{1/2}\|^2)\Delta_0.$$

$C_{g,2}$ basically collects the bounds in eqs. (5.71)–(5.74) and uses the estimates from eqs. (5.62) and (5.63) on this bound.

Then, on (5.61), we use eqs. (5.62), (5.63), (5.66) and (5.69)–(5.74), definition of primal-dual gap function in (5.11), and Jensen's inequality to conclude. ■

Theorem 5.7. Let Assumption 5.1 hold. We use the same parameters θ,τ,σ and the definitions for x_K^{av} and y_K^{av} as Theorem 5.5. We consider two cases separately:
▷ If $h(\cdot)=\delta_{\{b\}}(\cdot)$, we obtain

$$\mathbb{E}\left[f(x_K^{av})+g(x_K^{av})-f(x_\star)-g(x_\star)\right]\le\frac{C_o}{\underline{p}K}.$$
$$\mathbb{E}\left[\|Ax_K^{av}-b\|\right]\le\frac{C_f}{\underline{p}K}.$$

▷ If h is L_h-Lipschitz continuous, we obtain

$$\mathbb{E}\Big[f(x_K^{av})+g(x_K^{av})+h(Ax_K^{av})-f(x_\star)-g(x_\star)-h(Ax_\star)\Big]\le\frac{C_l}{\underline{p}K},$$

where $C_f=2c_2\sqrt{\|y_\star-y_0\|^2_{\sigma^{-1}\pi^{-1}}+C_sc_2^{-1}+2c_1c_2^{-1}\|x_0-x_\star\|^2_{\tau^{-1}P^{-1}}}+2c_2\|y_\star-y_0\|_{\sigma^{-1}\pi^{-1}}$,
$C_o=C_s+\|y_\star\|_{\sigma^{-1}\pi^{-1}}C_f+c_1\|x_0-x_\star\|^2_{\tau^{-1}P^{-1}}+c_2\|y_\star-y_0\|^2_{\sigma^{-1}\pi^{-1}}$,
$C_l=C_s+c_1\|x_\star-x_0\|^2_{\tau^{-1}P^{-1}}+4c_2L_h^2$,
$c_1=\frac{3p}{2}+\underline{p}\|(2P-\underline{p})^{1/2}\|\|\pi^{-1}-I\|$,
$c_2=\frac{3p}{2}+(1-\underline{p})\|(2P-\underline{p})^{1/2}\|$, $C_s=C_{g,2}+C_{g,5}+C_{g,6}$, with $C_{g,2}$ as defined in Theorem 5.5 and $C_{g,5},C_{g,6}$ are defined in the proof in (5.81), (5.82).

Proof. We will use the first result of Lemma 5.14 on the result of Lemma 5.13, similar to (5.61). The difference is that we take supremum and expectation after using Lemma 5.14 and we process the terms $(1-\underline{p})\left(D_p(x_k,z)-D_p(x_{k+1},z)\right)+\underline{p}\left(D_d^{\pi^{-1}-I}(\check{y}_k,z)-D_d^{\pi^{-1}-I}(\check{y}_{k+1},z)\right)$ with small differences. In particular,

$$\sum_{k=0}^{K-1}(1-\underline{p})\left(D_p(x_k;z)-D_p(x_{k+1};z)\right)=(1-\underline{p})\left(D_p(x_0;z)-D_p(x_K;z)\right)$$
$$=(1-\underline{p})\left(f(x_0)+g(x_0)-f(x_K)-g(x_K)+\langle A^\top y,x_0-x_K\rangle\right).$$

For the final term, we estimate using the step size rule (5.9)

$$\begin{aligned}
\langle A^\top y, x_0 - x_K\rangle &= \sum_{i=1}^n \sum_{j=1}^m A_{j,i} y^{(j)} (x_0^{(i)} - x_K^{(i)}) \leq \sum_{i=1}^n \sqrt{\sum_{j=1}^m A_{j,i}^2 (x_0^{(i)} - x_K^{(i)})^2 \sigma_j \pi_j} \sqrt{\sum_{j=1}^m (y^{(j)})^2 \sigma_j^{-1} \pi_j^{-1}} \\
&\leq \sum_{i=1}^n \sqrt{\underline{p}(2p_i - \underline{p})(x_0^{(i)} - x_K^{(i)})^2 \tau_i^{-1} p_i^{-1}} \sqrt{\sum_{j=1}^m (y^{(j)})^2 \sigma_j^{-1} \pi_j^{-1}} \\
&\leq \sum_{i=1}^n \left(\frac{1}{2} \sqrt{2p_i - \underline{p}} (x_0^{(i)} - x_K^{(i)})^2 \tau_i^{-1} p_i^{-1} + \frac{\underline{p}\sqrt{2p_i - \underline{p}}}{2} \|y\|^2_{\sigma^{-1}\pi^{-1}} \right) \\
&\leq \frac{\|(2P - \underline{p})^{1/2}\|}{2} \|x_0 - x_K\|^2_{\tau^{-1}P^{-1}} + \frac{\|(2P - \underline{p})^{1/2}\|}{2} \|y\|^2_{\sigma^{-1}\pi^{-1}} \\
&\leq \frac{\|(2P - \underline{p})^{1/2}\|}{2} \|x_0 - x_K\|^2_{\tau^{-1}P^{-1}} + \|(2P - \underline{p})^{1/2}\| \left(\|y - y_0\|^2_{\sigma^{-1}\pi^{-1}} + \|y_0\|^2_{\sigma^{-1}\pi^{-1}} \right). \qquad (5.76)
\end{aligned}$$

We estimate similarly to obtain

$$\sum_{k=0}^{K-1} \underline{p} \left(D_d^{\pi^{-1}-I}(\check{y}_k, z) - D_d^{\pi^{-1}-I}(\check{y}_{k+1}, z) \right) = \underline{p} \left(h^*_{\pi^{-1}-I}(\check{y}_0) - h^*_{\pi^{-1}-I}(\check{y}_K) - \langle Ax, (\pi^{-1} - I)(\check{y}_0 - \check{y}_K)\rangle \right),$$

and

$$\begin{aligned}
-\langle Ax, (\pi^{-1} - I)(\check{y}_0 - \check{y}_K)\rangle &\leq \frac{1}{2} \|(2P - \underline{p})^{1/2}\| \|\pi^{-1} - I\| \left(\|x\|^2_{\tau^{-1}P^{-1}} + \|\check{y}_0 - \check{y}_K\|^2_{\sigma^{-1}\pi^{-1}} \right) \\
&\leq \|(2P - \underline{p})^{1/2}\| \|\pi^{-1} - I\| \left(\|x - x_0\|^2_{\tau^{-1}P^{-1}} + \|x_0\|^2_{\tau^{-1}P^{-1}} + \frac{1}{2} \|\check{y}_0 - \check{y}_K\|^2_{\sigma^{-1}\pi^{-1}} \right). \qquad (5.77)
\end{aligned}$$

We sum the result of Lemma 5.13, use eqs. (5.76) and (5.77), use the first result of Lemma 5.14 (see (5.57)), move the terms depending on (x, y) to LHS, take the supremum over $\mathcal{Z}$ and expectation. With these steps, instead of (5.61), we get

$$\begin{aligned}
&\mathbb{E}\left[\sup_{z\in\mathcal{Z}} \sum_{k=1}^K \underline{p} \left(D_p(x_k, z) + D_d(\check{y}_k, z) \right) - c_1 \|x_0 - x\|^2_{\tau^{-1}P^{-1}} - c_2 \|y_0 - y\|^2_{\sigma^{-1}\pi^{-1}} \right] \\
&\qquad \leq \underline{p} \left(h^*_{\pi^{-1}-I}(\check{y}_0) - h^*_{\pi^{-1}-I}(\check{y}_K) \right) + (1 - \underline{p}) \left(f(x_0) + g(x_0) - f(x_K) - g(x_K) \right) \\
&\quad + \sum_{k=1}^K \frac{\underline{p}}{2} \mathbb{E} \|x_k - x_{k+1}\|^2_{\tau^{-1}P^{-1}} + \mathbb{E}\|y_k - y_{k+1}\|^2_{\sigma^{-1}\pi^{-1}} + \sum_{k=1}^K \frac{1}{2\underline{p}} \mathbb{E} \|\sigma\pi A(x_k - x_{k+1})\|^2_{\sigma^{-1}\pi^{-1}} \\
&\quad + \sum_{k=1}^K \frac{\underline{p}}{2} \mathbb{E} \|\tau P A^\top \pi^{-1} (\check{y}_k - \check{y}_{k+1})\|^2_{\tau^{-1}P^{-1}} + \|(2P - \underline{p})^{1/2}\| \left(\frac{1}{2} \mathbb{E}\|x_0 - x_K\|^2_{\tau^{-1}P^{-1}} + \|y_0\|^2_{\sigma^{-1}\pi^{-1}} \right) \\
&\qquad + \|(2P - \underline{p})^{1/2}\| \|\pi^{-1} - I\| \left(\|x_0\|^2_{\tau^{-1}P^{-1}} + \frac{1}{2} \mathbb{E}\|\check{y}_0 - \check{y}_K\|^2_{\sigma^{-1}\pi^{-1}} \right), \qquad (5.78)
\end{aligned}$$

where $c_1 = \frac{3\underline{p}}{2} + \underline{p}\|(2P - \underline{p})^{1/2}\| \|\pi^{-1} - I\|$, $c_2 = \frac{3\underline{p}}{2} + (1 - \underline{p}) \|(2P - \underline{p})^{1/2}\|$.

We divide both sides by $\underline{p}$ and use Jensen's inequality to obtain the smoothed gap function [TDFC18]

$$\mathcal{G}_{\frac{2c_1}{\underline{p}K},\frac{2c_2}{\underline{p}K}}(x_K^{av}, y_K^{av}, x_0, y_0) = \sup_{z\in\mathcal{Z}} D_p(x_K^{av}, z) + D_d(y_K^{av}; z) - \frac{c_1}{\underline{p}K}\|x-x_0\|^2_{\tau^{-1}P^{-1}} - \frac{c_2}{\underline{p}K}\|y-y_0\|^2_{\sigma^{-1}\pi^{-1}}.$$

Then, we have, as in the proof of Theorem 5.5 that (see eqs. (5.61) and (5.75))

$$\begin{aligned}\underline{p}K\mathbb{E}[\mathcal{G}_{\frac{2c_1}{\underline{p}K},\frac{2c_2}{\underline{p}K}}(x_K^{av}, y_K^{av}; x_0, y_0)] &\le C_{g,2} + \underline{p}\left(h^*_{\pi^{-1}-I}(\check{y}_0) - h^*_{\pi^{-1}-I}(\check{y}_K)\right)\\ &+ (1-\underline{p})\left(f(x_0)+g(x_0)-f(x_K)-g(x_K)\right)\\ &+ (1-\underline{p})\|(2P-\underline{p})^{1/2}\|\left(\frac{1}{2}\|x_0-x_K\|^2_{\tau^{-1}P^{-1}} + \|y_0\|^2_{\sigma^{-1}\pi^{-1}}\right)\\ &+ \underline{p}\|(2P-\underline{p})^{1/2}\|\|\pi^{-1}-I\|\left(\|x_0\|^2_{\tau^{-1}P^{-1}} + \frac{1}{2}\|\check{y}_0-\check{y}_K\|^2_{\sigma^{-1}\pi^{-1}}\right) \qquad (5.79)\end{aligned}$$

By using (5.65) and (5.68) we obtain the bound

$$(1-\underline{p})\left(f(x_0)+g(x_0)-f(x_K)-g(x_K)\right) + \underline{p}\left(h^*_{\pi^{-1}-I}(\check{y}_0) - h^*_{\pi^{-1}-I}(\check{y}_K)\right) \le C_{g,5}, \qquad (5.80)$$

where

$$\begin{aligned}C_{g,5} = (1-\underline{p})&\left(f(x_0)+g(x_0)-f(x_\star)-g(x_\star)+\|A^\top y_\star\|_{\tau P}\sqrt{\frac{2\Delta_0}{\underline{p}}}\right)\\ &+\underline{p}\left(h^*_{\pi^{-1}-I}(y_0) - \sum_{j=1}^m(\pi_j^{-1}-1)h_j^*(y_\star^j) + \frac{1}{2}\|Ax_\star\|^2_{\sigma\pi^{-1}} + \frac{\Delta_0}{\underline{p}} + \frac{\Delta_0\|2P-\underline{p}\|}{\underline{p}^2 C_{\tau,\tilde{V}}}\right). \qquad (5.81)\end{aligned}$$

Next, we bound the last two terms in (5.79) using $\mathbb{E}\left[V(z_k - z_\star)\right] \le \Delta_0$ from (5.37) and we denote the bound as $C_{g,6}$

$$\begin{aligned}C_{g,6} = (1-\underline{p})\|(2P-\underline{p})^{1/2}\|&\left(\frac{4\Delta_0}{\underline{p}} + 2\|y_\star\|^2_{\sigma^{-1}\pi^{-1}}\right)\\ &+\underline{p}\|(2P-\underline{p})^{1/2}\|\|\pi^{-1}-I\|\left(\frac{4\Delta_0}{\underline{p}} + 2\|x_\star\|^2_{\tau^{-1}P^{-1}}\right), \qquad (5.82)\end{aligned}$$

so that RHS of (5.79) is upper bounded by $C_{g,2}+C_{g,5}+C_{g,6}$.

We consider two cases:

- If h is L_h Lipschitz continuous in norm $\|\cdot\|_{\sigma\pi}$, then $\|y-y_0\|^2_{\sigma^{-1}\pi^{-1}} \le 4L_h^2$. Then, we argue as in [FB19, Theorem 11] to get

$$\begin{aligned}\mathbb{E}\left[f(x_K^{av}) + g(x_K^{av}) + h(Ax_K^{av}) - f(x_\star) - g(x_\star) - h(Ax_\star)\right] &\le \mathbb{E}\left[\mathcal{G}_{\frac{2c_1}{\underline{p}K},\frac{2c_2}{\underline{p}K}}(x_K^{av}, y_K^{av}; x_0, y_0)\right]\\ &+\frac{c_1}{\underline{p}K}\|x_\star - x_0\|^2_{\tau^{-1}P^{-1}} + \frac{4c_2}{\underline{p}K}L_h^2.\end{aligned}$$

- If $h(\cdot) = \delta_b(\cdot)$, we use [TDFC18, Lemma 1] to obtain

$$\begin{aligned}\mathbb{E}\left[f(x_K^{av}) + g(x_K^{av}) - f(x_\star) - g(x_\star)\right] \leq \mathbb{E}\left[\mathcal{G}_{\frac{2c_1}{\underline{p}K},\frac{2c_2}{\underline{p}K}}(x_K^{av}, y_K^{av}; x_0, y_0)\right] + \frac{c_1}{\underline{p}K}\|x_0 - x_\star\|^2_{\tau^{-1}P^{-1}} \\ + \frac{c_2}{\underline{p}K}\|y_\star - y_0\|^2_{\sigma^{-1}\pi^{-1}} + \mathbb{E}\left[\|y_\star\|_{\sigma^{-1}\pi^{-1}}\|Ax_K^{av} - b\|_{\sigma\pi}\right],\end{aligned} \tag{5.83}$$

$$\begin{aligned}\mathbb{E}\left[\|Ax - b\|_{\sigma\pi}\right] \leq \frac{2c_2}{\underline{p}K}\|y_\star - y_0\|_{\sigma^{-1}\pi^{-1}} \\ + \frac{2c_2}{\underline{p}K}\sqrt{\|y_\star - y_0\|^2_{\sigma^{-1}\pi^{-1}} + \frac{\underline{p}K}{c_2}\left(\mathbb{E}\left[\mathcal{G}_{\frac{2c_1}{\underline{p}K},\frac{2c_2}{\underline{p}K}}(x_K^{av}, y_K^{av}; x_0, y_0)\right] + \frac{c_1}{\underline{p}K}\|x_0 - x_\star\|^2_{\tau^{-1}P^{-1}}\right)}.\end{aligned}$$

We plug in the bound of $\mathbb{E}\left[\mathcal{G}_{\frac{2c_1}{\underline{p}K},\frac{2c_2}{\underline{p}K}}(x_K^{av}, y_K^{av}; x_0, y_0)\right]$ to obtain the final results. ■

Simplification of the constants. As mentioned before Theorem 5.5, we now give the inequalities we use to obtain the bounds we have in the main text for Theorem 5.5 and Theorem 5.7 compared to the ones we have in the appendix. It is easy to see by using coarse inequalities, we first $p_i\underline{p}^{-1} \leq \underline{p}^{-1}$, second, $2p_i - \underline{p} \leq 2$, third, $\pi_j^{-1} - 1 \leq \underline{p}^{-1}$ as $\pi_j \geq \underline{p}$. Finally, by the definition of τ_i in (5.9), we can derive $\|\tau^{1/2}P^{1/2}A^\top\pi^{-1/2}\sigma^{1/2}\|^2 \leq 2\underline{p}^{-1}$. By using these constants in the bounds of Theorem 5.5 and Theorem 5.7 in the appendix, we arrive at the bounds given for these theorems in the main text.

6 Stochastic variance reduction for variational inequalities

In this chapter, we focus on monotone variational inequalities (VI) with finite sum structure, which is a generalization of convex-concave min-max problems. We study variance reduced methods, that also provide alternative ways to PDCD for solving problems we focused in Chapters 4 and 5.

We introduce variance reduced extragradient, forward-backward-forward and forward-reflected-backward methods. Similar to minimization, the new methods potentially improve the complexity of the deterministic algorithms depending on the Lipschitz constants. Our results reinforce the correspondence between variance reduction in min-max problems and minimization. A recent work showed optimality of our algorithms by establishing matching lower bounds.

This chapter is based on joint works with Yura Malitsky and Volkan Cevher [AM21, AMC21].

6.1 Introduction

We are interested in solving variational inequalities (VI)

$$\text{Find } z_\star \in \mathcal{Z} : \langle F(z_\star), z - z_\star \rangle + g(z) - g(z_\star) \geq 0, \quad \forall z \in \mathcal{Z}, \tag{6.1}$$

where g is a proper lower semicontinuous convex function and F is a monotone operator also given as the finite sum $F = \frac{1}{n}\sum_{i=1}^n F_i$.

A special case of monotone VIs is the structured saddle point problem

$$\min_x \max_y \Psi(x, y) + f(x) - h(y), \tag{6.2}$$

where f, h are proper lower semicontinuous convex functions and Ψ is a smooth convex-concave function. Indeed, problem (6.2) can be formulated as (6.1) by setting

$$z = (x, y), \quad F(z) = \begin{bmatrix} \nabla_x \Psi(x, y) \\ -\nabla_y \Psi(x, y) \end{bmatrix}, \quad g(z) = f(x) + h(y),$$

and $F(z) = \frac{1}{n} \sum_{i=1}^{n} F_i(z)$ (see [BB16, Section 2], [CGFLJ19, CJST19] for examples).

Another related problem is the monotone inclusion where the aim is to

$$\text{find} \quad z_\star \in \mathcal{Z} \quad \text{such that} \quad 0 \in (A + F)(x),$$

where $A\colon \mathcal{Z} \rightrightarrows \mathcal{Z}$ and $F\colon \mathcal{Z} \to \mathcal{Z}$ are maximally monotone operators and F is Lipschitz continuous with finite sum form. Monotone inclusions generalize (6.1) and our results also extend to this setting as will be shown in Section 6.4.1. Due to convenient abstraction, it is the problem (6.1) that will be our main concern.

The case when Ψ in (6.2) is convex-concave and, in particular when it is bilinear, has found numerous applications in machine learning, image processing and operations research, resulting in efficient methods being developed in the respective areas [CP11, EZC10, SSZ13, HA21]. As VI methods solve the formulation (6.1), they seamlessly apply to solve instances of (6.2) with nonbilinear Ψ.

In addition to the potentially complex structure of Ψ, the size of the data in modern learning tasks lead to development of stochastic variants of VI methods [NJLS09, BMSV19, IJOT17]. An important technique on this front is stochastic variance reduction [JZ13] which exploits the finite sum structures in problems to match the convergence rates of deterministic algorithms.

In the specific case of convex minimization, variance reduction has been transformative over the last decade [JZ13, DBLJ14, HLLJM15, KHR20]. As a result, there has been several works on developing variance reduced versions of the standard VI methods, including forward-backward [BB16], extragradient [Kor76, CGFLJ19], and mirror-prox [Nem04, CJST19]. Despite recent remarkable advances in this field, these methods rely on strong assumptions such as strong monotonicity [BB16, CGFLJ19] or boundedness of the domain [CJST19] and have complicated structures for handling the cases with non-bilinear Ψ [CJST19].

Such a dichotomy does not exist in minimization: variance reduction comes with no extra assumptions. This points out to a fundamental lack of understanding for its use in saddle point problems. This chapter shows that there is indeed a natural correspondence between variance reduction in variational inequalities and minimization.

6.1.1 Contributions

We design variance reduced extragradient and forward-reflected-backward methods, with Euclidean and Bregman setups. Our algorithms either match or improve complexity of existing

	Assumptions	**μ-adaptivity**	**Complexity**
[Kor76, Nem04, MT20b] EG/MP, FoRB	F is monotone	✓	$\mathcal{O}\left(\frac{NL_F}{\varepsilon}\right)$
[BB16, CGFLJ19, SZY17] FB, EG/MP	F is strongly monotone	×	N/A
[CJST19] EG/MP	F is monotone + $z \mapsto \langle F(z) + \tilde{\nabla} g(z), z - u\rangle$ is cvx. $\forall u$; or bounded domains	×	$\mathcal{O}\left(N + \frac{\sqrt{N}L}{\varepsilon}\right)$
This chapter EG/MP, FoRB	F is monotone	✓	$\mathcal{O}\left(N + \frac{\sqrt{N}L}{\varepsilon}\right)$

Table 6.1 – Table of algorithms with $F(z) = \sum_{i=1}^N F_i(z)$. EG: Extragradient, MP: Mirror-Prox, FB: forward-backward, FoRB: forward-reflected-backward. Complexity column refers to complexity under mere monotonicity.

methods and also have unique properties compared to previous works:

▷ Our methods are the first variance reduced VI algorithms that converge almost surely, under mere monotonicity.

▷ For bilinear problems, we match the best-known complexity.

▷ For nonbilinear, convex-concave finite-sum problems, we improve the best-known complexity by a log factor and remove the boundedness assumption of the previous work, with simpler algorithms.

▷ We also show application of our techniques for solving monotone inclusions and strongly monotone problems. Our results for monotone inclusions potentially improve the rate of deterministic methods (depending on the Lipschitz constants).

▷ Our linear rate of convergence for strongly convex problems do not require parameters depending on the strong convexity constants.

▷ Recent work [HXZ21] proved matching lower bounds for the problem class we consider.

6.1.2 Related works

Most of the research in variance reduction has focused on convex minimization [JZ13, DBLJ14, KHR20, HLLJM15], leading to efficient methods in both theory and practice. On the other hand, variance reduction for solving VIs is started to be investigated recently. One common technique for reducing the variance in stochastic VIs, is to use increasing mini-batch sizes, which leads to high per iteration costs and slower convergence rates in practice [IJOT17, BMSV19, CS21].

A different approach used in [MKS$^+$20] was to use the same sample in both steps of stochastic extragradient method [JNT11] to reduce the variance, which results in a slower $O(1/\sqrt{k})$ rate. The results of [MKS$^+$20] for bilinear problems on the other hand are limited to the case when the matrix is full rank. The most related to our work, in the sense how variance reduction is used, are [BB16, CJST19, CGFLJ19] (see Table 2.1).

For the specific case of strongly monotone operators, [BB16] proposed algorithms based on SVRG and SAGA, with linear convergence rates. Two major questions for future work are posed in [BB16]: *(i)* obtaining convergence without strong monotonicity assumption and *(ii)* proving linear convergence without using strong monotonicity constant in the algorithm as a parameter.

The work by [CGFLJ19] proposed an algorithm based on extragradient method [Kor76] and under strong monotonicity assumption, proved linear convergence of the method. The step size in this work depends on cocoercivity constant, which might depend on strong monotonicity constant as discussed in [CGFLJ19, Table 1]. Thus, the result of [CGFLJ19] gave a partial answer to the second question of [BB16] while leaving the first one unanswered.

An elegant recent work of [CJST19] focused on matrix games and proposed a method based on the mirror prox [Nem04]. The extension of the method of [CJST19] for general min-max problems is also considered there. Unfortunately, this extension not only features a three loop structure, but also uses the bounded domain assumption actively and requires domain diameter as a parameter in the algorithm [CJST19, Corollary 2]. This result has been an important step towards an answer for the first question of [BB16].

Finally, this chapter answers an open problem posed in [MT20b] regarding a stochastic extensions of the forward-reflected-backward method. Our result improves the preliminary result in [MT20b, Section 6], which still requires evaluating the full operator every iteration.

6.1.3 Preliminaries and notation

We work in Euclidean space $\mathcal{Z} = \mathbb{R}^d$ with scalar product $\langle \cdot, \cdot \rangle$ and induced norm $\|\cdot\|$. Domain of a function $g\colon \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ is defined as $\operatorname{dom} g = \{z \in \mathcal{Z}\colon\ g(z) < +\infty\}$. Proximal operator of g is defined as

$$\operatorname{prox}_g(u) = \operatorname*{argmin}_{z \in \mathcal{Z}} \Big\{ g(z) + \frac{1}{2}\|z-u\|^2 \Big\}.$$

We call an operator $F\colon \mathcal{K} \to \mathcal{Z}$, where $\mathcal{K} \subseteq \mathcal{Z}$,

- L-Lipschitz, for $L > 0$, if $\|F(u) - F(v)\| \le L\|u - v\|, \quad \forall u, v \in \mathcal{K}$.
- monotone, if $\langle F(u) - F(v), u - v \rangle \ge 0, \quad \forall u, v \in \mathcal{K}$.
- ν-cocoercive, for $\nu > 0$, if $\langle F(u) - F(v), u - v \rangle \ge \nu\|F(u) - F(v)\|^2, \quad \forall u, v \in \mathcal{K}$.
- μ-strongly monotone, for $\mu > 0$, if $\langle F(u) - F(v), u - v \rangle \ge \mu\|u - v\|^2$, $\forall u, v \in \mathcal{K}$.

For example, in the context of (6.2) and (6.1), F is (strongly) monotone when Ψ is (strongly) convex- (strongly) concave. However, it is worth noting that both cocoercivity and strong monotonicity fail even for the simple bilinear case when $\Psi(x, y) = \langle Ax, y\rangle$ in (6.2).

Given iterates $\{z_k\}_{k\geq 1}$, $\{w_k\}_{k\geq 1}$ and the filtration $\mathcal{F}_k = \sigma\{z_1,\dots,z_k, w_1,\dots,w_{k-1}\}$, we define $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_k]$ as the conditional expectations with respect to $\mathcal{F}_k$.

Finally, we state our common assumptions for (6.1).

Assumption 6.1.

(a) $g\colon \mathcal{Z} \to \mathbb{R}\cup\{+\infty\}$ is proper lower semicontinuous convex.

(b) $F\colon \operatorname{dom} g \to \mathcal{Z}$ is monotone.

(c) $F = \frac{1}{n}\sum_{i=1}^{n} F_i$, with average Lipschitzness: $\mathbb{E}\|F_\xi(x) - F_\xi(y)\|^2 \leq L^2\|x-y\|^2$

(d) The solution set of (6.1), denoted by $\mathcal{Z}_\star$, is nonempty.

6.2 Algorithm

Our algorithm is a careful mixture of a recent deterministic algorithm for VIs, proposed by [MT20b], with a special technique of using variance reduction in finite sum minimization given in [HLLJM15] and [KHR20].

It is clear that for $n = 1$ any stochastic variance reduced algorithm for VI reduces to some deterministic one. As a consequence, this immediately rules out the most obvious choice — the well-known *forward-backward* method (FB)

$$z_{k+1} = \operatorname{prox}_{\tau g}(z_k - \tau F(z_k)), \tag{6.3}$$

since its convergence requires either strong monotonicity or cocoercivity of F. The classical algorithms that work under mere monotonicity [Kor76, Pop80, Tse00] have a more complicated structure, and thus, it is not clear how to meld them with a variance reduction technique for finite sum problems. Instead, we chose the recent *forward-reflected-backward* method (FoRB) [MT20b]

$$z_{k+1} = \operatorname{prox}_{\tau g}(z_k - \tau(2F(z_k) - F(z_{k-1}))), \tag{6.4}$$

which converges under Assumption 2.2 with $n = 1$.

When $g = 0$, this method takes its origin in the Popov's algorithm [Pop80]. In this specific case, FoRB is also equivalent to optimistic gradient ascent algorithm [DISZ18, RS13] which became increasingly popular in machine learning literature recently [DISZ18, DP18, MOP20, MLZ+19].

Among many variance reduced methods for solving finite sum problems $\min_z f(z) := \frac{1}{n}\sum_{i=1}^{n} f_i(z)$

Algorithm 6.1 Variance reduced forward-reflected-backward (VR-FoRB)

```
Input: Probability $p \in (0,1]$, step size $\tau = \frac{p}{4L}$. Let $z_0 = w_0 = z_{-1} = w_{-1} \in \mathcal{Z}$
for $k = 0, 1 \dots$ do
    Draw an index $i_k \in \{1, \dots, n\}$ uniformly at random
    $z_{k+1} = \operatorname{prox}_{\tau g}(z_k - \tau F(w_k) - \tau(F_{i_k}(z_k) - F_{i_k}(w_{k-1})))$
    $w_{k+1} = \begin{cases} z_{k+1}, & \text{with probability } p \\ w_k, & \text{with probability } 1-p \end{cases}$
end for
```

one of the simplest is the Loopless-SVRG method [KHR20] (see also [HLLJM15]),

$$
\begin{aligned}
z_{k+1} &= z_k - \tau \nabla f(w_k) - \tau(\nabla f_{i_k}(z_k) - \nabla f_{i_k}(w_k)) \\
w_{k+1} &= \begin{cases} z_k, & \text{with probability } p, \\ w_k, & \text{with probability } 1-p, \end{cases}
\end{aligned}
$$

which can be seen as a randomized version of the gradient and hence forward-backward methods. The latter is the exact reason why we cannot extend this method directly to the variational inequality setting, without cocoercivity or strong monotonicity.

An accurate blending of [MT20b] and [KHR20], described above, results in Algorithm 6.1. Compared to Loopless-SVRG, the last evaluation of the operator at step 4 of Algorithm 6.1 is done at w_{k-1}, instead of w_k. In the deterministic case when $n = 1$ or $p = 1$, this modification reduces the method to FoRB (6.4) and not FB (6.3). The other change is that we use the most recent iterate z_{k+1} in the update of w_{k+1}, instead of z_k in the Loopless-SVRG. Surprisingly, these two small distinctions result in the method which converges for general VIs without the restrictive assumptions of the previous works.

We note that we use uniform sampling for choosing i_k in Algorithm 6.1 for simplicity. Our arguments directly extend to arbitrary sampling as in [BB16, CJST19] which is used for obtaining tighter Lipschitz constants.

6.3 Convergence analysis

We start with a key lemma that appeared in [MT20b] for analyzing a general class of VI methods. The proof of this lemma is given in the appendix for completeness. The only change from [MT20b] is that we consider the proximal operator, instead of a more general resolvent.

Lemma 6.1. *[MT20b, Proposition 2.3] Let $g\colon \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ be proper lower semicontinuous convex and let $x_1, U_0, U_1, V_1 \in \mathcal{Z}$ be arbitrary points. Define x_2 as*

$$x_2 = \operatorname{prox}_g(x_1 - U_1 - (V_1 - U_0)). \tag{6.5}$$

Then for all $x \in \mathcal{Z}$ and $V_2 \in \mathcal{Z}$, it holds

$$\begin{aligned}&\|x_2 - x\|^2 + 2\langle V_2 - U_1, x - x_2\rangle + 2\langle V_2, x_2 - x\rangle + 2g(x_2) - 2g(x) \\ &\le \|x_1 - x\|^2 + 2\langle V_1 - U_0, x - x_1\rangle + 2\langle V_1 - U_0, x_1 - x_2\rangle - \|x_1 - x_2\|^2. \end{aligned} \tag{6.6}$$

The benefit of Lemma 6.1 is that it gives a candidate for a Lyapunov function that can be used to prove convergence. We will need a slight modification in this function due to randomization in Algorithm 6.1.

6.3.1 Convergence of the iterates

We start by proving the almost sure convergence of the iterates. Such a result states that the trajectories of the iterates generated by our algorithm converge to a point in the solution set. This type of result is the analogue of sequential convergence results for deterministic methods [MT20b].

For the iterates $\{z_k\}$, $\{w_k\}$ of Algorithm 6.1 and any $z \in \operatorname{dom} g$, $\beta > 0$ we define

$$\begin{aligned}\Phi_{k+1}(z) &:= \|z_{k+1} - z\|^2 + 2\tau\langle F(z_{k+1}) - F(w_k), z - z_{k+1}\rangle + \frac{\beta}{2}\|z_k - w_k\|^2 + \frac{1}{2}\|z_{k+1} - z_k\|^2 \\ \Theta_{k+1}(z) &:= \langle F(z_{k+1}), z_{k+1} - z\rangle + g(z_{k+1}) - g(z).\end{aligned}$$

The first equation plays the role of a Lyapunov function and the second is essential for the rate.

Lemma 6.2. *Let Assumption 6.1 hold and F_ξ be Lipschitz for all ξ, $\tau < \frac{1-\sqrt{1-p}}{2L}$, $\beta = \frac{1}{\sqrt{1-p}} - 1$, and the iterates $\{z_k\}$ are generated by Algorithm 6.1. Then for any $z \in \operatorname{dom} g$,*

$$\mathbb{E}_k[\Phi_{k+1}(z) + 2\tau\Theta_{k+1}(z)] \le \Phi_k(z). \tag{6.7}$$

This lemma is essential in establishing the convergence of iterates and sublinear convergence rates that we will derive in the next section. We now continue with the proof.

Proof. We set in Lemma 6.1 $U_0 = \tau F_i(w_{k-1})$, $U_1 = \tau F(w_k)$, $V_1 = \tau F_i(z_k)$, $V_2 = \tau F(z_{k+1})$, and $x_1 = z_k$, with $i_k = i$. Then by (6.5) and step 4 of Algorithm 6.1, $x_2 = z_{k+1}$, thus, by (6.6)

$$\begin{aligned}\|z_{k+1} - z\|^2 &+ 2\tau\langle F(z_{k+1}) - F(w_k), z - z_{k+1}\rangle + 2\tau\big(\langle F(z_{k+1}), z_{k+1} - z\rangle \\ &+ g(z_{k+1}) - g(z)\big) \le \|z_k - z\|^2 + 2\tau\langle F_i(z_k) - F_i(w_{k-1}), z - z_k\rangle \\ &+ 2\tau\langle F_i(z_k) - F_i(w_{k-1}), z_k - z_{k+1}\rangle - \|z_{k+1} - z_k\|^2.\end{aligned} \tag{6.8}$$

First, note that by Lipschitzness of F_i, Cauchy-Schwarz and Young's inequalities,

$$2\tau\langle F_i(z_k) - F_i(w_{k-1}), z_k - z_{k+1}\rangle \le 2\tau^2 L^2\|z_k - w_{k-1}\|^2 + \frac{1}{2}\|z_k - z_{k+1}\|^2. \tag{6.9}$$

Thus, it follows that

$$\begin{aligned}&\|z_{k+1}-z\|^2+2\tau\langle F(z_{k+1})-F(w_k),z-z_{k+1}\rangle+\frac{1}{2}\|z_{k+1}-z_k\|^2+2\tau\Theta_{k+1}(z)\\&\le\|z_k-z\|^2+2\tau\langle F_i(z_k)-F_i(w_{k-1}),z-z_k\rangle+2\tau^2L^2\|z_k-w_{k-1}\|^2.\end{aligned}\tag{6.10}$$

Taking expectation conditioning on the knowledge of z_k, w_{k-1} and using that $\mathbb{E}_kF_i(z_k)=F(z_k)$, $\mathbb{E}_kF_i(w_{k-1})=F(w_{k-1})$, we obtain

$$\begin{aligned}&\mathbb{E}_k\|z_{k+1}-z\|^2+2\tau\mathbb{E}_k\langle F(z_{k+1})-F(w_k),z-z_{k+1}\rangle+\frac{1}{2}\mathbb{E}_k\|z_{k+1}-z_k\|^2\\&\qquad+2\tau\mathbb{E}_k\Theta_{k+1}(z)\le\|z_k-z\|^2+2\tau\langle F(z_k)-F(w_{k-1}),z-z_k\rangle+2\tau^2L^2\|z_k-w_{k-1}\|^2.\end{aligned}\tag{6.11}$$

Adding

$$\frac{\beta}{2}\mathbb{E}_k\|z_k-w_k\|^2=\frac{\beta(1-p)}{2}\|z_k-w_{k-1}\|^2,\tag{6.12}$$

which follows from the definition of w_k, to (6.11), we obtain

$$\begin{aligned}&\mathbb{E}_k[\Phi_{k+1}(z)+2\tau\Theta_{k+1}(z)]\le\Phi_k(z)\\&\qquad+\Big(2\tau^2L^2+\frac{\beta(1-p)}{2}\Big)\|z_k-w_{k-1}\|^2-\frac{1}{2}\|z_k-z_{k-1}\|^2-\frac{\beta}{2}\|z_{k-1}-w_{k-1}\|^2.\end{aligned}\tag{6.13}$$

The proof will be complete, if we can show that the expression in the second line is nonpositive. Due to our choice of β and τ this is a matter of a simple algebra. As $\beta+1=\frac{1}{\sqrt{1-p}}$, $\frac{\beta}{1+\beta}=1-\sqrt{1-p}$, and $2\tau L<1-\sqrt{1-p}=\frac{\beta}{1+\beta}$, we have

$$2\tau^2L^2+\frac{\beta(1-p)}{2}\le\frac{1}{2}\Big(\frac{\beta^2}{(1+\beta)^2}+\frac{\beta}{(1+\beta)^2}\Big)=\frac{\beta}{2(1+\beta)}.\tag{6.14}$$

Then we must show that

$$\frac{\beta}{1+\beta}\|z_k-w_{k-1}\|^2\le\|z_k-z_{k-1}\|^2+\beta\|z_{k-1}-w_{k-1}\|^2,$$

which is a direct consequence of $\|u+v\|^2\le(1+\frac{1}{\beta})\|u\|^2+(1+\beta)\|v\|^2$. The proof is complete. ■

Theorem 6.3. *Let Assumption 6.1 hold and let $\tau<\frac{1-\sqrt{1-p}}{2L}$. Then for the iterates $\{z_k\}$ of Algorithm 6.1, almost surely there exists $z_\star\in\mathcal{Z}_\star$ such that $z_k\to z_\star$.*

Remark 6.4. For $p=1$, i.e., when the algorithm becomes deterministic, the bound for the stepsize is $\tau<\frac{1}{2L}$, which coincides with the one in [MT20b] and is known to be tight. In this case analysis will be still valid if for convenience we assume that $\infty\cdot 0=0$.

For small p we might use a simpler bound for the stepsize, as the following corollary suggests.

Corollary 6.5. *Suppose that $p=\frac{1}{n}$ and $\tau\le\frac{p}{4L}=\frac{1}{4Ln}$. Then the statement of Theorem 6.3 holds.*

Proof. We only have to check that $\frac{p}{2} \le 1 - \sqrt{1-p}$, which follows from $\sqrt{1-p} \le 1 - \frac{p}{2}$. ■

6.3.2 Convergence rate for the general case

In this section, we prove that the average of the iterates of the algorithm exhibits $O(1/k)$ convergence rate which is optimal for solving monotone VIs [Nem04]. The standard quantity to show sublinear rates for VIs is gap function which is defined as

$$G(\bar{z}) = \sup_{z \in \mathcal{Z}} \langle F(z), \bar{z} - z \rangle + g(\bar{z}) - g(z).$$

As this quantity requires taking a supremum over the whole space $\mathcal{Z}$ which is potentially unbounded, restricted versions of gap functions are used, for example in [Nes07, Mal19]

$$G_{\mathcal{C}}(\bar{z}) = \sup_{z \in \mathcal{C}} \langle F(z), \bar{z} - z \rangle + g(\bar{z}) - g(z), \tag{6.15}$$

where $\mathcal{C} \subseteq \operatorname{dom} g$ is an arbitrary bounded set. It is known that $G_{\mathcal{C}}(\bar{z})$ is a valid merit function, as proven by [Nes07, Lemma 1]. As we are concerned with randomized algorithms, we derive the rate of convergence for the expected gap function $\mathbb{E}\left[G_{\mathcal{C}}(z_k)\right]$.

Theorem 6.6. *Given $\{z_k\}$ generated by Algorithm 6.1, we define the averaged iterate $z_K^{av} = \frac{1}{K}\sum_{k=1}^{K} z_k$. Let $\mathcal{C} \subset \operatorname{dom} g$ be an arbitrary bounded set. Then under the hypotheses of Theorem 6.3 it holds that*

$$\mathbb{E}\left[G_{\mathcal{C}}(z_K^{av})\right] \le \frac{1}{K}\left[\frac{1}{\tau}\sup_{z \in \mathcal{C}} \|z_0 - z\|^2 + \frac{2\tau L^2(1+\beta)}{\delta\beta}\operatorname{dist}(z_0, \mathcal{Z}_\star)^2\right],$$

where $\delta = \frac{\beta}{1+\beta} - \frac{4\tau^2 L^2(1+\beta)}{\beta}$.

Remark 6.7. If we set $p = \frac{1}{n}$, $\tau = \frac{p}{3\sqrt{2}L}$, and $\beta = \frac{1}{\sqrt{1-p}} - 1$, the rate will be bounded by $\frac{nL}{K}\left(3\sqrt{2}\sup_{z \in \mathcal{C}} \|z_0 - z\|^2 + 12\sqrt{2}\operatorname{dist}(z_0, \mathcal{Z}_\star)^2\right)$, hence it is $O(\frac{nL}{K})$.

The high level idea of the proof is that on top of Lemma 6.2 we sum the resulting inequality and accumulate terms $\Theta_k(z)$. Then we use Jensen's inequality to obtain the result.

There are two intricate points in these kind of results. First, the convergence measure is the expected duality gap $\mathbb{E}[G_{\mathcal{C}}(z_K^{av})]$ that includes the expectation of the supremum. In a standard analysis, it is easy to obtain a bound for the supremum of expectation, however obtaining the former requires a technique, which is common in the literature for saddle point problems [NJLS09, AFC21]. Roughly, the idea is to use an auxiliary iterate to characterize the difference two quantities, and show that the error term does not degrade the rate.

Second, as duality gap requires taking a supremum over the domain, the rate might contain a diameter term as in [CJST19]. The standard way to adjust this result for unbounded domains is to utilize a restricted merit function as in (6.15) on which the rate is obtained [Nes07]. We note that the result in [CJST19] not only involves the domain diameter in the final bound, but it also requires the domain diameter as a parameter for the algorithm in the general monotone case [CJST19, Corollary 2].

It is worth mentioning that even though our method is simple and the convergence rate is $O(1/k)$ as in [CJST19], our complexity result has a worse dependence on n, compared to [CJST19]. In particular, our complexity is $O(n/\epsilon)$ instead of the $O(\sqrt{n}/\epsilon)$ of [CJST19]. This is because our step size has the factor of p which is of the order $\frac{1}{n}$ in general and it appears to be tight based on numerical experiments. We see in Section 6.4.3 one way to get over this issue and derive a method that works under our general assumptions and features favorable complexity guarantees as in [CJST19].

6.3.3 Convergence rate for strongly monotone case

We show that linear convergence is attained when strong monotonicity is assumed.

Theorem 6.8. *Let Assumption 6.1 hold and let F be μ-strongly monotone. Let $z_\star$ be the unique solution of* (6.1)*. Then for Algorithm 6.1 with $\tau = \frac{p}{4\sqrt{2}L}$, it holds that*

$$\mathbb{E}\|z_k - z_\star\|^2 \le \left(1 - \frac{\mu p}{8\sqrt{2}L}\right)^k \|z_0 - z_\star\|^2. \tag{6.16}$$

Remark 6.9. We analyzed the case when F is strongly monotone, however, the same analysis would go through when F is monotone and g is strongly convex. One can *transfer* strong convexity of g to make F strongly monotone.

A key characteristic of our result is that strong monotonicity constant is not required in the algorithm as a parameter to obtain the rate. This has been raised as an open question by [BB16] and a partial answer is studied by [CGFLJ19] (see Table 2.1). Our result gives a full answer to this question without using strong monotonicity constant in all cases.

We next discuss the dependence of μ in the convergence rate. Our rate has a dependence of $\frac{1}{\mu}$ compared to $\frac{1}{\mu^2}$ of non-accelerated methods of [BB16] and the method of [CGFLJ19]. This difference is important especially when μ is small. On the other hand, in terms of n, our complexity has a worse dependence compared to [CJST19] and accelerated method of [BB16] as discussed before. Using the analysis we have in Section 6.4.3, one can improve this complexity.

6.3.4 Beyond monotonicity

Lastly, we illustrate that our method has convergence guarantees for a class of non-monotone problems. There exist several relaxations of monotonicity that are used in the literature [DL15a, MLZ$^+$19, IJOT17, Mal19]. Among these, we assume the existence of the solutions to Minty variational inequality given as

$$\exists \hat{z} \in \mathcal{Z}: \quad \langle F(z), z - \hat{z} \rangle + g(z) - g(\hat{z}) \geq 0, \quad \forall z \in \mathcal{Z}. \tag{6.17}$$

Under (6.17), we can drop the monotonicity assumption and show almost sure subsequential convergence of the iterates of our method. Naturally, in this case one can no longer show sequential convergence as with monotonicity (see Theorem 6.3).

Theorem 6.10. *Suppose that Assumption 6.1 (a), (c), (d) and the condition* (6.17) *hold. Then almost surely all cluster points of the sequence* $\{z_k\}$ *generated by Algorithm 6.1 are in* $\mathcal{Z}_\star$.

Proof. We will proceed as in Theorem 6.3 and [Mal19, Theorem 6]. We note that Lemma 6.2 does not use monotonicity of F, thus its follows in this case. In the inequality

$$\mathbb{E}_k[\Phi_{k+1}(z) + 2\tau\Theta_{k+1}(z)] \leq \Phi_k(z).$$

we plug in $z = \hat{z}$ for a point satisfying (6.17).

Then, by (6.17), we have

$$\Theta_{k+1}(\hat{z}) = \langle F(z_{k+1}), z_{k+1} - \hat{z} \rangle + g(z_{k+1}) - g(\hat{z}) \geq 0.$$

We then argue the same way as in Theorem 6.3 to conclude that almost surely, $\{z_k\}$ is bounded and cluster points of $\{z_k\}$ are in $\mathcal{Z}_\star$.

Note that the steps in Theorem 6.3 for showing sequential convergence relies on the choice of z as an arbitrary point in $\mathcal{Z}_\star$, which is not the case here, therefore, we can only use the arguments from Theorem 6.3 for showing subsequential convergence. ■

6.4 Extensions

We illustrate extensions of our results to monotone inclusions and Bregman projections. We also show how to improve the complexity bounds derived in the previous section. The proofs for this section are given in the appendix in Section 6.6.

6.4.1 Monotone inclusions

We choose to focus on monotone VIs in the main part of the chapter for being able to derive sublinear rates for the gap function. In this section, we show that our analysis extends directly for solving monotone inclusions. In this case, we are interested in finding z such that $0 \in (A + F)(z)$, where A, F are monotone operators and each F_i is Lipschitz with the form $F = \frac{1}{n}\sum_{i=1}^n F_i$.

In this case, one changes the prox operator in the algorithm, to resolvent operator of A which is defined as $J_{\tau A}(z) = (I + \tau A)^{-1}(z)$. Then, one can use Lemma 6.1 as directly given in [MT20b, Proposition 2.3] to prove an analogous result of Theorem 6.3 for solving monotone inclusions. Moreover, when $A+F$ is strongly monotone, one can prove an analogue of Theorem 6.8. We prove the former result and we note that the latter can be shown by applying the steps in Theorem 6.8 on top of Theorem 6.11, which we do not repeat for brevity.

Theorem 6.11. *Let $A\colon \mathcal{Z} \rightrightarrows \mathcal{Z}$ be maximally monotone and $F\colon \mathcal{Z} \to \mathcal{Z}$ be monotone with $F = \frac{1}{n}\sum_{i=1}^n F_i$, where F_i is L-Lipschitz for all i. Assume that $(A+F)^{-1}(0)$ is nonempty and let the iterates $\{z_k\}$ be generated by Algorithm 6.1 with the update for z_{k+1}*

$$z_{k+1} = J_{\tau A}(z_k - \tau F(w_k) - \tau(F_{i_k}(z_k) - F_{i_k}(w_{k-1}))). \tag{6.18}$$

Then, for $\tau < \frac{1-\sqrt{1-p}}{2L}$, almost surely there exist $z_\star \in (A+F)^{-1}(0)$ such that $z_k \to z_\star$.

6.4.2 Bregman distances

We developed our analysis in the Euclidean setting, relying on ℓ_2-norm for simplicity. However, we can also generalize it to proximal operators involving Bregman distances. In this setting, we have a distance generating function $h\colon \mathcal{Z} \to \mathbb{R}$, which is 1-strongly convex and continuous. We follow the standard convention to assume that subdifferential of h admits a continuous selection, which means that there exists a continuous function ∇h such that $\nabla h(x) \in \partial h(x)$ for all $x \in \operatorname{dom} \partial h$. We define the Bregman distance as $D(z,\bar{z}) = h(z) - h(\bar{z}) - \langle \nabla h(\bar{z}), z - \bar{z}\rangle$. Then, we will change the proximal step 4 of Algorithm 6.1 with

$$z_{k+1} = \operatorname*{argmin}_z \Big\{ g(z) + \langle F(w_k) + F_{i_k}(z_k) - F_{i_k}(w_{k-1}), z - z_k\rangle + \frac{1}{\tau} D(z, z_k) \Big\}. \tag{6.19}$$

We prove an analogue of Lemma 6.2 with Bregman distances from which the convergence rate results will follow.

Lemma 6.12. *Let Assumption 6.1 hold and*

$$\Phi_{k+1}(z) := D(z, z_{k+1}) + \tau\langle F(z_{k+1}) - F(w_k), z - z_{k+1}\rangle + \frac{\beta}{4}\|z_k - w_k\|^2 + \frac{1}{2} D(z_{k+1}, z_k).$$

Moreover, suppose $\tau < \frac{1-\sqrt{1-p}}{2L}$, $\beta = \frac{1}{\sqrt{1-p}} - 1$, and the iterates $\{z_k\}$ are generated by Algorithm 6.1 with the update (6.19) *for z_{k+1}. Then for any $z \in \operatorname{dom} g$,*

$$\mathbb{E}_k[\Phi_{k+1}(z) + \tau\Theta_{k+1}(z)] \le \Phi_k(z).$$

6.4.3 Improving complexity

We introduce a variance reduced extragradient algorithm in Algorithm 6.2, building on the analysis techniques presented earlier in the chapter. This algorithm is able to use a bigger step size that carefully balances the complexity bounds. For running algorithm in practice, we suggest $p = \frac{2}{N}$, $\alpha = 1-p$, and $\tau = \frac{0.99\sqrt{p}}{L}$. However, specific problem may require a more careful

Algorithm 6.2 Extragradient with variance reduction

```
Input: Probability $p \in (0,1]$, probability distribution $Q$, step size $\tau$, $\alpha \in (0,1)$. Let $z_0 = w_0$
for $k = 0, 1, \dots$ do
    $\bar{z}_k = \alpha z_k + (1-\alpha) w_k$
    $z_{k+1/2} = \operatorname{prox}_{\tau g}(\bar{z}_k - \tau F(w_k))$
    Draw an index $\xi_k$ according to $Q$
    $z_{k+1} = \operatorname{prox}_{\tau g}(\bar{z}_k - \tau[F(w_k) + F_{\xi_k}(z_{k+1/2}) - F_{\xi_k}(w_k)])$
    $w_{k+1} = \begin{cases} z_{k+1}, & \text{with probability } p \\ w_k, & \text{with probability } 1-p \end{cases}$
end for
```

selection. As before, by eliminating all randomness, Algorithm 6.2 reduces to the classical extragradient method in [Kor76, Nem04].

Analysis for the Euclidean case

For the iterates (z_k), (w_k) of Algorithm 6.2 and any $z \in \operatorname{dom} g$, we define

$$\Phi_k(z) := \alpha \|z_k - z\|^2 + \frac{1-\alpha}{p} \|w_k - z\|^2.$$

Lemma 6.13. *Let Assumption 6.1 hold, $\alpha \in [0,1)$, $p \in (0,1]$, and $\tau = \frac{\sqrt{1-\alpha}}{L}\gamma$, for $\gamma \in (0,1)$. Then for (z_k) generated by Algorithm 6.2 and any $z_\star \in \mathsf{Sol}$, it holds that*

$$\mathbb{E}_k\left[\Phi_{k+1}(z_\star)\right] \le \Phi_k(z_\star) - (1-\gamma)\Big((1-\alpha)\|z_{k+1/2} - w_k\|^2 + \mathbb{E}_k\|z_{k+1} - z_{k+1/2}\|^2\Big).$$

Moreover, it holds that $\sum_{k=0}^{\infty}\Big((1-\alpha)\mathbb{E}\|z_{k+1/2} - w_k\|^2 + \mathbb{E}\|z_{k+1} - z_{k+1/2}\|^2\Big) \le \frac{1}{1-\gamma}\Phi_0(z_\star)$.

Proof. By convexity of g and the definitions of z_{k+1} and $z_{k+1/2}$, we have that for all z,

$$\begin{aligned} \langle z_{k+1} - \bar{z}_k + \tau[F(w_k) + F_{\xi_k}(z_{k+1/2}) - F_{\xi_k}(w_k)], z - z_{k+1}\rangle &\ge \tau g(z_{k+1}) - \tau g(z), \\ \langle z_{k+1/2} - \bar{z}_k + \tau F(w_k), z_{k+1} - z_{k+1/2}\rangle &\ge \tau g(z_{k+1/2}) - \tau g(z_{k+1}). \end{aligned} \tag{6.20}$$

We sum two inequalities and arrange to get

$$\begin{aligned} &\langle z_{k+1} - \bar{z}_k, z - z_{k+1}\rangle + \langle z_{k+1/2} - \bar{z}_k, z_{k+1} - z_{k+1/2}\rangle + \tau\langle F_{\xi_k}(w_k) - F_{\xi_k}(z_{k+1/2}), z_{k+1} - z_{k+1/2}\rangle \\ &\qquad + \tau\langle F(w_k) + F_{\xi_k}(z_{k+1/2}) - F_{\xi_k}(w_k), z - z_{k+1/2}\rangle \ge \tau[g(z_{k+1/2}) - g(z)]. \end{aligned} \tag{6.21}$$

For the first inner product we use definition of $\bar{z}_k$ and identity $2\langle a, b\rangle = \|a+b\|^2 - \|a\|^2 - \|b\|^2$

$$\begin{aligned} &2\langle z_{k+1} - \bar{z}_k, z - z_{k+1}\rangle = 2\alpha\langle z_{k+1} - z_k, z - z_{k+1}\rangle + 2(1-\alpha)\langle z_{k+1} - w_k, z - z_{k+1}\rangle \\ &\quad = \alpha\left(\|z_k - z\|^2 - \|z_{k+1} - z\|^2 - \|z_{k+1} - z_k\|^2\right) + (1-\alpha)\left(\|w_k - z\|^2 - \|z_{k+1} - z\|^2 - \|z_{k+1} - w_k\|^2\right) \end{aligned}$$

$$= \alpha\|z_k - z\|^2 - \|z_{k+1} - z\|^2 + (1-\alpha)\|w_k - z\|^2 - \alpha\|z_{k+1} - z_k\|^2 - (1-\alpha)\|z_{k+1} - w_k\|^2. \tag{6.22}$$

Similarly, for the second inner product in (6.21) we deduce

$$\begin{aligned} 2\langle z_{k+1/2} - \bar{z}_k, z_{k+1} - z_{k+1/2}\rangle = \alpha\|z_{k+1} - z_k\|^2 - \|z_{k+1} - z_{k+1/2}\|^2 + (1-\alpha)\|z_{k+1} - w_k\|^2 \\ - \alpha\|z_{k+1/2} - z_k\|^2 - (1-\alpha)\|z_{k+1/2} - w_k\|^2. \end{aligned} \tag{6.23}$$

For the remaining terms in (6.21), we plug in $z = z_\star$, use that $z_{k+1/2}, w_k$ is deterministic under the conditioning of $\mathbb{E}_k$ and $\mathbb{E}_k\left[F(w_k) + F_{\xi_k}(z_{k+1/2}) - F_{\xi_k}(w_k)\right] = F(z_{k+1/2})$ to obtain

$$\begin{aligned} &\mathbb{E}_k\left[\langle F(w_k) + F_{\xi_k}(z_{k+1/2}) - F_{\xi_k}(w_k), z_\star - z_{k+1/2}\rangle + g(z_\star) - g(z_{k+1/2})\right] \\ &\quad = \langle F(z_{k+1/2}), z_\star - z_{k+1/2}\rangle + g(z_\star) - g(z_{k+1/2}) \\ &\quad \le \langle F(z_\star), z_\star - z_{k+1/2}\rangle + g(z_\star) - g(z_{k+1/2}) \le 0 \end{aligned} \tag{6.24}$$

where the last step is due to monotonicity and definition of $z_\star$. Next, we estimate

$$\begin{aligned} \mathbb{E}_k\big[2\tau\langle F_{\xi_k}(w_k) - F_{\xi_k}(z_{k+1/2}), z_{k+1} - z_{k+1/2}\rangle\big] &\le \mathbb{E}_k\left[2\tau\|F_{\xi_k}(w_k) - F_{\xi_k}(z_{k+1/2})\|\|z_{k+1} - z_{k+1/2}\|\right] \\ &\le \frac{\tau^2}{\gamma}\mathbb{E}_k\left[\|F_{\xi_k}(z_{k+1/2}) - F_{\xi_k}(w_k)\|^2\right] + \gamma\mathbb{E}_k\left[\|z_{k+1} - z_{k+1/2}\|^2\right] \\ &\le (1-\alpha)\gamma\|z_{k+1/2} - w_k\|^2 + \gamma\mathbb{E}_k\left[\|z_{k+1} - z_{k+1/2}\|^2\right], \end{aligned} \tag{6.25}$$

by Cauchy-Scwarz, Young's inequalities and Lipschitzness. We use (6.22), (6.23), (6.24), and (6.25) in (6.21), after taking expectation $\mathbb{E}_k$ and letting $z = z_\star$, to deduce

$$\begin{aligned} \mathbb{E}_k\left[\|z_{k+1} - z_\star\|^2\right] \le \alpha\|z_k - z_\star\|^2 + (1-\alpha)\|w_k - z_\star\|^2 - (1-\alpha)(1-\gamma)\|z_{k+1/2} - w_k\|^2 \\ - \big(1-\gamma\big)\mathbb{E}_k\left[\|z_{k+1} - z_{k+1/2}\|^2\right]. \end{aligned} \tag{6.26}$$

By the definition of w_{k+1} and $\mathbb{E}_{k+1/2}$, it follows that

$$\frac{1-\alpha}{p}\mathbb{E}_{k+1/2}\left[\|w_{k+1} - z_\star\|^2\right] = (1-\alpha)\|z_{k+1} - z_\star\|^2 + (1-\alpha)\left(\frac{1}{p} - 1\right)\|w_k - z_\star\|^2. \tag{6.27}$$

We add (6.27) to (6.26) and apply the tower property $\mathbb{E}_k\left[\mathbb{E}_{k+1/2}[\cdot]\right] = \mathbb{E}_k[\cdot]$ to deduce

$$\begin{aligned} \alpha\mathbb{E}_k\left[\|z_{k+1} - z_\star\|^2\right] + \frac{1-\alpha}{p}\mathbb{E}_k\left[\|w_{k+1} - z_\star\|^2\right] \le \alpha\|z_k - z_\star\|^2 + \frac{1-\alpha}{p}\|w_k - z_\star\|^2 \\ - (1-\gamma)\Big((1-\alpha)\|z_{k+1/2} - w_k\|^2 + \mathbb{E}_k\left[\|z_{k+1} - z_{k+1/2}\|^2\right]\Big). \end{aligned}$$

Using the definition of $\Phi_k(z)$, we obtain the first result. Applyting total expectation and summing the inequality yields the second result. ■

Almost sure convergence of this method can be proven by using the same arguments as Theorem 6.3. For brevity, we omit the proof.

For the convergence rate, we use a similar technique as Theorem 6.6. We start with a simple lemma for "switching" the order of maximum and expectation, which is required for showing convergence of expected gap. Such a lemma is standard for such purpose [NJLS09].

Lemma 6.14. *Let $\mathcal{F} = (\mathcal{F}_k)_{k\geq 0}$ be a filtration and (u_k) a stochastic process adapted to $\mathcal{F}$ with $\mathbb{E}[u_{k+1}|\mathcal{F}_k] = 0$. Then for any $K \in \mathbb{N}$, $x_0 \in \mathcal{Z}$, and any compact set $\mathcal{C} \subset \mathcal{Z}$,*

$$\mathbb{E}\left[\max_{x\in\mathcal{C}} \sum_{k=0}^{K-1} \langle u_{k+1}, x\rangle\right] \leq \max_{x\in\mathcal{C}} \frac{1}{2}\|x_0 - x\|^2 + \frac{1}{2}\sum_{k=0}^{K-1} \mathbb{E}\|u_{k+1}\|^2.$$

Theorem 6.15. *Let Assumption 6.1 hold, $p \in (0,1]$, $\alpha = 1-p$, and $\tau = \frac{\sqrt{1-\alpha}}{L}\gamma$, for $\gamma \in (0,1)$. Then, for $z^K = \frac{1}{K}\sum_{k=0}^{K-1} z_{k+1/2}$, for the iterates of Algorithm 6.2, it follows that*

$$\mathbb{E}\left[\operatorname{Gap}(z^K)\right] = \mathcal{O}\left(\frac{L}{\sqrt{p}K}\right).$$

In particular, for $\tau = \frac{\sqrt{p}}{2L}$, the rate is $\mathbb{E}\left[\operatorname{Gap}(z^K)\right] \leq \frac{17.5L}{\sqrt{p}K}\max_{z\in\mathcal{C}}\|z_0 - z\|^2$.

Recall that we denote the cost of computing one $F_\xi(\cdot)$ as Cost, and the cost of computing $F(\cdot)$ as Cost × N. For a finite sum example, as in Section 7.2, this is the most natural assumption.

Corollary 6.16. *Let the conjecture of Theorem 6.15 hold. Then the average total complexity (see Remark 6.17) of Algorithm 6.2 to reach ε-accuracy is $\mathcal{O}\left(\mathsf{Cost} \times (pN+2)\left(1+\frac{L}{\sqrt{p}\varepsilon}\right)\right)$. In particular, for $p = \frac{2}{N}$ it is $\mathcal{O}\left(\mathsf{Cost}\times\left(N + \frac{\sqrt{N}L}{\varepsilon}\right)\right)$.*

Remark 6.17. For Algorithm 6.2, since per iteration cost is random, the result is "average" total complexity: *expected number of iterations to get a small expected gap.* On the other hand, Algorithm 6.3 has a fixed cost per iteration, thus, it gives a more standard notion of complexity: *number of iterations to get a small expected gap.*

Remark 6.18. To see why we let $\alpha = 1-p$, consider the proof with any choice of α. The resulting bound will be $\mathcal{O}\left(\frac{1}{\sqrt{1-\alpha}} + \frac{\sqrt{1-\alpha}}{p}\right)$. Then $\alpha = 1-p$ optimizes it in terms of p dependence.

Analysis for Bregman case

In this section, we use the same setup as Section 6.4.2, with primal-dual norm pair $\|\cdot\|$ and $\|\cdot\|_*$. We recall the three point identity which can be seen as the analogue of the standard Euclidean identity $2\langle a,b\rangle = \|a+b\|_2^2 - \|a\|_2^2 - \|b\|_2^2$:

$$\langle\nabla h(x) - \nabla h(y), z - x\rangle = D(z,y) - D(z,x) - D(x,y) \qquad \forall x,y,z \in \mathcal{Z}. \tag{6.28}$$

Note that since h is 1-strongly convex with respect to norm $\|\cdot\|$, we have $D(u,v) \geq \frac{1}{2}\|u-v\|^2$.

Naturally, we say that $F\colon \operatorname{dom} g \to \mathcal{Z}^*$ is L_F-Lipschitz, if $\|F(u) - F(v)\|_* \leq L_F\|u-v\|$, $\forall u, v$. However, Lipschitzness for a stochastic oracle this time is more involved. We prefer stochastic oracles F_ξ of F with as small L as possible. Moreover, the proof of Lemma 6.13 indicates that in k-th iteration we need Lipschitzness only for already known two iterates. Hence, following [GK95, CJST19], in contrast to Algorithm 6.2, we do not fix distribution Q in the beginning, but allow it to vary from iteration to iteration.

Definition 6.19. We say that F has a stochastic oracle F_ξ that is *variable* L-Lipschitz in mean, if for any $u, v \in \operatorname{dom} g$ there exists a distribution $Q_{u,v}$ such that

(i) F is unbiased: $F(z) = \mathbb{E}_{\xi \sim Q_{u,v}}\left[F_\xi(z)\right] \quad \forall z \in \operatorname{dom} g$;

(ii) $\mathbb{E}_{\xi \sim Q_{u,v}}\left[\|F_\xi(u) - F_\xi(v)\|_*^2\right] \le L^2\|u - v\|^2$.

Note that the second condition holds only for given u, v, but the constant L is universal for all u, v. Changing u, v also changes a distribution, hence the name "variable". Without loss of generality, we denote any distribution that realizes the above Lipschitz bound for given u, v by $Q_{u,v}$. This definition resembles the one in [CJST19, Definition 2]. It is easy to see when $Q_{u,v} = Q$ for all u, v, we get the same definition as before in Assumption 6.1.

For brevity we introduce the new set of assumptions. It is important to remark that Assumption 6.2 is not a restriction of Assumption 6.1: every item is either the same or more general.

Assumption 6.2.

(i) The solution set $\mathcal{Z}_\star$ is nonempty.

(ii) The function $g \in \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ is proper convex lsc.

(iii) The operator $F\colon \operatorname{dom} g \to \mathcal{Z}^*$ is monotone.

(iv) The operator F has a stochastic oracle F_ξ that is variable L-Lipschitz in mean (see Definition 6.19).

In this setting, we can simply adjust the steps of Algorithm 6.2 and correspondingly the analysis of Lemma 6.13, as in Section 6.4.2. However, to show a convergence rate, double randomization in Algorithm 6.2 causes technical complications. Therefore, in the Bregman setup we propose a double loop variant of Algorithm 6.2 (see Algorithm 6.3), similar to the classical SVRG [JZ13]. Our algorithm can be seen as a variant of Mirror-Prox [Nem04] with variance reduction.

Compared to Algorithm 6.2, w^s serves the same purpose as w_k: the snapshot point in the language of SVRG [JZ13]. Since we have two loops in this case, we get w^s by averaging, again, similar to SVRG for non-strongly convex optimization [RHS$^+$16, AZY16]. The difference due to Bregman setup is that we have the additional point $\bar{w}^s$ that averages in the dual space. This operation does not incur additional cost. For running algorithm in practice, we suggest $K = \frac{N}{2}$, $\alpha = 1 - \frac{1}{K}$, and $\tau = \frac{0.99\sqrt{p}}{L}$.

Algorithm 6.3 Mirror-prox with variance reduction

```
Input: Step size $\tau$, $\alpha \in (0,1)$, $K > 0$. Let $z_j^{-1} = z_0^0 = w^0 = z_0, \forall j \in [K]$
for $s = 0, 1 \ldots$ do
    for $k = 0, 1 \ldots K-1$ do
        $z^s_{k+1/2} = \operatorname{argmin}_z \Big\{ g(z) + \langle F(w^s), z\rangle + \frac{\alpha}{\tau} D(z, z_k^s) + \frac{1-\alpha}{\tau} D(z, \bar{w}^s) \Big\}$.
        Fix distribution $Q_{z^s_{k+1/2}, w^s}$ and sample $\xi_k^s$ according to it
        $z^s_{k+1} = \operatorname{argmin}_z \Big\{ g(z) + \langle F(w^s) + F_{\xi_k^s}(z^s_{k+1/2}) - F_{\xi_k^s}(w^s), z\rangle + \frac{\alpha}{\tau} D(z, z_k^s) + \frac{1-\alpha}{\tau} D(z, \bar{w}^s) \Big\}$.
    end for
    $w^{s+1} = \frac{1}{K}\sum_{k=1}^K z_k^s$
    $\nabla h(\bar{w}^{s+1}) = \frac{1}{K}\sum_{k=1}^K \nabla h(z_k^s)$
    $z_0^{s+1} = z_K^s$
end for
```

Similar to Euclidean case, we define for the iterates (z_k^s) of Algorithm 6.3 and any $z \in \operatorname{dom} g$,

$$\Phi^s(z) := \alpha D(z, z_0^s) + (1-\alpha)\sum_{j=1}^m D(z, z_j^{s-1}),$$

where $\Phi^0(z) = (\alpha + K(1-\alpha))D(z, z_0)$, due to the definition of z^{-1} from Algorithm 6.3. Since we have two indices s, k in Algorithm 6.3, we define $\mathcal{F}_k^s = \sigma(z^0_{1/2}, \ldots, z^0_{K-1/2}, \ldots, z^s_{1/2}, \ldots, z^s_{k+1/2})$ and $\mathbb{E}_{s,k}[\cdot] = \mathbb{E}\left[\cdot | \mathcal{F}_k^s\right]$. We now introduce some definitions to be used in the proofs of this section.

$$\Theta^s_{k+1/2}(z) = \langle F(z^s_{k+1/2}), z^s_{k+1/2} - z\rangle + g(z^s_{k+1/2}) - g(z), \tag{6.29}$$

$$e(z, s, k) = \tau\langle F(z^s_{k+1/2}) - F_{\xi_k^s}(z^s_{k+1/2}) - F(w^s) + F_{\xi_k^s}(w^s), z^s_{k+1/2} - z\rangle. \tag{6.30}$$

$$\delta(s,k) = \tau\langle F_{\xi_k^s}(w^s) - F_{\xi_k^s}(z^s_{k+1/2}), z^s_{k+1} - z^s_{k+1/2}\rangle - \frac{1}{2}\|z^s_{k+1} - z^s_{k+1/2}\|^2 - \frac{1-\alpha}{2}\|z^s_{k+1/2} - w^s\|^2 \tag{6.31}$$

The first expression is for deriving the rate, the second $e(z, s, k)$ for controlling the error caused by $\max_{z\in\mathcal{C}} \mathbb{E}[\cdot] \neq \mathbb{E}\max_{z\in\mathcal{C}}[\cdot]$, and the third term $\delta(s,k)$ is nonpositive after taking expectation.

Lemma 6.20. *Let Assumption 6.2 hold, $\alpha \in [0,1)$, and $\tau = \frac{\sqrt{1-\alpha}}{L}\gamma$ for $\gamma \in (0,1)$. We have:*

(i) For any $z \in \mathcal{Z}$ and $s, K \in \mathbb{N}$, it holds that

$$\sum_{k=0}^{K-1} \tau\Theta^s_{k+1/2}(z) + \alpha D(z, z_0^{s+1}) + (1-\alpha)\sum_{j=1}^K D(z, z_j^s)$$
$$\leq \alpha D(z, z_0^s) + (1-\alpha)\sum_{j=1}^K D(z, z_j^{s-1}) + \sum_{k=0}^{K-1}[e(z,s,k) + \delta(s,k)].$$

(ii) For any solution $z_\star$, it holds that

$$\mathbb{E}_{s,0}\left[\Phi^{s+1}(z_\star)\right] \le \Phi^s(z_\star) - \frac{(1-\alpha)(1-\gamma^2)}{2}\sum_{k=0}^{K-1}\mathbb{E}_{s,0}\left[\|z^s_{k+1/2}-w^s\|^2\right].$$

(iii) It holds that $\sum_{s=0}^{\infty}\sum_{k=0}^{K-1}\mathbb{E}\|z^s_{k+1/2}-w^s\|^2 \le \frac{2}{(1-\alpha)(1-\gamma^2)}\Phi^0(z_\star)$.

In order to prove the convergence rate, we need the Bregman version of Lemma 6.14.

Lemma 6.21. *Let* $\mathcal{F} = (\mathcal{F}^s_k)_{s\ge 0, k\in[0,K-1]}$ *be a filtration and* (u^s_k) *a stochastic process adapted to* $\mathcal{F}$ *with* $\mathbb{E}[u^s_{k+1}|\mathcal{F}^s_k] = 0$. *Given* $x_0 \in \mathcal{Z}$, *for any* $S \in \mathbb{N}$ *and any compact set* $\mathcal{C} \subset \operatorname{dom} g$

$$\mathbb{E}\left[\max_{x\in\mathcal{C}}\sum_{s=0}^{S-1}\sum_{k=0}^{K-1}\langle u^s_{k+1}, x\rangle\right] \le \max_{x\in\mathcal{C}} D(x,x_0) + \frac{1}{2}\sum_{s=0}^{S-1}\sum_{k=0}^{K-1}\mathbb{E}\|u^s_{k+1}\|^2_*.$$

Theorem 6.22. *Let Assumption 6.2 hold,* $\alpha \in [0,1)$, *and* $\tau = \frac{\sqrt{1-\alpha}}{L}\gamma$ *for* $\gamma \in (0,1)$. *Then, for* $z^S = \frac{1}{KS}\sum_{s=0}^{S-1}\sum_{k=0}^{K-1} z^s_{k+1/2}$, *with the iterates of Algorithm 6.3, it follows that*

$$\mathbb{E}\left[\operatorname{Gap}(z^S)\right] \le \frac{1}{\tau KS}\left(1 + \left(1+\frac{8\gamma^2}{1-\gamma^2}\right)(\alpha + K(1-\alpha)\right)\max_{z\in\mathcal{C}} D(z,z_0).$$

Corollary 6.23. *Let* $K = \frac{N}{2}$ *and* $\alpha = 1-\frac{1}{K} = 1-\frac{2}{N}$, *and* $\tau = \frac{\sqrt{1-\alpha}}{L}\gamma$ *for* $\gamma \in (0,1)$. *Then the total complexity of Algorithm 6.3 to reach* ε*-accuracy is* $\mathcal{O}\left(\mathsf{Cost}\times\left(N + \frac{L\sqrt{N}}{\varepsilon}\right)\right)$. *In particular, if* $\tau = \frac{\sqrt{1-\alpha}}{3L} = \frac{\sqrt{2}}{3\sqrt{NL}}$, *the total complexity is* $\mathsf{Cost}\times\left(2N + \frac{43\sqrt{NL}}{\varepsilon}\max_{z\in\mathcal{C}} D(z,z_0)\right)$.

6.5 Numerical verification

In this section, we include preliminary experimental results for our algorithms. We would like to note that these results are mainly for verifying our theoretical results and are not intended to serve as complete benchmarks. More experimental results can also be found in [AM21].

First, we apply Algorithm 6.1 to the *unconstrained* bilinear problem. It was shown in [CGFLJ19] that this simple problem is particularly challenging for stochastic methods, due to unboundedness of the domain, where the standard methods, such as stochastic extragradient method [JNT11], diverges. Our assumptions are general enough to cover this case and we now verify in practice that our method indeed converges for this problem by setting $d = n = 100$ and generating $A_i \in \mathbb{R}^{d\times d}$ randomly with distribution $\mathcal{N}(0,1)$

$$\min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^d}\frac{1}{n}\sum_{i=1}^{n}\langle A_i x, y\rangle. \tag{6.32}$$

For this experiment, we test the tightness of our step size rule by progressively increasing it. Recall that our step size is $\tau = \frac{p}{cL}$, where $c = 4$ is suggested in our analysis, see Corollary 6.5. We try the values of $c = \{0.5, 1, 2, 4\}$ and observe that for $c = 0.5$ the algorithm diverges, see the

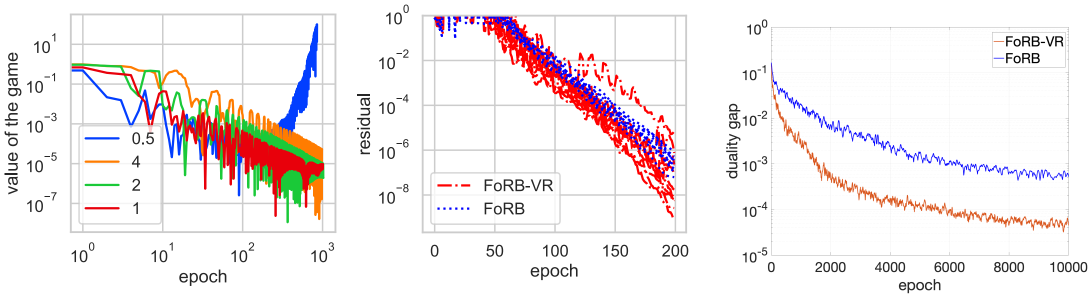

Figure 6.1 – Left: bilinear problem. Middle: Constrained minimization with data generated by normal distribution. Right: Constrained minimization with data generated by uniform distribution.

first plot in Figure 6.1. The message of this experiment is that even though slightly higher step sizes than what is allowed in our theory might work, it is not possible to significantly increase it.

The second problem we consider is constrained minimization, which is an instance where the dual domain is not necessarily bounded. We want to solve

$$\min_{x\in C} f(x) \quad \text{s.t.} \quad h_i(x) \leq 0, \quad i = 1,\dots,m,$$

where $f(x) = \frac{1}{2}\|x-u\|^2$ for some $u \in \mathcal{Z}$ and $h_i(x) = \|A_i x - b_i\|^2 - \delta_i$ for $A_i \in \mathbb{R}^{d\times d}$, $b_i \in \mathbb{R}^d$, $\delta_i \in \mathbb{R}_{++}$, and C is a unit ball. In other words, we want to find a projection of u onto the intersection given by C and the constraint inequalities $\{x\colon h_i(x) \leq 0\}$. Introducing Lagrange multipliers y_i for each constraint, we obtain (see Section 5.7 for further details)

$$\min_{x\in\mathbb{R}^d} \max_{y\in\mathbb{R}^m_+} f(x) + \sum_{i=1}^m y_i h_i(x).$$

As the Lipschitz constant in this problem does not admit a closed-form expression, we first estimate the Lipschitz constant by finding an L such that deterministic method converges. Next, we note that even though we analyzed Algorithm 6.1 with a single step size τ for both primal and dual variables x, y, one can use different step sizes for primal and dual variables (see [Mal19, Section 4.1]). Therefore, we tuned the scaling of primal and dual step sizes for both methods with one random instance and we used the same scaling for all tests for both methods.

We set $p = 1/m$. Every iteration, the deterministic method needs to go through all m constraints to compute $\sum_{i=1}^m y_i \nabla h_i(x)$, whereas our method computes $\nabla h_i(x)$ for only one i. The setup is with $m = 400$, $d = 100$, and the data is generated with the normal distribution $\mathcal{N}(0,1)$. We ran 10 different instances of randomly generated data and plotted all results, see the second plot in Figure 6.1. We observe that practical performance is similar with Algorithm 6.1 and deterministic methods.

In the third plot of Figure 6.1, we implement Algorithm 6.2 for solving simplex constrained matrix games and compare with deterministic algorithms. We see that as predicted by the theory of Algorithm 6.2, the practical performance improves that of deterministic methods.

6.6 Proofs

6.6.1 Proofs for Section 6.3

Proof of Lemma 6.1. By using the definition of x_2 and convexity of g, we have for all $x \in \mathcal{Z}$

$$\begin{aligned} g(x) &\geq g(x_2) + \langle x_1 - U_1 - (V_1 - U_0) - x_2, x - x_2 \rangle \\ &= g(x_2) + \langle x_1 - x_2, x - x_2 \rangle - \langle U_1, x - x_2 \rangle - \langle V_1 - U_0, x - x_2 \rangle. \end{aligned} \tag{6.33}$$

Since $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a-b\|^2$, $\forall a, b$, it follows that

$$2\langle x_1 - x_2, x - x_2 \rangle = \|x_1 - x_2\|^2 + \|x - x_2\|^2 - \|x - x_1\|^2.$$

Simple rearrangements give

$$-\langle U_1, x - x_2 \rangle = \langle V_2 - U_1, x - x_2 \rangle - \langle V_2, x - x_2 \rangle$$

and

$$-\langle V_1 - U_0, x - x_2 \rangle = -\langle V_1 - U_0, x - x_1 \rangle - \langle V_1 - U_0, x_1 - x_2 \rangle.$$

Using the last three equalities in (6.33) completes the result. ■

Proof of Theorem 6.3. From Lemma 6.2 we have for any $z \in \operatorname{dom} g$

$$\mathbb{E}_k[\Phi_{k+1}(z) + 2\tau\Theta_{k+1}(z)] \leq \Phi_k(z).$$

First, we show that $\Phi_{k+1}(z)$ is nonnegative for all $z \in \operatorname{dom} g$. This is straightforward but tedious. Recall that $1 - \sqrt{1-p} = \frac{\beta}{1+\beta}$ and hence $2\tau L \leq \frac{\beta}{1+\beta}$. Then by Cauchy-Schwarz and Young's inequalities,

$$\begin{aligned} -2\tau\langle F(z_{k+1}) - F(w_k), z - z_{k+1}\rangle &\leq 2\tau L \|z_{k+1} - w_k\| \|z_{k+1} - z\| \\ &\leq \frac{\beta}{2(1+\beta)}\Big(\|z_{k+1} - w_k\|^2 + \|z_{k+1} - z\|^2\Big) \\ &\leq \frac{\beta}{2(1+\beta)}\|z_{k+1} - z\|^2 + \frac{\beta}{2(1+\beta)}\Big(\big(1 + \frac{1}{\beta}\big)\|z_{k+1} - z_k\|^2 + (1+\beta)\|z_k - w_k\|^2\Big) \\ &= \frac{\beta}{2(1+\beta)}\|z_{k+1} - z\|^2 + \frac{1}{2}\|z_{k+1} - z_k\|^2 + \frac{\beta}{2}\|z_k - w_k\|^2. \end{aligned} \tag{6.34}$$

Therefore, we deduce

$$\Phi_{k+1}(z) \geq \|z_{k+1} - z\|^2 - \frac{\beta}{2(1+\beta)}\|z_{k+1} - z\|^2 \geq \frac{1}{2}\|z_{k+1} - z\|^2. \tag{6.35}$$

Now let $z = \bar{z} \in \mathcal{Z}_\star$. Then by monotonicity of F and (6.1),

$$\Theta_{k+1}(\bar{z}) = \langle F(z_{k+1}), z_{k+1} - \bar{z}\rangle + g(z_{k+1}) - g(\bar{z}) \geq \langle F(\bar{z}), z_{k+1} - \bar{z}\rangle + g(z_{k+1}) - g(\bar{z}) \geq 0. \tag{6.36}$$

Summing up, we have that $\Theta_{k+1}(\bar{z}) \geq 0$, $\Phi_k(\bar{z}) \geq 0$ and $\mathbb{E}_k \Phi_{k+1}(\bar{z}) \leq \Phi_k(\bar{z})$. Unfortunately, this is still not sufficient for us, so we are going to strengthen this inequality by reexamining the proof of Lemma 6.2. In estimating the second line of inequality (6.13) we used that $2\tau L \leq 1 - \sqrt{1-p}$, however, both in the statements of Lemma 6.2 and Theorem 6.3 we assumed a strict inequality. Let

$$\delta = \frac{\beta}{1+\beta} - \frac{4\tau^2L^2(1+\beta)}{\beta} \iff 4\tau^2L^2 = \frac{\beta^2}{(1+\beta)^2} - \frac{\delta\beta}{1+\beta}. \tag{6.37}$$

From $2\tau L < 1 - \sqrt{1-p} = \frac{\beta}{1+\beta}$ it follows that $\delta > 0$. Now, inequality (6.14) can be improved to equality as

$$2\tau^2L^2 + \frac{\beta(1-p)}{2} = \frac{1}{2}\Big(\frac{\beta^2}{(1+\beta)^2} - \frac{\delta\beta}{(1+\beta)} + \frac{\beta}{(1+\beta)^2}\Big) = \frac{\beta(1-\delta)}{2(1+\beta)}. \tag{6.38}$$

This change results in a slightly stronger version of (6.7)

$$\mathbb{E}_k[\Phi_{k+1}(\bar{z}) + 2\tau\Theta_{k+1}(\bar{z})] \leq \Phi_k(\bar{z}) - \frac{\delta}{2}\Big(\|z_k - z_{k-1}\|^2 + \beta\|z_{k-1} - w_{k-1}\|^2\Big). \tag{6.39}$$

As $\Phi_{k+1}(\bar{z}) \geq 0$ and $\Theta_{k+1}(\bar{z}) \geq 0$, we can apply Robbins-Siegmund lemma [RS71] to conclude that $\{\Phi_{k+1}(\bar{z})\}$ converges almost surely and that

$$\sum_{k=1}^{\infty} \mathbb{E}\left[\|z_k - z_{k-1}\|^2 + \|z_{k-1} - w_{k-1}\|^2\right] < \infty. \tag{6.40}$$

It then follows that almost surely, $\|z_k - z_{k-1}\|^2 \to 0$ and $\|z_{k-1} - w_{k-1}\|^2 \to 0$. Moreover, due to (6.35), $\{z_k\}$ is almost surely bounded and therefore by the definition of Φ_k, continuity of F, and (6.40), we have that $\|z_k - \bar{z}\|^2$ converges almost surely.

More specifically, this means that for every $\bar{z} \in \mathcal{Z}_\star$, there exists $\Omega_{\bar{z}}$ with $\mathbb{P}(\Omega_{\bar{z}}) = 1$ such that $\forall \omega \in \Omega_{\bar{z}}$, $\|z_k(\omega) - \bar{z}\|^2$ converges. We can strengthen this result by using the arguments from [Ber11, Proposition 9], [CP15, Proposition 2.3] to obtain that there exists Ω with $\mathbb{P}(\Omega) = 1$ such that for every $\bar{z} \in \mathcal{Z}_\star$ and for every $\omega \in \Omega$, $\|z_k(\omega) - \bar{z}\|^2$ converges.

We now pick a realization $\omega \in \Omega$ and note that $z_k(\omega) - z_{k-1}(\omega) \to 0$ and $z_{k-1}(\omega) - w_{k-1}(\omega) \to 0$. Let us denote by $\tilde{z}$ a cluster point of the bounded sequence $z_k(\omega)$. By using the definition of

z_k and convexity of g, as in the proof of Lemma 6.1, we have for any $z \in \mathcal{Z}$

$$g(z) \geq g(z_k(\omega)) + \frac{1}{\tau}\langle z_{k-1}(\omega) - z_k(\omega), z - z_k(\omega)\rangle - \langle F(w_{k-1}(\omega)), z - z_k(\omega)\rangle \\ - \langle F_{i_{k-1}}(z_{k-1}(\omega)) - F_{i_{k-1}}(w_{k-2}(\omega)), z - z_k(\omega)\rangle.$$

Taking the limit as $k \to \infty$ and using that g is l.s.c. and $\forall i$, F_i is Lipschitz, $z_k(\omega) - z_{k-1}(\omega) \to 0$ and $z_{k-1}(\omega) - w_{k-1}(\omega) \to 0$, we get that $\tilde{z} \in \mathcal{Z}_\star$. Then, as we have that $\|z_k(\omega) - \tilde{z}\|^2$ converges and we have shown that $\|z_k(\omega) - \tilde{z}\|^2$ converges to 0 at least on one sebsequence, we conclude that the sequence $(z_k(\omega))$ converges to some point $\tilde{z}$, where $\tilde{z} \in \mathcal{Z}_\star$. ■

Proof of Theorem 6.6. First, we collect some useful bounds. Consider (6.39) with a specific choice $\bar{z} = P_{\mathcal{Z}^\star}(z_0)$. Taking a full expectation and then summing that inequality, we get

$$\frac{\delta}{2}\sum_{k=0}^{\infty} \mathbb{E}\left[\|z_k - z_{k-1}\|^2 + \beta\|z_{k-1} - w_{k-1}\|^2\right] \leq \|z_0 - P_{\mathcal{Z}_\star}(z_0)\|^2 = \operatorname{dist}(z_0, \mathcal{Z}_\star)^2, \tag{6.41}$$

which also implies by Young's inequality that

$$\frac{\beta\delta}{2(1+\beta)}\sum_{k=0}^{\infty} \mathbb{E}\|z_k - w_{k-1}\|^2 \leq \operatorname{dist}(z_0, \mathcal{Z}_\star)^2. \tag{6.42}$$

Next, we rewrite (6.10) as

$$2\tau\Theta_{k+1}(z) + \|z_{k+1} - z\|^2 + 2\tau\langle F(z_{k+1}) - F(w_k), z - z_{k+1}\rangle + \frac{1}{2}\|z_{k+1} - z_k\|^2 \\ \leq \|z_k - z\|^2 + 2\tau\langle F(z_k) - F(w_{k-1}), z - z_k\rangle + 2\tau^2 L^2\|z_k - w_{k-1}\|^2 \\ + 2\tau\langle F_{i_k}(z_k) - F_{i_k}(w_{k-1}) - (F(z_k) - F(w_{k-1})), z - z_k\rangle. \tag{6.43}$$

Let $v_k = \tau(F_{i_k}(z_k) - F_{i_k}(w_{k-1}) - (F(z_k) - F(w_{k-1})))$, then $\mathbb{E}_k[v_k] = 0$. We define the process $\{\hat{z}_k\}$ by $\hat{z}_0 = z_0$ and

$$\hat{z}_{k+1} = \hat{z}_k + v_k. \tag{6.44}$$

Note that for $\mathcal{F}_k = \sigma\{z_1, \dots, z_k, w_1, \dots, w_{k-1}\}$, $\hat{z}_k$ is $\mathcal{F}_k$-measurable. It also follows that $\forall z \in \mathcal{Z}$

$$\|\hat{z}_{k+1} - z\|^2 = \|\hat{z}_k - z\|^2 + 2\langle v_k, \hat{z}_k - z\rangle + \|v_k\|^2, \tag{6.45}$$

which after summation over $k = 0, \dots, K-1$ yields

$$\sum_{k=0}^{K-1} 2\langle v_k, z - \hat{z}_k\rangle \leq \|z_0 - z\|^2 + \sum_{k=0}^{K-1}\|v_k\|^2. \tag{6.46}$$

With the definition of v_k we can rewrite (6.43) as

$$2\tau\Theta_{k+1}(z) + \|z_{k+1} - z\|^2 + 2\tau\langle F(z_{k+1}) - F(w_k), z - z_{k+1}\rangle + \frac{1}{2}\|z_{k+1} - z_k\|^2 \\ \leq \|z_k - z\|^2 + 2\tau\langle F(z_k) - F(w_{k-1}), z - z_k\rangle + 2\tau^2 L^2\|z_k - w_{k-1}\|^2$$

$$+2\langle v_k, z-\hat{z}_k\rangle + 2\langle v_k, \hat{z}_k - z_k\rangle.$$

We use (6.12), the definition of Φ_k, and the arguments in Lemma 6.2 to show that the last line of (6.13) is nonpositive, to obtain

$$\begin{aligned}&2\tau\Theta_{k+1}(z) + \Phi_{k+1}(z) + \frac{\beta}{2}\left(\mathbb{E}_k\|z_k - w_k\|^2 - \|z_k - w_k\|^2\right)\\ &\qquad \le \Phi_k(z) + 2\langle v_k, z-\hat{z}_k\rangle + 2\langle v_k, \hat{z}_k - z_k\rangle. \qquad (6.47)\end{aligned}$$

Summing this inequality over $k = 0, \ldots, K-1$ and using bound (6.46) yields

$$\begin{aligned}&2\tau\sum_{k=0}^{K-1}\Theta_{k+1}(z) + \Phi_K(z) + \frac{\beta}{2}\sum_{k=0}^{K-1}\left(\mathbb{E}_k\|z_k - w_k\|^2 - \|z_k - w_k\|^2\right)\\ &\le \Phi_0(z) + 2\sum_{k=0}^{K-1}\langle v_k, z-\hat{z}_k\rangle + 2\sum_{k=0}^{K-1}\langle v_k, \hat{z}_k - z_k\rangle\\ &\overset{(6.46)}{\le} \Phi_0(z) + \|z_0 - z\|^2 + 2\sum_{k=0}^{K-1}\|v_k\|^2 + 2\sum_{k=0}^{K-1}\langle v_k, \hat{z}_k - z_k\rangle\\ &= 2\|z_0 - z\|^2 + 2\sum_{k=0}^{K-1}\|v_k\|^2 + 2\sum_{k=0}^{K-1}\langle v_k, \hat{z}_k - z_k\rangle. \qquad (6.48)\end{aligned}$$

We now take the supremum of this inequality over $z \in \mathcal{C}$ and then take a full expectation. As $\hat{z}_k$ is $\mathcal{F}_k$-measurable, $\mathbb{E}[\mathbb{E}_k[\cdot]] = \mathbb{E}[\cdot]$, and $\mathbb{E}_k v_k = 0$, we have $\mathbb{E}_k[\langle v_k, \hat{z}_k - z_k\rangle] = 0$. Using this and that $\Phi_K(z) \ge 0$ by (6.35), we arrive at

$$\tau\mathbb{E}\left[\sup_{z\in\mathcal{C}}\sum_{k=0}^{K-1}\Theta_{k+1}(z)\right] \le \sup_{z\in\mathcal{C}}\|z_0 - z\|^2 + \sum_{k=0}^{K-1}\mathbb{E}\|v_k\|^2. \qquad (6.49)$$

It remains to estimate the last term $\sum_{k=0}^{K-1}\mathbb{E}\|v_k\|^2$. For this, we use a standard inequality $\mathbb{E}\|X - \mathbb{E}X\|^2 \le \mathbb{E}\|X\|^2$ and Lipschitzness of F_{i_k}

$$\begin{aligned}\sum_{k=0}^{K-1}\mathbb{E}\|v_k\|^2 &= \sum_{k=0}^{K-1}\mathbb{E}\tau^2\|F_{i_k}(z_k) - F_{i_k}(w_{k-1}) - (F(z_k) - F(w_{k-1}))\|^2\\ &\le \tau^2\sum_{k=0}^{K-1}\mathbb{E}\|F_{i_k}(z_k) - F_{i_k}(w_{k-1})\|^2 \le \tau^2 L^2\sum_{k=0}^{K-1}\mathbb{E}\|z_k - w_{k-1}\|^2\\ &\overset{(6.42)}{\le} \frac{2\tau^2L^2(1+\beta)}{\delta\beta}\operatorname{dist}(z_0, \mathcal{Z}_\star)^2. \qquad (6.50)\end{aligned}$$

Plugging this bound into (6.49), we obtain

$$\tau\mathbb{E}\left[\sup_{z\in\mathcal{C}}\sum_{k=0}^{K-1}\Theta_{k+1}(z)\right] \le \sup_{z\in\mathcal{C}}\|z_0 - z\|^2 + \frac{2\tau^2L^2(1+\beta)}{\delta\beta}\operatorname{dist}(z_0, \mathcal{Z}_\star)^2. \qquad (6.51)$$

Finally, using monotonicity of F, followed by Jensen inequality, we deduce

$$\sup_{z\in\mathcal{C}} \sum_{k=0}^{K-1} \Theta_{k+1}(z) \geq \sup_{z\in\mathcal{C}} \sum_{k=1}^{K} \Big(\langle F(z), z^k - z\rangle + g(z^k) - g(z) \Big) \geq K G_{\mathcal{C}}(z_K^{av}),$$

which combined with (6.51) finishes the proof. ■

Proof of Theorem 6.8. We start from (6.8) with $i_k = i$,

$$\begin{aligned} \|z_{k+1} - z\|^2 + 2\tau\langle F(z_{k+1}) - F(w_k), z - z_{k+1}\rangle + 2\tau g(z_{k+1}) - 2\tau g(z) \\ + 2\tau\langle F(z_{k+1}), z_{k+1} - z\rangle \leq \|z_k - z\|^2 + 2\tau\langle F_i(z_k) - F_i(w_{k-1}), z - z_k\rangle \\ + 2\tau\langle F_i(z_k) - F_i(w_{k-1}), z_k - z_{k+1}\rangle - \|z_{k+1} - z_k\|^2 \end{aligned}$$

Setting $z = z_\star$ and using strong monotonicity of F,

$$\begin{aligned} \langle F(z_{k+1}), z_{k+1} - z_\star\rangle + g(z_{k+1}) - g(z_\star) \geq \langle F(z_\star), z_{k+1} - z_\star\rangle + \mu\|z_{k+1} - z_\star\|^2 \\ + g(z_{k+1}) - g(z_\star) \geq \mu\|z_{k+1} - z_\star\|^2. \end{aligned}$$

Hence, we have

$$\begin{aligned} (1 + 2\tau\mu)\|z_{k+1} - z_\star\|^2 + 2\tau\langle F(z_{k+1}) - F(w_k), z_\star - z_{k+1}\rangle + \|z_{k+1} - z_k\|^2 \\ \leq \|z_k - z_\star\|^2 + 2\tau\langle F_i(z_k) - F_i(w_{k-1}), z_\star - z_k\rangle \\ + 2\tau\langle F_i(z_k) - F_i(w_{k-1}), z_k - z_{k+1}\rangle. \end{aligned}$$

Then, we continue as in the proof of Theorem 6.3 until we obtain a stronger version of (6.39) due to the strong monotonicity term

$$\begin{aligned} \mathbb{E}_k\Bigg[(1 + 2\mu\tau)\|z_{k+1} - z_\star\|^2 + 2\tau\langle F(z_{k+1}) - F(w_k), z_\star - z_{k+1}\rangle \\ + \frac{\beta}{2}\|z_k - w_k\|^2 + \frac{1}{2}\|z_{k+1} - z_k\|^2 \Bigg] \leq \|z_k - z_\star\|^2 + 2\tau\langle F(z_k) - F(w_{k-1}), z_\star - z_k\rangle \\ + \frac{\beta}{2}\|z_{k-1} - w_{k-1}\|^2 + \frac{1}{2}\|z_k - z_{k-1}\|^2 - \frac{\delta}{2}\Big(\|z_k - z_{k-1}\|^2 + \beta\|z_{k-1} - w_{k-1}\|^2 \Big). \end{aligned} \tag{6.52}$$

Let $a_{k+1} = \frac{1}{2}\|z_{k+1} - z_\star\|^2$ and

$$b_{k+1} = \frac{1}{2}\|z_{k+1} - z_\star\|^2 + 2\tau\langle F(z_{k+1}) - F(w_k), z_\star - z_{k+1}\rangle + \frac{1}{2}\|z_{k+1} - z_k\|^2 + \frac{\beta}{2}\|z_k - w_k\|^2.$$

Note that we have $b_{k+1} + \frac{1}{2}\|z_{k+1} - z_\star\|^2 = \Phi_{k+1}(z_\star) \geq \frac{1}{2}\|z_{k+1} - z_\star\|^2$ by (6.35), hence $b_{k+1} \geq 0$.

Using the definitions of a_k and b_k in (6.52), it follows that for any $\varepsilon \leq \delta$,

$$\mathbb{E}_k\big[(1 + 4\mu\tau)a_{k+1} + b_{k+1} \big] \leq a_k + b_k - \frac{\varepsilon}{2}\left(\|z_k - z_{k-1}\|^2 + \beta\|z_{k-1} - w_{k-1}\|^2 \right), \tag{6.53}$$

Next, we derive

$$\text{RHS of (6.53)} = a_k + b_k - \frac{\varepsilon}{2}\|z_k - z_{k-1}\|^2 - \frac{\varepsilon}{2}\beta\|z_{k-1} - w_{k-1}\|^2 \tag{6.54}$$

$$= \left(1+\frac{\varepsilon}{2}\right)a_k + \left(1-\frac{\varepsilon}{2}\right)b_k - \frac{\varepsilon}{4}\|z_k - z_{k-1}\|^2 - \frac{\varepsilon\beta}{4}\|z_{k-1} - w_{k-1}\|^2$$

$$+ \varepsilon\tau\langle F(z_k) - F(w_{k-1}), z_\star - z_k\rangle \le \left(1+\frac{3\varepsilon}{2}\right)a_k + \left(1-\frac{\varepsilon}{2}\right)b_k, \tag{6.55}$$

where the last inequality follows from (6.34) with a shifted index k. Then, (6.53) becomes

$$\mathbb{E}_k\left[(1+4\mu\tau)a_{k+1} + b_{k+1}\right] \le \left(1+\frac{3\varepsilon}{2}\right)a_k + \left(1-\frac{\varepsilon}{2}\right)b_k. \tag{6.56}$$

Since $\varepsilon \le \delta$ is arbitrary, we can choose ε such that $1+4\mu\tau > 1+\frac{3\varepsilon}{2}$. For instance, we can set

$$\varepsilon = \min\{\delta, 2\mu\tau\}, \tag{6.57}$$

that results in

$$\begin{aligned}\mathbb{E}_k\left[(1+4\mu\tau)a_{k+1}+b_{k+1}\right] &\le (1+3\mu\tau)a_k + \left(1-\frac{\varepsilon}{2}\right)b_k\\ &= \left(1-\frac{\mu\tau}{1+4\mu\tau}\right)(1+4\mu\tau)a_k + \left(1-\frac{\varepsilon}{2}\right)b_k\\ &\le \left(1-\min\left\{\frac{\mu\tau}{1+4\mu\tau}, \frac{\varepsilon}{2}\right\}\right)\big((1+4\mu\tau)a_k + b_k\big).\end{aligned} \tag{6.58}$$

Taking a full expectation and using that $\frac{\varepsilon}{2} = \min\{\frac{\delta}{2}, \mu\tau\}$ and $b_0 = 0$, we obtain

$$\begin{aligned}\mathbb{E}\left[(1+4\mu\tau)a_{k+1}+b_{k+1}\right] &\le \left(1-\min\left\{\frac{\mu\tau}{1+4\mu\tau}, \frac{\delta}{2}\right\}\right)\mathbb{E}\left[(1+4\mu\tau)a_k + b_k\right]\\ &\le \left(1-\min\left\{\frac{\mu\tau}{1+4\mu\tau}, \frac{\delta}{2}\right\}\right)^{k+1}(1+4\mu\tau)a_0.\end{aligned}$$

Now it only remains to compute the contraction factor. By our choice of τ, we have $\tau L = \frac{p}{4\sqrt{2}} \le \frac{1-\sqrt{1-p}}{2\sqrt{2}} = \frac{\beta}{2\sqrt{2}(1+\beta)}$, and hence,

$$\delta = \frac{\beta}{1+\beta} - \frac{4\tau^2L^2(1+\beta)}{\beta} \ge \frac{\beta}{2(1+\beta)} \ge \frac{1-\sqrt{1-p}}{2} \ge \frac{p}{4}. \tag{6.59}$$

From $\mu \le L$ it follows that $4\mu\tau = \frac{\mu p}{\sqrt{2}L} \le \frac{p}{\sqrt{2}} < 1$ and, hence, $\frac{\mu\tau}{1+4\mu\tau} \ge \frac{\mu\tau}{2} = \frac{\mu p}{8\sqrt{2}L}$. Thus, we obtain

$$\min\left\{\frac{\mu\tau}{1+4\mu\tau}, \frac{\delta}{2}\right\} \ge \min\left\{\frac{\mu p}{8\sqrt{2}L}, \frac{p}{8}\right\} = \frac{\mu p}{8\sqrt{2}L},$$

which finally implies the result. ■

6.6.2 Proofs for Section 6.4

We first need a generalized version of Lemma 6.1. In fact, this is the exact form proven in [MT20b], therefore we do not provide its proof.

Lemma 6.24. *[MT20b, Proposition 2.3] Let $A\colon \mathcal{Z} \rightrightarrows \mathcal{Z}$ be maximally monotone and let $x_1, U_0, U_1, V_1 \in \mathcal{Z}$ be arbitrary points. Define x_2 as*

$$x_2 = J_A(x_1 - U_1 - (V_1 - U_0)). \tag{6.60}$$

Then for all $x \in \mathcal{Z}$, $V_2 \in \mathcal{Z}$, and $U \in -A(x)$, we have

$$\begin{aligned} &\|x_2 - x\|^2 + 2\langle V_2 - U_1, x - x_2\rangle + 2\langle V_2 - U, x_2 - x\rangle \\ &\leq \|x_1 - x\|^2 + 2\langle V_1 - U_0, x - x_1\rangle + 2\langle V_1 - U_0, x_1 - x_2\rangle - \|x_1 - x_2\|^2. \end{aligned} \tag{6.61}$$

Proof of Theorem 6.11

Proof. We will start similar to Lemma 6.2. After setting $U_0 = \tau F_i(w_{k-1})$, $U_1 = \tau F(w_k)$, $V_1 = \tau F_i(z_k)$, $V_2 = \tau F(z_{k+1})$, $x_1 = z_k$, $x_2 = z_{k+1}$ with $i_k = i$ and plugging into Lemma 6.24, we have

$$\begin{aligned} \|z_{k+1} - z\|^2 + 2\tau\langle F(z_{k+1}) - F(w_k), z - z_{k+1}\rangle &\leq \|z_k - z\|^2 \\ + 2\tau\langle F_i(z_k) - F_i(w_{k-1}), z - z_k\rangle + 2\tau\langle F_i(z_k) &- F_i(w_{k-1}), z_k - z_{k+1}\rangle - \|z_{k+1} - z_k\|^2 \\ &- 2\tau\langle F(z_{k+1}) - F(z), z_{k+1} - z\rangle. \end{aligned} \tag{6.62}$$

We use monotonicity for the last term and get

$$\begin{aligned} \|z_{k+1} - z\|^2 + 2\tau\langle F(z_{k+1}) - F(w_k), z - z_{k+1}\rangle &\leq \|z_k - z\|^2 - \|z_{k+1} - z_k\|^2 \\ + 2\tau\langle F_i(z_k) - F_i(w_{k-1}), z - z_k\rangle &+ 2\tau\langle F_i(z_k) - F_i(w_{k-1}), z_k - z_{k+1}\rangle. \end{aligned} \tag{6.63}$$

The rest of Lemma 6.2 follows in this case the same way with the lack of the terms with $\Theta_{k+1}(z)$. Then, similar arguments as in Theorem 6.3 with the changes of i) not having $\Theta_{k+1}(z)$, ii) using the definition of resolvent instead of proximal operator to show cluster points are solutions, will give the result (see also [MT20b, Theorem 2.5]). ■

We now present a version of Lemma 6.1 with the proximal operator using Bregman distance.

Lemma 6.25. *Let $g\colon \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ be proper lower semicontinuous convex and let $x_1, U_0, U_1, V_1 \in \mathcal{Z}$ be arbitrary points. Define x_2 as*

$$x_2 = \operatorname*{argmin}_{z \in \mathcal{Z}} \Big\{ g(z) + \langle U_1 + (V_1 - U_0), z - x_1\rangle + D(z, x_1) \Big\}. \tag{6.64}$$

Then, for all $x \in \mathcal{Z}$, $V_2 \in \mathcal{Z}$ we have

$$D(x, x_2) + \langle V_2 - U_1, z - x_2\rangle + \langle V_2, x_2 - x\rangle + g(x_2) - g(x)$$

$$\leq D(x,x_1)+\langle V_1-U_0,x-x_1\rangle+\langle V_1-U_0,x_1-x_2\rangle-D(x_2,x_1). \tag{6.65}$$

Proof. By the definition of x_2, it follows from [Tse08, Property 1] that

$$g(x)\geq g(x_2)-\langle U_1+V_1-U_0,x-x_2\rangle-D(x,x_1)+D(x,x_2)+D(x_2,x_1).$$

For the bilinear term, we argue the same as Lemma 6.1. ■

Proof of Lemma 6.12

Proof. We will follow the proof of Lemma 6.2 with suitable changes for Bregman distances.

First, set $U_0=\tau F_i(w_{k-1})$, $U_1=\tau F(w_k)$, $V_1=\tau F_i(z_k)$, $V_2=\tau F(z_{k+1})$, $x_1=z_k$, then $x_2=z_{k+1}$ with $i_k=i$ and we plug these into (6.65) to get

$$\begin{aligned}D(z,z_{k+1})+\tau\langle F(z_{k+1})-F(w_k),z-z_{k+1}\rangle+\tau(\langle F(z_{k+1}),z_{k+1}-z\rangle\\+g(z_{k+1})-g(z))\leq D(z,z_k)+\tau\langle F_i(z_k)-F_i(w_{k-1}),z-z_k\rangle\\+\tau\langle F_i(z_k)-F_i(w_{k-1}),z_k-z_{k+1}\rangle-D(z_{k+1},z_k).\end{aligned}$$

First, note that by Lipschitzness of F_i, Cauchy-Schwarz, Young's inequalities, and since $\frac{1}{2}\|z_k-z_{k-1}\|^2\leq D(z_k,z_{k-1})$,

$$\begin{aligned}\tau\langle F_i(z_k)-F_i(w_{k-1}),z_k-z_{k+1}\rangle&\leq\tau^2L^2\|z_k-w_{k-1}\|^2+\frac{1}{4}\|z_k-z_{k+1}\|^2\\&\leq\tau^2L^2\|z_k-w_{k-1}\|^2+\frac{1}{2}D(z_{k+1},z_k)\end{aligned} \tag{6.66}$$

Thus, it follows that

$$\begin{aligned}&D(z,z_{k+1})+\tau\langle F(z_{k+1})-F(w_k),z-z_{k+1}\rangle+\frac{1}{2}D(z_{k+1},z_k)+\tau\Theta_{k+1}(z)\\&\leq D(z,z_k)+\tau\langle F_i(z_k)-F_i(w_{k-1}),z-z_k\rangle+\tau^2L^2\|z_k-w_{k-1}\|^2.\end{aligned} \tag{6.67}$$

Taking expectation conditioning on the knowledge of z_k, w_{k-1} and using that $\mathbb{E}_kF_i(z_k)=F(z_k)$, $\mathbb{E}_kF_i(w_{k-1})=F(w_{k-1})$, we obtain

$$\begin{aligned}\mathbb{E}_kD(z,z_{k+1})+\tau\mathbb{E}_k\langle F(z_{k+1})-F(w_k),z-z_{k+1}\rangle+\frac{1}{2}\mathbb{E}_kD(z_{k+1},z_k)\\+\tau\mathbb{E}_k\Theta_{k+1}(z)\leq D(z,z_k)+\tau\langle F(z_k)-F(w_{k-1}),z-z_k\rangle+\tau^2L^2\|z_k-w_{k-1}\|^2.\end{aligned} \tag{6.68}$$

Adding

$$\frac{\beta}{4}\mathbb{E}_k\|z_k-w_k\|^2=\frac{\beta(1-p)}{4}\|z_k-w_{k-1}\|^2, \tag{6.69}$$

which follows by the definition of w_k, to (6.68), we obtain

$$\begin{aligned}\mathbb{E}_k[\Phi_{k+1}(z)+\Theta_{k+1}(z)] &\le \Phi_k(z)\\ &+\left(\tau^2L^2+\frac{\beta(1-p)}{4}\right)\|z_k-w_{k-1}\|^2-\frac{1}{2}D(z_k,z_{k-1})-\frac{\beta}{4}\|z_{k-1}-w_{k-1}\|^2. \quad (6.70)\end{aligned}$$

To show that the last line is nonpositive, we use (6.14), Young's inequality as in Lemma 6.2 and $\frac{1}{2}\|z_k-z_{k-1}\|^2 \le D(z_k,z_{k-1})$.

Nonnegativity of Φ_k follows as in Theorem 6.3 after using $\frac{1}{2}\|z_k-z_{k-1}\|^2 \le D(z_k,z_{k-1})$. ■

6.6.3 Proofs Section 6.4.3

Proof of Theorem 6.15

Proof of Lemma 6.14. First, we define the sequence $x_{k+1}=x_k+u_{k+1}$. It is easy to see that x_k is $\mathcal{F}_k$-measurable. Next, by using the definition of (x_k), we have

$$\|x_{k+1}-x\|^2=\|x_k-x\|^2+2\langle u_{k+1},x_k-x\rangle+\|u_{k+1}\|^2.$$

Summing over $k=0,\dots,K-1$, we obtain

$$\sum_{k=0}^{K-1}2\langle u_{k+1},x-x_k\rangle \le \|x_0-x\|^2+\sum_{k=0}^{K-1}\|u_{k+1}\|^2.$$

Next, we take maximum of both sides and then expectation

$$\mathbb{E}\left[\max_{x\in\mathcal{C}}\sum_{k=0}^{K-1}\langle u_k,x\rangle\right] \le \max_{x\in\mathcal{C}}\frac{1}{2}\|x_0-x\|^2+\frac{1}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\|u_{k+1}\|^2\right]+\sum_{k=0}^{K-1}\mathbb{E}\left[\langle u_{k+1},x_k\rangle\right].$$

We use the tower property, $\mathcal{F}_k$-measurability of x_k, and $\mathbb{E}[u_{k+1}|\mathcal{F}_k]=0$ to finish the proof, since $\sum_{k=0}^{K-1}\mathbb{E}[\langle u_{k+1},x_k\rangle]=\sum_{k=0}^{K-1}\mathbb{E}[\langle\mathbb{E}[u_{k+1}|\mathcal{F}_k],x_k\rangle]=0$. ■

Proof of Theorem 6.15. Let

$$\Theta_{k+1/2}(z)=\langle F(z_{k+1/2}),z_{k+1/2}-z\rangle+g(z_{k+1/2})-g(z).$$

We proceed as Lemma 6.13 until getting (6.26): using (6.22) and (6.23) in (6.21) gives

$$\begin{aligned}2\tau\Theta_{k+1/2}(z)+\|z_{k+1}-z\|^2 &\le \alpha\|z_k-z\|^2+(1-\alpha)\|w_k-z\|^2\\ &+2\tau\langle F_{\xi_k}(w_k)-F_{\xi_k}(z_{k+1/2}),z_{k+1}-z_{k+1/2}\rangle\\ &-(1-\alpha)\|z_{k+1/2}-w_k\|^2-\|z_{k+1}-z_{k+1/2}\|^2\\ &+\underbrace{2\tau\langle F(z_{k+1/2})-F_{\xi_k}(z_{k+1/2})-F(w_k)+F_{\xi_k}(w_k),z_{k+1/2}-z\rangle}_{e_1(z,k)}, \quad (6.71)\end{aligned}$$

where we call the last term by $e_1(z,k)$.

Now, we set $\alpha = 1-p$. We want to rewrite (6.71) using $\Phi_k(z) = (1-p)\|z_k - z\|^2 + \|w_k - z\|^2$. For this, we need to add $\|w_{k+1} - z\|^2 - \|w_k - z\|^2$ to both sides. Then, we define the error

$$\begin{aligned} e_2(z,k) &= p\|w_k - z\|^2 + \|w_{k+1} - z\|^2 - \|w_k - z\|^2 - p\|z_{k+1} - z\|^2 \\ &= 2\langle p z_{k+1} + (1-p) w_k - w_{k+1}, z\rangle - p\|z_{k+1}\|^2 - (1-p)\|w_k\|^2 + \|w_{k+1}\|^2. \end{aligned}$$

With this at hand, we can cast (6.71) as

$$\begin{aligned} 2\tau\Theta_{k+1/2}(z) + \Phi_{k+1}(z) \le \Phi_k(z) &+ e_1(z,k) + e_2(z,k) \\ &+ 2\tau\langle F_{\xi_k}(w_k) - F_{\xi_k}(z_{k+1/2}), z_{k+1} - z_{k+1/2}\rangle \\ &- p\|z_{k+1/2} - w_k\|^2 - \|z_{k+1} - z_{k+1/2}\|^2. \end{aligned}$$

We sum this inequality over $k = 0, \dots, K-1$, take maximum of both sides over $z \in \mathcal{C}$, and then take total expectation to obtain

$$\begin{aligned} 2\tau K \mathbb{E}\left[\mathrm{Gap}(z^K)\right] \le \max_{z\in\mathcal{C}} \Phi_0(z) &+ \mathbb{E}\left[\max_{z\in\mathcal{C}} \sum_{k=0}^{K-1} \left(e_1(z,k) + e_2(z,k)\right)\right] \\ &- \mathbb{E}\sum_{k=0}^{K-1}\left(\|z_{k+1} - z_{k+1/2}\|^2 + p\|z_{k+1/2} - w_k\|^2\right) \\ &+ 2\tau\mathbb{E}\sum_{k=0}^{K-1}\left[\langle F_{\xi_k}(w_k) - F_{\xi_k}(z_{k+1/2}), z_{k+1} - z_{k+1/2}\rangle\right] \end{aligned} \tag{6.72}$$

where we used $\mathbb{E}\max_{z\in\mathcal{C}} \sum_{k=0}^{K-1} \Theta_{k+1/2}(z) \ge K\mathbb{E}\left[\mathrm{Gap}(z^K)\right]$, which follows from monotonicity of F, linearity of $z_{k+1/2} \mapsto \langle F(z), z_{k+1/2} - z\rangle$, and convexity of g.

The tower property, the estimation from (6.25), and $1-\alpha = p$ applied on (6.72) imply

$$2\tau K \mathbb{E}\left[\mathrm{Gap}(z^K)\right] \le \max_{z\in\mathcal{C}} \Phi_0(z) + \mathbb{E}\left[\max_{z\in\mathcal{C}} \sum_{k=0}^{K-1} \left(e_1(z,k) + e_2(z,k)\right)\right]. \tag{6.73}$$

Therefore, the proof will be complete upon deriving an upper bound for the second term on RHS. We instantiate Lemma 6.14 twice for this bound. First, for $e_1(z,k)$ we set in Lemma 6.14,

$$\mathcal{F}_k = \sigma(\xi_0, \dots, \xi_{k-1}, w_k), \quad \tilde{x}_0 = z_0, \quad u_{k+1} = 2\tau\left([F_{\xi_k}(z_{k+1/2}) - F_{\xi_k}(w_k)] - [F(z_{k+1/2}) - F(w_k)]\right),$$

where by definition we set $\mathcal{F}_0 = \sigma(\xi_0, \xi_{-1}, w_0) = \sigma(\xi_0)$. With this, we obtain the bound

$$\begin{aligned} \mathbb{E}\left[\max_{z\in\mathcal{C}} \sum_{k=0}^{K-1} e_1(z,k)\right] &= \mathbb{E}\left[\max_{z\in\mathcal{C}} \sum_{k=0}^{K-1} \langle u_{k+1}, z\rangle\right] - \mathbb{E}\left[\sum_{k=0}^{K-1} \langle u_{k+1}, z_{k+1/2}\rangle\right] = \mathbb{E}\left[\max_{z\in\mathcal{C}} \sum_{k=0}^{K-1} \langle u_{k+1}, z\rangle\right] \\ &\le \max_{z\in\mathcal{C}} \frac{1}{2}\|z_0 - z\|^2 + \frac{1}{2}\sum_{k=0}^{K-1} \mathbb{E}\|u_{k+1}\|^2 \\ &\le \max_{z\in\mathcal{C}} \frac{1}{2}\|z_0 - z\|^2 + 2\tau^2 L^2 \sum_{k=0}^{K-1} \mathbb{E}\|z_{k+1/2} - w_k\|^2, \end{aligned} \tag{6.74}$$

where the second equality follows by the tower property, $\mathbb{E}_k[u_{k+1}] = 0$, and $\mathcal{F}_k$-measurability of $z_{k+1/2}$. The last inequality is due to

$$\mathbb{E}\,\|u_{k+1}\|^2 = \mathbb{E}\left[\mathbb{E}_k\,\|u_{k+1}\|^2\right] \le 4\tau^2\,\mathbb{E}\left[\mathbb{E}_k\,\|F_{\xi_k}(z_{k+1/2}) - F_{\xi_k}(w_k)\|^2\right] \le 4\tau^2 L^2\,\mathbb{E}\,\|z_{k+1/2} - w_k\|^2,$$

where we use the tower property, $\mathbb{E}\,\|X - \mathbb{E}\,X\|^2 \le \mathbb{E}\,\|X\|^2$, and Assumption 6.1(iv).

Second, we set in Lemma 6.14

$$\mathcal{F}_k = \sigma(\xi_0,\dots,\xi_k, w_k), \qquad \tilde{x}_0 = z_0, \qquad u_{k+1} = p z_{k+1} + (1-p) w_k - w_{k+1},$$

and use $\mathbb{E}\left[\mathbb{E}_{k+1/2}[\|w_{k+1}\|^2 - p\|z_{k+1}\|^2 - (1-p)\|w_k\|^2]\right] = 0$, to obtain the bound

$$\begin{aligned}\mathbb{E}\max_{z\in\mathcal{C}} \sum_{k=0}^{K-1} e_2(z,k) = 2\,\mathbb{E}\max_{z\in\mathcal{C}} \sum_{k=0}^{K-1} \langle u_{k+1}, z\rangle &\le \max_{z\in\mathcal{C}} \|z_0 - z\|^2 + \sum_{k=0}^{K-1} \mathbb{E}\,\|u_{k+1}\|^2 \\ &= \max_{z\in\mathcal{C}} \|z_0 - z\|^2 + p(1-p) \sum_{k=0}^{K-1} \mathbb{E}\,\|z_{k+1} - w_k\|^2, \end{aligned} \tag{6.75}$$

where the inequality follows from Lemma 6.14 and the second equality from the derivation

$$\begin{aligned}\mathbb{E}\,\|u_{k+1}\|^2 &= \mathbb{E}\left[\mathbb{E}_{k+1/2}\,\|u_{k+1}\|^2\right] = \mathbb{E}\left[\mathbb{E}_{k+1/2}\,\|\mathbb{E}_{k+1/2}\,[w_{k+1}] - w_{k+1}\|^2\right] \\ &= \mathbb{E}\left[\mathbb{E}_{k+1/2}\,\|w_{k+1}\|^2 - \|\,\mathbb{E}_{k+1/2}\,[w_{k+1}]\,\|^2\right] \\ &= \mathbb{E}\left[p\|z_{k+1}\|^2 + (1-p)\|w_k\|^2 - \|p z_{k+1} + (1-p) w_k\|^2\right] = p(1-p)\,\mathbb{E}\,\|z_{k+1} - w_k\|^2,\end{aligned}$$

which uses $\mathbb{E}\,\|X - \mathbb{E}\,X\|^2 = \mathbb{E}\,\|X\|^2 - \|\,\mathbb{E}\,X\|^2$.

Combining (6.74), (6.75), and (6.73), we finally arrive at

$$\begin{aligned}2\tau K\mathbb{E}\left[\mathrm{Gap}(z^K)\right] &\le \max_{z\in\mathcal{C}} \Phi_0(z) + \max_{z\in\mathcal{C}} \frac{1}{2}\|z_0 - z\|^2 + 2\tau^2 L^2 \sum_{k=0}^{K-1} \mathbb{E}\,\|z_{k+1/2} - w_k\|^2 \\ &\quad + \max_{z\in\mathcal{C}} \|z_0 - z\|^2 + p(1-p) \sum_{k=0}^{K-1} \mathbb{E}\,\|z_{k+1} - w_k\|^2 \end{aligned} \tag{6.76}$$

We have to estimate terms under the sum:

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=0}^{K-1} \left(2\tau^2 L^2 \|z_{k+1/2} - w_k\|^2 + p(1-p)\|z_{k+1} - w_k\|^2\right)\right] \\ &\le p\mathbb{E}\left[\sum_{k=0}^{K-1} \left(2\|z_{k+1/2} - w_k\|^2 + \|z_{k+1} - w_k\|^2\right)\right] \\ &\le p\mathbb{E}\left[\sum_{k=0}^{K-1} \left((2+\sqrt{2})\|z_{k+1/2} - w_k\|^2 + (2+\sqrt{2})\|z_{k+1} - z_{k+1/2}\|^2\right)\right] \\ &\le \frac{2+\sqrt{2}}{1-\gamma}\Phi_0(z_*) \le \frac{3.5}{1-\gamma} \max_{z\in\mathcal{C}} \Phi_0(z), \end{aligned} \tag{6.77}$$

where the first inequality in (6.77) uses Lemma 6.13 and $1-\alpha = p$.

Now we use that $w_0 = z_0$ and, hence, $\Phi_0(z) = (2-p)\|z_0 - z\|^2 \le 2\|z_0 - z\|^2$ in (6.76). This yields

$$2\tau K\mathbb{E}\left[\mathrm{Gap}(z^K)\right] \le \left(2 + \frac{3}{2} + \frac{7}{1-\gamma}\right)\max_{z\in\mathcal{C}}\|z_0 - z\|^2 = 7\left(\frac{1}{2} + \frac{1}{1-\gamma}\right)\max_{z\in\mathcal{C}}\|z_0 - z\|^2.$$

Finally, using $\tau = \frac{\sqrt{p}\gamma}{L}$, we obtain

$$\mathbb{E}\left[\mathrm{Gap}(z^K)\right] \le \frac{7L}{2\sqrt{p}\gamma K}\left(\frac{1}{2} + \frac{1}{1-\gamma}\right)\max_{z\in\mathcal{C}}\|z_0 - z\|^2 = \mathcal{O}\left(\frac{L}{\sqrt{p}K}\right).$$

In particular, with a stepsize $\tau = \frac{\sqrt{p}}{2L}$, the right-hand side reduces to $\frac{17.5L}{\sqrt{p}K}\max_{z\in\mathcal{C}}\|z_0 - z\|^2$. ■

Proof of Corollary 6.16. In average each iteration costs $pN\mathsf{Cost} + 2\mathsf{Cost} = (pN+2)\mathsf{Cost}$. To reach ε-accuracy we need $\left\lceil \mathcal{O}\left(\frac{L}{\sqrt{p}\varepsilon}\right)\right\rceil$ iterations. Hence, the total average complexity is $\mathcal{O}\left(\frac{\mathsf{Cost}\times(pN+2)L}{\sqrt{p}\varepsilon}\right)$. Finally, the optimal choice $p = \frac{2}{N}$ results in $\mathcal{O}\left(\frac{\mathsf{Cost}\times\sqrt{N}L}{\varepsilon}\right)$ complexity. ■

Proof of Theorem 6.22

Proof of Lemma 6.21. Define for each $s \ge 0$ and for $k \in \{0,\dots,K-1\}$,

$$x^s_{k+1} = \operatorname*{argmin}_{x\in\operatorname{dom} g}\{\langle -u^s_{k+1}, x\rangle + D(x, x^s_k)\}, \text{ and let } x^{s+1}_0 = x^s_m.$$

First, we observe x^s_k is $\mathcal{F}^s_k$-measurable. By the definition of x^s_{k+1}, we have for all $x \in \operatorname{dom} g$,

$$\langle \nabla h(x^s_{k+1}) - \nabla h(x^s_k) - u^s_{k+1}, x - x^s_{k+1}\rangle \ge 0.$$

We apply three point identity to obtain

$$D(x, x^s_k) - D(x, x^s_{k+1}) - D(x^s_{k+1}, x^s_k) - \langle u^s_{k+1}, x - x^s_{k+1}\rangle \ge 0.$$

We estimate the inner product by Hölder's, Young's inequalities, and strong convexity of h,

$$\begin{aligned}\langle u^s_{k+1}, x - x^s_{k+1}\rangle &= \langle u^s_{k+1}, x - x^s_k\rangle + \langle u^s_{k+1}, x^s_k - x^s_{k+1}\rangle\\ &\ge \langle u^s_{k+1}, x - x^s_k\rangle - \frac{1}{2}\|u^s_{k+1}\|^2_* - \frac{1}{2}\|x^s_{k+1} - x^s_k\|^2\\ &\ge \langle u^s_{k+1}, x - x^s_k\rangle - \frac{1}{2}\|u^s_{k+1}\|^2_* - D(x^s_{k+1}, x^s_k),\end{aligned}$$

which, combined with the previous inequality gives

$$\langle u^s_{k+1}, x\rangle \le D(x, x^s_k) - D(x, x^s_{k+1}) + \langle u^s_{k+1}, x^s_k\rangle + \frac{1}{2}\|u^s_{k+1}\|^2_*.$$

We sum this inequality over k, s, take maximum, use $x_0^{s+1} = x_K^s$ and the same derivations as at the end of the proof of Lemma 6.14 to show $\sum_{s=0}^{S-1}\sum_{k=0}^{K-1}\mathbb{E}\left[\langle u_{k+1}^s, x_k^s\rangle\right] = 0$. ■

Proof of Lemma 6.20. By the definition of $z_{k+1/2}^s$, using convexity of g and three point identity (for example see [AM21, Lemma 3.2]), we have

$$\begin{aligned}\tau\left(g(z_{k+1}^s) - g(z_{k+1/2}^s) + \langle F(w^s), z_{k+1}^s - z_{k+1/2}^s\rangle\right) &\geq D(z_{k+1}^s, z_{k+1/2}^s)\\ &+ \alpha\Big(D(z_{k+1/2}^s, z_k^s) - D(z_{k+1}^s, z_k^s)\Big) + (1-\alpha)\Big(D(z_{k+1/2}^s, \bar{w}^s) - D(z_{k+1}^s, \bar{w}^s)\Big). \end{aligned}\tag{6.78}$$

With the same reasoning as the previous inequality, by using z_{k+1}^s, we have for any $z \in \mathcal{Z}$

$$\begin{aligned}\tau\Big(g(z) - g(z_{k+1}^s) + \langle F(w^s) + F_{\xi_k^s}(z_{k+1/2}^s) - F_{\xi_k^s}(w^s), z - z_{k+1}^s\rangle\Big) &\geq D(z, z_{k+1}^s)\\ &+ \alpha\Big(D(z_{k+1}^s, z_k^s) - D(z, z_k^s)\Big) + (1-\alpha)\Big(D(z_{k+1}^s, \bar{w}^s) - D(z, \bar{w}^s)\Big).\end{aligned}\tag{6.79}$$

Note that for any u, v, the expression $D(u, \bar{w}^s) - D(v, \bar{w}^s)$ is linear in terms of $\nabla h(\bar{w}^s)$, that is

$$D(u, \bar{w}^s) - D(v, \bar{w}^s) = \frac{1}{K}\sum_{j=1}^K \Big(D(u, z_j^{s-1}) - D(v, z_j^{s-1})\Big).\tag{6.80}$$

Summing up (6.78) and (6.79) and using (6.80), we obtain

$$\begin{aligned}\tau\Big(g(z) - g(z_{k+1/2}^s) + \langle F(w^s) + F_{\xi_k^s}(z_{k+1/2}^s) - F_{\xi_k^s}(w^s), z - z_{k+1/2}^s\rangle\Big) &\geq D(z, z_{k+1}^s) - \alpha D(z, z_k^s)\\ &+ \frac{1-\alpha}{K}\sum_{j=1}^K D(z_{k+1/2}^s, z_j^{s-1}) - \frac{1-\alpha}{K}\sum_{j=1}^K D(z, z_j^{s-1}) + D(z_{k+1}^s, z_{k+1/2}^s)\\ &+ \tau\langle F_{\xi_k^s}(z_{k+1/2}^s) - F_{\xi_k^s}(w^s), z_{k+1}^s - z_{k+1/2}^s\rangle.\end{aligned}\tag{6.81}$$

By $D(u,v) \geq \frac{1}{2}\|u - v\|^2$ and Jensen's inequality, we have

$$\frac{1-\alpha}{K}\sum_{j=1}^K D(z_{k+1/2}^s, z_j^{s-1}) \geq \frac{1-\alpha}{K}\sum_{j=1}^K \frac{1}{2}\|z_{k+1/2}^s - z_j^{s-1}\|^2 \geq \frac{1-\alpha}{2}\|z_{k+1/2}^s - w^s\|^2,\tag{6.82}$$

$$D(z_{k+1}^s, z_{k+1/2}^s) \geq \frac{1}{2}\|z_{k+1}^s - z_{k+1/2}^s\|^2.\tag{6.83}$$

By using (6.29), (6.82), and (6.83) in (6.81), we deduce

$$\begin{aligned}\tau\Theta_{k+1/2}^s(z) + D(z, z_{k+1}^s) &\leq \alpha D(z, z_k^s) + \frac{1-\alpha}{K}\sum_{j=1}^K D(z, z_j^{s-1})\\ &+ \tau\langle F_{\xi_k^s}(w^s) - F_{\xi_k^s}(z_{k+1/2}^s), z_{k+1}^s - z_{k+1/2}^s\rangle - \frac{1}{2}\|z_{k+1}^s - z_{k+1/2}^s\|^2 - \frac{1-\alpha}{2}\|z_{k+1/2}^s - w^s\|^2,\\ &\underbrace{+\tau\langle F(z_{k+1/2}^s) - F_{\xi_k^s}(z_{k+1/2}^s) - F(w^s) + F_{\xi_k^s}(w^s), z_{k+1/2}^s - z\rangle}_{e(z,s,k)},\end{aligned}$$

where we have defined the last term as $e(z,s,k)$ (see (6.30)). We sum this inequality over k to obtain the result in (i).

Next, similar to (6.25), we estimate by Assumption 6.2(iv) and Young's inequality

$$\begin{aligned}
&\tau\mathbb{E}_{s,k}\langle F_{\xi_k^s}(w^s) - F_{\xi_k^s}(z_{k+1/2}^s), z_{k+1}^s - z_{k+1/2}^s\rangle \\
&\qquad\le \mathbb{E}_{s,k}\left[\frac{\tau^2}{2}\|F_{\xi_k^s}(w^s) - F_{\xi_k^s}(z_{k+1/2}^s)\|_*^2 + \frac{1}{2}\|z_{k+1}^s - z_{k+1/2}^s\|^2\right] \\
&\qquad\qquad\le \frac{(1-\alpha)\gamma^2}{2}\|z_{k+1/2}^s - w^s\|^2 + \frac{1}{2}\mathbb{E}_{s,k}\|z_{k+1}^s - z_{k+1/2}^s\|^2, \quad (6.84)
\end{aligned}$$

since $\tau^2L^2 = (1-\alpha)\gamma^2$. We take expectation of (6.81), plug in $z = z_*$; use (6.24), (6.84), (6.82), and (6.83) to get

$$\mathbb{E}_{s,k}\left[D(z_*, z_{k+1}^s)\right] \le \alpha D(z_*, z_k^s) + \frac{1-\alpha}{K}\sum_{j=1}^{K} D(z_*, z_j^{s-1}) + \frac{(1-\alpha)(\gamma^2-1)}{2}\|z_{k+1/2}^s - w^s\|^2. \quad (6.85)$$

By using $\mathbb{E}_{s,0}[\cdot] = \mathbb{E}_{s,0}\left[\mathbb{E}_{s,k}[\cdot]\right]$, we have

$$\mathbb{E}_{s,0}D(z_*, z_{k+1}^s) \le \mathbb{E}_{s,0}\left[\alpha D(z_*, z_k^s) + \frac{1-\alpha}{K}\sum_{j=1}^{K} D(z_*, z_j^{s-1}) - \frac{(1-\alpha)(1-\gamma^2)}{2}\|z_{k+1/2}^s - w^s\|^2\right]. \quad (6.86)$$

Summing the inequality over $k = 0,\dots,K-1$ and using the definition of $\Phi^s(z_*)$ with $z_0^{s+1} = z_K^s$, we get *(ii)*. Finally, we take total expectation of *(ii)* and sum over s to obtain *(iii)*. ■

Proof of Theorem 6.22. We start with the result of Lemma 6.20 and proceed similar to Theorem 6.15. Since $z_0^{s+1} = z_K^s$, we use definition of $\Phi^s(z)$, and sum the inequality in Lemma 6.20(i) over s to obtain

$$\sum_{s=0}^{S-1}\sum_{k=0}^{K-1}\tau\Theta_{k+1/2}^s(z) + \Phi^S(z) \le \Phi^0(z) + \sum_{s=0}^{S-1}\sum_{k=0}^{K-1}[e(z,s,k) + \delta(s,k)]$$

We take maximum and expectation, use $\mathbb{E}\left[\max_{z\in\mathcal{C}}\sum_{s=0}^{S-1}\sum_{k=0}^{K-1}\tau\Theta_{k+1/2}^s(z)\right] \ge \tau KS\mathbb{E}\left[\mathrm{Gap}(z^S)\right]$ to deduce

$$\tau KS\mathbb{E}\left[\mathrm{Gap}(z^S)\right] \le \max_{z\in\mathcal{C}}\Phi^0(z) + \mathbb{E}\left[\max_{z\in\mathcal{C}}\sum_{s=0}^{S-1}\sum_{k=0}^{K-1}e(z,s,k)\right] + \mathbb{E}\left[\sum_{s=0}^{S-1}\sum_{k=0}^{K-1}\delta(s,k)\right].$$

The term $\mathbb{E}\sum_{s=0}^{S-1}\sum_{k=0}^{K-1}\delta(s,k)$ is nonpositive by the tower property, Lipschitzness, Young's inequality, and $\tau < \frac{\sqrt{p}}{L}$ (the same arguments used in (6.84) can be applied here with $\delta(s,k)$ defined as (6.31)). Therefore,

$$\tau KS\mathbb{E}\left[\mathrm{Gap}(z^S)\right] \le \max_{z\in\mathcal{C}}\Phi^0(z) + \mathbb{E}\left[\max_{z\in\mathcal{C}}\sum_{s=0}^{S-1}\sum_{k=0}^{K-1}e(z,s,k)\right].$$

We next bound the second term on RHS, similar to the proof of Theorem 6.15. For $s \in \{0,\dots,S-1\}$ and $k \in \{0,\dots,K-1\}$, set $\mathcal{F}_k^s = \sigma(z^0_{1/2},\dots,z^0_{K-1/2},\dots,z^s_{1/2},\dots,z^s_{k+1/2})$, $u^s_{k+1} = \tau[F(w^s) - F_{\xi^s_k}(w^s) - F(z^s_{k+1/2}) + F_{\xi^s_k}(z^s_{k+1/2})]$, which help us write

$$\begin{aligned}\mathbb{E}\max_{z\in\mathcal{C}}\sum_{s=0}^{S-1}\sum_{k=0}^{K-1} e(z,k) &= \mathbb{E}\max_{z\in\mathcal{C}}\sum_{s=0}^{S-1}\sum_{k=0}^{K-1}\tau\langle F(w^s) - F_{\xi^s_k}(w^s) - F(z^s_{k+1/2}) + F_{\xi^s_k}(z^s_{k+1/2}), z - z^s_{k+1/2}\rangle \\ &= \mathbb{E}\max_{z\in\mathcal{C}}\sum_{s=0}^{S-1}\sum_{k=0}^{K-1}\langle u^s_{k+1}, z\rangle - \sum_{s=0}^{S-1}\sum_{k=0}^{K-1}\mathbb{E}\langle u^s_{k+1}, z^s_{k+1/2}\rangle = \mathbb{E}\max_{z\in\mathcal{C}}\sum_{s=0}^{S-1}\sum_{k=0}^{K-1}\langle u^s_{k+1}, z\rangle,\end{aligned}$$

where the last equality is by the tower property, $\mathcal{F}^s_k$-measurability of $z^s_{k+1/2}$ and $\mathbb{E}_{s,k}[u^s_{k+1}] = 0$.

We apply Lemma 6.21 with the specified $\mathcal{F}^s_k$, u^s_{k+1} to obtain

$$\begin{aligned}\mathbb{E}\max_{z\in\mathcal{C}}\sum_{s=0}^{S-1}\sum_{k=0}^{K-1} e(z,k) &\le \max_{z\in\mathcal{C}} D(z,z_0) + \sum_{s=0}^{S-1}\sum_{k=0}^{K-1}\tau^2\mathbb{E}\|F_{\xi^s_k}(z^s_{k+1/2}) - F_{\xi^s_k}(w^s) + F(w^s) - F(z^s_{k+1/2})\|_*^2 \\ &\le \max_{z\in\mathcal{C}} D(z,z_0) + \sum_{s=0}^{S-1}\sum_{k=0}^{K-1} 4\tau^2\mathbb{E}\|F_{\xi^s_k}(z^s_{k+1/2}) - F_{\xi^s_k}(w^s)\|_*^2 \qquad (6.87) \\ &\le \max_{z\in\mathcal{C}} D(z,z_0) + \sum_{s=0}^{S-1}\sum_{k=0}^{K-1} 4\tau^2L^2\mathbb{E}\|z^s_{k+1/2} - w^s\|^2 \qquad (6.88) \\ &\le \max_{z\in\mathcal{C}} D(z,z_0) + \frac{8\tau^2L^2}{(1-\alpha)(1-\gamma^2)}\Phi^0(z_\star), \qquad (6.89)\end{aligned}$$

where (6.87) is due to the tower property and $\mathbb{E}\|X - \mathbb{E}X\|_*^2 \le 2\mathbb{E}\|X\|_*^2 + 2\|\mathbb{E}X\|_*^2 \le 4\mathbb{E}\|X\|_*^2$, which follows from triangle inequality, Young's inequality, and Jensen's inequality. Moreover, (6.88) is by variable Lipschitzness of F_ξ, and the last step is by Lemma 6.20. Consequently, by $\Phi^0(z_\star) \le \max_{z\in\mathcal{C}}\Phi^0(z) = (\alpha + K(1-\alpha))\max_{z\in\mathcal{C}} D(z,z_0)$ and $\tau^2L^2 = (1-\alpha)\gamma^2$ we have

$$\begin{aligned}\tau KS\mathbb{E}\left[\operatorname{Gap}(z^S)\right] &\le \max_{z\in\mathcal{C}}\left[D(z,z_0) + \left(1 + \frac{8\tau^2L^2}{(1-\alpha)(1-\gamma^2)}\right)\Phi^0(z)\right] \\ &= \left(1 + \left(1 + \frac{8\gamma^2}{1-\gamma^2}\right)(\alpha + K(1-\alpha)\right)\max_{z\in\mathcal{C}} D(z,z_0). \qquad ■\end{aligned}$$

Proof of Corollary 6.23. As $\alpha = 1 - \frac{1}{K}$, it holds that $\alpha + K(1-\alpha) = 1 - \frac{1}{K} + 1 \le 2$. With this, from Theorem 6.22 it follows

$$\begin{aligned}\mathbb{E}\left[\operatorname{Gap}(z^S)\right] &\le \frac{1}{\tau KS}\left(1 + \left(1 + \frac{8\gamma^2}{1-\gamma^2}\right)(\alpha + K(1-\alpha)\right)\max_{z\in\mathcal{C}} D(z,z_0) \\ &\le \frac{L}{\sqrt{K}\gamma S}\left(3 + \frac{16\gamma^2}{1-\gamma^2}\right)\max_{z\in\mathcal{C}} D(z,z_0) = \mathcal{O}\left(\frac{L}{\sqrt{N}S}\right). \qquad (6.90)\end{aligned}$$

One epoch requires one evaluation of F and $2K$ of F_ξ, thus in total we have $(N + 2K)\mathrm{Cost} = 2N\mathrm{Cost}$. To reach ε accuracy, we need $\left\lceil\mathcal{O}\left(\frac{L}{\sqrt{N}\varepsilon}\right)\right\rceil$ epochs. Hence, the final complexity is

$\mathcal{O}\left(\text{Cost} \times \left(N + \frac{L\sqrt{N}}{\varepsilon}\right)\right)$. By setting $\gamma = \frac{1}{3}$ in (6.90), we get specific constants. We have

$$\mathbb{E}\left[\operatorname{Gap}(z^S)\right] \le \frac{15L}{\sqrt{K}S} \max_{z\in\mathcal{C}} D(z,z_0) = \frac{15\sqrt{2}L}{\sqrt{N}S} \max_{z\in\mathcal{C}} D(z,z_0).$$

Since $30\sqrt{2} < 43$, the final complexity is $\text{Cost} \times \left(2N + \frac{43\sqrt{N}L}{\varepsilon} \max_{z\in\mathcal{C}} D(z,z_0)\right)$. ■

6.7 Bibliographic note

For the results in Section 6.4.3, the main contributions of the author of this dissertation are derivation of the improved complexity results in Euclidean case and designing and analyzing the algorithm for Bregman case. The design of the algorithm in the Euclidean case (Algorithm 6.2) (in particular, the idea of using $\bar{z}_k$ for improving the step size) and the initial proof for the Euclidean case (in particular, Lemma 6.13) are due to Yura Malitsky.

7 Sample complexity of two-player zero sum Markov Games

In this chapter, we focus on two player zero-sum Markov games, with applications in competitive reinforcement learning (RL). Unlike the previous chapters, this problem can be nonconvex-nonconcave, however it has a special structure that ensures tractability. We provide some preliminary results on improving the sample complexity of *policy gradient methods* for this problem. Our analysis have intimate connections to the techniques presented in Chapters 4–6.

Therefore, we believe that this chapter is a good illustration of importance of fundamental techniques for convex-concave problems, even when solving nonconvex problems that arise in modern machine learning.

This chapter is based on the unpublished joint work with Niao He, Luca Viano and Volkan Cevher.

7.1 Introduction

Markov game framework is introduced by [Sha53] as *stochastic games* and popularized in RL with [Lit94]. In the basic form of the model, two agents with competing interests interact in an environment where the reward and the state transition depend on the actions of both players. Even with this simplicity, such systems have seen impressive success for example in game-playing and robotics [KBP13, SSS$^+$17, MKS$^+$15, VBC$^+$19, BS19].

While value-based methods [SWYY20, BJY20, BJ20, XCWY20, TWYS20] offer near-optimal guarantees, the policy gradient (PG) methods, including actor-critic (AC) algorithms have found limited use in the zero-sum Markov games despite their model-free and easy-to-implement structure, their flexibility and generality [SLA$^+$15, SWD$^+$17, WBH$^+$17].

The PG methods [Kak01, SMSM00] directly optimize the value function in the policy space— a non-convex optimization problem even in the basic tabular, single agent setting. Intriguingly, recent results [AKLM20, CCC$^+$20, MXSS20, BR19, BR21, XWL20b, Lan21, KDMR21, HWWY20, XWL20a, KDMR21] demonstrate globally optimal convergence of PG methods by identifying hidden convexity, including extensions to the multi-agent setting [DFG20, WLZL21, ZTLD21].

The existing results on PG methods for tabular two-player zero-sum Markov games mostly focus on decentralized algorithms with sample complexities $\tilde{\mathcal{O}}(\epsilon^{-12.5})$ [DFG20], $\tilde{\mathcal{O}}(\epsilon^{-8})$ [WLZL21], and even $\tilde{\mathcal{O}}(\epsilon^{-4})$ with more restrictions [WLZL21]; see Section 7.1.2 for the details. With function approximation, [ZTLD21] obtains $\tilde{\mathcal{O}}(\epsilon^{-6})$ sample complexity when given access to unbiased sampling oracles for the value functions.

On the other hand, the best-known sample complexity for converging to a globally optimal policy in the single agent problem is $\tilde{\mathcal{O}}(\epsilon^{-2})$ in the tabular case [Lan21]. As this complexity is achieved by value-based/model-based methods in the multi-agent setting [SWYY20, ZKBY20], one expects a similar complexity to be attainable for policy-based methods. Our work precisely bridges this gap and develops policy gradient methods whose performance for multi-agent RL is closer to their single agent counterparts.

7.1.1 Contributions

We propose an algorithm based on actor-critic framework for solving two-player zero-sum Markov games in the tabular case, that match the best-known sample complexity results to find a globally optimal policy in the single agent setting [Lan21, KDMR21, HWWY20, XWL20b].

Surprisingly, we achieve these results—to our knowledge, for the first time with policy gradient methods—mostly by a careful adaptation of the recent results for policy gradient methods in single agent setting, temporal difference learning, error propagation framework of policy iteration, and by employing techniques from stochastic primal-dual optimization in the two-stage framework of [PSPP15].

These developments require a careful algorithm design and analysis. In particular, two-stage nature of the algorithm incurs biases between the stages that we have to control carefully. Obtaining $\tilde{\mathcal{O}}(\epsilon^{-2})$ complexity requires a tighter analysis for both stages of the algorithm, with strict control on the aforementioned bias. Therefore, it requires more advanced techniques and algorithms, inspired from the stochastic primal-dual optimization literature. We explicitly highlight our important new techniques as *insights* in the sequel. The full proofs are included in the appendices.

7.1.2 Related works

Policy gradient methods. Recently, there is growing interest in global convergence of policy gradient methods in the single agent setting. In particular, several papers have shown convergence rates of natural policy gradient (NPG) methods in the tabular setting with assuming access to exact value function oracle [AKLM20, CCC$^+$20, MXSS20, BR19, BR21] and when value functions are estimated from data [SEM20, XWL20b, Lan21, KDMR21, HWWY20, XWL20a, KDMR21]. To our knowledge, the best sample complexity for NPG methods with inner loop for policy evaluation is $\tilde{\mathcal{O}}(\epsilon^{-2})$ and is due to [Lan21]. For two time-scale NAC, the best sample complexity is $\tilde{\mathcal{O}}(\epsilon^{-4})$ as obtained in [KDMR21, HWWY20, XWL20b] (see also [KCM21, Table 1]). For a general overview of results in multi agent RL we refer to [ZYB21] and here we only cover the results most related to ours.

Policy gradient methods for two-player zero-sum Markov games. With the positive results on global convergence of PG methods for single agent problems, translating these results to the competitive multi-agent setting has been the goal of many recent works. In particular, independent policy gradient methods where the agents are interacting symmetrically has been considered in [DFG20, WLZL21]. The work of [DFG20] built on [AKLM20] by using REINFORCE gradient estimator [Wil92] and obtained sample complexity of $\mathcal{O}(\epsilon^{-12.5})$ for reaching to one-sided Nash equilibrium.

The algorithm of [WLZL21] built on optimistic gradient descent-ascent (GDA) method combined with a running estimate of the value function, obtaining $\tilde{\mathcal{O}}(\epsilon^{-8})$ sample complexity for obtaining a policy pair with small duality gap. In addition, [WLZL21] showed that one can improve this complexity to $\tilde{\mathcal{O}}(\epsilon^{-4})$ when restricted to Euclidean setup with metric subregularity assumption. There are two drawbacks of this result: First, as pointed out in [DFG20], metric subregularity constant can be arbitrarily small, resulting in degradation of the rate. Second, as also pointed out by [WLZL21], this result is limited to Euclidean setting and cannot be extended to the NPG with softmax policy update, which is non-Euclidean. The algorithm can be seen similar to the gradient ascent algorithm in [AKLM20]. As shown in [AKLM20] for single agent problems, NPG methods have much better convergence properties than Euclidean projected gradient ascent methods. For comparison with the works in [DFG20, WLZL21], we also refer to Remark 7.1.

Another very related work to ours is [ZTLD21] which considered *(i)* tabular setting with exact value functions and *(ii)* online setting with function approximation, also using the error propagation scheme of [PSPP15]. Building on [AKLM20], this work showed $\tilde{\mathcal{O}}(\epsilon^{-6})$ sample complexity with function approximation, with access to unbiased samples of the value functions. However, the sample complexity in the tabular case is not characterized in this paper and transferring the result obtained for function approximation would give $\tilde{\mathcal{O}}(\epsilon^{-6})$ sample complexity with access to unbiased samples for the value functions. In contrast, we focus on the tabular setting and analyze the sample complexity when we do not have access to unbiased value function oracles. Indeed, lack of unbiased samples for value functions required us to use new *insights* described in the sequel, to derive the tighter complexities $\tilde{\mathcal{O}}(\epsilon^{-2})$.

7.2 Preliminaries

Notation. We consider the tabular setting with finite state and action spaces denoted by S, A, B. The discount factor is $\gamma < 1$. The policy of the min agent is denoted as x and the max agent as y with actions sets A, B, respectively. Interaction of the agents is as follows: At state s, both agents take actions independently of each other $a \sim x(\cdot|s)$ and $b \sim y(\cdot|s)$. Based on the actions, the environment transitions to the next state $s' \sim P(\cdot|s, a, b)$ and the agents receive reward $r(s, a, b)$. Given a policy pair x, y, we denote the stationary state distribution induced by the pair as $\rho^{x,y}$. We use $e(s_t) \in \mathbb{R}^{|\mathcal{S}|}$ to denote the unit vector such that $e(s) = 1$ if $s = s_t$ and $e(s) = 0$, if $s \neq s_t$. We use the same notation for $e(s_t, a_t)$. Given a policy x, we sometimes use the notation x^s for $x(\cdot|s)$. The value function for state s is defined as

$$V^{x,y}(s) = \mathbb{E}_{x,y}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, b_t)|s_0 = s\right],$$

where $\mathbb{E}_{x,y}$ is taking over random variables s_t, a_t, b_t for all $t \geq 0$ as $a_t \sim x(\cdot|s_t)$, $b_t \sim y(\cdot|s_t)$ and $s_{t+1} \sim P(\cdot|s_t, a_t, b_t)$. Similarly, the action value function is defined as $Q^{x,y}(s, a, b) = \mathbb{E}_{x,y}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, b_t)|s_0 = s, a_0 = a, b_0 = b\right]$.

With these definitions, we can now state the formal problem. For all s, we aim to solve

$$\min_{x(\cdot|s)\in\Delta} \max_{y(\cdot|s)\in\Delta} V^{x,y}(s).$$

Nash equilibrium. We assume the existence of a pair of policies $x^\star, y^\star$ that are Nash equilibrium, namely, for all s, $V^{x^\star,y}(s) \leq V^{x^\star,y^\star}(s) \leq V^{x,y^\star}(s)$. We refer to $V^{x^\star,y^\star}$ as $V^\star$ for lighter notation. In this chapter, we are interested in finding a one-sided Nash equilibrium, similar to [DFG20, ZTLD21, ZYB19, BRM19]. As mentioned in [DFG20], for the other player, one can rerun the algorithm by switching roles to have the guarantee for both players. In particular, for an initial state distribution μ, we seek for x_{out} such that

$$\mathbb{E}_{s_0\sim\mu}[\max_y V^{x_{out},y}(s_0) - V^\star(s_0)] \leq \epsilon.$$

Interaction procedure. We use the interactions of the agents with the environment to estimate the value functions and related oracles for the running of the algorithm. At each interaction episode, agents have access to $(s_i, a_i, r(s_i, a_i, b_i), s_{i+1})$ and $(s_i, b_i, r(s_i, a_i, b_i), s_{i+1})$, respectively. In terms of access of agents, our oracle model is similar to [DFG20, WLZL21]. However, one difference in our case is that we require a *game etiquette*: Our algorithms have two stages where the agents have to behave differently (see Section 7.2). In particular, in the second stage of our algorithms, one agent fixes its policy as the other agent tries to find an approximate best response. In the first stage, both policies are updated simultaneously. As long as this etiquette is respected by the agents, they do not need further communication.

Bregman distances and softmax update rule. For convenience, we use the formalism of Bregman distances [NY83]. Given Bregman distance $D(\cdot,\cdot)$ and vector $g(s,\cdot)$, the update rule

$$x_{t+1}(\cdot|s) = P(x_t(\cdot|s), g(s,\cdot)) = \arg\min_{x(\cdot|s)\in\Delta} \langle g(s,\cdot), x(\cdot|s)\rangle + D(x(\cdot|s), x_t(\cdot|s)), \tag{7.1}$$

corresponds to the softmax update when D is chosen to be the KL divergence. When $g = Q^{x_t}$, this update is known as NPG with softmax parameterization [AKLM20, Lemma 5.1]. To simplify our bounds, we instantiate the constants throughout the chapter when D is KL divergence. However, arguments in our developments would also hold for arbitrary Bregman distances as in [ZCH$^+$21].

Assumption 7.1.
(i) For given state distributions μ, σ, the concentrability coefficients are bounded [PSPP15]:

$$\sup_j \sup_{x_1, y_1, \ldots, x_j, y_j} \left\| \frac{\mu P_{x_1,y_1} \ldots P_{x_j,y_j}}{\sigma} \right\|_\infty =: C_{\mu,\sigma} < +\infty.$$

(ii) There exists $\underline{\rho}$ such that, for any policy pair x, y, $\rho^{x,y} \geq \underline{\rho} > 0$, where $\rho^{x,y}$ is the stationary state distribution induced by the policy pair.

(iii) There exists $\underline{x}, \underline{y}$ such that, for any policy pair x, y, $x \geq \underline{x} > 0, y \geq \underline{y} > 0$.

(iv) $r(s,a,b) \leq 1$.

Our rationale on the assumptions. Assumption 7.1(ii) and (iii) essentially mean positive definiteness of the sampling matrices in policy evaluation. To our knowledge, some form of this assumption is required in most of the existing work with best-complexity on TD-type methods [BRS18, XWL20b, KCM21, Lan21, HWWY20, XWL20a, WZXG20, ZXL19] in the single-agent setting. An alternative to Assumption 7.1(iii), as proposed in [KDMR21], introduces ϵ-greedy exploration with a certain deterioration in the rate. From an analysis perspective, the use of ϵ is the same as the use of Assumption 7.1(iii). Therefore, to be consistent with most of the literature, we use Assumption 7.1(iii) (See also [Lan21, Remark 1, Section 5.2]). The assumption (iv) is for simplicity. In the sequel, we will use the parameters $\lambda^\theta_{\min}, \lambda^v_{\min}$ depending on $\underline{x}, \underline{y}, \underline{\rho}$ from Assumption 7.1 and $\lambda^\omega_{\min}$ only depends on $\underline{\rho}$. These are the minimum eigenvalues of the sampling matrices in policy evaluation routines.

Remark 7.1. Among the related works for policy gradient methods in multi agent setting, Assumption 7.1(ii) is required in [WLZL21] but not Assumption 7.1(iii). A different assumption is made in [DFG20] regarding the minimum probability of the game stopping at any state action pair being nonzero: therefore one should be careful while comparing complexities. These works use ϵ-greedy exploration instead of Assumption 7.1(iii). To avoid Assumption 7.1(iii), we can also use greedy exploration [KDMR21], [Lan21, Remark 1], with degradation in ϵ dependence, that we omit for brevity. To compare with single agent complexities, we keep Assumption 7.1(iii). One could also do an algorithm-specific analysis, similar to [MXSS20, Lemma 5, Lemma 9] to characterize when Assumption 7.1(iii) holds.

A relaxed form of concentrability coefficient is used in [DFG20]. In [WLZL21], the sample complexity bound does not have dependence on concentrability coefficient, however, the bounds in [WLZL21, Theorem 1, 2] have $|S|$ dependence even with access to true value functions. Indeed, concentrability coefficient can be bounded by $|S|$ by picking σ accordingly [Mun03]. As the existing results on policy gradient methods for Markov games already have a pessimistic dependence on $|S|$ [DFG20, WLZL21], it seems this additional dependence on our results is not too problematic in terms of final dependence on $|S|$.

Error propagation for approximate dynamic programming In [PSPP15], an error propagation analysis is conducted for an approximate version of generalized policy iteration, for zero-sum Markov games. In particular, the authors showed that the following two-stage algorithm will converge:

∘ *Greedy step:* Given a fixed value function V_{k-1}, find the policy pair which is an ϵ-equilibrium.

$$\min_{x^s\in\Delta}\max_{y^s\in\Delta}\sum_{a,b} x(a|s)y(a|s)Q_{k-1}(s,a,b) =: x^s Q^s_{k-1} y^s, \tag{7.2}$$

where $Q_{k-1}(s,a,b) = r(s,a,b) + \gamma\sum_{s'} P(s'|s,a,b)V_{k-1}(s')$. When it is clear from the context, we drop the subscript of Q_{k-1}. This problem is a matrix game and is notably the sample-complexity bottleneck [PSPP15]. Let us denote by ϵ_g the accuracy for this step and x_k as the output at iteration k:

$$\mathbb{E}\mathbb{E}_{s\sim\sigma}[\max_{y^s\in\Delta} x^s_k Q^s_{k-1} y^s - \min_{x^s\in\Delta}\max_{y^s\in\Delta} x^s Q^s_{k-1} y^s] = \epsilon_{g,k}(s),$$

where the expectation is over the randomness of the specific algorithm used to generate x_k.

∘ *Evaluation step:* This step consists of finding an approximate best response. As one policy is fixed (x_k), one can view the fixed policy as a part of the environment. Denote y_k as the approximate best-response computed in this step. The resulting value function $V_k = V^{x_k,y_k}$ is fed to the greedy step in the next iteration. Let us denote by ϵ_e the accuracy for this step and y^*_k the best response:

$$\mathbb{E}_{s\sim\sigma}[V^{x_k,y^*_k}(s) - V^{x_k,y_k}(s)] = \epsilon_{e,k}(s),$$

where the expectation is taking over the randomness of the algorithm used to generate y_k. Then, [PSPP15, Theorem 1], [ZTLD21] have shown that the following inequality holds.

$$\mathbb{E}\mathbb{E}_{s\sim\mu}[\max_{y\in\Delta} V^{x_K,y}(s) - V^\star(s)] \le \frac{C_{\mu,\sigma}K}{1-\gamma}\tilde{\mathcal{O}}\left(\sup_{k\le K}\epsilon_{g,k} + \sup_{k\le K}\epsilon_{e,k}\right) + \mathcal{O}\left(\frac{C_{\mu,\sigma}\gamma^K}{1-\gamma}\right). \tag{7.3}$$

Natural policy gradient. As we work in the tabular setting, in this chapter, we focused on the natural policy gradients [Kak01] in softmax parameterization which admits a simple update rule. In particular, the update rule for NPG in single agent setting is [AKLM20, Lemma 5.1]

$$\pi_{t+1}(\cdot|s) \propto \pi_t(\cdot|s)\exp(\eta Q^{\pi_t}(s,\cdot))$$

which corresponds to the more general update (7.1) when Bregman distance D in (7.1) is chosen to be the KL divergence. To get a sample-based version of this algorithm, one needs to learn $Q^{\pi_t}(s,a) = \mathbb{E}_{\pi_t}\left[\sum_{t=0}^\infty \gamma^t r(s_t,a_t)|s_0 = s, a_0 = a\right]$ typically in an inner policy evaluation loop as in [Lan21]. This approach is sometimes referred to as natural actor critic (NAC). Note that the update rule in [AKLM20, Lemma 5.1] is written with the advantage function, however, due to softmax parameterization, it is equivalent to the form we give.

Temporal difference learning. For constructing action value function from samples, we will use temporal difference learning and in particular TD(0) [Sut88, BRS18, TVR97]. This algorithm can be seen as a stochastic approximation scheme for solving a linear equation [TVR97, Lan21]. In particular, by denoting the stationary state distribution under π as ρ^π, we define

$$F^\pi(\theta)(s,a) = \rho^\pi(s)\pi(a|s)\Big(\theta(s,a) - r(s,a) - \gamma \sum_{s',a'} P(s'|s,a)\pi(a'|s')\theta(s',a')\Big).$$

First, we note that $F^\pi(\theta^\star)=0$ where $\theta^\star = Q^\pi$. Under Assumption 7.1(ii, iii), it is well-known that F^π is strongly monotone (see [BRS18, Lemma 3], [Lan21, Section 5.2]. The main tools to show this are Assumption 7.1(ii, iii) and Bellman operator being γ-contraction. Then, one can use for example [BC11, Example 22.6, Example 20.7].

One can sample $s_t \sim \rho^\pi$, $a_t \sim \pi(\cdot|s_t)$ and $s_{t+1} \sim P(\cdot|s_t,a_t)$, $a_{t+1} \sim \pi(\cdot|s_{t+1})$ and form the stochastic oracle $\tilde{F}(\theta_t) = e(s_t,a_t)\left(\theta_t(s_t,a_t) - r(s_t,a_t) - \gamma\theta_t(s_{t+1},a_{t+1})\right)$ for TD(0). Note that under i.i.d. assumption, $\tilde{F}(\theta_t)$ is an unbiased estimate of $F^\pi(\theta_t)$. The results for TD(0) can be extended to Markovian setting without the i.i.d. assumption by using a uniform mixing assumption [BRS18].

Our approach. We introduce an algorithm in natural actor critic (NAC) framework with inner loops for policy evaluation [KT00, PS08]. By solving the abovementioned two steps, we will obtain an approximate Nash equilibrium. We will leverage forward-reflected-backward algorithm to solve the matrix game in the evaluation step efficiently [MT20b]. For estimating value functions used as oracles in these algorithms, we employ TD(0) [TVR97, BRS18, Sut88]. To our knowledge, the best complexities in the single agent setting are obtained with this approach [Lan21, HWWY20, KCM21].

Markovian bias. We assume that we can sample from the steady state distribution of a given policy. With normal interaction with the environment, this is not the case and we obtain a single stream of data. As a result, TD(0) update is biased—commonly referred to as the Markovian bias. A large body of literature in the single agent literature showed that the effect of this bias in TD(0) update is essentially additive and can be handled by assuming uniform mixing of the induced Markov chain [WZXG20, BRS18, ZXL19, KDMR21, XWL20b, XWL20a]. These analyses apply to our policy evaluation routines, extending them to the Markovian setting. For simplicity, we show our techniques with i.i.d. assumption, which can be extended to Markovian data with the uniform mixing assumption.

7.3 A reflected natural actor-critic algorithm with a game etiquette

Greedy step. In this step, at iteration k, we compute an approximate equilibrium of the matrix game (7.2). As V_{k-1} is fixed, this is a standard matrix game and throughout this loop, we omit the dependence of Q on k and leave it implicit. Let us denote the oracle of x-player for solving (7.2): $\theta^x_{\star,t}(s,a) = \mathbb{E}_{b\sim y_t(\cdot|s)}Q(s,a,b) = \sum_b y_t(b|s)r(s,a,b) + \gamma\sum_{s',b} y_t(b|s)P(s'|s,a,b)V_{k-1}(s')$. Note that unlike standard stochastic setting, we do not have an unbiased estimate of $\theta^x_{\star,t}$ but instead, we have inner loops to learn this oracle. With this oracle (that we will see next how to obtain), the update of FoRB will be

$$x_{t+1}(\cdot|s) = P(x_t(\cdot|s), \eta\left(2\theta^x_{t+1}(s,\cdot) - \theta^x_t(s,\cdot)\right)). \tag{7.4}$$

We will give an analysis of FoRB without unbiased oracles and we will take special care for the stochastic dependency to make sure to decompose bias and variance (See Insight 1).

We now show how to compute this oracle for x player without accessing to actions or policy of y. The same reasoning applies to y update. Similar to [Lan21], using the sampling matrix $\operatorname{diag}(\rho^{x_t,y_t}) \otimes \operatorname{diag}(x_t)$, we define the operator

$$F^x_t(\theta^x)(s,a) = \rho^{x_t,y_t}(s)x_t(a|s)\big(\theta^x(s,a) - \sum_b y_t(b|s)r(s,a,b) - \gamma \sum_{s',b} y_t(b|s)P(s'|s,a,b)V_{k-1}(s')\big). \tag{7.5}$$

First, it holds that $F^x_t(\theta^x_{\star,t}) = 0$. Moreover, F^x_t is strongly monotone as $\min_{s,a} \rho^{x_t,y_t}(s)x_t(a|s)$ is separated from 0, by Assumption 7.1(ii), (iii). One important point here is that we do not have access to an unbiased sample of V_{k-1} as it is the value function depending on x_{k-1}, y_{k-1}. Instead, we will use a potentially biased estimate $\hat{V}_{k-1}$. After obtaining a sample $\xi_n = (s_n, a_n, b_n, s_{n+1})$, we define the stochastic operator for x-player

$$\tilde{F}^x_t(\theta^x_n, \xi_n) = e(s_n,a_n)\Big(\theta^x_n(s_n,a_n) - r(s_n,a_n,b_n) - \gamma \hat{V}_{k-1}(s_{n+1})\Big). \tag{7.6}$$

Assuming access to i.i.d. samples, we see that expectation of this operator w.r.t. ξ_n gives

$$\mathbb{E}_{\xi_n}[\tilde{F}^x_t(\theta^x_n, \xi_n)] = F^x_t(\theta^x_n) + \delta_{v,t}, \tag{7.7}$$

with $\delta_{v,t} = \gamma \sum_{s,a,b,s'} \rho^{x_t,y_t}(s)x_t(a|s)y_t(b|s)P(s'|s,a,b)e(s,a)\left(V_{k-1}(s') - \hat{V}_{k-1}(s')\right) = \gamma P_{x_t,y_t}(V_{k-1} - \hat{V}_{k-1})$.

Formulation (7.6) ensures that x agent only accesses $s_n, r(s_n,a_n,b_n), s_{t+n}$ and its own action a_n to form the stochastic oracle and y accesses $s_n, r(s_n,a_n,b_n), s_{n+1}$ and its own action b_n. We have additional bias coming from the approximation of V_{k-1} by $\hat{V}_{k-1}$, the estimation of which is important for getting our complexity results. Markovian data would bring additional bias as mentioned before.

Evaluation step. In this step, x player fixes its policy and y computes an approximate best response. This step will output V_{k-1} will be the value function of the policy pair that will be the output of this step. In our algorithm, the players will remember these policies because they will need to compute an estimate of V_{k-1} with small bias in the greedy step.

In particular, at iteration t, agents will be interacting with policies x_k and y_t, since x player keeps its policy fixed. The natural policy gradient for the y-player is $v_{\star,t} = \mathbb{E}_{a\sim x_k(\cdot|s)} Q^{x_k,\bar{y}_t}(\cdot,a,\cdot)$:

$$v_{\star,t}(s,b) = \sum_a x_k(a|s)r(s,a,b) + \gamma \sum_{s',a,b'} P(s'|s,a,b)x_k(a|s)\bar{y}_t(b'|s')v_{\star,t}(s',b')$$

Algorithm 7.1 Reflected NAC with a game etiquette

```
Require: $P$ defined in (7.1) in Section 7.1. Subroutine Policy-Eval (see Algorithm 7.2).
  Initial policies $x_0, y_0, \bar{y}_0$
  for $k = 0, 1, \dots$ do
    Greedy step
    for $t = 0, 1, \dots, T-1$ do
      $[\hat{V}^x_{k-1}, \hat{V}^y_{k-1}] = [\texttt{Policy-Eval}(x_{k-1}, y_{k-1}, N, \beta^\omega_n), \texttt{Policy-Eval}(x_{k-1}, y_{k-1}, N, \beta^\omega_n)]$
      $[\theta^x_{t+1}, \theta^y_{t+1}] = [\texttt{Policy-Eval}(x_t, y_t, N, \hat{V}^x_{k-1}, \beta^\theta_n), \texttt{Policy-Eval}(x_t, y_t, N, \hat{V}^y_{k-1}, \beta^\theta_n)]$
      $x_{t+1}(\cdot|s) = P(x_t(\cdot|s), \eta\left(2\theta^x_{t+1}(s,\cdot) - \theta^x_t(s,\cdot)\right))$
      $y_{t+1}(\cdot|s) = P(y_t(\cdot|s), -\eta\left(2\theta^y_{t+1}(s,\cdot) - \theta^y_t(s,\cdot)\right))$
    end for
    Output $x_k = \frac{1}{T}\sum_{t=1}^T x_t$.
    Evaluation step
    for $t = 0, 1, \dots, T-1$ do
      $v_{t+1} = \texttt{Policy-Eval}(x_k, \bar{y}_t, N, \beta = \beta^v_n)$
      $\bar{y}_{t+1}(\cdot|s) = P(\bar{y}_t(\cdot|s), -\eta v_{t+1}(s,\cdot))$
    end for
    Output $y_k = \bar{y}_{\hat{t}}$, where $\hat{t} \in [T]$ is selected uniformly at random.
  end for
```

We use the sampling matrix (as [Lan21, Sec. 5.2]) $D(\rho^{x_k,\bar{y}_t}) \otimes D(\bar{y}_t)$ and define the operator

$$F^v_t(v_t)(s,b) = \rho^{x_k,\bar{y}_t}\bar{y}_t(b|s)\Big[v_t(s,b) - \sum_a x_k(a|s)r(s,a,b) \\ - \gamma \sum_{s',a,b'} x_k(a|s)P(s'|s,a,b)\bar{y}_t(b'|s')v_t(s',b')\Big],$$

such that $F^v_t(v_{\star,t}) = 0$. Strong monotonicity of F_t follows from Assumption 7.1(ii), (iii), and that the operator $Tv(s,b) = \sum_a x_k(a|s)r(s,a,b) + \gamma\sum_{s',a,b'} x_k(a|s)P(s'|s,a,b)\bar{y}_t(b'|s')v(b',s')$ is γ contraction in ℓ_∞ norm [BC11, Example 22.6 and 20.7]. We define the stochastic operator after sampling $s_n \sim \rho^{x_k,\bar{y}_t}$, $a_n \sim x_k(\cdot|s_n)$, $b_n \sim \bar{y}_t(\cdot|s_n)$, $s_{n+1} \sim P(\cdot|s_n,a_n,b_n)$, $b_{n+1} \sim \bar{y}_t(\cdot|s_{n+1})$

$$\tilde{F}^v_t(v_n, \xi_n) = e(s_n,b_n)\left(v_n(s_n,b_n) - r(s_n,a_n,b_n) - \gamma v_n(s_{n+1},b_{n+1})\right),$$

and as we assume we sample $s_n \sim \rho^{x_k,\bar{y}_t}$, $\mathbb{E}_{\xi_n}[\tilde{F}^v_t(v_n,\xi_n)] = F^v_t(v_n)$. In particular, we see that as long as $s_n, a_n, b_n, s_{n+1}, b_{n+1}$ are estimated in the prescribed way, there is no need for $\bar{y}_t$ update to see the actions or policy of x_k for $\tilde{F}^v_t(v_n,\xi_n)$ to be unbiased estimate of $F^v_t(v_n)$.

Remark 7.2. At the evaluation step, policy and the value function at a random iterate is outputted. We do not need to run for T iterations and store all the variables, instead, we can compute $\hat{t}$ before starting the algorithm, which is standard, see [FB19, Remark 5].

Remark 7.3. For the best complexity, our analysis requires fresh estimates of $\hat{V}_{k-1}$ at every iteration (see Algorithm 7.1 and Insight 4). This allows us to obtain a tight bound for the bias and get the $\mathcal{O}(\epsilon^{-2})$ complexity. Without fresh samples, Algorithm 7.1 would have a $\mathcal{O}(\epsilon^{-3})$

Algorithm 7.2 `Policy-Eval` (y-player)

```
Require: Policy pair $x, y$, iteration counter $N$, oracle $\hat{V}_{k-1}$, step size $\beta$
  for $n = 0, 1, \dots, N-1$ do
    Sample $s_n \sim \rho^{x,y}(\cdot)$, $a_n \sim x(\cdot|s_n)$, $b_n \sim y(\cdot|s_n)$, $s_{n+1} \sim P(\cdot|s_n, a_n, b_n)$.
    if $\beta = \beta_n^{\omega}$ then
      $\tilde{F}(\varphi_n, \xi_n) = e(s_n)\big(\varphi_n(s_n) - r(s_n, a_n, b_n) - \gamma\varphi_n(s_{n+1})\big)$
    else if $\beta = \beta_n^{\theta}$ then
      $\tilde{F}(\varphi_n, \xi_n) = e(s_n, b_n)\big(\varphi_n(s_n, b_n) - r(s_n, a_n, b_n) - \gamma\hat{V}_{k-1}(s_{n+1})\big)$
    else if $\beta = \beta_n^{\nu}$ then
      Sample also $b_{n+1} \sim y(\cdot|s_{n+1})$.
      $\tilde{F}(\varphi_n, \xi_n) = e(s_n, b_n)\big(\varphi_n(s_n, b_n) - r(s_n, a_n, b_n) - \gamma\varphi_n(s_{n+1}, b_{n+1})\big)$
    end if
    $\varphi_{n+1} = \varphi_n - \beta_n\tilde{F}(\varphi_n, \xi_n)$
  end for
Ensure: $\varphi_N$
```

complexity. This insight is in contrast to the black box view of [PSPP15], which uses an estimate of V_{k-1} from the evaluation step within the greedy step. Our analysis behooves both agents to remember the output policies of evaluation step instead, so that they can recompute V_{k-1} with a lower bias in the greedy step.

7.3.1 Convergence of Reflected NAC with a game etiquette

Theorem 7.4. *(See Theorem 7.14) Let Assumption 7.1 hold. For Algorithm 7.1, for the output of x-player*

$$\mathbb{E}\mathbb{E}_{s_0\sim\mu}[\max_y V^{x_k,y}(s_0) - V^{\star}(s_0)] \leq kC_{\rho,\sigma}\tilde{O}\left(\frac{1}{T(1-\gamma)^3} + \frac{|S|^2(|A|^2 \vee |B|^2)}{N(1-\gamma)^5(\lambda^{\theta}_{\min}\lambda^{\omega}_{\min}\lambda^{\nu}_{\min})^2}\right) + O(\gamma^k)$$

which gives $\tilde{O}(\frac{C^2_{\mu,\sigma}|S|^2(|A|^2\vee|B|^2)}{\epsilon^2(1-\gamma)^8(\lambda^{\theta}_{\min}\lambda^{\omega}_{\min}\lambda^{\nu}_{\min})^2})$ *sample complexity.*

A critical point to derive the fastest rate as observed by [Lan21] in the single agent setting is to characterize the bias and variance separately. As the algorithm in [Lan21] would correspond to GDA when applied in our setting, we extend the ideas there to the FoRB algorithm.

Insight 1. The existing analyses for stochastic versions of FoRB are not suitable for us. In the stochastic variant in [MT20b], deterministic oracle is computed at each iteration. [BSCB20] uses unbiased oracles with bounded variance and decreasing step size. In our case, we will have biased oracles and we will use inner loops to decrease bias and variance. Therefore, we need to develop an analysis with constant step size and that characterizes the bias and variance explicitly in the next lemma.

We drop the superscripts from $\theta, \hat{V}_{k-1}$ (see Alg. 7.1) as estimations are symmetric. Define $e_{1,t} + e_{2,t} = \eta\langle\theta_{t+1}(\cdot|s) - \mathbb{E}[\theta_{t+1}(\cdot|s)|x_t], x(\cdot|s) - x_t(\cdot|s)\rangle - \eta\langle\theta^y_{t+1}(\cdot|s) - \mathbb{E}[\theta^y_{t+1}(\cdot|s)|y_t], y(\cdot|s) - y_t(\cdot|s)\rangle$.

Lemma 7.5. *(See Lemma 7.15) Let Assumption 7.1 hold. Denote* $x_{out} = \frac{1}{T}\sum_{t=1}^T x_t$ *and* $y_{out} = \frac{1}{T}\sum_{t=1}^T y_t$ *and let* $\eta = \frac{1-\gamma}{8}$

$$\mathbb{E}\mathbb{E}_{s\sim\mu}\left[\max_{x^s,y^s} x^s_{out}Q^s y^s - x^s Q^s y_{out}\right] = \tilde{\mathcal{O}}\left(\frac{1}{\eta T}\right) + \mathcal{O}\left(\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\mathbb{E}[\theta_{t+1}|x_t] - \theta_{\star,t}\|\right)$$
$$+ \mathcal{O}\left(\frac{\eta}{T}\sum_{t=1}^T \mathbb{E}\|\theta_{t+1} - \theta_{\star,t}\|^2 + \mathbb{E}\|\theta_t - \theta_{\star,t-1}\|^2\right) + \frac{1}{T\eta}\mathbb{E}\mathbb{E}_{s\sim\mu}\max_z \sum_{t=1}^T [e_{1,t} + e_{2,t}].$$

Remark 7.6. Even though the first term is of the correct order, we have to estimate the three remaining terms. The second and third terms are the bias and variance coming from inexact estimation of the oracles of both players.

Insight 2. The last error term in the lemma involving $e_{1,t}, e_{2,t}$ is due to the coupling between the free variables x^s, y^s and randomness of the algorithm. For this error, we adapt the “ghost iterate” trick from [NJLS09] which we adapted to coordinate methods in Chapter 4.

Next is the the variance estimation, which is similar to [Lan21], except handling the error term coming from $\hat{V}_{k-1}$ as in Insight 3.

Lemma 7.7. *Let Assumption 7.1 hold. Let* $\beta^\theta_n = \frac{2}{\lambda^\theta_{\min}(n+n_0)}$ *for* $n_0 \geq 1$. *Then, for Algorithms 7.1 and 7.2,*

$$\mathbb{E}\|\theta_N - \theta_{\star,t}\|_2^2 \leq \mathcal{O}\left(\frac{|S||A|}{(1-\gamma)^2N^2} + \frac{1}{N(\lambda^\theta_{\min})^2(1-\gamma)^2} + \frac{|S||A|}{(\lambda^\theta_{\min})^2}\mathbb{E}\|\hat{V}_{k-1} - V_{k-1}\|^2_\infty\right).$$

Insight 3. Different from the standard critic analyses [HWWY20, KDMR21], we account for the additional bias coming from having $\hat{V}_{k-1}$ instead of real V_{k-1} (see (7.7)). We exploit strong monotonicity of the operator F_t in (7.5) to make sure the error term appears as $\mathbb{E}\|\hat{V}_{k-1} - V_{k-1}\|^2_\infty$ in the bound instead of $\mathbb{E}\|\hat{V}_{k-1} - V_{k-1}\|_\infty$, which would deteriorate the rate.

The next estimation is critical for obtaining the complexity result. In particular, we will see how to bound the bias of θ_{t+1}. Since in Lemma 7.5, we need a tight bound for $\|\mathbb{E}[\theta_{t+1}|x_t] - \theta_{\star,t}\| = \|\mathbb{E}[\theta_N|x_t] - \theta_{\star,t}\|$, we have to be careful with the additional bias from $\hat{V}_{k-1}$.

Lemma 7.8. *Let Assumption 7.1 hold,* $\beta^\theta_n = \frac{2}{\lambda^\theta_{\min}(n+n_0)}$, $n_0 = \frac{6\lambda^2_{\max}}{(\lambda^\theta_{\min})^2}$. *For Algorithms 7.1 and 7.2*

$$\|\mathbb{E}[\theta_N|x_t] - \theta_{\star,t}\|^2 \leq \mathcal{O}\left(\frac{|S||A|}{(1-\gamma)^2N^2} + \frac{10|S||A|}{(\lambda^\theta_{\min})^2}\|\mathbb{E}[\hat{V}_{k-1}|x_t] - V_{k-1}\|^2_\infty\right). \tag{7.8}$$

Insight 4. The reason to use fresh estimates for $\hat{V}_{k-1}$ at each t as in Algorithm 7.1 is the result of this lemma (see Remark 7.3). Since the bias term in the algorithm's analysis is $\|\mathbb{E}[\theta_N|x_t]-\theta_{\star,t}\|$ in Lemma 7.5, we take the square root of the result of Lemma 7.8. If $\hat{V}_{k-1}$ is estimated before x_t, then we will have $\mathbb{E}\|\hat{V}_{k-1}-V_{k-1}\|$ in the bound of Lemma 7.5, which will have the rate $\mathcal{O}(1/\sqrt{N})$. On the other hand, if we estimate $\hat{V}_{k-1}$ freshly as in Algorithm 7.1, then we will be able to use the improved bias bound $\|\mathbb{E}[\hat{V}_{k-1}|x_t]-V_{k-1}\|\le\mathcal{O}(1/N)$ as in the next lemma.

Lemma 7.9. *Let Assumption 7.1 hold and $\beta_n^\omega=\frac{2}{\lambda_{\min}^\omega(n+n_0)}$, with $n_0=\frac{6\lambda_{\max}^2}{(\lambda_{\min}^\omega)^2}$. The variance and bias of $\hat{V}_{k-1}$, computed as in Algorithm 7.1 satisfies*

$$\|\mathbb{E}[\hat{V}_{k-1}|x_t,y_t]-V_{k-1}\|_2^2\le\mathcal{O}\left(\frac{|S||A|}{(1-\gamma)^2N^2}\right),\quad \mathbb{E}\|\hat{V}_{k-1}-V_{k-1}\|_2^2\le\mathcal{O}\left(\frac{1}{N(1-\gamma)^2\lambda_{\min}^2}\right).$$

Unlike the greedy step, the evaluation step (finding the best response) mirrors the single agent analysis closely. We defer the details to Section 7.4.2. Combining Lemma 7.5 with the result of the evaluation step (which is of the same order) in (7.3) gives Theorem 7.4.

7.4 Proofs

7.4.1 Basic results on RL

Lemma 7.10. *Define θ_t recursively as $\theta_{t+1}=\theta_t-\beta_t\tilde{F}(\theta_t,\xi_t)$ where $r(s,a,b)\le1$ and $\tilde{F}(\theta_t,\xi_t)=e(s',a')(\theta_t(s',a')-r(s,a,b)-\gamma\theta_t(s'',a''))$ and recall the definition of $Q_k(s,a,b)=r(s,a,b)+\gamma\sum_{s'}P(s'|s,a,b)V_k(s')$. Then, it follows for any t,k*

$$\|\theta_t\|_\infty\le\frac{1}{1-\gamma},\qquad \|\theta_t\|_2\le\frac{\sqrt{|S||A|}}{1-\gamma},\qquad \|\hat{V}_{k-1}\|_\infty\le\frac{1}{1-\gamma},$$
$$\|\tilde{F}(\theta_t,\xi_t)\|_2\le\frac{3}{1-\gamma},\qquad \|Q_k(s,a,b)\|_\infty\le\frac{2}{1-\gamma}.$$

Proof. The first inequality is proven by induction, for example see [KDMR21, Lemma C.10]. Following inequalities are either basic consequences of the first inequality or directly follow from definition. ■

A classical result that we use frequently in the proofs is performance difference lemma [KL02]. The statement of the lemma is slightly different due to multi agent setting, but since one policy is held fixed while changing the other one, the original proof of the lemma extends straightforwardly. The proof for this case is given in [DFG20].

Lemma 7.11 (Performance difference lemma. See [KL02, DFG20]). *For any policies x,y_1,y_2*

and any state s_0

$$V^{x,y_1}(s_0) - V^{x,y_2}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{x,y_1}} \langle \mathbb{E}_{a\sim x(\cdot|s)} Q^{x,y_2}(s,a,\cdot), y_1(\cdot|s) - y_2(\cdot|s)\rangle$$

A standard result that we use is Lipschitzness of $y \mapsto V^{x,y}(s_0)$.

Lemma 7.12. *For any policies x, y_1, y_2,*

$$\|V^{x,y_1} - V^{x,y_2}\|_\infty \leq \frac{2}{(1-\gamma)^2}\max_s \|y_1(\cdot|s) - y_2(\cdot|s)\|_1.$$

Proof. By performance difference lemma [KL02] and Cauchy-Schwarz inequality, for any s_0,

$$\begin{aligned} V^{x,y_1}(s_0) - V^{x,y_2}(s_0) &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{x,y_1}} \langle \mathbb{E}_{a\sim x(\cdot|s)} Q^{x,y_2}(s,a,\cdot), y_1(\cdot|s) - y_2(\cdot|s)\rangle \\ &\leq \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{x,y_1}} \|\mathbb{E}_{a\sim x(\cdot|s)} Q^{x,y_2}(s,a,\cdot)\|_\infty \|y_1(\cdot|s) - y_2(\cdot|s)\|_1. \end{aligned}$$

Next, we are going to further upper bound the right hand side using Lemma 7.10

$$\begin{aligned} V^{x,y_1}(s_0) - V^{x,y_2}(s_0) &\leq \frac{1}{1-\gamma}\max_s \|\mathbb{E}_{a\sim x(\cdot|s)} Q^{x,y_2}(s,a,\cdot)\|_\infty \|y_1(\cdot|s) - y_2(\cdot|s)\|_1 \\ &\leq \frac{2}{(1-\gamma)^2}\max_s \|y_1(\cdot|s) - y_2(\cdot|s)\|_1. \end{aligned}$$

We take maximum over s_0 to conclude. ■

Lemma 7.13. *We have*

$$\|Q^{x,y_1} - Q^{x,y_2}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2}\max_s \|y_1(\cdot|s) - y_2(\cdot|s)\|_1.$$

Proof. We note that by the definition of Q^{x,y_1} it follows that for all s, a, b

$$|Q^{x,y_1}(s,a,b) - Q^{x,y_2}(s,a,b)| = \gamma \left| \sum_{s'} P(s'|s,a,b)\left(V^{x,y_1}(s') - V^{x,y_2}(s')\right)\right|.$$

Jensen's inequality, and the previous lemma gives the result. ■

7.4.2 Proofs for Reflected NAC with a game etiquette

Proofs for greedy step of Reflected NAC with a game etiquette

We recall the definition of F_t when $\theta_{\star,t}(s,b) = \mathbb{E}_{a\sim x_t(\cdot|s)} Q(s,a,b)$

$$F_t(\theta)(s,b) = \rho^{x_t,y_t}(s) y_t(b|s)\Big[\theta(s,b) - \sum_a x_t(a|s) r(s,a,b) - \gamma \sum_{s',a} x_t(a|s) P(s'|s,a,b) V_{k-1}(s')\Big].$$

We recall that F_t is strongly monotone with $\lambda^{\theta}_{\min}$ under Assumption 7.1. Moreover F_t is Lipschitz with $\lambda_{\max}$. We refer to previous section for how the oracles in the algorithm can be computed without accessing to other agent's policy or actions. Moreover, we do not put subscripts θ^x, θ^y as the estimations will be symmetric.

Theorem 7.14. *[See Theorem 7.4] Let Assumption 7.1 hold. For Algorithm 7.1*

$$\begin{aligned}\mathbb{E}\mathbb{E}_{s_0\sim\mu}[\max_y V^{x_k,y}(s_0) - V^{\star}(s_0)] \leq \frac{C_{\mu,\sigma} k}{(1-\gamma)} \tilde{\mathcal{O}}\Bigg\{ &\frac{1}{T(1-\gamma)^2} + \frac{|S|(|A| \vee |B|)}{\lambda^{\theta}_{\min}(1-\gamma)^2 N} \\ &+ \frac{|S|^2(|A|^2 \vee |B|^2)}{(\lambda^{\theta}_{\min})^2 N^2 (1-\gamma)^2} + \frac{|S||A|}{(\lambda^{\theta}_{\min})^2(\lambda^{\omega}_{\min})^2(1-\gamma)^2 N} \\ &\frac{|S||B|}{(1-\gamma)^4 N^2} + \frac{1}{N(1-\gamma)^4(\lambda^{\nu}_{\min})^2} + \frac{\sqrt{|S||B|}}{(1-\gamma)^2 N}\Bigg\} + \mathcal{O}\left(\frac{C_{\mu,\sigma}\gamma^k}{(1-\gamma)}\right),\end{aligned}$$

which gives $\tilde{\mathcal{O}}(\frac{C^2_{\mu,\sigma}|S|^2(|A|^2 \vee |B|^2)}{\epsilon^2(1-\gamma)^8(\lambda^{\theta}_{\min}\lambda^{\omega}_{\min}\lambda^{\nu}_{\min})^2})$ *sample complexity.*

Proof. We combine Lemmas 7.7–7.9 and 7.15 and corollary 7.20. ■

Our theoretical results here bring together ideas from single agent NPG analysis of [Lan21] and stochastic primal-dual optimization techniques from [MT20b, NJLS09]. In particular, we will be using ideas from [MT20a, NJLS09] in the analysis we develop for extending ideas of [Lan21] to the greedy step of the multi agent algorithm we have.

We first analyze the policy evaluation routine in Algorithm 7.1. In particular, we will bound the variance and bias of θ_{t+1} as an estimate of $\theta_{\star,t}(s,b) = \mathbb{E}_{a\sim x_t(\cdot|s)} Q_{k-1}(s,a,b)$. As this routine is in an inner loop (indexed by n), the policies we sample, consequentlyF_t is fixed, therefore we drop the subscript. The proofs of these lemmas will be similar to [Lan21], except the additional bias we have due to $\hat{V}_{k-1}$.

Proof of Lemma 7.7. By the definition of θ_n,

$$\|\theta_{n+1} - \theta_{\star,t}\|_2^2 = \|\theta_n - \theta_{\star,t}\|_2^2 - 2\beta_n\langle \tilde{F}(\theta_n,\xi_n), \theta_n - \theta_{\star,t}\rangle + \beta_n^2 \|\tilde{F}^{\theta}(\theta_n,\xi_n)\|_2^2.$$

We take expectation $\mathbb{E}_{\xi_n}$ where $\xi_n = (s_n, a_n, b_n, s_{n+1})$ is the sample at iteration n of Algorithm 7.2

$$\mathbb{E}_{\xi_n}\tilde{F}(\theta_n,\xi_n) = F(\theta_n) + \gamma P_{x_t,y_t}(V_{k-1} - \hat{V}_{k-1}),$$

as in (7.7) where P_{x_t,y_t} was also defined. As we stated, we omit the dependence of F_t to t as t is fixed throughout this loop. Thus,

$$\mathbb{E}_{\xi_n}\|\theta_{n+1} - \theta_{\star,t}\|_2^2 = \|\theta_n - \theta_{\star,t}\|_2^2 - 2\beta_n\langle F(\theta_n), \theta_n - \theta_{\star,t}\rangle$$

$$-2\beta_n\gamma\langle P_{x_t,y_t}(V_{k-1}-\hat{V}_{k-1}), \theta_n-\theta_{\star,t}\rangle+\beta_n^2\|\tilde{F}(\theta_n,\xi_n)\|_2^2.$$

We use strong monotonicity (with $F(\theta_{\star,t})=0$) for the first inner product and Cauchy-Schwarz and Young's inequalities for the second inner product

$$\begin{aligned}\mathbb{E}_{\xi_n}\|\theta_{n+1}-\theta_{\star,t}\|_2^2 &\le \left(1-2\beta_n\lambda^\theta_{\min}\right)\|\theta_n-\theta_{\star,t}\|_2^2+\frac{\beta_n\gamma^2}{\lambda^\theta_{\min}}\|P_{x_t,y_t}(V_{k-1}-\hat{V}_{k-1})\|_2^2\\ &\quad+\beta_n\lambda^\theta_{\min}\|\theta_n-\theta_{\star,t}\|_2^2+\beta_n^2\mathbb{E}_{\xi_n}\|\tilde{F}(\theta_n,\xi_n)\|_2^2\\ &\le \left(1-\beta_n\lambda^\theta_{\min}\right)\|\theta_n-\theta_{\star,t}\|_2^2+\frac{\beta_n\gamma^2|S||A|}{\lambda^\theta_{\min}}\|V_{k-1}-\hat{V}_{k-1}\|_\infty^2+\beta_n^2\mathbb{E}_{\xi_n}\|\tilde{F}(\theta_n,\xi_n)\|_2^2, \qquad (7.9)\end{aligned}$$

where we estimated $\|P_{x_t,y_t}(\hat{V}_{k-1}-V_{k-1})\|_2^2$. We will use Lemma 7.10 to upper bound $\|\tilde{F}(\theta_n,\xi_n)\|_2^2 \le \frac{2}{(1-\gamma)^2}$. We define Θ_n such that $\Theta_n(1-\beta_n\lambda_{\min})\le\Theta_{n-1}$ with $\Theta_0=\Theta_1=1$. We multiply both sides of the inequality with Θ_n after taking total expectation, to get

$$\Theta_n\mathbb{E}\|\theta_{n+1}-\theta_{\star,t}\|_2^2\le\Theta_{n-1}\mathbb{E}\|\theta_n-\theta_{\star,t}\|_2^2+\frac{\Theta_n\beta_n\gamma^2|S||A|}{\lambda^\theta_{\min}}\mathbb{E}\|\hat{V}_{k-1}-V_{k-1}\|_\infty^2+\frac{2\Theta_n\beta_n^2}{(1-\gamma)^2}.$$

Summing the inequality gives

$$\Theta_N\mathbb{E}\|\theta_{N+1}-\theta_{\star,t}\|_2^2\le\Theta_0\|\theta_1-\theta_{\star,t}\|_2^2+\sum_{n=1}^N\frac{\Theta_n\beta_n\gamma^2|S||A|}{\lambda^\theta_{\min}}\mathbb{E}\|\hat{V}_{k-1}-V_{k-1}\|_\infty^2+\sum_{n=1}^N\frac{2\Theta_n\beta_n^2}{(1-\gamma)^2}.$$

Using the definition of β_n and setting $\Theta_n(1-\beta_n\lambda_{\min})=\Theta_{n-1}$ gives $\Theta_n=\Theta_1\frac{(n+n_0)(n+n_0-1)}{(n_0)(n_0+1)}$. Let us use $\Theta_0=\Theta_1=1$ and bounds from Lemma 7.10 for $\|\theta_1-\theta_{\star,t}\|_2^2$,

$$\begin{aligned}\mathbb{E}\|\theta_N-\theta_{\star,t}\|_2^2\le\frac{2n_0(n_0+1)|S||A|}{(1-\gamma)^2(N+n_0)(N+n_0-1)}&+\frac{8N}{(N+n_0)(N+n_0-1)(1-\gamma)^2(\lambda^\theta_{\min})^2}\\ &+\frac{3\gamma^2|S||A|}{(\lambda^\theta_{\min})^2}\mathbb{E}\|\hat{V}_{k-1}-V_{k-1}\|_\infty^2.\end{aligned}$$

■

Proof of Lemma 7.8. We are going to take expectation of the recursion

$$\theta_{n+1}=\theta_n-\beta_n\tilde{F}(\theta_n,\xi_n),$$

first w.r.t. sample ξ_n,

$$\begin{aligned}\mathbb{E}_{\xi_n}\theta_{n+1}&=\theta_n-\beta_n\mathbb{E}_{\xi_n}\tilde{F}(\theta_n,\xi_n)\\ &=\theta_n-\beta_nF(\theta_n)-\beta_n\gamma P_{x_t,y_t}(\hat{V}_{k-1}-V_{k-1}),\end{aligned}$$

where we used (7.7) where P_{x_t,y_t} was also defined.

We will now take expectation $\mathbb{E}[\cdot|x_t]$. We note F and P_{x_t,y_t} are linear

$$\mathbb{E}[\theta_{n+1}|x_t] = \mathbb{E}[\theta_n|x_t] - \beta_n F_t(\mathbb{E}[\theta_n|x_t]) - \beta_n \gamma P_{x_t,y_t}(\mathbb{E}[\hat{V}_{k-1}|x_t] - V_{k-1}).$$

We denote $\bar{\theta}_n = \mathbb{E}[\theta_n|x_t]$ and $\bar{\delta} = \gamma P_{x_t,y_t}(\mathbb{E}[\hat{V}_{k-1}|x_t] - V_{k-1})$ in the above equality which makes the recursion $\bar{\theta}_{n+1} = \bar{\theta}_n - \beta_n F(\bar{\theta}_n) - \beta_n \bar{\delta}$. We then have

$$\|\bar{\theta}_{n+1} - \theta_{\star,t}\|_2^2 = \|\bar{\theta}_n - \theta_{\star,t}\|_2^2 - 2\beta_n \langle F(\bar{\theta}_n), \bar{\theta}_n - \theta_{\star,t}\rangle - 2\beta_n \langle \bar{\delta}, \bar{\theta}_n - \theta_{\star,t}\rangle + \frac{3\beta_n^2}{2}\|F(\bar{\theta}_n)\|_2^2 + 3\beta_n^2 \|\bar{\delta}\|_2^2, \quad (7.10)$$

where we also used Young's inequality to split the term $\beta_n^2 \|F(\bar{\theta}_n) + \bar{\delta}\|^2$.

By strong monotonicity and Lipschitzness of F along with $F(\theta_{\star,t}) = 0$,

$$2\beta_n \langle F(\bar{\theta}_n), \bar{\theta}_n - \theta_{\star,t}\rangle = 2\beta_n \langle F(\bar{\theta}_n) - F(\theta_{\star,t}), \bar{\theta}_n - \theta_{\star,t}\rangle \geq 2\beta_n \lambda^\theta_{\min} \|\bar{\theta}_n - \theta_{\star,t}\|_2^2,$$
$$\beta_n^2 \|F(\bar{\theta}_n)\|_2^2 = \beta_n^2 \|F(\bar{\theta}_n) - F(\theta_{\star,t})\|_2^2 \leq \beta_n^2 \lambda_{\max}^2 \|\bar{\theta}_n - \theta_{\star,t}\|_2^2.$$

By Cauchy-Schwarz and Young's inequalities, it follows that $2\beta_n \langle \bar{\delta}, \bar{\theta}_n - \theta_{\star,t}\rangle \leq \frac{\beta_n \lambda^\theta_{\min}}{2} \|\bar{\theta}_n - \theta_{\star,t}\|_2^2 + \frac{2\beta_n}{\lambda^\theta_{\min}} \|\bar{\delta}\|_2^2$. Using these three inequalities in (7.10) gives

$$\|\bar{\theta}_{n+1} - \theta_{\star,t}\|_2^2 \leq \left(1 - \frac{3}{2}\beta_n \lambda^\theta_{\min} + \frac{3}{2}\beta_n^2 \lambda_{\max}^2\right) \|\bar{\theta}_n - \theta_{\star,t}\|_2^2 + \frac{2\beta_n}{\lambda^\theta_{\min}} \|\bar{\delta}\|_2^2 + 3\beta_n^2 \|\bar{\delta}\|_2^2.$$

We now use $n_0 = \frac{6\lambda_{\max}^2}{(\lambda^\theta_{\min})^2}$ and $\beta_n = \frac{2}{\lambda^\theta_{\min}(n+n_0)}$ to estimate

$$\frac{3\beta_n}{2}\left(\lambda^\theta_{\min} - \beta_n \lambda_{\max}^2\right) = \frac{3\beta_n}{2}\left(\lambda^\theta_{\min} - \frac{2\lambda_{\max}^2}{\lambda^\theta_{\min}(n+n_0)}\right) \geq \frac{3\beta_n}{2}\left(\lambda^\theta_{\min} - \frac{2\lambda_{\max}^2}{\lambda^\theta_{\min} n_0}\right) = \beta_n \lambda^\theta_{\min}.$$

Therefore, the recursion is

$$\|\bar{\theta}_{n+1} - \theta_{\star,t}\|_2^2 \leq \left(1 - \beta_n \lambda^\theta_{\min}\right) \|\bar{\theta}_n - \theta_{\star,t}\|_2^2 + \frac{2\beta_n}{\lambda^\theta_{\min}} \|\bar{\delta}\|^2 + 3\beta_n^2 \|\bar{\delta}\|_2^2.$$

This recursion is similar to (7.9), in particular, by noting $\beta_n \leq \frac{1}{\lambda^\theta_{\min}}$, and bounding $\|\bar{\delta}\|_2^2$ as $\|P_{x_t,y_t}(\mathbb{E}[\hat{V}_{k-1}|x_t] - V_{k-1})\|_2^2 \leq |S||A|\|P_{x_t,y_t}(\mathbb{E}[\hat{V}_{k-1}|x_t] - V_{k-1})\|_\infty^2 \leq |S||A|\|\mathbb{E}[\hat{V}_{k-1}|x_t] - V_{k-1}\|_\infty^2$

$$\|\bar{\theta}_{n+1} - \theta_{\star,t}\|^2 \leq \left(1 - \beta_n \lambda^\theta_{\min}\right) \|\bar{\theta}_n - \theta_{\star,t}\|^2 + \frac{5\beta_n |S||A|}{\lambda^\theta_{\min}} \|\mathbb{E}[\hat{V}_{k-1}|x_t] - V_{k-1}\|_\infty^2.$$

We finally define Θ_n as $\Theta_n(1 - \beta_n \lambda_{\min}) = \Theta_{n-1}$ gives $\Theta_n = \Theta_1 \frac{(n+n_0)(n+n_0-1)}{n_0(n_0-1)}$, where $\Theta_0 = \Theta_1 = 1$. We multiply both sides of the inequality with Θ_n and sum to get the result. ■

We estimate the bias and variance of the estimation of $\hat{V}_{k-1}$ in Algorithm 7.1, very similar to [Lan21]. Unlike [Lan21] that derived $\mathcal{O}(1/N^3)$ bound for the bias, we derive a $\mathcal{O}(1/N^2)$ bound which will be sufficient. We note that the previous two lemmas had additional bias not present in [Lan21], however the next result does not have this bias and therefore the arguments in [Lan21] would be enough. We provide a brief proof to be self-contained.

Let us recall that $V_{k-1} = V^{x_{k-1},y_{k-1}}$ and by sampling $s_n \sim \rho^{x_{k-1},y_{k-1}}$, $a_n \sim x_{k-1}(\cdot|s_n)$, $b_n \sim y_{k-1}(\cdot|s_n)$, $s_{n+1} \sim P(\cdot|s_n, a_n, b_n)$, the oracle

$$\tilde{F}^{\omega}(\omega_n, \xi_n) = e(s_n)\left(\omega_n(s_n) - r(s_n, a_n, b_n) - \gamma\omega_n(s_{n+1})\right),$$

satisfies $\mathbb{E}_{\xi_n}\tilde{F}^{\omega}(\omega_n,\xi_n) = F^{\omega}(\omega_n)$, where F^{ω} is defined as

$$\begin{aligned}F^{\omega}_{k-1}(\omega)(s) = \rho^{x_{k-1},y_{k-1}}(s)\Big(\omega(s) - \sum_{a,b} x_{k-1}(a|s)y_{k-1}(b|s)r(s,a,b)\\ -\gamma\sum_{s',a,b} x_{k-1}(a|s)y_{k-1}(b|s)P(s'|s,a,b)\omega(s')\Big),\end{aligned}$$

where $F^{\omega}_{k-1}(V_{k-1}) = 0$ and also as before F^{ω}_{k-1} is strongly monotone with $\lambda^{\omega}_{\min}$. We will drop the subscript of F^{ω} since k is fixed in this loop.

Proof of Lemma 7.9. For the variance, we have by taking expectation w.r.t. ξ_n

$$\mathbb{E}_{\xi_n}\|\omega_{n+1} - V_{k-1}\|_2^2 = \|\omega_n - V_{k-1}\|_2^2 - 2\beta_n\langle\mathbb{E}_{\xi_n}[\tilde{F}^{\omega}(\omega_n,\xi_n)], \omega_n - V_{k-1}\rangle + \beta_n^2\mathbb{E}_{\xi_n}\|\tilde{F}^{\omega}(\omega_n,\xi_n)\|_2^2.$$

By $\mathbb{E}_{\xi_n}\tilde{F}^{\omega}(\omega_n,\xi_n) = F^{\omega}(\omega_n)$, $F^{\omega}(V_{k-1}) = 0$, and strong monotonicity of F^{ω}, similar to our previous proofs for policy evaluation,

$$\mathbb{E}\|\omega_{n+1} - V_{k-1}\|_2^2 = \left(1 - 2\beta_n\lambda^{\omega}_{\min}\right)\mathbb{E}\|\omega_n - V_{k-1}\|_2^2 + \beta_n^2\mathbb{E}\|\tilde{F}^{\omega}(\omega_n,\xi_n)\|_2^2.$$

The end of the proof is the same as Lemma 7.7, except that we do not have here the additional bias term in Lemma 7.7. Therefore, the result follows.

For the bias, we will argue as in Lemma 7.8. Taking expectation of the recursion w.r.t. ξ_n gives

$$\mathbb{E}_{\xi_n}\omega_{n+1} = \omega_n - \beta_n F^{\omega}(\omega_n).$$

We now unroll the expectation until x_t and use linearity of F^{ω}

$$\mathbb{E}[\omega_{n+1}|x_t] = \mathbb{E}[\omega_n|x_t] - \beta_n F^{\omega}(\mathbb{E}[\omega_n|x_t]).$$

Denoting $\bar{\omega}_n = \mathbb{E}[\omega_n|x_t]$ gives the recursion $\bar{\omega}_{n+1} = \bar{\omega}_n - \beta_n F^{\omega}(\bar{\omega}_n)$, and therefore

$$\|\bar{\omega}_{n+1} - V_{k-1}\|_2^2 = \|\bar{\omega}_n - V_{k-1}\|_2^2 - 2\beta_n\langle F^{\omega}_{k-1}(\bar{\omega}_n), \bar{\omega}_n - V_{k-1}\rangle + \beta_n^2\|F^{\omega}(\bar{\omega}_n)\|_2^2.$$

We will now use Lipschitzness and strong monotonicity of F^{ω} and that $F^{\omega}(V_{k-1}) = 0$ and similar to Lemma 7.8, we obtain the recursion

$$\|\bar{\omega}_{n+1} - V_{k-1}\|_2^2 = \left(1 - 2\beta_n\lambda_{\min} + \beta_n^2\lambda_{\max}^2\right)\|\bar{\omega}_n - V_{k-1}\|_2^2.$$

By the choice of n_0 and β_n, similar to Lemma 7.8, it holds that $2\beta_n\lambda_{\min} - \beta_n^2\lambda_{\max}^2 \geq \beta_n\lambda_{\min}$. By defining Θ_n the same way as Lemma 7.8 and summing the inequality gives the result. ■

We analyze the outer algorithm for solving the matrix game in greedy step. The algorithm is based on FoRB from [MT20b] due to its simple update with one projection and one oracle.

Similar to [MT20b], let us define the "Lyapunov-like" function

$$\begin{aligned}\Phi_{t+1}^s = D(x(\cdot|s), x_{t+1}(\cdot|s)) + \eta\langle\theta_{\star,t+1}(s,\cdot) - \theta_{t+1}(s,\cdot), x(\cdot|s) - x_{t+1}(\cdot|s)\rangle \\ + \frac{1}{2}D(x_{t+1}(\cdot|s), x_t(\cdot|s)).\end{aligned} \tag{7.11}$$

We call this "Lyapunov-like" since it is not non-increasing. Moreover, unlike [MT20b], Φ_t is not necessarily nonnegative. However, it is sufficient for our purposes as it is bounded. Note that we will also use the following error functions

$$\begin{aligned}e_{1,t} &= \eta\langle\theta_{t+1}(\cdot|s) - \mathbb{E}[\theta_{t+1}(\cdot|s)|x_t], x(\cdot|s) - x_t(\cdot|s)\rangle \\ e_{2,t} &= \eta\langle\theta_{t+1}^y(\cdot|s) - \mathbb{E}[\theta_{t+1}^y(\cdot|s)|y_t], y_t(\cdot|s) - y(\cdot|s)\rangle.\end{aligned}$$

Lemma 7.15. *[See Lemma 7.5] Let Assumption 7.1 hold. Denote $x_{out} = \frac{1}{T}\sum_{t=1}^T x_t$ and $y_{out} = \frac{1}{T}\sum_{t=1}^T y_t$ and let $\eta = \frac{1-\gamma}{8}$*

$$\begin{aligned}\mathbb{E}\mathbb{E}_{s\sim\sigma}\left[\max_{x^s,y^s} x_{out}^s Q^s y^s - x^s Q^s y_{out}\right] &= \mathcal{O}\left(\frac{\Phi_0^s - \Phi_T^s}{\eta T}\right) + \mathcal{O}\left(\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\mathbb{E}[\theta_{t+1}|x_t] - \theta_{\star,t}\|\right) \\ &+ \mathcal{O}\left(\frac{1}{T}\sum_{t=1}^T \eta\mathbb{E}\|\theta_{t+1} - \theta_{\star,t}\|^2 + \mathbb{E}\|\theta_t - \theta_{\star,t-1}\|^2\right) + \frac{1}{T\eta}\mathbb{E}\mathbb{E}_{s\sim\sigma}\max_z\sum_{t=1}^T[e_{1,t} + e_{2,t}]).\end{aligned}$$

Remark 7.16. By Lemma 7.7 and Lemma 7.8, the second and third term will bring dependence $\mathcal{O}\left(\frac{1}{N} + \mathbb{E}\|\mathbb{E}[\hat{V}_{k-1}|x_t] - V_{k-1}\|\right)$. We will see in the next lemma how to handle error terms e_1, e_2 and will use the bound derived earlier for $\mathbb{E}\|\mathbb{E}[\hat{V}_{k-1}|x_t] - V_{k-1}\|$ in Lemma 7.9.

Proof. By the update rule, it follows for all s and $x(\cdot|s) \in \Delta$,

$$\langle\nabla D(x_{t+1}(\cdot|s), x_t(\cdot|s)) + \eta(2\theta_{t+1}(s,\cdot) - \theta_t(s,\cdot)), x(\cdot|s) - x_{t+1}(\cdot|s)\rangle \geq 0.$$

By three point identity,

$$\begin{aligned}D(x(\cdot|s), x_{t+1}(\cdot|s)) \leq D(x(\cdot|s), x_t(\cdot|s)) - D(x_{t+1}(\cdot|s), x_t(\cdot|s)) \\ + \eta\langle(2\theta_{t+1}(s,\cdot) - \theta_t(s,\cdot)), x(\cdot|s) - x_{t+1}(\cdot|s)\rangle.\end{aligned} \tag{7.12}$$

We now manipulate the inner product by adding and subtracting $\theta_{\star,t+1}$

$$\begin{aligned}\eta\langle(2\theta_{t+1}(s,\cdot)-\theta_t(s,\cdot)),x(\cdot|s)-x_{t+1}(\cdot|s)\rangle &= \eta\langle\theta_{t+1}-\theta_{\star,t+1}(s,\cdot),x(\cdot|s)-x_{t+1}(\cdot|s)\rangle\\ &\quad+\eta\langle\theta_{t+1}(s,\cdot)-\theta_t(s,\cdot)+\theta_{\star,t+1}(s,\cdot),x(\cdot|s)-x_{t+1}(\cdot|s)\rangle\\ &= \eta\langle\theta_{t+1}(s,\cdot)-\theta_{\star,t+1}(s,\cdot),x(\cdot|s)-x_{t+1}(\cdot|s)\rangle+\eta\langle\theta_{t+1}(s,\cdot)-\theta_t(s,\cdot),x(s,\cdot)-x_t(\cdot|s)\rangle\\ &\quad+\eta\langle\theta_{t+1}(s,\cdot)-\theta_t(s,\cdot),x_t(\cdot|s)-x_{t+1}(\cdot|s)\rangle+\eta\langle\theta_{\star,t+1}(s,\cdot),x(\cdot|s)-x_{t+1}(\cdot|s)\rangle. \quad (7.13)\end{aligned}$$

The first two inner products in the final inequality will telescope if we can replace θ_{t+1} with $\theta_{\star,t}$ in the second one. For this we have to be careful with bias and variance. Let us take the second inner product

$$\begin{aligned}\eta\langle\theta_{t+1}(s,\cdot)-\theta_t(s,\cdot),x(\cdot|s)-x_t(\cdot|s)\rangle &= \eta\langle\theta_{\star,t}(s,\cdot)-\theta_t(s,\cdot),x(\cdot|s)-x_t(\cdot|s)\rangle\\ &\quad+\eta\langle\theta_{t+1}(s,\cdot)-\theta_{\star,t}(s,\cdot),x(\cdot|s)-x_t(\cdot|s)\rangle.\end{aligned}$$

Now in this estimation, we will add and subtract terms involving $\mathbb{E}[\theta_{t+1}(s,\cdot)|x_t]$ to obtain

$$\begin{aligned}\eta\langle\theta_{t+1}(s,\cdot)-\theta_t(s,\cdot),x(\cdot|s)-x_t(\cdot|s)\rangle &= \eta\langle\theta_{\star,t}(s,\cdot)-\theta_t(s,\cdot),x(\cdot|s)-x_t(\cdot|s)\rangle\\ &\quad+\eta\langle\mathbb{E}[\theta_{t+1}(s,\cdot)|x_t]-\theta_{\star,t},x(\cdot|s)-x_t(\cdot|s)\rangle+\eta\langle\theta_{t+1}(s,\cdot)-\mathbb{E}[\theta_{t+1}(s,\cdot)|x_t],x(\cdot|s)-x_t(\cdot|s)\rangle\\ &\le \eta\langle\theta_{\star,t}(s,\cdot)-\theta_t(s,\cdot),x(\cdot|s)-x_t(\cdot|s)\rangle+2\eta\|\mathbb{E}[\theta_{t+1}|x_t]-\theta_{\star,t}\|_\infty+e_{1,t}, \quad (7.14)\end{aligned}$$

where the inequality is due to Cauchy-Schwarz and we use the definition of $e_{1,t}$ for the last term. Next, we use Cauchy-Schwarz and Young's inequalities for the third inner product in RHS of (7.13) to derive

$$\begin{aligned}\eta\langle\theta_{t+1}(s,\cdot)-\theta_t(\cdot|s),x_t(\cdot|s)-x_{t+1}(\cdot|s)\rangle &\le \eta^2\|\theta_{t+1}(s,\cdot)-\theta_t(\cdot|s)\|_\infty^2+\frac{1}{4}\|x_t(\cdot|s)-x_{t+1}(\cdot|s)\|_1^2\\ &\le 4\eta^2\left[\|\theta_{t+1}(s,\cdot)-\theta_{\star,t}(s,\cdot)\|_\infty^2+\|\theta_{\star,t}(s,\cdot)-\theta_{\star,t-1}(s,\cdot)\|_\infty^2+\|\theta_{\star,t-1}(s,\cdot)-\theta_t(s,\cdot)\|_\infty^2\right]\\ &\quad+\frac{1}{4}\|x_t(\cdot|s)-x_{t+1}(\cdot|s)\|_1^2. \quad (7.15)\end{aligned}$$

As $\theta_{\star,t}(s,a)=\mathbb{E}_{b\sim y_t(\cdot|s)}Q(s,a,b)$, we have

$$\begin{aligned}\|\theta_{\star,t}(s,\cdot)-\theta_{\star,t-1}(s,\cdot)\|_\infty &\le \max_{b,a}|Q(s,a,b)|\|y_t(\cdot|s)-y_{t-1}(\cdot|s)\|_1\\ &\le \frac{2}{1-\gamma}\|y_t(\cdot|s)-y_{t-1}(\cdot|s)\|_1, \quad (7.16)\end{aligned}$$

where the second inequality is by Lemma 7.10 and the first by Jensen. We join (7.14), (7.15), and (7.16) in (7.13)

$$\begin{aligned}\eta\langle(2\theta_{t+1}(s,\cdot)-\theta_t(s,\cdot)),x(\cdot|s)-x_{t+1}(\cdot|s)\rangle &\le \eta\langle\theta_{t+1}(s,\cdot)-\theta_{\star,t+1}(s,\cdot),x(\cdot|s)-x_{t+1}(\cdot|s)\rangle\\ &\quad\eta\langle\theta_{\star,t}(s,\cdot)-\theta_t(s,\cdot),x(\cdot|s)-x_t(\cdot|s)\rangle+2\eta\|\mathbb{E}[\theta_{t+1}|x_t]-\theta_{\star,t}\|_\infty+e_{1,t}\end{aligned}$$

$$+4\eta^2\left[\|\theta_{t+1}-\theta_{\star,t}\|_\infty^2+\|\theta_{\star,t-1}-\theta_t\|_\infty^2\right]+\frac{16\eta^2}{(1-\gamma)^2}\|y_t(\cdot|s)-y_{t-1}(\cdot|s)\|_1^2$$
$$+\frac{1}{4}\|x_t(\cdot|s)-x_{t+1}(\cdot|s)\|_1^2+\eta\langle\theta_{\star,t+1}(s,\cdot),x(\cdot|s)-x_{t+1}(\cdot|s)\rangle. \quad (7.17)$$

We note that by strong convexity of Bregman distance w.r.t. ℓ_1 norm, $\frac{1}{4}\|x_t(\cdot|s)-x_{t+1}(\cdot|s)\|_1^2 \le \frac{1}{2}D(x_{t+1}(\cdot|s),x_t(\cdot|s))$ and similarly for the term involving difference of y_t and y_{t-1}.

We insert (7.17) into (7.12) by using the definition of Φ_t

$$\eta\langle\theta_{\star,t+1}(s,\cdot),x_{t+1}(\cdot|s)-x(\cdot|s)\rangle+\Phi_{t+1}^s \le \Phi_t^s+e_{1,t}+2\eta\|\mathbb{E}[\theta_{t+1}|x_t]-\theta_{\star,t}\|_\infty$$
$$+4\eta^2\left[\|\theta_{t+1}-\theta_{\star,t}\|_\infty^2+\|\theta_{\star,t-1}-\theta_t\|_\infty^2\right]$$
$$+\frac{32\eta^2}{(1-\gamma)^2}D(y_t(\cdot|s),y_{t-1}(\cdot|s))-\frac{1}{2}D(x_t(\cdot|s),x_{t-1}(\cdot|s)).$$

We sum this inequality and use the definition of $\theta_{\star,t+1}$ to obtain

$$\frac{\eta}{T}\sum_{t=0}^{T-1}\langle\mathbb{E}_{b\sim y_{t+1}(\cdot|s)}Q(s,\cdot,b),x_{t+1}(\cdot|s)-x(\cdot|s)\rangle \le \frac{\Phi_0^s-\Phi_T^s}{T}+\frac{1}{T}\sum_{t=0}^{T-1}e_{1,t}$$
$$+\frac{1}{T}\sum_{t=0}^{T-1}2\eta\|\mathbb{E}[\theta_{t+1}|x_t]-\theta_{\star,t}\|_\infty+\frac{4\eta^2}{T}\sum_{t=1}^{T}\left[\|\theta_{t+1}-\theta_{\star,t}\|_\infty^2+\|\theta_{\star,t-1}-\theta_t\|_\infty^2\right]$$
$$+\frac{1}{T}\sum_{t=0}^{T}\frac{32\eta^2}{(1-\gamma)^2}D(y_t(\cdot|s),y_{t-1}(\cdot|s))-\frac{1}{2}D(x_t(\cdot|s),x_{t-1}(\cdot|s)).$$

We estimate the error terms in the last line. The terms in the second line will be the bias and variance arising from using θ_{t+1} instead of the true oracle. By the symmetric estimation on the y player, we can obtain the similar inequality. For making the comparison, we will denote the corresponding oracle as θ^y (θ in the previous estimations correspond to θ^x). In particular $\theta^y_{\star,t+1}(s,b)=\mathbb{E}_{a\sim x_{t+1}(\cdot|s)}Q(s,a,b)$, and the corresponding Lyapunov-like function as $\Phi^s_{t,y}$

$$\frac{\eta}{T}\sum_{t=0}^{T-1}\langle\mathbb{E}_{a\sim x_{t+1}(\cdot|s)}Q(s,a,\cdot),y(\cdot|s)-y_{t+1}(\cdot|s)\rangle \le \frac{\Phi_{0,y}^s-\Phi_{y,T}^s}{T}+\frac{1}{T}\sum_{t=0}^{T-1}e_{2,t}$$
$$+\frac{1}{T}\sum_{t=0}^{T-1}2\eta\|\mathbb{E}[\theta^y_{t+1}|x_t]-\theta^y_{\star,t}\|_\infty+\frac{4\eta^2}{T}\sum_{t=1}^{T}\left[\|\theta^y_{t+1}-\theta^y_{\star,t}\|_\infty^2+\|\theta^y_{\star,t-1}-\theta^y_t\|_\infty^2\right]$$
$$+\frac{1}{T}\sum_{t=0}^{T}\frac{32\eta^2}{(1-\gamma)^2}D(x_t(\cdot|s),x_{t-1}(\cdot|s))-\frac{1}{2}D(y_t(\cdot|s),y_{t-1}(\cdot|s)).$$

After summing up the two inequalities and recalling that we bound the RHS of:

$$x^s_{\text{out}}Q^sy^s-x^sQ^sy^s_{\text{out}}=\frac{1}{T}\sum_{t=1}^{T}\langle\mathbb{E}_{a\sim x_t(\cdot|s)}Q(s,a,\cdot),y(\cdot|s)\rangle-\frac{1}{T}\sum_{t=1}^{T}\langle\mathbb{E}_{b\sim y_t(\cdot|s)}Q(s,\cdot,b),x(\cdot|s)\rangle$$
$$=\frac{1}{T}\sum_{t=1}^{T}\left[\langle\mathbb{E}_{a\sim x_t(\cdot|s)}Q(s,a,\cdot),y(\cdot|s)-y_t(\cdot|s)\rangle-\langle\mathbb{E}_{b\sim y_t(\cdot|s)}Q(s,\cdot,b),x(\cdot|s)-x_t(\cdot|s)\rangle\right], \quad (7.18)$$

we pick $\eta \leq \frac{1-\gamma}{8}$ to cancel the last terms in the last lines of the estimations. Since we estimate θ_t and θ_t^y in the same way, their bounds as we derived in Lemma 7.8, Lemma 7.7 will be the same, therefore in the bound we do not include both and simply put them under big-Oh. Next, we take maximum over x, y, take expectation w.r.t. state distribution σ and total expectation w.r.t. randomness in the algorithm and use the definitions of x_{out} and y_{out} to conclude. ■

For error terms $e_{1,t}, e_{2,t}$, we use the technique to change the order of maximum and expectation from the literature of stochastic primal-dual methods [NJLS09, Lemma 3.1]. Recall:

$$e_{1,t} = \eta\langle\theta_{t+1}(\cdot|s) - \mathbb{E}[\theta_{t+1}(\cdot|s)|x_t], x(\cdot|s) - x_t(\cdot|s)\rangle$$
$$e_{2,t} = \eta\langle\theta^y_{t+1}(\cdot|s) - \mathbb{E}[\theta^y_{t+1}(\cdot|s)|y_t], y_t(\cdot|s) - y(\cdot|s)\rangle$$

We will derive the bound or $e_{1,t}$ and the bound for $e_{2,t}$ is symmetrical.

Lemma 7.17. *We have*

$$\frac{1}{T}\mathbb{E}\mathbb{E}_{s\sim\sigma}\max_x\sum_{t=1}^T e_{1,t} \leq \frac{\log|A|}{T} + \frac{1}{T}\sum_{t=1}^T 4\eta^2\mathbb{E}\|\theta_{t+1} - \theta_{\star,t}\|^2_\infty.$$

Proof. First note that $\langle\theta_{t+1}(\cdot|s) - \mathbb{E}[\theta_{t+1}(\cdot|s)|x_t], x_t(\cdot|s)\rangle$ does not depend on x and by the tower property of conditional expectation,

$$\sum_{t=1}^T \mathbb{E}\mathbb{E}_{s\sim\sigma}\eta\langle\theta_{t+1}(\cdot|s) - \mathbb{E}[\theta_{t+1}(\cdot|s)|x_t], x_t(\cdot|s)\rangle$$
$$= \mathbb{E}\mathbb{E}_{s\sim\sigma}\eta\langle\mathbb{E}[\theta_{t+1}(\cdot|s)|x_t] - \mathbb{E}[\theta_{t+1}(\cdot|s)|x_t], x_t(\cdot|s)\rangle = 0.$$

Therefore, we have to estimate

$$\mathbb{E}\mathbb{E}_{s\sim\sigma}\max_x\sum_{t=1}^T \eta\langle\theta_{t+1}(\cdot|s) - \mathbb{E}[\theta_{t+1}(\cdot|s)|x_t], x(\cdot|s)\rangle.$$

Let $n_t(s,\cdot) = -\eta(\theta_{t+1}(\cdot|s) - \mathbb{E}[\theta_{t+1}(\cdot|s)|x_t])$. First, we note that $\mathbb{E}[n_t(s,\cdot)|x_t] = 0$. Next, we define the auxiliary "ghost" process

$$\tilde{x}_{t+1}(\cdot|s) = \arg\min_x\langle n_t(s,\cdot), x(\cdot|s)\rangle + D(x(\cdot|s), \tilde{x}_t(\cdot|s)).$$

Note that $\tilde{x}_t$ and x_t depend on the same randomness by definition of $\tilde{x}_t$, therefore conditioned on x_t, $\tilde{x}_t$ is deterministic. Standard mirror descent analysis gives for any x

$$\langle n_t(s,\cdot), x(\cdot|s)\rangle \leq D(x(\cdot|s), \tilde{x}_t(\cdot|s)) - D(x(\cdot|s), \tilde{x}_{t+1}(\cdot|s)) + \langle n_t(s,\cdot), \tilde{x}_t(\cdot|s)\rangle + \|n_t(\cdot|s)\|^2_*.$$

We sum the inequality take maximum and then expectation

$$\mathbb{E}\mathbb{E}_{s\sim\sigma}\max_x\sum_{t=1}^T\langle -n_t(s,\cdot), x(\cdot|s)\rangle \leq \mathbb{E}_{s\sim\sigma}D(x(\cdot|s), \tilde{x}_1(\cdot|s)) + \sum_{t=1}^T\mathbb{E}\mathbb{E}_{s\sim\sigma}\langle -n_t(s,\cdot), \tilde{x}_t(\cdot|s)\rangle$$

$$+\sum_{t=1}^{T}\mathbb{E}\mathbb{E}_{s\sim\rho}\|n_t(\cdot|s)\|_\infty^2.$$

By tower property and that $\tilde{x}_t$ is deterministic conditioned on x_t, we have $\sum_{t=1}^{T}\mathbb{E}\langle n_t(s,\cdot),\tilde{x}_t(\cdot|s)\rangle = \sum_{t=1}^{T}\mathbb{E}\langle\mathbb{E}[n_t(s,\cdot)|x_t],\tilde{x}_t(\cdot|s)\rangle = 0$. Recall the definition of n_t and use Young's inequality with Jensen's inequality to get

$$\begin{aligned}\mathbb{E}\|n_t(s,\cdot)\|^2 &= \mathbb{E}\eta^2\|\theta_{t+1}(\cdot|s)-\mathbb{E}[\theta_{t+1}(\cdot|s)|x_t]\|_\infty^2\\ &\leq 2\mathbb{E}\eta^2\|\theta_{t+1}(\cdot|s)-\theta_{\star,t}\|_\infty^2+2\mathbb{E}\eta^2\|\theta_{\star,t}-\mathbb{E}[\theta_{t+1}(\cdot|s)|x_t]\|_\infty^2 \leq 4\mathbb{E}\eta^2\|\theta_{t+1}(\cdot|s)-\theta_{\star,t}\|_\infty^2.\end{aligned}$$

■

Proofs for evaluation step of Reflected NAC with a game etiquette

This part mirror closely the analyses for single agent setting, as the best response step is like a single agent problem where the other agent (fixed) can be seen as part of the environment. Therefore, the development in this part will be similar to [Lan21]. Let us restate that the main concern in this part was to make sure that $\bar{y}_t$ updates do not require seeing the policy x_k or the actions of x-player. As we showed that it is the case, we will only provide the proofs here, with mostly using the arguments of [Lan21]. Therefore, the proofs in this part are brief and are included for being self-contained and for easy navigation. At the point of view of this loop (runs from $n=0,\cdots,N-1$), $v_{\star,t}$ is fixed.

Lemma 7.18. *Let Assumption 7.1 hold, $\beta_n = \frac{2}{\lambda_{\min}^{v}(n+n_0)}$. Evaluation step in Alg. 7.1 satisfies*

$$\mathbb{E}\|v_N - v_{\star,t}\|_2^2 \leq \mathcal{O}\left(\frac{|S||B|}{(1-\gamma)^2N^2}+\frac{1}{N(1-\gamma)^2(\lambda_{\min}^{v})^2}\right),\quad \|\mathbb{E}[v_N|\bar{y}_t]-v_{\star,t}\|_2^2\leq\mathcal{O}\left(\frac{2|S||B|}{(1-\gamma)^2N^2}\right).$$

Proof. For the variance, we have by taking expectation w.r.t. $\xi_n=(s_n,a_n,b_n,s_{n+1},b_{n+1})$

$$\mathbb{E}_{\xi_n}\|v_{n+1}-v_{\star,t}\|_2^2 = \|v_n-v_{\star,t}\|_2^2 - 2\beta_n\langle\mathbb{E}_{\xi_n}[\tilde{F}_t^v(v_n,\xi_n)],v_n-v_{\star,t}\rangle+\beta_n^2\mathbb{E}_{\xi_n}\|\tilde{F}_t^v(v_n,\xi_n)\|_2^2.$$

By $\mathbb{E}_{\xi_n}\tilde{F}_t^v(v_n,\xi_n)=F_t^v(v_n)$, $F_t^v(v_{\star,t})=0$, and strong monotonicity of F_t^v,

$$\mathbb{E}\|v_{n+1}-v_{\star,t}\|_2^2 = \left(1-2\beta_n\lambda_{\min}^{v}\right)\mathbb{E}\|v_n-v_{\star,t}\|_2^2+\beta_n^2\mathbb{E}\|\tilde{F}_t^v(v_n,\xi_n)\|_2^2.$$

The end of the proof is the same as Lemma 7.9.

For the bias, we will argue as in Lemma 7.9. Taking expectation of the recursion w.r.t. ξ_n gives

$$\mathbb{E}_{\xi_n}v_{n+1}=v_n-\beta_nF_t^v(v_n).$$

We now unroll the expectation until y_t and use linearity of F_t^v

$$\mathbb{E}[v_{n+1}|\bar{y}_t]=\mathbb{E}[v_n|\bar{y}_t]-\beta_nF_t^v(\mathbb{E}[v_n|\bar{y}_t]).$$

Denoting $\bar{v}_n = \mathbb{E}[v_n|\bar{y}_t]$ gives

$$\|\bar{v}_{n+1} - v_{\star,t}\|_2^2 = \|\bar{v}_n - v_{\star,t}\|_2^2 - 2\beta_n\langle F_t^v(\bar{v}_n), \bar{v}_n - v_{\star,t}\rangle + \beta_n^2\|F_t^v(\bar{v}_n)\|_2^2.$$

We will now use Lipschitzness and strong monotonicity of F_t^v and that $F_t^v(v_{\star,t}) = 0$ and similar to Lemma 7.9, we obtain the recursion

$$\|\bar{v}_{n+1} - v_{\star,t}\|_2^2 = \left(1 - 2\beta_n\lambda_{\min}^v + \beta_n^2\lambda_{\max}^2\right)\|\bar{v}_n - v_{\star,t}\|_2^2.$$

By the choice of n_0 and β_n, similar to Lemma 7.9 the result follows. ■

We will now give a proof similar to [Lan21, Theorem 2] [AKLM20] regarding the NPG algorithm for finding the best response.

Theorem 7.19. *Let Assumption 7.1 hold and $\eta > 0$. For the evaluation step of Algorithm 7.1.*

$$\frac{1}{T}\sum_{t=1}^{T} V^{x_k,y_k^*}(s_0) - V^{x_k,\bar{y}_t}(s_0) \le \frac{\eta}{(1-\gamma)T}\mathbb{E}\left[\mathbb{E}_{s\sim d_{s_0}^{x_k,y_k^*}} D(y_k^*(\cdot|s), y_1(\cdot|s)) - V^{x_k,\bar{y}_1}(s) + V^{x_k,\bar{y}_{t+1}}(s)\right]$$
$$+ \frac{\eta}{2(1-\gamma)^2}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|v_{\star,t} - v_{t+1}\|_\infty^2 + \frac{2}{1-\gamma}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\mathbb{E}[v_{t+1}|\bar{y}_t] - v_{\star,t}\|_\infty$$

Corollary 7.20. *We use the bound from Lemma 7.18 to obtain*

$$\tilde{\mathcal{O}}\left(\frac{1}{T(1-\gamma)^2}\right) + \mathcal{O}\left(\frac{|S||B|}{(1-\gamma)^4N^2} + \frac{1}{N(1-\gamma)^4(\lambda_{\min}^v)^2}\right) + \mathcal{O}\left(\frac{\sqrt{|S||B|}}{(1-\gamma)^2N}\right). \tag{7.19}$$

Proof. By the update rule of $\bar{y}_{t+1}$, it follows for any $s, \bar{y}$ [Tse08, Property 1]

$$D(\bar{y}(\cdot|s), \bar{y}_{t+1}(\cdot|s)) \le D(\bar{y}(\cdot|s), \bar{y}_t(\cdot|s)) - D(\bar{y}_{t+1}(\cdot|s), \bar{y}_t(\cdot|s)) - \langle\eta v_{t+1}(s,\cdot), \bar{y}(\cdot|s) - \bar{y}_{t+1}(\cdot|s)\rangle. \tag{7.20}$$

We manipulate the inner product

$$\begin{aligned}
-\eta\langle v_{t+1}(s,\cdot), \bar{y}(\cdot|s) - \bar{y}_{t+1}(\cdot|s)\rangle &= -\eta\langle v_{t+1}(s,\cdot), \bar{y}(\cdot|s) - \bar{y}_t(\cdot|s)\rangle - \eta\langle v_{t+1}(s,\cdot), \bar{y}_t(\cdot|s) - \bar{y}_{t+1}(\cdot|s)\rangle \\
&= -\eta\langle v_{\star,t}(s,\cdot), \bar{y}(\cdot|s) - \bar{y}_t(\cdot|s)\rangle - \eta\langle v_{t+1}(s,\cdot), \bar{y}_t(\cdot|s) - \bar{y}_{t+1}(\cdot|s)\rangle \\
&\quad - \eta\langle v_{t+1}(s,\cdot) - v_{\star,t}(s,\cdot), \bar{y}(\cdot|s) - \bar{y}_t(\cdot|s)\rangle.
\end{aligned} \tag{7.21}$$

By the performance difference lemma and using the definition of $v_{\star,t} = \mathbb{E}_{a\sim x_k}Q^{x_k,\bar{y}_t}(\cdot,a,\cdot)$.

$$\begin{aligned}
V^{x_k,\bar{y}_{t+1}}(s_0) - V^{x_k,\bar{y}_t}(s_0) &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{x_k,\bar{y}_{t+1}}}\langle\mathbb{E}_{a\sim x_k(\cdot|s)}Q^{x_k,\bar{y}_t}(s,a,\cdot), \bar{y}_{t+1}(\cdot|s) - \bar{y}_t(\cdot|s)\rangle \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{x_k,\bar{y}_{t+1}}}\langle v_{\star,t}(s,\cdot), \bar{y}_{t+1}(\cdot|s) - \bar{y}_t(\cdot|s)\rangle \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{x_k,\bar{y}_{t+1}}}\Big[\langle v_{t+1}(s,\cdot), \bar{y}_{t+1}(\cdot|s) - \bar{y}_t(\cdot|s)\rangle
\end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad + \langle v_{\star,t}(s,\cdot) - v_{t+1}(s,\cdot), \bar{y}_{t+1}(\cdot|s) - \bar{y}_t(\cdot|s)\rangle \Big] \\
&\geq \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{x_k,\bar{y}_{t+1}}}\Big[\langle v_{t+1}(s,\cdot), \bar{y}_{t+1}(\cdot|s) - \bar{y}_t(\cdot|s)\rangle - \frac{\eta}{2}\|v_{\star,t}(s,\cdot) - v_{t+1}(s,\cdot)\|_\infty^2 \\
&\qquad\qquad - \frac{1}{2\eta}\|\bar{y}_{t+1}(\cdot|s) - \bar{y}_t(\cdot|s)\|_1^2\Big], \qquad (7.22)
\end{aligned}
$$

where the last step uses Cauchy-Schwarz and Young's inequalities.

Plugging in $\bar{y} = \bar{y}_t$ in (7.20) and using strong convexity of D gives

$$
\begin{aligned}
-\eta\langle v_{t+1}(s,\cdot), \bar{y}_t(\cdot|s) - \bar{y}_{t+1}(\cdot|s) &\geq D(\bar{y}_t(\cdot|s), \bar{y}_{t+1}(\cdot|s)) + D(\bar{y}_{t+1}(\cdot|s), \bar{y}_t(\cdot|s)) \\
&\geq \|\bar{y}_{t+1}(\cdot|s) - \bar{y}_t(\cdot|s)\|_1^2,
\end{aligned}
$$

which implies that $\langle v_{t+1}(s,\cdot), \bar{y}_{t+1}(\cdot|s) - \bar{y}_t(\cdot|s)\rangle - \frac{1}{2\eta}\|\bar{y}_{t+1}(\cdot|s) - \bar{y}_t(\cdot|s) \geq 0$.

Recall that $d_{s_0}^{x_k,\bar{y}t+1}(s) = (1-\gamma)\sum_{t=0}^\infty \gamma^t \Pr^{x_k,\bar{y}_{t+1}}(s_t = s|s_0)$, therefore $1-\gamma \leq d_{s_0}^{x_k,\bar{y}t+1}(s_0) \leq 1$. Using the two previous inequalities in (7.22) gives

$$
\begin{aligned}
\langle v_{t+1}(s,\cdot), \bar{y}_{t+1}(\cdot|s) - \bar{y}_t(\cdot|s)\rangle \leq V^{x_k,\bar{y}_{t+1}}(s) - V^{x_k,\bar{y}_t}(s) + \frac{1}{2\eta}\|\bar{y}_{t+1}(\cdot|s) - \bar{y}_t(\cdot|s)\|_1^2 \\
+ \frac{\eta}{2(1-\gamma)}\|v_{\star,t} - v_{t+1}\|_\infty^2.
\end{aligned}
$$

We use the final inequality, (7.21), and strong convexity of D in (7.20) to get

$$
\begin{aligned}
&\eta\langle v_{\star,t}(s,\cdot), \bar{y}(\cdot|s) - \bar{y}_t(\cdot|s)\rangle + D(\bar{y}(\cdot|s), y_{t+1}(\cdot|s)) - \eta V^{x_k,\bar{y}_{t+1}}(s) \leq D(\bar{y}(\cdot|s), y_t(\cdot|s)) \\
&\quad - \eta V^{x_k,\bar{y}_t}(s) + \frac{\eta^2}{2(1-\gamma)}\|v_{\star,t} - v_{t+1}\|_\infty^2 + \eta\langle v_{\star,t}(s,\cdot) - v_{t+1}(s,\cdot), \bar{y}(\cdot|s) - \bar{y}_t(\cdot|s)\rangle. \quad (7.23)
\end{aligned}
$$

In view of the definition $v_{\star,t} = \mathbb{E}_{a\sim x_k(\cdot|s)}Q^{x_k,\bar{y}_t}(\cdot,a,\cdot)$, performance difference lemma gives $(1-\gamma)(V^{x_k,y_k^*}(s_0) - V^{x_k,\bar{y}_t}(s_0)) = \mathbb{E}_{s\sim d_{s_0}^{x_k,y_k^*}}\langle v_{\star,t}(s,\cdot), y_k^*(\cdot|s) - y_t(\cdot|s)\rangle$. Plugging in $y = y_k^*$ in (7.23) and taking $\mathbb{E}_{s\sim d_{s_0}^{x_k,y_k^*}}$ of both sides give

$$
\begin{aligned}
&\eta(1-\gamma)(V^{x_k,y_k^*}(s_0) - V^{x_k,\bar{y}_t}(s_0)) + \mathbb{E}_{s\sim d_{s_0}^{x_k,y_k^*}}\Big[D(y_k^*(\cdot|s), \bar{y}_{t+1}(\cdot|s)) - \eta V^{x_k,\bar{y}_{t+1}}(s)\Big] \\
&\quad\leq \mathbb{E}_{s\sim d_{s_0}^{x_k,y_k^*}}\Big[D(y_k^*(\cdot|s), \bar{y}_t(\cdot|s)) - \eta V^{x_k,\bar{y}_t}(s) + \frac{\eta^2}{2(1-\gamma)}\|v_{\star,t} - v_{t+1}\|_\infty^2 \\
&\qquad\qquad + \eta\langle v_{\star,t}(s,\cdot) - v_{t+1}(s,\cdot), \bar{y}_k^*(\cdot|s) - \bar{y}_t(\cdot|s)\rangle\Big]. \quad (7.24)
\end{aligned}
$$

We take expectation w.r.t. the randomness in the algorithm, use tower property, the fact that conditioned on $\bar{y}_t$, $y_k^*(\cdot|s) - \bar{y}_t(\cdot|s)$ is deterministic, Cauchy-Schwarz inequality, $\bar{y}(\cdot|s) \in \Delta$ for any $\bar{y}, s$. Then, we note that $d_{s_0}^{x_k,y_k^*}$ does not depend on t and sum the inequality over t. ■

8 Conclusions and Future Directions

In this dissertation, we designed and analyzed stochastic algorithms for solving structured nonsmooth problems given in eqs. (1.2) and (1.3). Our results enhance the toolkit of continuous optimization via adaptive and practical algorithms with optimal rate and complexity guarantees, either matching or improving the state-of-the-art.

We summarize the contributions of each chapter with potential future directions.

• In Chapter 2, we studied adaptive gradient methods for solving nonsmooth stochastic optimization problems with convex and nonconvex objectives. First, we introduced a regret analysis framework for the more general online convex optimization (OCO) template. This framework addresses an important theory-practice gap on the choice of exponential moving average (EMA) parameters, which are of paramount importance on the empirical success of these methods. Next, we analyze an Adam-type algorithm for solving constrained weakly convex problems. This problem template generalizes unconstrained smooth minimization, which is the only setting studied in the previous literature for adaptive methods.

Role of momentum. Even though our developments made it possible to obtain regret guarantees with constant EMA (a.k.a. momentum) parameters, our bounds still suggest not incorporation momentum. This is common in all the analyses that we are aware for Adam-type algorithms. Recently, some works investigate the role of momentum to idenfity when momentum improves SGD [Def20, GLZX19, CM20]. On this line of work, we believe that it is interesting to study adaptivity along with momentum to understand why or when Adam-type algorithms can improve SGD.

Convergence for nonsmooth nonconvex optimization. Even though our results for nonconvex case generalize the existing convergence results, they are still not sufficient to explain why these methods work in complicated neural network structures with nonsmooth and nonconvex objective functions. The hardness of this problem template is recently established along with some positive results for SGD [ZLSJ20, Sha20]. Studying convergence properties of Adam-type algorithms for these general problems is a natural future direction due to their practical use.

• In Chapter 3, we designed algorithms coupling Nesterov's smoothing with SGD and accelerated proximal coordinate descent to solve linearly constrained problems with optimal rates. Our results also stochastic optimization problems with stochastic constraints that hold almost surely.

Stochastic constraints. One natural future direction concerning our developments is to consider more general classes of stochastic constraints. Some examples are constraints that hold in expectation or probabilistic constraints with potential use in the popular field of distributionally robust optimization [RM19].

Theoretical understanding of restart. To enhance the empirical performance of our smoothing based accelerated proximal CD framework, we found out that restarting strategies were essential. We showed in a followup work that smoothing with deterministic accelerated methods still have the same worst-case guarantees [TDAFC19]. However, this does not explain the improved practical performance we observe with restarts. Some recent progress on restarting in minimization [FQ20] and deterministic primal-dual optimization [Fer21, HL20] can be good starting points.

• In Chapter 4, we analyzed the stochastic primal-dual hybrid gradient (SPDHG) algorithm which had empirical success but weak theoretical guarantees. To explain the favorable practical performance, we proved almost sure convergence of the sequence to a solution, the optimal $\mathcal{O}(1/k)$ rate on the expected primal-dual gap and adaptive linear convergence with an error bound condition.

Tighter analysis for adaptive linear convergence. Despite these strengthening of the guarantees, our results still do not completely characterize the adaptive linear convergence behavior of SPDHG in practice. In particular, one of the three main contributions of Chapter 4 was to show that SPDHG obtains linear rate of convergence under general assumptions that hold for a large body of problems, with an agnostic step size selection. A natural question is: How does this rate translate to practice? For this purpose, we perform a controlled experiment on a simple problem

$$\min_{x\in\mathbb{R}^d} \frac{\mu}{2}\|x\|^2 : Ax = b,$$

with $d = n = 10$. Upon writing the KKT conditions for the problem, we obtain $F = \begin{bmatrix} \mu I & A^\top \\ A & 0 \end{bmatrix}$ and metric subregularity constant η is the smallest eigenvalue of F in absolute value.

For simplicity, we focus on PDHG, which is a specific case of SPDHG, and plot the predicted rate and the empirical rate in Figure 8.1.

We observe that the empirical rate of convergence is much faster than the worst case rate predicted by theory. We point out several explanations for this phenomenon: ∘ Metric subreg-

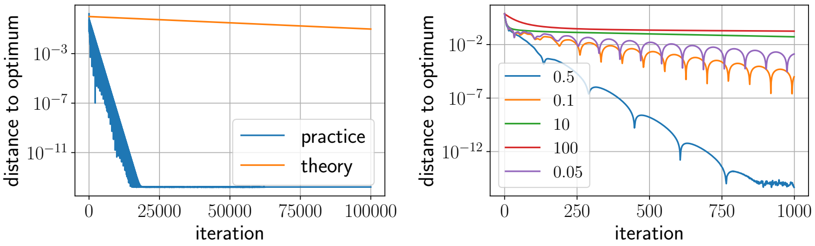

Figure 8.1 – left: empirical and theoretical linear rates, right: empirical rates with different μ.

ularity is too general to capture structures observed in practice.

○ Our step size choice is independent of metric subregularity constant, preventing optimizing the theoretical rate with respect to these quantities.

In fact, this phenomenon is not specific to our analysis and it seems like a drawback of the existing approaches utilizing metric subregularity [LFP19, LFP16]. On this front, we observe that in our example, as μ increases, metric subregularity constant η degrades. However, as we see in the plot, the practical performance degrades when μ is either too big or too small (see Figure 8.1). This observation suggests that there might exist better regularity measures beyond metric subregularity that would help us derive better rates. We believe that this is a promising future direction. A recent work by Fercoq [Fer21] made progress on this question.

- In Chapter 5 we designed the first PDCD algorithm which is efficient with both sparse and dense data. Moreover, we also proved the best-known convergence guarantees for our new method PURE-CD, using some of the techniques we developed in Chapter 4. In our numerical experiments with varying levels of sparsity, PURE-CD showed the fastest convergence among the the state-of-the-arts, as predicted by our theory.

Strong convexity and better bounds. We believe that there are two interesting future directions on improving the theoretical understanding of PURE-CD. The most immediate one is to exploit strong convexity when it exists, to improve the convergence rate. Strongly convex setting is much more studied in the literature due to its simplicity and we expect PURE-CD to recover the best-known theoretical results. The second future direction is a more careful study on the dimension dependence of our bound. In particular, we expect PURE-CD to obtain a better overall complexity than deterministic methods and PDCD methods without adaptation to sparsity.

Linear programming solver. As also mentioned before, metric subregularity that we used for showing linear convergence is connected to Hoffman's Lemma for linear programming (LP). There is a recent trend in the literature to design fast first-order algorithms for LPs [AHLL21, YZH$^+$15]. Our practical experience with PURE-CD was highly positive due to its adaptivity to metric subregularity and sparsity simultaneously. An interesting future direction is to particularize PURE-CD for LPs both theoretically and practically by implementing it as an LP solver.

• In Chapter 6, we designed variance reduced algorithms for solving convex-concave min-max problems with finite sum structure. Our algorithms have almost sure convergence under convexity, and they improve the existing complexity results. Recent work also showed matching lower bounds to our developments.

Sparsity. The recent work by [CJST20] built on the algorithm in [CJST19] and improved the complexity for matrix games in Euclidean setup, for sparse data, by using specialized data structures. We believe that these techniques can also be used in our algorithms.

Stochastic oracles. As we have seen for bilinear and nonbilinear problems, harnessing the structure is very important for devising suitable stochastic oracles with small Lipschitz constants. On top of our algorithms, an interesting direction is to study important nonbilinear min-max problems and devise particular Bregman distances and stochastic oracles to obtain complexity improvements.

• In Chapter 7, we analyzed a policy optimization method for solving two player zero-sum Markov games. Our sample complexity result improved the best-known complexity of policy gradient methods in this setting.

Simpler algorithms. To get the tightest estimates, we needed to use an asymmetric algorithm with inner loops for policy evaluation. We think that an interesting future direction is to design symmetric algorithms and/or algorithms with less inner loops with the same complexity estimate. These aspects would enhance the practical impact of the method.

Bibliography

[AAEF07] Fouad Ben Abdelaziz, Belaid Aouni, and Rimeh El Fayedh. Multi-objective stochastic programming for portfolio selection. *European Journal of Operational Research*, 177(3):1811–1823, 2007.

[AAF+20] Mahmoud Assran, Arda Aytekin, Hamid Reza Feyzmahdavian, Mikael Johansson, and Michael G Rabbat. Advances in asynchronous parallel and distributed optimization. *Proceedings of the IEEE*, 108(11):2013–2031, 2020.

[AAHU58] Kenneth Joseph Arrow, Hirofumi Azawa, Leonid Hurwicz, and Hirofumi Uzawa. *Studies in linear and non-linear programming*, volume 2. Stanford University Press, 1958.

[ACBG02] Peter Auer, Nicolo Cesa-Bianchi, and Claudio Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75, 2002.

[ADFC17] Ahmet Alacaoglu, Quoc Tran Dinh, Olivier Fercoq, and Volkan Cevher. Smooth primal-dual coordinate descent algorithms for nonsmooth convex optimization. In *Advances in Neural Information Processing Systems*, pages 5852–5861, 2017.

[AFC20] Ahmet Alacaoglu, Olivier Fercoq, and Volkan Cevher. Random extrapolation for primal-dual coordinate descent. In *International conference on machine learning*, pages 191–201. PMLR, 2020.

[AFC21] Ahmet Alacaoglu, Olivier Fercoq, and Volkan Cevher. On the convergence of stochastic primal-dual hybrid gradient. *SIAM Journal on Optimization*, forthcoming, 2021.

[AHLL21] David Applegate, Oliver Hinder, Haihao Lu, and Miles Lubin. Faster first-order primal-dual methods for linear programming using restarts and sharpness. *arXiv preprint arXiv:2105.12715*, 2021.

[AKLM20] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pages 64–66. PMLR, 2020.

[AKSV18] Sanjeev Arora, Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. A compressed sensing view of unsupervised text embeddings, bag-of-n-grams, and LSTMs. In *International Conference on Learning Representations*, 2018.

[AM21] Ahmet Alacaoglu and Yura Malitsky. Stochastic variance reduction for variational inequality methods. *arXiv preprint arXiv:2102.08352*, 2021.

[AMC20] Ahmet Alacaoglu, Yura Malitsky, and Volkan Cevher. Convergence of adaptive algorithms for weakly convex constrained optimization. *arXiv preprint arXiv:2006.06650*, 2020.

[AMC21] Ahmet Alacaoglu, Yura Malitsky, and Volkan Cevher. Forward-reflected-backward method with variance reduction. *Computational Optimization and Applications*, 2021.

Bibliography

[AMMC20] Ahmet Alacaoglu, Yura Malitsky, Panayotis Mertikopoulos, and Volkan Cevher. A new regret analysis for Adam-type algorithms. In *International Conference on Machine Learning*, 2020.

[AWBR09] Alekh Agarwal, Martin J Wainwright, Peter Bartlett, and Pradeep Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. *Advances in Neural Information Processing Systems*, 22:1–9, 2009.

[AZ17] Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):8194–8244, 2017.

[AZY16] Zeyuan Allen-Zhu and Yang Yuan. Improved SVRG for non-strongly-convex or sum-of-non-convex objectives. In *International conference on machine learning*, pages 1080–1089. PMLR, 2016.

[BB61] Edwin F Beckenbach and Richard Bellman. *Inequalities*, volume 30. Springer Science & Business Media, 1961.

[BB16] Palaniappan Balamurugan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems*, pages 1416–1424, 2016.

[BB20] Anas Barakat and Pascal Bianchi. Convergence rates of a momentum algorithm with bounded adaptive step size for nonconvex optimization. In *Asian Conference on Machine Learning*, pages 225–240. PMLR, 2020.

[BC11] Heinz H Bauschke and Patrick L Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011.

[Ber99] Dimitri Bertsekas. *Nonlinear programming*. Athena Scientific, 1999.

[Ber11] Dimitri P Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical Programming*, 129(2):163, 2011.

[BEYG04] Allan Borodin, Ran El-Yaniv, and Vincent Gogan. Can we learn to beat the best stock. In *Advances in Neural Information Processing Systems*, pages 345–352, 2004.

[BHS19] Pascal Bianchi, Walid Hachem, and Adil Salim. A constant step forward-backward algorithm involving random maximal monotone operators. *Journal of Convex Analysis*, 26(2):397–436, 2019.

[Bia15] Pascal Bianchi. A stochastic proximal point algorithm: convergence and application to convex optimization. In *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2015 IEEE 6th International Workshop on*, pages 1–4. IEEE, 2015.

[BJ20] Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *International Conference on Machine Learning*, pages 551–560. PMLR, 2020.

[BJY20] Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. *Advances in Neural Information Processing Systems*, 33, 2020.

[BLNZ95] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[BMSV19] Radu Ioan Boţ, Panayotis Mertikopoulos, Mathias Staudigl, and Phan Tu Vuong. Forward-backward-forward methods with variance reduction for stochastic variational inequalities. *arXiv preprint arXiv:1902.03355*, 2019.

[BR19] Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

[BR21] Jalaj Bhandari and Daniel Russo. On the linear convergence of policy gradient methods for finite mdps. In *International Conference on Artificial Intelligence and Statistics*, pages 2386–2394. PMLR, 2021.

[BRM19] Jingjing Bu, Lillian J Ratliff, and Mehran Mesbahi. Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games. *arXiv preprint arXiv:1911.04672*, 2019.

[BRS18] Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference On Learning Theory*, pages 1691–1692. PMLR, 2018.

[BS19] Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365(6456):885–890, 2019.

[BSCB20] Axel Böhm, Michael Sedlmayer, Ernö Robert Csetnek, and Radu Ioan Boţ. Two steps at a time–taking gan training in stride with tseng's method. *arXiv preprint arXiv:2006.09033*, 2020.

[BT09] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[BTSK17] Felix Berkenkamp, Matteo Turchetta, Angela P Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 908–919, 2017.

[CCC$^+$20] Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *arXiv preprint arXiv:2007.06558*, 2020.

[CDS01] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

[CERS18] Antonin Chambolle, Matthias J Ehrhardt, Peter Richtárik, and Carola-Bibiane Schonlieb. Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. *SIAM Journal on Optimization*, 28(4):2783–2808, 2018.

[CGFLJ19] Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in GAN training with variance reduced extragradient. In *Advances in Neural Information Processing Systems*, pages 391–401, 2019.

[CJST19] Yair Carmon, Yujia Jin, Aaron Sidford, and Kevin Tian. Variance reduction for matrix games. In *Advances in Neural Information Processing Systems*, pages 11377–11388, 2019.

[CJST20] Yair Carmon, Yujia Jin, Aaron Sidford, and Kevin Tian. Coordinate methods for matrix games. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 283–293. IEEE, 2020.

[CL11a] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

[CL11b] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[CLSH19] Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. In *International Conference on Learning Representations*, 2019.

[CLX⁺19] Xiangyi Chen, Sijia Liu, Kaidi Xu, Xingguo Li, Xue Lin, Mingyi Hong, and David Cox. ZO-AdaMM: Zeroth-order adaptive momentum method for black-box optimization. In *Advances in Neural Information Processing Systems*, pages 7202–7213, 2019.

[CM20] Ashok Cutkosky and Harsh Mehta. Momentum improves normalized sgd. In *International Conference on Machine Learning*, pages 2260–2268. PMLR, 2020.

[Con13] Laurent Condat. A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 158(2):460–479, 2013.

[CP11] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[CP15] Patrick L Combettes and Jean-Christophe Pesquet. Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. *SIAM Journal on Optimization*, 25(2):1221–1248, 2015.

[CP16a] Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016.

[CP16b] Antonin Chambolle and Thomas Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1-2):253–287, 2016.

[CP19] Patrick L Combettes and Jean-Christophe Pesquet. Stochastic quasi-fejér block-coordinate fixed point iterations with random sweeping ii: mean-square and linear convergence. *Mathematical Programming*, 174(1-2):433–451, 2019.

[CS21] Shisheng Cui and Uday V Shanbhag. On the analysis of variance-reduced and randomized projection variants of single projection schemes for monotone stochastic variational inequality problems. *Set-Valued and Variational Analysis*, pages 1–47, 2021.

[CSN⁺19] Dami Choi, Christopher J Shallue, Zachary Nado, Jaehoon Lee, Chris J Maddison, and George E Dahl. On empirical comparisons of optimizers for deep learning. *arXiv preprint arXiv:1910.05446*, 2019.

[CZT⁺20] Jinghui Chen, Dongruo Zhou, Yiqi Tang, Ziyan Yang, Yuan Cao, and Quanquan Gu. Closing the generalization gap of adaptive gradient methods in training deep neural networks. In *29th International Joint Conference on Artificial Intelligence*, 2020.

[DBBU20] Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. On the convergence of Adam and Adagrad. *arXiv preprint arXiv:2003.02395*, 2020.

[DBLJ14] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

[DD19] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[DDKL20] Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics*, 20(1):119–154, 2020.

[DDP20] Damek Davis, Dmitriy Drusvyatskiy, and Courtney Paquette. The nonsmooth landscape of phase retrieval. *IMA Journal of Numerical Analysis*, 40(4):2652–2695, 2020.

[Def20] Aaron Defazio. Understanding the role of momentum in non-convex optimization: Practical insights from a lyapunov analysis. *arXiv preprint arXiv:2010.00406*, 2020.

[DFG20] Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[DG19] Damek Davis and Benjamin Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019.

[DGTV14] Elvis Dopgima Dohmatob, Alexandre Gramfort, Bertrand Thirion, and Gael Varoquaux. Benchmarking solvers for tv-ℓ_1 least-squares and logistic regression in brain imaging. In *Pattern Recognition in Neuroimaging, 2014 International Workshop on*, pages 1–4. IEEE, 2014.

[DHS10] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. Technical Report UCB/EECS-2010-24, EECS Department, University of California, Berkeley, Mar 2010.

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[DISZ18] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with Optimism. In *International Conference on Learning Representations*, 2018.

[DL14] Cong Dang and Guanghui Lan. Randomized first-order methods for saddle point optimization. *arXiv preprint arXiv:1409.8625*, 2014.

[DL15a] Cong D Dang and Guanghui Lan. On the convergence properties of non-Euclidean extragradient methods for variational inequalities with generalized monotone operators. *Computational Optimization and applications*, 60(2):277–310, 2015.

[DL15b] Cong D Dang and Guanghui Lan. Stochastic block mirror descent methods for nonsmooth and stochastic optimization. *SIAM Journal on Optimization*, 25(2):856–881, 2015.

[DL18] Dmitriy Drusvyatskiy and Adrian S Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

[Don06] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[DP18] Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems*, pages 9236–9246, 2018.

[DP19] Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1-2):503–558, 2019.

[DR09] Asen L Dontchev and R Tyrrell Rockafellar. Implicit functions and solution mappings. *Springer Monographs in Mathematics. Springer*, 208, 2009.

[DR18] John C Duchi and Feng Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, 28(4):3229–3259, 2018.

[DR19] John C Duchi and Feng Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Information and Inference: A Journal of the IMA*, 8(3):471–529, 2019.

[EM14] Yonina C Eldar and Shahar Mendelson. Phase retrieval: Stability and recovery guarantees. *Applied and Computational Harmonic Analysis*, 36(3):473–494, 2014.

[EMC$^+$17] Matthias J Ehrhardt, Pawel Markiewicz, Antonin Chambolle, Peter Richtárik, Jonathan Schott, and Carola-Bibiane Schönlieb. Faster pet reconstruction with a stochastic primal-dual hybrid gradient method. In *Wavelets and Sparsity XVII*, volume 10394, page 103941O. International Society for Optics and Photonics, 2017.

[EZC10] Ernie Esser, Xiaoqun Zhang, and Tony F Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences*, 3(4):1015–1046, 2010.

[FANC19] Olivier Fercoq, Ahmet Alacaoglu, Ion Necoara, and Volkan Cevher. Almost surely constrained convex optimization. In *International Conference on Machine Learning*, pages 1910–1919, 2019.

[FB19] Olivier Fercoq and Pascal Bianchi. A coordinate-descent primal-dual algorithm with large step size and possibly nonseparable functions. *SIAM Journal on Optimization*, 29(1):100–134, 2019.

[Fer19] Olivier Fercoq. A generic coordinate descent solver for non-smooth convex optimisation. *Optimization Methods and Software*, pages 1–21, 2019.

[Fer21] Olivier Fercoq. Quadratic error bound of the smoothed gap and the restarted averaged primal-dual hybrid gradient. preprint: hal-03228252, May 2021.

[FK19] Biyi Fang and Diego Klabjan. Convergence analyses of online ADAM algorithm in convex setting and two-layer relu neural network. *arXiv preprint arXiv:1905.09356*, 2019.

[FQ20] Olivier Fercoq and Zheng Qu. Restarting the accelerated coordinate descent method with a rough strong convexity estimate. *Computational Optimization and Applications*, 75(1):63–91, 2020.

[FR15] Olivier Fercoq and Peter Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015.

[GB14] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx, March 2014.

[GG09] Pierre Garrigues and Laurent E Ghaoui. An homotopy algorithm for the lasso with online observations. In *Advances in neural information processing systems*, pages 489–496, 2009.

[GGBHD05] Isabelle Guyon, Steve Gunn, Asa Ben-Hur, and Gideon Dror. Result analysis of the nips 2003 feature selection challenge. In *Advances in neural information processing systems*, pages 545–552, 2005.

[GK95] Michael D Grigoriadis and Leonid G Khachiyan. A sublinear-time randomized approximation algorithm for matrix games. *Operations Research Letters*, 18(2):53–58, 1995.

[GL13] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[GLZ16] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305, 2016.

[GLZX19] Igor Gitman, Hunter Lang, Pengchuan Zhang, and Lin Xiao. Understanding the role of momentum in stochastic gradient methods. *Advances in Neural Information Processing Systems*, 32:9633–9643, 2019.

[GPAM$^+$14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[GS18] Tom Goldstein and Christoph Studer. Phasemax: Convex phase retrieval via basis pursuit. *IEEE Transactions on Information Theory*, 64(4):2675–2689, 2018.

[GXZ19] Xiang Gao, Yang-Yang Xu, and Shu-Zhong Zhang. Randomized primal–dual proximal block coordinate updates. *Journal of the Operations Research Society of China*, 7(2):205–250, 2019.

[HA21] Erfan Yazdandoost Hamedani and Necdet Serhat Aybat. A primal-dual algorithm with line search for general convex-concave saddle point problems. *SIAM Journal on Optimization*, 31(2):1299–1329, 2021.

[HAK07] Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

[Haz16] Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

[Hil57] Clifford Hildreth. A quadratic programming procedure. *Naval Research Logistics Quarterly*, 4(1):79–85, 1957.

[HL20] Oliver Hinder and Miles Lubin. A generic adaptive restart scheme with applications to saddle point algorithms. *arXiv preprint arXiv:2006.08484*, 2020.

[HLLJM15] Thomas Hofmann, Aurelien Lucchi, Simon Lacoste-Julien, and Brian McWilliams. Variance reduced stochastic gradient descent with neighbors. In *Advances in Neural Information Processing Systems*, pages 2305–2313, 2015.

[Hof52] Alan J Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49(4), 1952.

[HRTZ04] Trevor Hastie, Saharon Rosset, Robert Tibshirani, and Ji Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5(Oct):1391–1415, 2004.

[HSNL18] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

[HWD19] Haiwen Huang, Chang Wang, and Bin Dong. Nostalgic Adam: Weighting more of the past gradients when designing the adaptive learning rate. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 2556–2562, 2019.

[HWWY20] Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170*, 2020.

[HXZ21] Yuze Han, Guangzeng Xie, and Zhihua Zhang. Lower complexity bounds of finite-sum optimization problems: The results and construction. *arXiv preprint arXiv:2103.08280*, 2021.

[HY12] Bingsheng He and Xiaoming Yuan. Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective. *SIAM Journal on Imaging Sciences*, 5(1):119–149, 2012.

[IBCH13] Franck Iutzeler, Pascal Bianchi, Philippe Ciblat, and Walid Hachem. Asynchronous distributed optimization using a randomized alternating direction method of multipliers. In *52nd IEEE conference on decision and control*, pages 3671–3676. IEEE, 2013.

[IEAL18] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning*, pages 2137–2146, 2018.

[IJOT17] Alfredo N Iusem, Alejandro Jofré, Roberto Imbuzeiro Oliveira, and Philip Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 27(2):686–724, 2017.

[JNT11] Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[JOV09] Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pages 433–440, 2009.

[JZ13] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

[Kak01] Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

[Kar37] S Karczmarz. Angenaherte auflosung von systemen linearer glei-chungen. *Bull. Int. Acad. Pol. Sic. Let., Cl. Sci. Math. Nat.*, pages 355–357, 1937.

[KB15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[KBP13] Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

[KCM21] Sajad Khodadadian, Zaiwei Chen, and Siva Theja Maguluri. Finite-sample analysis of off-policy natural actor-critic algorithm. In *International Conference on Machine Learning*. PMLR, 2021.

[KDMR21] Sajad Khodadadian, Thinh T Doan, Siva Theja Maguluri, and Justin Romberg. Finite sample analysis of two-time-scale natural actor-critic algorithm. *arXiv preprint arXiv:2101.10506*, 2021.

[KHR20] Dmitry Kovalev, Samuel Horvath, and Peter Richtárik. Don't Jump Through Hoops and Remove Those Loops: SVRG and Katyusha are Better Without the Outer Loop. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, pages 451–467, 2020.

[KL02] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

[Kor76] GM Korpelevich. The extragradient method for finding saddle points and other problems. *Ekon. Mat. Metody*, 12:747–756, 1976.

[KPB$^+$19] Daniil Kazantsev, Edoardo Pasca, Mark Basham, Martin Turner, Matthias J Ehrhardt, Kris Thielemans, Benjamin A Thomas, Evgueni Ovtchinnikov, Philip J Withers, and Alun W Ashton. Versatile regularisation toolkit for iterative image reconstruction with proximal splitting algorithms. In *15th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine*, volume 11072, page 110722D. International Society for Optics and Photonics, 2019.

[KT00] Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014. Citeseer, 2000.

[Lan21] Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *arXiv preprint arXiv:2102.00135*, 2021.

[LFP16] Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Convergence rates with inexact non-expansive operators. *Mathematical Programming*, 159(1-2):403–434, 2016.

[LFP19] Puya Latafat, Nikolaos M Freris, and Panagiotis Patrinos. A new randomized block-coordinate primal-dual proximal algorithm for distributed optimization. *IEEE Transactions on Automatic Control*, 64(10):4050–4065, 2019.

[Lic13] M. Lichman. UCI machine learning repository, 2013.

[Lit94] Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

[LL19] Pan Liu and Xin Yang Lu. Real order (an)-isotropic total variation in image processing-part ii: Learning of optimal structures. *arXiv preprint arXiv:1903.08513*, 2019.

[LM18] D Russell Luke and Yura Malitsky. Block-coordinate primal-dual method for nonsmooth minimization over linear constraints. In *Large-Scale and Distributed Optimization*, pages 121–147. Springer, 2018.

[LO19] Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 983–992, 2019.

[LP18] Puya Latafat and Panagiotis Patrinos. Primal-dual proximal algorithms for structured convex optimization: A unifying framework. In *Large-Scale and Distributed Optimization*, pages 97–120. Springer, 2018.

[LS13] Yin Tat Lee and Aaron Sidford. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 147–156. IEEE, 2013.

[LXL19] Liangchen Luo, Yuanhao Xiong, and Yan Liu. Adaptive gradient methods with dynamic bound of learning rate. In *International Conference on Learning Representations*, 2019.

[LYRL04] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397, 2004.

[LYZZ18] Yongchao Liu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. Partial error bound conditions and the linear convergence rate of the alternating direction method of multipliers. *SIAM Journal on Numerical Analysis*, 56(4):2095–2123, 2018.

[Mal19] Yura Malitsky. Golden ratio algorithms for variational inequalities. *Mathematical Programming*, pages 1–28, 2019.

[MB11] Eric Moulines and Francis R Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

[MJ20] Vien V Mai and Mikael Johansson. Convergence of a stochastic gradient method with momentum for nonsmooth nonconvex optimization. In *International Conference on Machine Learning*, 2020.

[MKS$^+$15] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[MKS$^+$20] Konstantin Mishchenko, Dmitry Kovalev, Egor Shulgin, Peter Richtárik, and Yura Malitsky. Revisiting stochastic extragradient. In *The 23rd International Conference on Artificial Intelligence and Statistics*, pages 4573–4582. PMLR, 2020.

[MLZ$^+$19] Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. In *International Conference on Learning Representations*, 2019.

[MMS$^+$18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[MNSF17] Pablo Marquez Neila, Mathieu Salzmann, and Pascal Fua. Imposing hard constraints on deep networks: Promises and limitations. In *CVPR Workshop on Negative Results in Computer Vision*, number CONF, 2017.

[MOP20] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507. PMLR, 2020.

[MT20a] Yura Malitsky and Matthew K Tam. A forward-backward splitting method for monotone inclusions without cocoercivity. *SIAM Journal on Optimization*, 30(2):1451–1472, 2020.

[MT20b] Yura Malitsky and Matthew K Tam. A forward-backward splitting method for monotone inclusions without cocoercivity. *SIAM Journal on Optimization*, 30(2):1451–1472, 2020.

[Mun03] Rémi Munos. Error bounds for approximate policy iteration. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, pages 560–567, 2003.

[MXSS20] Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR, 2020.

[MYJ13] Mehrdad Mahdavi, Tianbao Yang, and Rong Jin. Stochastic convex optimization with multiple objectives. In *Advances in Neural Information Processing Systems*, pages 1115–1123, 2013.

[Nem04] Arkadi Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[Nes03] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[Nes05] Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

[Nes07] Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007.

[Nes12] Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[NJLS09] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[NNG18] Ion Necoara, Yurii Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, doi: 10.1007/s10107-018-1232-1, 2018.

[NRP19] Ion Necoara, Peter Richtárik, and Andrei Patrascu. Randomized projection methods for convex feasibility: Conditioning and convergence rates. *SIAM Journal on Optimization*, 29(4):2814–2852, 2019.

[NY83] Arkadi Nemirovski and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

[OC15] Brendan O'Donoghue and Emmanuel Candes. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.

[OG12] Hua Ouyang and Alexander Gray. Stochastic smoothing for nonsmooth minimizations: Accelerating SGD by exploiting structure. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 33–40, 2012.

[PAD$^+$21] Evangelos Papoutsellis, Evelina Ametova, Claire Delplancke, Gemma Fardell, Jakob S Jørgensen, Edoardo Pasca, Martin Turner, Ryan Warr, William RB Lionheart, and Philip J Withers. Core imaging library–part ii: Multichannel reconstruction for dynamic and spectral tomography. *arXiv preprint arXiv:2102.06126*, 2021.

[PJ92] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[PN17] Andrei Patrascu and Ion Necoara. Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization. *Journal of Machine Learning Research*, 18:198–1, 2017.

[Pop80] Leonid Denisovich Popov. A modification of the Arrow-Hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

[PR15] Jean-Christophe Pesquet and Audrey Repetti. A class of randomized primal-dual algorithms for distributed optimization. *Journal of Nonlinear and Convex Analysis*, 16(12):2453–2490, 2015.

[PS08] Jan Peters and Stefan Schaal. Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190, 2008.

[PSPP15] Julien Perolat, Bruno Scherrer, Bilal Piot, and Olivier Pietquin. Approximate dynamic programming for two-player zero-sum markov games. In *International Conference on Machine Learning*, pages 1321–1329. PMLR, 2015.

[QR16] Zheng Qu and Peter Richtárik. Coordinate descent with arbitrary sampling i: Algorithms and complexity. *Optimization Methods and Software*, 31(5):829–857, 2016.

[RHS$^+$16] Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323. PMLR, 2016.

[RKK18] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

[RM51] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[RM19] Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

[RS71] Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier, 1971.

[RS13] Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, pages 3066–3074, 2013.

[RT14] Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, 2014.

[RT16] Peter Richtárik and Martin Takáč. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156(1-2):433–484, 2016.

[RVV20] Lorenzo Rosasco, Silvia Villa, and Bang Công Vũ. Convergence of stochastic proximal gradient algorithm. *Applied Mathematics & Optimization*, 82(3):891–917, 2020.

[RW09] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[Sal18] Adil Salim. *Random monotone operators and application to stochastic optimization*. PhD thesis, Université Paris-Saclay, 2018.

[Sav19] Pedro Savarese. On the convergence of adabound and its connection to sgd. *arXiv preprint arXiv:1908.04457*, 2019.

[SEM20] Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5668–5675, 2020.

[Sha53] Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

[Sha20] Ohad Shamir. Can we find near-approximately-stationary points of nonsmooth nonconvex functions? *arXiv preprint arXiv:2002.11962*, 2020.

[SLA+15] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[SMSM00] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural information Processing Systems*, 12:1057–1063, 2000.

[SMV+20] Prabhu Teja Sivaprasad, Florian Mai, Thijs Vogels, Martin Jaggi, and François Fleuret. Optimizer benchmarking needs to account for hyperparameter tuning. In *International Conference on Machine Learning*, pages 9036–9045. PMLR, 2020.

[SSS+17] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[SSSSC11] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30, 2011.

[SSZ13] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599, 2013.

[SSZ14] Shai Shalev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In *International Conference on Machine Learning*, pages 64–72, 2014.

[Sut88] Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

[SWD+17] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[SWYY20] Aaron Sidford, Mengdi Wang, Lin Yang, and Yinyu Ye. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pages 2992–3002. PMLR, 2020.

[SZY17] Zhan Shi, Xinhua Zhang, and Yaoliang Yu. Bregman divergence for stochastic variance reduction: Saddle-point and adversarial prediction. In *NIPS*, pages 6031–6041, 2017.

[TDAFC19] Quoc Tran-Dinh, Ahmet Alacaoglu, Olivier Fercoq, and Volkan Cevher. An adaptive primal-dual framework for nonsmooth convex minimization. *Mathematical Programming Computation*, Oct 2019.

[TDFC18] Quoc Tran-Dinh, Olivier Fercoq, and Volkan Cevher. A smooth primal-dual optimization framework for nonsmooth composite convex minimization. *SIAM Journal on Optimization*, 28(1):96–134, 2018.

[TH12] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-RMSprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[TQMZ20] Conghui Tan, Yuqiu Qian, Shiqian Ma, and Tong Zhang. Accelerated dual-averaging primal-dual method for composite convex minimization. *Optimization Methods and Software*, 0(0):1–26, 2020.

[Tse00] Paul Tseng. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization*, 38(2):431–446, 2000.

[Tse08] Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 1, 2008.

[TSR$^+$05] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[TVR97] John N Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE transactions on automatic control*, 42(5):674–690, 1997.

[TWYS20] Yi Tian, Yuanhao Wang, Tiancheng Yu, and Suvrit Sra. Provably efficient online agnostic learning in markov games. *arXiv preprint arXiv:2010.15020*, 2020.

[VBC$^+$19] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[vN28] J von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

[VNFC17] Quang Van Nguyen, Olivier Fercoq, and Volkan Cevher. Smoothing technique for nonsmooth composite minimization with linear operator. *arXiv preprint arXiv:1706.05837*, 2017.

[VPG19] Nino Vieillard, Olivier Pietquin, and Matthieu Geist. On connections between constrained optimization and reinforcement learning. *arXiv preprint arXiv:1910.08476*, 2019.

[Vũ13] Bang Công Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3):667–681, 2013.

[Wan20] Mengdi Wang. Randomized linear programming solves the markov decision problem in nearly linear (sometimes sublinear) time. *Mathematics of Operations Research*, 45(2):517–546, 2020.

[WBH$^+$17] Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Remi Munos, Koray Kavukcuoglu, and Nando de Freitas. Sample efficient actor-critic with experience replay. In *International Conference on Learning Representations*, 2017.

[WCLG15] Mengdi Wang, Yichen Chen, Jialin Liu, and Yuantao Gu. Random multi-constraint projection: Stochastic gradient methods for convex optimization with many constraints. *arXiv preprint arXiv:1511.03760*, 2015.

[Wil92] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[WLC+20] Guanghui Wang, Shiyin Lu, Quan Cheng, Weiwei Tu, and Lijun Zhang. {SA}dam: A variant of adam for strongly convex functions. In *International Conference on Learning Representations*, 2020.

[WLZL21] Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. In *Conference on Learning Theory*, 2021.

[WWB19] Rachel Ward, Xiaoxia Wu, and Leon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. In *International Conference on Machine Learning*, pages 6677–6686, 2019.

[WZXG20] Yue Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite time analysis of two time-scale actor critic methods. In *Advances in neural information processing systems*, 2020.

[XCWY20] Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on Learning Theory*, pages 3674–3682. PMLR, 2020.

[Xu20] Yangyang Xu. Primal-dual stochastic gradient method for convex programs with many functional constraints. *SIAM Journal on Optimization*, 30(2):1664–1692, 2020.

[XWL20a] Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for (natural) actor-critic algorithms. *Advances in Neural Information Processing Systems*, 33, 2020.

[XWL20b] Tengyu Xu, Zhe Wang, and Yingbin Liang. Non-asymptotic convergence analysis of two time-scale (natural) actor-critic algorithms. *arXiv preprint arXiv:2005.03557*, 2020.

[XZ14] Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

[XZ18] Yangyang Xu and Shuzhong Zhang. Accelerated primal–dual proximal block coordinate updating methods for constrained convex optimization. *Computational Optimization and Applications*, 70(1):91–128, 2018.

[YH16] Wei Hong Yang and Deren Han. Linear convergence of the alternating direction method of multipliers for a class of convex optimization problems. *SIAM journal on Numerical Analysis*, 54(2):625–640, 2016.

[YNW17] Hao Yu, Michael Neely, and Xiaohan Wei. Online convex optimization with stochastic constraints. In *Advances in Neural Information Processing Systems*, pages 1428–1438, 2017.

[YZH+15] Ian En-Hsu Yen, Kai Zhong, Cho-Jui Hsieh, Pradeep K Ravikumar, and Inderjit S Dhillon. Sparse linear programming via primal and dual augmented coordinate descent. In *Advances in Neural Information Processing Systems*, pages 2368–2376, 2015.

[ZCH+21] Wenhao Zhan, Shicong Cen, Baihe Huang, Yuxin Chen, Jason D Lee, and Yuejie Chi. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence. *arXiv preprint arXiv:2105.11066*, 2021.

[ZKBY20] Kaiqing Zhang, Sham Kakade, Tamer Basar, and Lin Yang. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. *Advances in Neural Information Processing Systems*, 33, 2020.

[ZL15] Yuchen Zhang and Xiao Lin. Stochastic primal-dual coordinate method for regularized empirical risk minimization. In *International Conference on Machine Learning*, pages 353–361. PMLR, 2015.

[ZLSJ20] Jingzhao Zhang, Hongzhou Lin, Suvrit Sra, and Ali Jadbabaie. On complexity of finding stationary points of nonsmooth nonconvex functions. In *International Conference on Machine Learning*, 2020.

[ZSC18] Kaiwen Zhou, Fanhua Shang, and James Cheng. A simple stochastic variance reduced algorithm with fast convergence rates. In *International Conference on Machine Learning*, pages 5975–5984, 2018.

[ZSJ$^+$18] Fangyu Zou, Li Shen, Zequn Jie, Ju Sun, and Wei Liu. Weighted adagrad with unified momentum. *arXiv preprint arXiv:1808.03408*, 2018.

[ZSJ$^+$19] Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of Adam and RMSprop. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11127–11135, 2019.

[ZTLD21] Yulai Zhao, Yuandong Tian, Jason D Lee, and Simon S Du. Provably efficient policy gradient methods for two-player zero-sum markov games. *arXiv preprint arXiv:2102.08903*, 2021.

[ZTY$^+$18] Dongruo Zhou, Yiqi Tang, Ziyan Yang, Yuan Cao, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018.

[ZX17] Yuchen Zhang and Lin Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *The Journal of Machine Learning Research*, 18(1):2939–2980, 2017.

[ZXL19] Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for sarsa with linear function approximation. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[ZYB19] Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games. *Advances in Neural Information Processing Systems*, 32, 2019.

[ZYB21] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384, 2021.