Stochastic Three-Composite Convex Minimization

Alp Yurtsever, Bằng Công Vũ, and Volkan Cevher

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne, Switzerland
alp.yurtsever@epfl.ch, bang.vu@epfl.ch, volkan.cevher@epfl.ch

Abstract

We propose a stochastic optimization method for the minimization of the sum of three convex functions, one of which has Lipschitz continuous gradient as well as restricted strong convexity. Our approach is most suitable in the setting where it is computationally advantageous to process smooth term in the decomposition with its stochastic gradient estimate and the other two functions separately with their proximal operators, such as doubly regularized empirical risk minimization problems. We prove the convergence characterization of the proposed algorithm in expectation under the standard assumptions for the stochastic gradient estimate of the smooth term. Our method operates in the primal space and can be considered as a stochastic extension of the three-operator splitting method. Numerical evidence supports the effectiveness of our method in real-world problems.

1 Introduction

We propose a stochastic optimization method for the three-composite minimization problem:

$$\underset{\mathbf{x}\in\mathbb{R}^d}{\text{minimize}}\; f(\mathbf{x}) + g(\mathbf{x}) + h(\mathbf{x}), \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$ are proper, lower semicontinuous convex functions that admit tractable proximal operators, and $h : \mathbb{R}^d \to \mathbb{R}$ is a smooth function with restricted strong convexity. We assume that we have access to unbiased, stochastic estimates of the gradient of h in the sequel, which is key to scale up optimization and to address streaming settings where data arrive in time.

Template (1) covers a large number of applications in machine learning, statistics, and signal processing by appropriately choosing the individual terms. Operator splitting methods are powerful in this setting, since they reduce the complex problem (1) into smaller subproblems. These algorithms are easy to implement, and they typically exhibit state-of-the-art performance.

To our knowledge, there is no operator splitting framework that can currently tackle template (1) using stochastic gradient of h and the proximal operators of f and g separately, which is critical to the scalability of the methods. This paper specifically bridges this gap.

Our basic framework is closely related to the deterministic three operator splitting method proposed in [11], but we avoid the computation of the gradient ∇h and instead work with its unbiased estimates. We provide rigorous convergence guarantees for our approach and provide guidance in selecting the learning rate under different scenarios.

Road map. Section 2 introduces the basic optimization background. Section 3 then presents the main algorithm and provides its convergence characterization. Section 4 places our contributions in light of the existing work. Numerical evidence that illustrates our theory appears in Section 5. We relegate the technical proofs to the supplementary material.

2 Notation and background

This section recalls a few basic notions from the convex analysis and the probability theory, and presents the notation used in the rest of the paper. Throughout, $\Gamma_0(\mathbb{R}^d)$ denotes the set of all proper, lower semicontinuous convex functions from $\mathbb{R}^d$ to $[-\infty, +\infty]$, and $\langle \cdot \mid \cdot \rangle$ is the standard scalar product on $\mathbb{R}^d$ with its associated norm $\|\cdot\|$.

Subdifferential. The subdifferential of $f \in \Gamma_0(\mathbb{R}^d)$ at a point $\mathbf{x} \in \mathbb{R}^d$ is defined as

$$\partial f(\mathbf{x}) = \{\mathbf{u} \in \mathbb{R}^d \mid f(\mathbf{y}) - f(\mathbf{x}) \geq \langle \mathbf{y} - \mathbf{x} \mid \mathbf{u} \rangle, \forall \mathbf{y} \in \mathbb{R}^d\}.$$

We denote the domain of ∂f as

$$\mathrm{dom}(\partial f) = \{\mathbf{x} \in \mathbb{R}^d \mid \partial f(\mathbf{x}) \neq \varnothing\}.$$

If $\partial f(\mathbf{x})$ is a singleton, then f is a differentiable function, and $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.

Indicator function. Given a nonempty subset $\mathcal{C}$ in $\mathbb{R}^d$, the indicator function of $\mathcal{C}$ is given by

$$\iota_{\mathcal{C}}(\mathbf{x}) = \begin{cases} 0 & \text{if } \mathbf{x} \in \mathcal{C}, \\ +\infty & \text{if } \mathbf{x} \notin \mathcal{C}. \end{cases} \tag{2}$$

Proximal operator. The proximal operator of a function $f \in \Gamma_0(\mathbb{R}^d)$ is defined as follows

$$\mathrm{prox}_f(\mathbf{x}) = \arg\min_{\mathbf{z} \in \mathbb{R}^d} \left\{ f(\mathbf{z}) + \frac{1}{2}\|\mathbf{z} - \mathbf{x}\|^2 \right\}. \tag{3}$$

Roughly speaking, the proximal operator is tractable when the computation of (3) is cheap. If f is the indicator function of a nonempty, closed convex subset $\mathcal{C}$, its proximity operator is the projection operator on $\mathcal{C}$.

Lipschitz continuos gradient. A function $f \in \Gamma_0(\mathbb{R}^d)$ has Lipschitz continuous gradient with Lipschitz constant $L > 0$ (or simply L-Lipschitz), if

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

Strong convexity. A function $f \in \Gamma_0(\mathbb{R}^d)$ is called strongly convex with some parameter $\mu > 0$ (or simply μ-strongly convex), if

$$\langle \boldsymbol{p} - \boldsymbol{q} \mid \mathbf{x} - \mathbf{y} \rangle \geq \mu\|\mathbf{x} - \mathbf{y}\|^2, \qquad \forall \mathbf{x}, \mathbf{y} \in \mathrm{dom}(\partial f), \, \forall \boldsymbol{p} \in \partial f(\mathbf{x}), \, \forall \boldsymbol{q} \in \partial f(\mathbf{y}).$$

Solution set. We denote optimum points of (1) by $\mathbf{x}^\star$, and the solution set by $\boldsymbol{\mathcal{X}}^\star$:

$$\mathbf{x}^\star \in \boldsymbol{\mathcal{X}}^\star = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{0} \in \nabla h(\mathbf{x}) + \partial g(\mathbf{x}) + \partial f(\mathbf{x})\}.$$

Throughout this paper, we assume that $\boldsymbol{\mathcal{X}}^\star$ is not empty.

Restricted strong convexity. A function $f \in \Gamma_0(\mathbb{R}^d)$ has restricted strong convexity with respect to a point $\mathbf{x}^\star$ in a set $\boldsymbol{\mathcal{M}} \subset \mathrm{dom}(\partial f)$, with parameter $\mu > 0$, if

$$\langle \boldsymbol{p} - \boldsymbol{q} \mid \mathbf{x} - \mathbf{x}^\star \rangle \geq \mu\|\mathbf{x} - \mathbf{x}^\star\|^2, \qquad \forall \mathbf{x} \in \boldsymbol{\mathcal{M}}, \, \forall \boldsymbol{p} \in \partial f(\mathbf{x}), \, \forall \boldsymbol{q} \in \partial f(\mathbf{x}^\star).$$

Let $(\Omega, \boldsymbol{\mathcal{F}}, \mathsf{P})$ be a probability space. An $\mathbb{R}^d$-valued random variable is a measurable function $\mathbf{x} \colon \Omega \to \mathbb{R}^d$, where $\mathbb{R}^d$ is endowed with the Borel σ-algebra. We denote by $\sigma(\mathbf{x})$ the σ-field generated by $\mathbf{x}$. The expectation of a random variable $\mathbf{x}$ is denoted by $\mathbf{E}[\mathbf{x}]$. The conditional expectation of $\mathbf{x}$ given a σ-field $\mathcal{A} \subset \boldsymbol{\mathcal{F}}$ is denoted by $\mathbf{E}[\mathbf{x}|\mathcal{A}]$. Given a random variable $\mathbf{y} \colon \Omega \to \mathbb{R}^d$, the conditional expectation of $\mathbf{x}$ given $\mathbf{y}$ is denoted by $\mathbf{E}[\mathbf{x}|\mathbf{y}]$. See [17] for more details on probability theory. An $\mathbb{R}^d$-valued random process is a sequence $(\mathbf{x}_n)_{n \in \mathbb{N}}$ of $\mathbb{R}^d$-valued random variables.

3 Stochastic three-composite minimization algorithm and its analysis

We present stochastic three-composite minimization method (S3CM) in Algorithm 1, for solving the three-composite template (1). Our approach combines the stochastic gradient of h, denoted as $\mathbf{r}$, and the proximal operators of f and g in essentially the same structrure as the three-operator splitting method [11, Algorithm 2]. Our technique is a nontrivial combination of the algorithmic framework of [11] with stochastic analysis.

Algorithm 1 Stochastic three-composite minimization algorithm (S3CM)

Input: An initial point $\mathbf{x}_{f,0}$, a sequence of learning rates $(\gamma_n)_{n\in\mathbb{N}}$, and a sequence of squared integrable $\mathbb{R}^d$-valued stochastic gradient estimates $(\mathbf{r}_n)_{n\in\mathbb{N}}$.

Initialization:
$\quad \mathbf{x}_{g,0} = \mathrm{prox}_{\gamma_0 g}(\mathbf{x}_{f,0})$
$\quad \mathbf{u}_{g,0} = \gamma_0^{-1}(\mathbf{x}_{f,0} - \mathbf{x}_{g,0})$

Main loop:
for $n = 0, 1, 2, \ldots$ **do**
$\quad \mathbf{x}_{g,n+1} = \mathrm{prox}_{\gamma_n g}(\mathbf{x}_{f,n} + \gamma_n \mathbf{u}_{g,n})$
$\quad \mathbf{u}_{g,n+1} = \gamma_n^{-1}(\mathbf{x}_{f,n} - \mathbf{x}_{g,n+1}) + \mathbf{u}_{g,n}$
$\quad \mathbf{x}_{f,n+1} = \mathrm{prox}_{\gamma_{n+1} f}(\mathbf{x}_{g,n+1} - \gamma_{n+1}\mathbf{u}_{g,n+1} - \gamma_{n+1}\mathbf{r}_{n+1})$
end for

Output: $\mathbf{x}_{g,n}$ as an approximation of an optimal solution $\mathbf{x}^\star$.

Theorem 1 *Assume that h is μ_h-strongly convex and has L-Lipschitz continuous gradient. Further assume that g is μ_g-strongly convex, where we allow $\mu_g = 0$. Consider the following update rule for the learning rate:*

$$\gamma_{n+1} = \frac{-\gamma_n^2 \mu_h \eta + \sqrt{(\gamma_n^2 \mu_h \eta)^2 + (1 + 2\gamma_n \mu_g)\gamma_n^2}}{1 + 2\gamma_n \mu_g}, \qquad \textit{for some } \gamma_0 > 0 \textit{ and } \eta \in]0,1[.$$

Define $\mathcal{F}_n = \sigma(\mathbf{x}_{f,k})_{0\le k\le n}$, and suppose that the following conditions hold for every $n \in \mathbb{N}$:

1. $\mathbf{E}[\mathbf{r}_{n+1}|\mathcal{F}_n] = \nabla h(\mathbf{x}_{g,n+1})$ *almost surely,*

2. *There exists $c \in [0, +\infty[$ and $t \in \mathbb{R}$, that satisfies $\sum_{k=0}^n \mathbf{E}[\|\mathbf{r}_k - \nabla h(\mathbf{x}_{g,k})\|^2] \le cn^t$.*

Then, the iterates of S3CM satisfy
$$\mathbf{E}[\|\mathbf{x}_{g,n} - \mathbf{x}^\star\|^2] = \mathcal{O}(1/n^2) + \mathcal{O}(1/n^{2-t}). \tag{4}$$

Remark 1 The variance condition of the stochastic gradient estimates in the theorems above is satisfied when $\mathbf{E}[\|\mathbf{r}_n - \nabla h(\mathbf{x}_{g,n})\|^2] \le c$ for all $n \in \mathbb{N}$ and for some constant $c \in [0, +\infty[$. See [15, 22, 26] for details.

Remark 2 When $\mathbf{r}_n = \nabla h(\mathbf{x}_n)$, S3CM reduces to the deterministic three-operator splitting scheme [11, Algorithm 2] and we recover the convergence rate $\mathcal{O}(1/n^2)$ as in [11]. When g is zero, S3CM reduces to the standard stochastic proximal point algorithm [2, 13, 26].

Remark 3 Learning rate sequence $(\gamma_n)_{n\in\mathbb{N}}$ in Theorem 1 depends on the strong convexity parameter μ_h, which may not be available a priori. Our next result avoids the explicit reliance on the strong convexity parameter, while providing essentially the same convergence rate.

Theorem 2 *Assume that h is μ_h-strongly convex and has L-Lipschitz continuous gradient. Consider a positive decreasing learning rate sequence $\gamma_n = \Theta(1/n^\alpha)$ for some $\alpha \in]0,1]$, and denote $\beta = \lim_{n\to\infty} 2\mu_h n^\alpha \gamma_n$.*

Define $\mathcal{F}_n = \sigma(\mathbf{x}_{f,k})_{0\le k\le n}$, and suppose that the following conditions hold for every $n \in \mathbb{N}$:

1. $\mathbf{E}[\mathbf{r}_{n+1}|\mathcal{F}_n] = \nabla h(\mathbf{x}_{g,n+1})$ *almost surely,*

2. $\mathbf{E}[\|\mathbf{r}_n - \nabla h(\mathbf{x}_{g,n})\|^2]$ *is uniformly bounded by some positive constant.*

3. $\mathbf{E}[\|\mathbf{u}_{g,n} - \mathbf{x}^\star\|^2]$ *is uniformly bounded by some positive constant.*

Then, the iterates of S3CM satisfy
$$\mathbf{E}[\|\mathbf{x}_{g,n} - \mathbf{x}^\star\|^2] = \begin{cases} \mathcal{O}(1/n^\alpha) & \textit{if } 0 < \alpha < 1 \\ \mathcal{O}(1/n^\beta) & \textit{if } \alpha = 1, \textit{ and } \beta < 1 \\ \mathcal{O}((\log n)/n) & \textit{if } \alpha = 1, \textit{ and } \beta = 1, \\ \mathcal{O}(1/n) & \textit{if } \alpha = 1, \textit{ and } \beta > 1. \end{cases}$$

Proof outline. We consider the proof of three-operator splitting method as a baseline, and we use the stochastic fixed point theory to derive the convergence of the iterates via the stochastic Fejér monotone sequence. See the supplement for the complete proof.

Remark 4 Note that $\mathbf{u}_{g,n} \in \partial g(\mathbf{x}_{g,n})$. Hence, we can replace condition 3 in Theorem 2 with the bounded subgradient assumption: $\|\boldsymbol{p}\| \leq c, \forall \boldsymbol{p} \in \partial g(\mathbf{x}_{g,n})$, for some positive constant c.

Remark 5 (Restricted strong convexity) Let $\mathcal{M}$ be a subset of $\mathbb{R}^d$ that contains $(\mathbf{x}_{g,n})_{n\in\mathbb{N}}$ and $\mathbf{x}^\star$. Suppose that h has restricted strong convexity on $\mathcal{M}$ with parameter μ_h. Then, Theorems 1 and 2 still hold. An example role of the restricted strong convexity assumption on algorithmic convergence can be found in [1, 21].

Remark 6 (Extension to arbitrary number of non-smooth terms.) Using the product space technique [5, Section 6.1], S3CM can be applied to composite problems with arbitrary number of non-smooth terms:

$$\underset{\mathbf{x}\in\mathbb{R}^d}{\text{minimize}} \sum_{i=1}^{m} f_i(\mathbf{x}) + h(\mathbf{x}),$$

where $f_i : \mathbb{R}^d \to \mathbb{R}$ are proper, lower semicontinuous convex functions, and $h : \mathbb{R}^d \to \mathbb{R}$ is a smooth function with restricted strong convexity. We present this variant in Algorithm 2. Theorems 1 and 2 hold for this variant, replacing $\mathbf{x}_{g,n}$ by $\overline{\mathbf{x}}_n$, and $\mathbf{u}_{g,n}$ by $\mathbf{u}_{i,n}$ for $i = 1, 2, \ldots, m$.

Algorithm 2 Stochastic m(ulti)-composite minimization algorithm (SmCM)

Input: Initial points $\{\mathbf{x}_{f_1,0}, \mathbf{x}_{f_2,0}, \ldots, \mathbf{x}_{f_m,0}\}$, a sequence of learning rates $(\gamma_n)_{n\in\mathbb{N}}$, and a sequence of squared integrable $\mathbb{R}^d$-valued stochastic gradient estimates $(\mathbf{r}_n)_{n\in\mathbb{N}}$

Initialization:
$\overline{\mathbf{x}}_0 = m^{-1}\sum_{i=1}^{m} \mathbf{x}_{f_i,0}$
for i=1,2,...,m **do**
    $\mathbf{u}_{i,0} = \gamma_0^{-1}(\mathbf{x}_{f_i,0} - \overline{\mathbf{x}}_0)$
end for

Main loop:
for $n = 0, 1, 2, \ldots$ **do**
    $\overline{\mathbf{x}}_{n+1} = m^{-1}\sum_{i=1}^{m}(\mathbf{x}_{f_i,n} + \gamma_n\mathbf{u}_{i,n})$
    for i=1,2,...,m **do**
        $\mathbf{u}_{i,n+1} = \gamma_n^{-1}(\mathbf{x}_{f_i,n} - \overline{\mathbf{x}}_{n+1}) + \mathbf{u}_{i,n}$
        $\mathbf{x}_{f_i,n+1} = \text{prox}_{\gamma_{n+1}mf_i}(\overline{\mathbf{x}}_{n+1} - \gamma_{n+1}\mathbf{u}_{i,n+1} - \gamma_{n+1}\mathbf{r}_{n+1})$
    end for
end for

Output: $\overline{\mathbf{x}}_n$ as an approximation of an optimal solution $\mathbf{x}^\star$.

Remark 7 With a proper learning rate, S3CM still converges even if h is not (restricted) strongly convex under mild assumptions. Suppose that h has L-Lipschitz continuous gradient. Set the learning rate such that $\varepsilon \leq \gamma_n \equiv \gamma \leq \alpha(2L^{-1} - \varepsilon)$, for some α and ε in $]0, 1[$. Define $\mathcal{F}_n = \sigma(\mathbf{x}_{f,k})_{0\leq k\leq n}$, and suppose that the following conditions hold for every $n \in \mathbb{N}$:

1. $\mathbf{E}[\mathbf{r}_{n+1}|\mathcal{F}_n] = \nabla \boldsymbol{h}(\mathbf{x}_{g,n+1})$ almost surely.

2. $\sum_{n\in\mathbb{N}} \mathbf{E}[\|\mathbf{r}_{n+1} - \nabla \boldsymbol{h}(\mathbf{x}_{g,n+1})\|^2|\mathcal{F}_n] < +\infty$ almost surely.

Then, $(\mathbf{x}_{g,n})_{n\in\mathbb{N}}$ converges to a $\mathcal{X}^\star$-valued random vector almost surely. See [7] for details.

Remark 8 All the results above hold for any separable Hilbert space, except that the strong convergence in Remark 7 is replaced by weak convergence. Note however that extending Remark 7 to variable metric setting as in [10, 27] is an open problem.

4

4 Contributions in the light of prior work

Recent algorithms in the operator splitting, such as generalized forward-backward splitting [24], forward-Douglas-Rachford splitting [5], and the three-operator splitting [11], apply to our problem template (1). These key results, however, are in the deterministic setting.

Our basic framework can be viewed as a combination of the three-operator splitting method in [11] with the stochastic analysis.

The idea of using unbiased estimates of the gradient dates back to [25]. Recent developments of this idea can be viewed as proximal based methods for solving the generic composite convex minimization template with a single non-smooth term [2, 9, 12, 13, 15, 16, 19, 26, 23]. This generic form arises naturally in regularized or constrained composite problems [3, 13, 20], where the smooth term typically encodes the data fidelity. These methods require the evaluation of the joint prox of f and g when applied to the three-composite template (1).

Unfortunately, evaluation of the joint prox is arguably more expensive compared to the individual prox operators. To make comparison stark, consider the simple example where f and g are indicator functions for two convex sets. Even if the projection onto the individual sets are easy to compute, projection onto the intersection of these sets can be challenging.

Related literature also contains algorithms that solve some specific instances of template (1). To point out a few, random averaging projection method [28] handles multiple constraints simultaneously but cannot deal with regularizers. On the other hand, accelerated stochastic gradient descent with proximal average [29] can handle multiple regularizers simultaneously, but the algorithm imposes a Lipschitz condition on regularizers, and hence, it cannot deal with constraints.

To our knowledge, our method is the first operator splitting framework that can tackle optimization template (1) using the stochastic gradient estimate of h and the proximal operators of f and g separately, without any restriction on the non-smooth parts except that their subdifferentials are maximally monotone. When h is strongly convex, under mild assumptions, and with a proper learning rate, our algorithm converges with $\mathcal{O}(1/n)$ rate, which is optimal for the stochastic methods under strong convexity assumption for this problem class.

5 Numerical experiments

We present numerical evidence to assess the theoretical convergence guarantees of the proposed algorithm. We provide two numerical examples from Markowitz portfolio optimization and support vector machines.

As a baseline, we use the deterministic three-operator splitting method [11]. Even though the random averaging projection method proposed in [28] does not apply to our template (1) with its all generality, it does for the specific applications that we present below. In our numerical tests, however, we observed that this method exhibits essentially the same convergence behavior as ours when used with the same learning rate sequence. For the clarity of the presentation, we omit this method in our results.

5.1 Portfolio optimization

Traditional Markowitz portfolio optimization aims to reduce risk by minimizing the variance for a given expected return. Mathematically, we can formulate this as a convex optimization problem [6]:

$$\underset{\mathbf{x}\in\mathbb{R}^d}{\text{minimize}} \quad \mathbf{E}\left[|\boldsymbol{a}_i^T\mathbf{x}-b|^2\right] \quad \text{subject to} \quad \mathbf{x}\in\Delta, \quad \boldsymbol{a}_{av}^T\mathbf{x}\geq b,$$

where Δ is the standard simplex for portfolios with no-short positions or a simple sum constraint, $\boldsymbol{a}_{av}=\mathbf{E}\left[\boldsymbol{a}_i\right]$ is the average returns for each asset that is assumed to be known (or estimated), and b encodes a minimum desired return.

This problem has a streaming nature where new data points arrive in time. Hence, we typically do not have access to the whole dataset, and the stochastic setting is more favorable. For implementation,

we replace the expectation with the empirical sample average:

$$\underset{\mathbf{x}\in\mathbb{R}^d}{\text{minimize}} \quad \frac{1}{p}\sum_{i=1}^{p}(\boldsymbol{a}_i^T\mathbf{x}-b)^2 \quad \text{subject to} \quad \mathbf{x}\in\Delta, \quad \boldsymbol{a}_{av}^T\,\mathbf{x}\geq b. \tag{5}$$

This problem fits into our optimization template (1) by setting

$$h(\mathbf{x})=\frac{1}{p}\sum_{i=1}^{p}(\boldsymbol{a}_i^T\mathbf{x}-b)^2, \quad g(\mathbf{x})=\iota_\Delta(\mathbf{x}), \quad \text{and} \quad f(\mathbf{x})=\iota_{\{\mathbf{x}\;|\;\boldsymbol{a}_{av}^T\mathbf{x}\geq b\}}(\mathbf{x}).$$

We compute the unbiased estimates of the gradient by $\mathbf{r}_n = 2(\boldsymbol{a}_{i_n}^T\mathbf{x}-b)\boldsymbol{a}_{i_n}$, where index i_n is chosen uniformly random.

We use 5 different real portfolio datasets: Dow Jones industrial average (DJIA, with 30 stocks for 507 days), New York stock exchange (NYSE, with 36 stocks for 5651 days), Standard & Poor's 500 (SP500, with 25 stocks for 1276 days), Toronto stock exchange (TSE, with 88 stocks for 1258 days) that are also considered in [4]; and one dataset by Fama and French (FF100, 100 portfolios formed on size and book-to-market, 23,647 days) that is commonly used in financial literature, e.g., [6, 14]. We impute the missing data in FF100 using nearest-neighbor method with Euclidean distance.
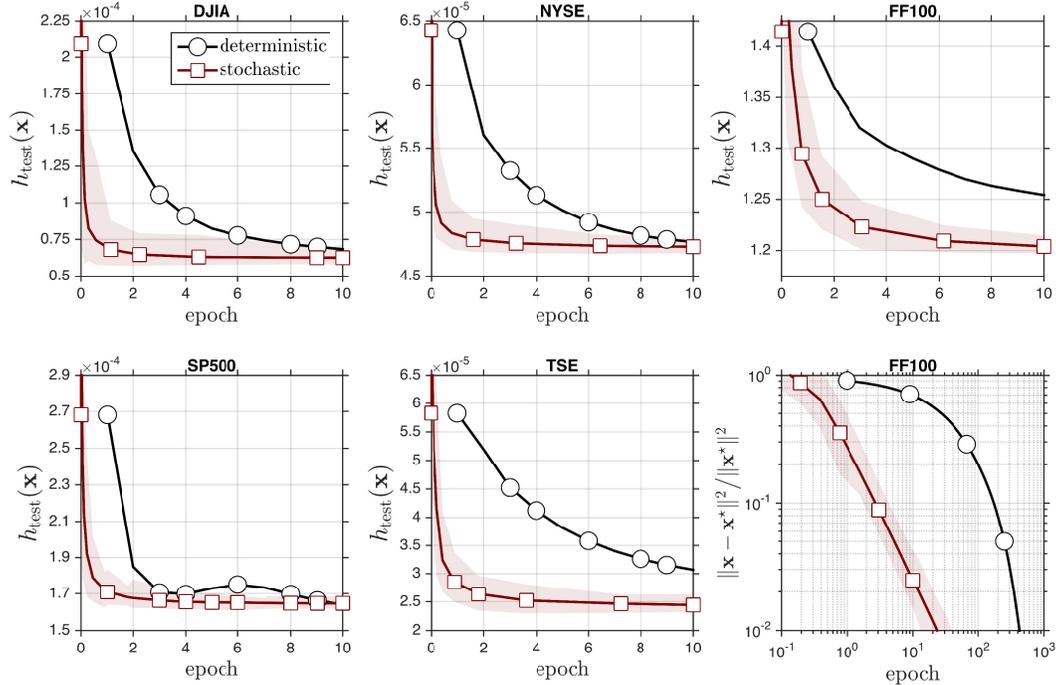

Figure 1: Comparison of the deterministic three-operators splitting method [11, Algorithm 2] and our stochastic three-composite minimization method (S3CM) for Markowitz portfolio optimization (5). Results are averaged over 100 Monte-Carlo simulations, and the boundaries of the shaded area are the best and worst instances.

For the deterministic algorithm, we set $\eta = 0.1$. We evaluate the Lipschitz constant L and the strong convexity parameter μ_h to determine the step-size. For the stochastic algorithm, we do not have access to the whole data, so we cannot compute these parameter. Hence, we adopt the learning rate sequence defined in Theorem 2. We simply use $\gamma_n = \gamma_0/(n+1)$ with $\gamma_0 = 1$ for FF100, and $\gamma_0 = 10^3$ for others.[1] We start both algorithms from the zero vector.

---

[1] Note that a fine-tuned learning rate with a more complex definition can improve the empirical performance, e.g., $\gamma_n = \gamma_0/(n+\zeta)$ for some positive constants γ_0 and ζ.

We split all the datasets into test (10%) and train (90%) partitions randomly. We set the desired return as the average return over all assets in the training set, $b = \text{mean}(\boldsymbol{a}_{av})$. Other b values exhibit qualitatively similar behavior.

The results of this experiment are compiled in Figure 1. We compute the objective function over the datapoints in the test partition, h_{test}. We compare our algorithm against the deterministic three-operator splitting method [11, Algorithm 2]. Since we seek statistical solutions, we compare the algorithms to achieve low to medium accuracy. [11] provides other variants of the deterministic algorithm, including two ergodic averaging schemes that feature improved theoretical rate of convergence. However, these variants performed worse in practice than the original method, and are omitted.

Solid lines in Figure 1 present the average results over 100 Monte-Carlo simulations, and the boundaries of the shaded area are the best and worst instances. We also assess empirical evidence of the $\mathcal{O}(1/n)$ convergence rate guaranteed in Theorem 2, by presenting squared relative distance to the optimum solution for FF100 dataset. Here, we approximate the ground truth by solving the problem to high accuracy with the deterministic algorithm for 10^5 iterations.

5.2 Nonlinear support vector machines classification

This section demonstrates S3CM on a support vector machines (SVM) for binary classification problem. We are given a training set $\mathcal{A} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_d\}$ and the corresponding class labels $\{b_1, b_2, \ldots, b_d\}$, where $\boldsymbol{a}_i \in \mathbb{R}^p$ and $b_i \in \{-1, 1\}$. The goal is to build a model that assigns new examples into one class or the other correctly.

As common in practice, we solve the dual soft-margin SVM formulation:

$$\underset{\mathbf{x} \in \mathbb{R}^d}{\text{minimize}} \quad \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} K(\boldsymbol{a}_i, \boldsymbol{a}_j) b_i b_j x_i x_j - \sum_{i=1}^{d} x_i \quad \text{subject to} \quad \mathbf{x} \in [0, C]^d, \quad \boldsymbol{b}^T \mathbf{x} = 0,$$

where $C \in [0, +\infty[$ is the penalty parameter and $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is a kernel function. In our example we use the Gaussian kernel given by $K_\sigma(\boldsymbol{a}_i, \boldsymbol{a}_j) = \exp(-\sigma \|\boldsymbol{a}_i - \boldsymbol{a}_j\|^2)$ for some $\sigma > 0$.

Define symmetric positive semidefinite matrix $\boldsymbol{M} \in \mathbb{R}^{d \times d}$ with entries $M_{ij} = K_\sigma(\boldsymbol{a}_i, \boldsymbol{a}_j) b_i b_j$. Then the problem takes the form

$$\underset{\mathbf{x} \in \mathbb{R}^d}{\text{minimize}} \quad \frac{1}{2} \mathbf{x}^T \boldsymbol{M} \mathbf{x} - \sum_{i=1}^{d} x_i \quad \text{subject to} \quad \mathbf{x} \in [0, C]^d, \quad \boldsymbol{b}^T \mathbf{x} = 0. \tag{6}$$

This problem fits into three-composite optimization template (1) with

$$h(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \boldsymbol{M} \mathbf{x} - \sum_{i=1}^{d} x_i, \quad g(\mathbf{x}) = \iota_{[0,C]^d}(\mathbf{x}), \quad \text{and} \quad f(\mathbf{x}) = \iota_{\{\mathbf{x} \mid \boldsymbol{b}^T \mathbf{x} = 0\}}(\mathbf{x}).$$

One can solve this problem using three-operator splitting method [11, Algorithm 1]. Note that prox_f and prox_g, which are projections onto the corresponding constraint sets, incur $\mathcal{O}(d)$ computational cost, whereas the cost of computing the gradient is $\mathcal{O}(d^2)$.

To compute an unbiased gradient estimate, we choose an index i_n uniformly random, and we form $\mathbf{r}_n = d\boldsymbol{M}_{i_n} x_{i_n} - \mathbf{1}$. Here $\boldsymbol{M}_{i_n}$ denotes i_n^{th} column of matrix $\boldsymbol{M}$, and $\mathbf{1}$ represents the vector of ones. We can compute $\mathbf{r}_n$ in $\mathcal{O}(d)$ computations, hence each iteration of S3CM costs an order cheaper compared to deterministic algorithm.

We use UCI machine learning dataset "a1a", with $d = 1605$ datapoints and $p = 123$ features [8, 18]. Note that our goal here is to demonstrate the optimization performance of our algorithm for a real world problem, rather than competing the prediction quality of the best engineered solvers. Hence, to keep experiments simple, we fix problem parameters $C = 1$ and $\sigma = 2^{-2}$, and we focus on the effects of algorithmic parameters on the convergence behavior.

Since $p < d$, $\boldsymbol{M}$ is rank deficient and h is not strongly convex. Nevertheless we use S3CM with the learning rate $\gamma_n = \gamma_0/(n+1)$ for various values of γ_0. We observe $\mathcal{O}(1/n)$ empirical convergence rate on the squared relative error for large enough γ_0, which is guaranteed under restricted strong convexity assumption. See Figure 2 for the results.
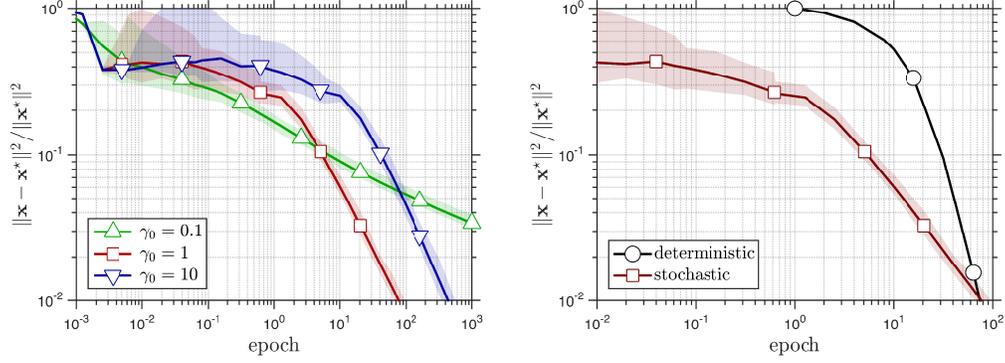
7

Figure 2: *[Left]* Convergence of S3CM in the squared relative error with learning rate $\gamma_n = \gamma_0/(n+1)$. *[Right]* Comparison of the deterministic three-operators splitting method [11, Algorithm 1] and S3CM with $\gamma_0 = 1$ for SVM classification problem. Results are averaged over 100 Monte-Carlo simulations. Boundaries of the shaded area are the best and worst instances.

References

[1] A. Agarwal, S. Negahban, and M. J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *Ann. Stat.*, 40(5):2452–2482, 2012.

[2] Y. F. Atchadé, G. Fort, and E. Moulines. On stochastic proximal gradient algorithms. arXiv:1402.2365v2, 2014.

[3] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer-Verlag, 2011.

[4] A. Borodin, R. El-Yaniv, and V. Gogan. Can we learn to beat the best stock. In *Advances in Neural Information Processing Systems 16*, pages 345–352. 2004.

[5] L. M. Briceño-Arias. Forward-Douglas–Rachford splitting and forward-partial inverse method for solving monotone inclusions. *Optimization*, 64(5):1239–1261, 2015.

[6] J. Brodie, I. Daubechies, C. de Mol, D. Giannone, and I. Loris. Sparse and stable Markowitz portfolios. *Proc. Natl. Acad. Sci.*, 106:12267–12272, 2009.

[7] V. Cevher, B. C. Vũ, and A. Yurtsever. Stochastic forward–Douglas–Rachford splitting for monotone inclusions. EPFL-Report-215759, 2016.

[8] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, 2011.

[9] P. L. Combettes and J.-C. Pesquet. Stochastic approximations and perturbations in forward-backward splitting for monotone operators. arXiv:1507.07095v1, 2015.

[10] P. L. Combettes and B. C. Vũ. Variable metric forward–backward splitting with applications to monotone inclusions in duality. *Optimization*, 63(9):1289–1318, 2014.

[11] D. Davis and W. Yin. A three-operator splitting scheme and its optimization applications. arXiv:1504.01032v1, 2015.

[12] O. Devolder. Stochastic first order methods in smooth convex optimization. Technical report, Center for Operations Research and Econometrics, 2011.

[13] J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *J. Mach. Learn. Res.*, 10:2899–2934, 2009.

[14] E. F. Fama and K. R. French. Multifactor explanations of asset pricing anomalies. *Journal of Finance,*, 51:55–84, 1996.

[15] C. Hu, W. Pan, and J. T. Kwok. Accelerated gradient methods for stochastic optimization and online learning. In *Advances in Neural Information Processing Systems 22*, pages 781–789. 2009.

[16] G. Lan. An optimal method for stochastic composite optimization. *Math. Program.*, 133(1):365–397, 2012.

[17] M. Ledoux and M. Talagrand. *Probability in Banach spaces: Isoperimetry and processes*. Springer-Verlag, 1991.

[18] M. Lichman. UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, 2013.

[19] Q. Lin, X. Chen, and J. Peña. A smoothing stochastic gradient method for composite optimization. *Optimization Methods and Software*, 29(6):1281–1301, 2014.

[20] S. Mosci, L. Rosasco, M. Santoro, A. Verri, and S. Villa. Solving structured sparsity regularization with proximal methods. In *European Conf. Machine Learning and Principles and Practice of Knowledge Discovery*, pages 418–433, 2010.

[21] S. Negahban, B. Yu, M. J. Wainwright, and P. K. Ravikumar. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems 22*, pages 1348–1356, 2009.

[22] A. 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 J. on Optimization*, 15(1):229–251, 2005.

[23] A. Nitanda. Stochastic proximal gradient descent with acceleration techniques. In *Advances in Neural Information Processing Systems 27*, pages 1574–1582. 2014.

[24] H. Raguet, J. Fadili, and G. Peyré. A generalized forward-backward splitting. *SIAM Journal on Imaging Sciences*, 6(3):1199–1226, 2013.

[25] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407, 1951.

[26] L. Rosasco, S. Villa, and B. C. Vũ. Convergence of stochastic proximal gradient algorithm. arXiv:1403.5074v3, 2014.

[27] B. C. Vũ. Almost sure convergence of the forward–backward–forward splitting algorithm. *Optimization Letters*, 10(4):781–803, 2016.

[28] M. Wang, Y. Chen, J. Liu, and Y. Gu. Random multi–constraint projection: Stochastic gradient methods for convex optimization with many constraints. arXiv:1511.03760v1, 2015.

[29] W. Zhong and J. Kwok. Accelerated stochastic gradient method for composite regularization. *J. Mach. Learn. Res.*, 33:1086–1094, 2014.

Appendix: Proof of the main result

In this supplement, we provide the proofs of Theorem 1 and Theorem 2.

Proof of Theorem 1. For every $n \in \mathbb{N}$, we have

$$
\begin{cases}
\mathbf{u}_{f,n} = \gamma_n^{-1}(\mathbf{x}_{g,n} - \mathbf{x}_{f,n}) - (\mathbf{u}_{g,n} + \mathbf{r}_n) \in \partial f(\mathbf{x}_{f,n}) \\
\mathbf{u}_{g,n} \in \partial g(\mathbf{x}_{g,n}) \\
\gamma_n(\mathbf{u}_{g,n+1} - \mathbf{u}_{g,n}) = \mathbf{x}_{f,n} - \mathbf{x}_{g,n+1} \\
\gamma_n(\mathbf{u}_{f,n} + \mathbf{u}_{g,n} + \mathbf{r}_n) = \mathbf{x}_{g,n} - \mathbf{x}_{f,n} \\
\gamma_n(\mathbf{u}_{g,n+1} + \mathbf{u}_{f,n} + \mathbf{r}_n) = \mathbf{x}_{g,n} - \mathbf{x}_{g,n+1}.
\end{cases}
$$

Now, let us define

$$
\begin{cases}
\chi_n &= 2\gamma_n \langle \mathbf{x}_{f,n} - \mathbf{x}^\star \mid \mathbf{u}_{f,n} + \mathbf{r}_n \rangle + 2\gamma_n \langle \mathbf{x}_{g,n+1} - \mathbf{x}^\star \mid \mathbf{u}_{g,n+1} \rangle \\
\chi_{1,n} &= 2 \langle \mathbf{x}_{g,n+1} - \mathbf{x}^\star \mid \mathbf{x}_{g,n} - \mathbf{x}_{g,n+1} \rangle \\
\chi_{2,n} &= 2 \langle \mathbf{x}_{f,n} - \mathbf{x}_{g,n+1} \mid \mathbf{x}_{g,n} - \mathbf{x}_{f,n} \rangle \\
\chi_{3,n} &= 2\gamma_n \langle \mathbf{x}_{g,n+1} - \mathbf{x}_{f,n} \mid \mathbf{u}_{g,n} - \mathbf{u}_g^\star \rangle = 2\gamma_n^2 \langle \mathbf{u}_{g,n} - \mathbf{u}_{g,n+1} \mid \mathbf{u}_{g,n} - \mathbf{u}_g^\star \rangle \\
\chi_{4,n} &= 2\gamma_n \langle \mathbf{x}_{g,n+1} - \mathbf{x}_{f,n} \mid \mathbf{u}_g^\star \rangle,
\end{cases}
$$

where $\mathbf{u}_g^\star \in \partial g(\mathbf{x}^\star)$. Then, by simple calculations we get

$$
\begin{cases}
\chi_{2,n} &= \|\mathbf{x}_{g,n} - \mathbf{x}_{g,n+1}\|^2 - \|\mathbf{x}_{f,n} - \mathbf{x}_{g,n+1}\|^2 - \|\mathbf{x}_{g,n} - \mathbf{x}_{f,n}\|^2 \\
\chi_{1,n} &= \|\mathbf{x}_{g,n} - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{g,n+1} - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{g,n} - \mathbf{x}_{g,n+1}\|^2 \\
\chi_{3,n} &= \gamma_n^2 \|\mathbf{u}_{g,n+1} - \mathbf{u}_{g,n}\|^2 + \gamma_n^2 \|\mathbf{u}_{g,n} - \mathbf{u}_g^\star\|^2 - \gamma_n^2 \|\mathbf{u}_{g,n+1} - \mathbf{u}_g^\star\|^2 \\
&= \gamma_n^2 \|\mathbf{u}_{g,n} - \mathbf{u}_g^\star\|^2 - \gamma_n^2 \|\mathbf{u}_{g,n+1} - \mathbf{u}_g^\star\|^2 + \|\mathbf{x}_{f,n} - \mathbf{x}_{g,n+1}\|^2.
\end{cases}
\tag{7}
$$

Furthermore, for every $n \in \mathbb{N}$, we can express χ_n as follows:

$$
\begin{aligned}
\chi_n &= 2\gamma_n \langle \mathbf{x}_{f,n} - \mathbf{x}_{g,n+1} \mid \mathbf{u}_{f,n} + \mathbf{r}_n \rangle + 2\gamma_n \langle \mathbf{x}_{g,n+1} - \mathbf{x}^\star \mid \mathbf{u}_{g,n+1} + \mathbf{u}_{f,n} + \mathbf{r}_n \rangle \\
&= \chi_{1,n} + 2\gamma_n \langle \mathbf{x}_{f,n} - \mathbf{x}_{g,n+1} \mid \mathbf{u}_{f,n} + \mathbf{r}_n \rangle \\
&= \chi_{1,n} + 2\gamma_n \big(\langle \mathbf{x}_{f,n} - \mathbf{x}_{g,n+1} \mid \mathbf{u}_{f,n} + \mathbf{r}_n + \mathbf{u}_{g,n} \rangle - \langle \mathbf{x}_{f,n} - \mathbf{x}_{g,n+1} \mid \mathbf{u}_{g,n} \rangle \big) \\
&= \chi_{1,n} + \chi_{2,n} + 2\gamma_n \langle \mathbf{x}_{g,n+1} - \mathbf{x}_{f,n} \mid \mathbf{u}_{g,n} \rangle \\
&= \chi_{1,n} + \chi_{2,n} + 2\gamma_n \langle \mathbf{x}_{g,n+1} - \mathbf{x}_{f,n} \mid \mathbf{u}_g^\star \rangle + 2\gamma_n \langle \mathbf{x}_{g,n+1} - \mathbf{x}_{f,n} \mid \mathbf{u}_{g,n} - \mathbf{u}_g^\star \rangle \\
&= \chi_{1,n} + \chi_{2,n} + \chi_{3,n} + \chi_{4,n}.
\end{aligned}
$$

Now, summing the equalities in (7), we obtain,

$$
\begin{aligned}
\chi_n &= \gamma_n^2 \big(\|\mathbf{u}_{g,n} - \mathbf{u}_g^\star\|^2 - \|\mathbf{u}_{g,n+1} - \mathbf{u}_g^\star\|^2 \big) + \|\mathbf{x}_{g,n} - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{g,n+1} - \mathbf{x}^\star\|^2 \\
&\quad - \|\mathbf{x}_{g,n} - \mathbf{x}_{f,n}\|^2 + \chi_{4,n}.
\end{aligned}
$$

Denote $\mathbf{u}_f^\star \in \partial f(\mathbf{x}^\star)$. We have $\left\langle \mathbf{x}_{f,n} - \mathbf{x}^\star \mid \mathbf{u}_{f,n} - \mathbf{u}_f^\star \right\rangle \geq 0$, since f is convex. Hence,

$$
\begin{aligned}
\chi_n &= 2\gamma_n \big(\langle \mathbf{x}_{f,n} - \mathbf{x}^\star \mid \mathbf{u}_{f,n} - \mathbf{u}_f^\star \rangle + \langle \mathbf{x}_{f,n} - \mathbf{x}^\star \mid \mathbf{u}_f^\star + \mathbf{r}_n \rangle + \langle \mathbf{x}_{g,n+1} - \mathbf{x}^\star \mid \mathbf{u}_{g,n+1} \rangle \big) \\
&\geq 2\gamma_n \big(\langle \mathbf{x}_{f,n} - \mathbf{x}^\star \mid \mathbf{u}_f^\star + \mathbf{r}_n \rangle + \langle \mathbf{x}_{g,n+1} - \mathbf{x}^\star \mid \mathbf{u}_{g,n+1} \rangle \big) \\
&= 2\gamma_n \big(\langle \mathbf{x}_{f,n} - \mathbf{x}^\star \mid \mathbf{u}_f^\star + \mathbf{r}_n \rangle + \langle \mathbf{x}_{g,n+1} - \mathbf{x}^\star \mid \mathbf{u}_g^\star \rangle + \langle \mathbf{x}_{g,n+1} - \mathbf{x}^\star \mid \mathbf{u}_{g,n+1} - \mathbf{u}_g^\star \rangle \big) \\
&\geq 2\gamma_n \big(\langle \mathbf{x}_{f,n} - \mathbf{x}^\star \mid \mathbf{u}_f^\star + \mathbf{r}_n \rangle + \mu_g \|\mathbf{x}_{g,n+1} - \mathbf{x}^\star\|^2 + \langle \mathbf{x}_{g,n+1} - \mathbf{x}^\star \mid \mathbf{u}_g^\star \rangle \big),
\end{aligned}
\tag{8}
$$

where the last inequality follows from the assumption that g is μ_g-strongly convex. Set

$$
\mathbf{x}_{f,n}^e = \mathrm{prox}_{\gamma_n f} \big((\mathbf{x}_{g,n} - \gamma_n \mathbf{u}_{g,n} - \gamma_n \nabla h(\mathbf{x}_{g,n})).
$$

Then, using the non-expansiveness of $\mathrm{prox}_{\gamma_n f}$, we get

$$
\|\mathbf{x}_{f,n}^e - \mathbf{x}_{f,n}\| \leq \gamma_n \|\nabla h(\mathbf{x}_{g,n}) - \mathbf{r}_n\|.
$$

Now, let us define

$$
\begin{cases}
\chi_{5,n} = \left\langle \mathbf{x}_{f,n} - \mathbf{x}_{f,n}^e \mid \mathbf{r}_n - \nabla h(\mathbf{x}_{g,n}) \right\rangle \\
\chi_{6,n} = \left\langle \mathbf{x}_{f,n}^e - \mathbf{x}^\star \mid \mathbf{r}_n - \nabla h(\mathbf{x}_{g,n}) \right\rangle \\
\chi_{7,n} = \chi_{5,n} + \chi_{6,n} = \left\langle \mathbf{x}_{f,n} - \mathbf{x}^\star \mid \mathbf{r}_n - \nabla h(\mathbf{x}_{g,n}) \right\rangle.
\end{cases}
$$

Then, we have

$$\chi_{5,n} = \langle \mathbf{x}_{f,n} - \mathbf{x}_{f,n}^e \mid \mathbf{r}_n - \nabla h(\mathbf{x}_{g,n}) \rangle$$
$$\leq \|\mathbf{x}_{f,n} - \mathbf{x}_{f,n}^e\| \cdot \|\mathbf{r}_n - \nabla h(\mathbf{x}_{g,n})\|$$
$$\leq \gamma_n \|\mathbf{r}_n - \nabla h(\mathbf{x}_{g,n})\|^2,$$

and since $\mathbf{x}_{f,n}^e$ is $\mathcal{F}_{n-1}$-measurable (by induction), we obtain

$$\mathbf{E}[\chi_{6,n}|\mathcal{F}_{n-1}] = \langle \mathbf{x}_{f,n}^e - \mathbf{x}^\star \mid \mathbf{E}[\mathbf{r}_n - \nabla h(\mathbf{x}_{g,n})|\mathcal{F}_{n-1}] \rangle = 0.$$

Furthermore, for any $\eta \in]0,1[$, since h is μ_h-strongly convex and has L-Lipschitz continuous gradient, we have

$$2 \langle \mathbf{x}_{f,n} - \mathbf{x}^\star \mid \mathbf{r}_n \rangle = 2 \langle \mathbf{x}_{f,n} - \mathbf{x}^\star \mid \nabla h(\mathbf{x}_{g,n}) \rangle + 2 \langle \mathbf{x}_{f,n} - \mathbf{x}^\star \mid \mathbf{r}_n - \nabla h(\mathbf{x}_{g,n}) \rangle$$
$$= 2 \langle \mathbf{x}_{f,n} - \mathbf{x}_{g,n} \mid \nabla h(\mathbf{x}_{g,n}) - \nabla h(\mathbf{x}^\star) \rangle + 2 \langle \mathbf{x}_{g,n} - \mathbf{x}^\star \mid \nabla h(\mathbf{x}_{g,n}) - \nabla h(\mathbf{x}^\star) \rangle$$
$$+ 2 \langle \mathbf{x}_{f,n} - \mathbf{x}^\star \mid \nabla h(\mathbf{x}^\star) \rangle + 2\chi_{7,n}$$
$$\geq \frac{-L}{2(1-\eta)} \|\mathbf{x}_{f,n} - \mathbf{x}_{g,n}\|^2 - 2\frac{(1-\eta)}{L} \|\nabla h(\mathbf{x}_{g,n}) - \nabla h(\mathbf{x}^\star)\|^2 + 2\eta\mu_h \|\mathbf{x}_{g,n} - \mathbf{x}^\star\|^2$$
$$+ 2\chi_{7,n} + 2\frac{(1-\eta)}{L} \|\nabla h(\mathbf{x}_{g,n}) - \nabla h(\mathbf{x}^\star)\|^2 + 2 \langle \mathbf{x}_{f,n} - \mathbf{x}^\star \mid \nabla h(\mathbf{x}^\star) \rangle$$
$$\geq \frac{-L}{2(1-\eta)} \|\mathbf{x}_{f,n} - \mathbf{x}_{g,n}\|^2 + 2\eta\mu_h \|\mathbf{x}_{g,n} - \mathbf{x}^\star\|^2 + 2\chi_{7,n} + 2 \langle \mathbf{x}_{f,n} - \mathbf{x}^\star \mid \nabla h(\mathbf{x}^\star) \rangle. \quad (9)$$

Now, inserting (9) into (8), we arrive at

$$\chi_n \geq 2\gamma_n \langle \mathbf{x}_{f,n} - \mathbf{x}^\star \mid \mathbf{u}_f^\star + \nabla h(\mathbf{x}^\star) \rangle + 2\gamma_n \langle \mathbf{x}_{g,n+1} - \mathbf{x}^\star \mid \mathbf{u}_g^\star \rangle + 2\gamma_n \chi_{7,n}$$
$$+ 2\eta\mu_h\gamma_n \|\mathbf{x}_{g,n} - \mathbf{x}^\star\|^2 + 2\mu_g\gamma_n \|\mathbf{x}_{g,n+1} - \mathbf{x}^\star\|^2 - \frac{\gamma_n L}{2(1-\eta)} \|\mathbf{x}_{f,n} - \mathbf{x}_{g,n}\|^2, \quad (10)$$

since it follows that

$$2\gamma_n \langle \mathbf{x}_{f,n} - \mathbf{x}^\star \mid \mathbf{u}_f^\star + \nabla h(\mathbf{x}^\star) \rangle + 2\gamma_n \langle \mathbf{x}_{g,n+1} - \mathbf{x}^\star \mid \mathbf{u}_g^\star \rangle) - \chi_{4,n}$$
$$= 2\gamma_n(\langle \mathbf{x}_{f,n} - \mathbf{x}^\star \mid \mathbf{u}_f^\star + \mathbf{u}_g^\star + \nabla h(\mathbf{x}^\star) \rangle$$
$$= 0.$$

We derive from (10) and (8) that

$$(1 + 2\gamma_n\mu_g)\|\mathbf{x}_{g,n+1} - \mathbf{x}^\star\|^2 + \gamma_n^2\|\mathbf{u}_{g,n+1} - \mathbf{x}^\star\|^2 + (1 - \frac{\gamma_n L}{2(1-\eta)})\|\mathbf{x}_{f,n} - \mathbf{x}_{g,n}\|^2$$
$$\leq (1 - 2\gamma_n\mu_h\eta)\|\mathbf{x}_{g,n} - \mathbf{x}^\star\|^2 + \gamma_n^2\|\mathbf{u}_{g,n} - \mathbf{x}^\star\|^2 - 2\gamma_n\chi_{7,n}.$$

Since $(\gamma_n)_{n\in\mathbb{N}}$ is a nonnegative sequence that converges to 0, there exists some positive integer n_0 such that $(1 - \frac{\gamma_n L}{2(1-\eta)}) \geq 0$ for any $n \geq n_0$. Hence,

$$(1 + 2\gamma_n\mu_g)\|\mathbf{x}_{g,n+1} - \mathbf{x}^\star\|^2 + \gamma_n^2\|\mathbf{u}_{g,n+1} - \mathbf{x}^\star\|^2$$
$$\leq (1 - 2\gamma_n\mu_h\eta)\|\mathbf{x}_{g,n} - \mathbf{x}^\star\|^2 + \gamma_n^2\|\mathbf{u}_{g,n} - \mathbf{x}^\star\|^2 - 2\gamma_n\chi_{7,n}, \qquad \forall n \geq n_0.$$

Now, taking the conditonal expectation with respect to $\mathcal{F}_{n-1}$, we obtain

$$(1 + 2\gamma_n\mu_g)\mathbf{E}[\|\mathbf{x}_{g,n+1} - \mathbf{x}^\star\|^2|\mathcal{F}_{n-1}] + \gamma_n^2\mathbf{E}[\|\mathbf{u}_{g,n+1} - \mathbf{x}^\star\|^2|\mathcal{F}_{n-1}]$$
$$\leq (1 - 2\gamma_n\mu_h\eta)\|\mathbf{x}_{g,n} - \mathbf{x}^\star\|^2 + \gamma_n^2\|\mathbf{u}_{g,n} - \mathbf{x}^\star\|^2 - 2\gamma_n\mathbf{E}[\chi_{7,n}|\mathcal{F}_{n-1}] \qquad (11)$$
$$= (1 - 2\gamma_n\mu_h\eta)\|\mathbf{x}_{g,n} - \mathbf{x}^\star\|^2 + \gamma_n^2\|\mathbf{u}_{g,n} - \mathbf{x}^\star\|^2 - 2\gamma_n\mathbf{E}[\chi_{5,n}|\mathcal{F}_{n-1}]$$
$$\leq (1 - 2\gamma_n\mu_h\eta)\|\mathbf{x}_{g,n} - \mathbf{x}^\star\|^2 + \gamma_n^2\|\mathbf{u}_{g,n} - \mathbf{x}^\star\|^2 + 2\gamma_n^2\mathbf{E}[\|\mathbf{r}_n - \nabla h(\mathbf{x}_{g,n})\|^2|\mathcal{F}_{n-1}], \quad \forall n \geq n_0.$$

As indicated in the proof of [11], we have

$$\gamma_n^{-2}(1 + 2\gamma_n\mu_g) = \gamma_{n+1}^{-2}(1 - 2\gamma_{n+1}\mu_h\eta).$$

and

$$\lim_{n\to\infty} (n+1)\gamma_n = (\eta\mu_h + \mu_g)^{-1}. \qquad (12)$$

Therefore, by dividing both sides of (11) by γ_n^2, and taking the expectations, we obtain

$$\gamma_{n+1}^{-2}(1 - 2\gamma_{n+1}\mu_h\eta)\mathbf{E}[\|\mathbf{x}_{g,n+1} - \mathbf{x}^\star\|^2] + \mathbf{E}[\|\mathbf{u}_{g,n+1} - \mathbf{x}^\star\|^2]$$
$$\leq \gamma_n^{-2}(1 - 2\gamma_n\mu_h\eta)\mathbf{E}[\|\mathbf{x}_{g,n} - \mathbf{x}^\star\|^2] + \mathbf{E}[\|\mathbf{u}_{g,n} - \mathbf{x}^\star\|^2] + 2\mathbf{E}[\|\mathbf{r}_n - \nabla h(\mathbf{x}_{g,n})\|^2].$$

Now, summing this inequality from $n = n_0$ to $n = N$, we get

$$\gamma_{N+1}^{-2}(1 - 2\gamma_{N+1}\mu_h\eta)\mathbf{E}[\|\mathbf{x}_{g,N+1} - \mathbf{x}^\star\|^2] \tag{13}$$

$$\leq \gamma_{n_0}^{-2}(1 - 2\gamma_{n_0}\mu_h\eta)\mathbf{E}[\|\mathbf{x}_{g,n_0} - \mathbf{x}^\star\|^2] + \mathbf{E}[\|\mathbf{u}_{g,n_0} - \mathbf{x}^\star\|^2] + \sum_{k=n_0}^{N}\mathbf{E}[\|\mathbf{r}_k - \nabla h(\mathbf{x}_{g,k})\|^2].$$

In view of (12), (4) follows from (13). $\qquad\square$

We now present the key lemma for the proof of Theorem 2. This lemma is a direct corollary from [26, Lemma 4.4], hence we omit the proof.

Lemma 1 *Let $\alpha \in]0, 1]$, let c and τ be in $]0, +\infty[$, and let n_0 be a positive integer. Let $(\theta_n)_{n\in\mathbb{N}}$ be a positive sequence defined by $\theta_n = cn^{-\alpha}$. Let $(s_n)_{n\in\mathbb{N}}$ be a sequence that satisfies*

$$0 \leq s_{n+1} \leq (1 - \theta_n)s_n + \tau\theta_n^2, \qquad \forall n \geq n_0.$$

Then, s_n satisfies

$$s_n = \begin{cases} \mathcal{O}(1/n^\alpha) & \text{if } 0 < \alpha < 1 \\ \mathcal{O}(1/n^c) & \text{if } \alpha = 1, \text{ and } 0 < c < 1 \\ \mathcal{O}((\log n)/n) & \text{if } \alpha = 1, \text{ and } c = 1 \\ \mathcal{O}(1/n) & \text{if } \alpha = 1, \text{ and } c > 1. \end{cases}$$

Proof of Theorem 2. Taking the expectations of both sides in (11), we get

$$\mathbf{E}[\|\mathbf{x}_{g,n+1} - \mathbf{x}^\star\|^2] \tag{14}$$
$$\leq (1 + 2\gamma_n\mu_g)\mathbf{E}[\|\mathbf{x}_{g,n+1} - \mathbf{x}^\star\|^2] + \gamma_n^2\mathbf{E}[\|\mathbf{u}_{g,n+1} - \mathbf{x}^\star\|^2]$$
$$\leq (1 - 2\gamma_n\mu_h\eta)\mathbf{E}[\|\mathbf{x}_{g,n} - \mathbf{x}^\star\|^2] + \gamma_n^2\mathbf{E}[\|\mathbf{u}_{g,n} - \mathbf{x}^\star\|^2] + 2\gamma_n^2\mathbf{E}[\|\mathbf{r}_n - \nabla h(\mathbf{x}_{g,n})\|^2], \quad \forall n \geq n_0.$$

Since the learning rate $\gamma_n = \Theta(n^{-\alpha})$, we can find two positive real numbers $c_0 \leq c_1$ and a positive integer $n_1 \geq n_0$, such that $c_0 n^{-\alpha} \leq \gamma_n \leq c_1 n^{-\alpha}$ for any $n \geq n_1$. Then, we obtain

$$\mathbf{E}[\|\mathbf{x}_{g,n+1} - \mathbf{x}^\star\|^2] \leq (1 - 2\mu_h\eta c_0 n^{-\alpha})\mathbf{E}[\|\mathbf{x}_{g,n} - \mathbf{x}^\star\|^2]$$
$$+ (c_1 n^{-\alpha})^2\big(\mathbf{E}[\|\mathbf{u}_{g,n} - \mathbf{x}^\star\|^2] + 2\mathbf{E}[\|\mathbf{r}_n - \nabla h(\mathbf{x}_{g,n})\|^2]\big), \quad \forall n \geq n_1.$$

$\mathbf{E}[\|\mathbf{r}_n - \nabla h(\mathbf{x}_{g,n})\|^2]$ and $\mathbf{E}[\|\mathbf{u}_{g,n} - \mathbf{x}^\star\|^2]$ are uniformly bounded by some positive constants by assumption. Denote these constants by τ_0 and τ_1, then we have

$$\mathbf{E}[\|\mathbf{x}_{g,n+1} - \mathbf{x}^\star\|^2] \leq (1 - 2\mu_h\eta c_0 n^{-\alpha})\mathbf{E}[\|\mathbf{x}_{g,n} - \mathbf{x}^\star\|^2] + (2\tau_0 + \tau_1)(c_1 n^{-\alpha})^2, \quad \forall n \geq n_1.$$

Setting $\theta_n = 2\mu_h\eta c_0 n^{-\alpha}$ and $\tau = c_1^2(2\tau_0 + \tau_1)(2\mu_h\eta c_0)^{-2}$, we get

$$\mathbf{E}[\|\mathbf{x}_{g,n+1} - \mathbf{x}^\star\|^2] \leq (1 - \theta_n)\mathbf{E}[\|\mathbf{x}_{g,n} - \mathbf{x}^\star\|^2] + \tau\theta_n^2, \quad \forall n \geq n_1.$$

Proof follows from Lemma 1. $\qquad\square$