FRANK-WOLFE WORKS FOR NON-LIPSCHITZ CONTINUOUS GRADIENT OBJECTIVES: SCALABLE POISSON PHASE RETRIEVAL

Gergely Odor[*], *Yen-Huan Li*[*], *Alp Yurtsever*[*], *Ya-Ping Hsieh*[*], *Quoc Tran-Dinh*[*†], *Marwa El Halabi*[*], and *Volkan Cevher*[*]

[*]Laboratory for Information and Inference Systems
École Polytechnique Fédérale de Lausanne, Switzerland

[†]Department of Statistics and Operations Research
The University of North Carolina at Chapel Hill, USA

ABSTRACT

We study a phase retrieval problem in the Poisson noise model. Motivated by the PhaseLift approach, we approximate the maximum-likelihood estimator by solving a convex program with a nuclear norm constraint. While the Frank-Wolfe algorithm, together with the Lanczos method, can efficiently deal with nuclear norm constraints, our objective function does not have a Lipschitz continuous gradient, and hence existing convergence guarantees for the Frank-Wolfe algorithm do not apply. In this paper, we show that the Frank-Wolfe algorithm works for the Poisson phase retrieval problem, and has a global convergence rate of $O(1/t)$, where t is the iteration counter. We provide rigorous theoretical guarantee and illustrating numerical results.

***Index Terms*—** Phase retrieval, Poisson noise, PhaseLift, Frank-Wolfe algorithm, non-Lipschitz continuous gradient

1. INTRODUCTION

Phase retrieval is the problem of estimating a complex-valued signal from intensity measurements, which arises in many applications such as X-ray crystallography, diffraction imaging, astronomical imaging, and many others [29].

We focus on the Poisson noise case in this paper. Formally speaking, we are interested in estimating a signal $x^\natural \in \mathbb{C}^p$, given $a_1, \ldots, a_n \in \mathbb{C}^p$ and measurement outcomes $y_1, \ldots, y_n$, modeled as independent random variables following the Poisson distribution:

$$\mathbb{P}\left\{y_i = y\right\} = \frac{\exp\left(-\lambda_i\right)\lambda_i^y}{y!}, \quad y \in \{0\} \cup \mathbb{N}$$

where $\lambda_i := \left|\left\langle a_i, x^\natural \right\rangle\right|^2$ for all i. In practice, each y_i represents the number of photons detected by the sensor [15].

The corresponding maximum-likelihood (ML) estimation yields a non-convex optimization problem which is difficult to solve. A recent approach to circumvent this computational issue is PhaseLift [7, 11]. The PhaseLift approach casts the phase retrieval problem as a low rank matrix recovery problem, and then we can apply any convex optimization-based estimator, such as the basis pursuit like estimator [27], the nuclear-norm penalized estimator [10], and the Lasso like estimator [13].

This work was supported in part by ERC Future Proof, SNF 200021-146750 and SNF CRSII2-147633.

Following the PhaseLift approach, we show in Section 2 that we can recover $x^\natural$ by solving

$$\hat{X} \in \arg\min_X \left\{f(X) : X \in \mathcal{X}\right\}, \tag{1}$$

where

$$f(X) := \sum_{i=1}^n \left\{-y_i \log\left[\mathrm{Tr}\left(A_i X\right)\right] + \mathrm{Tr}\left(A_i X\right)\right\}, \tag{2}$$

$$\mathcal{X} := \left\{X \geq 0,\ \|X\|_* \leq c,\ X \in \mathbb{C}^{p\times p}\right\}. \tag{3}$$

for some $c > 0$, $A_i := a_i a_i^H$. A rule of thumb for choosing c can be found in Section 3. We then find an eigenvector associated with the largest eigenvalue of $\hat{X}$ as our estimate of $x^\natural$.

It is easy to check that (1) is a convex optimization problem. Existing convex optimization tools, however, are not directly applicable to solving (1) due to two issues.

1. Most existing algorithms, such as [30], are computationally expensive for nuclear norm constraints, as they require computing the eigenvalue decomposition of a matrix in $\mathbb{C}^{p\times p}$ at each iteration.
2. While Frank-Wolfe-type algorithms can be relatively scalable for nuclear norm constraints [21], existing theoretical convergence guarantees for these Frank-Wolfe-type algorithms are not valid for our loss function in (1).

We will address the issues in detail in Section 4.

In this paper, we show that the standard Frank-Wolfe algorithm works for the optimization problem (1), with a properly chosen parameter to be explicitly specified in Theorem 5.1. Our theorem guarantees that the Frank-Wolfe algorithm converges at the rate $\mathcal{O}(1/t)$ globally, where t is the iteration counter. Numerical experiments show that the empirical convergence rate can be even faster. The algorithm shares the same merit of the standard Frank-Wolfe algorithm, in the sense that it is scalable when dealing with a nuclear norm constraint.

To the best of our knowledge, this is the first theoretical guarantee for the Frank-Wolfe algorithm applied to a non-Hölder (and hence non-Lipschitz) continuous gradient objective function.

2. POISSON PHASE RETRIEVAL BY CONVEX OPTIMIZATION

For the Poisson noise model, the ML estimator of $x^\natural$ is given by

$$\hat{x}_{\mathrm{ML}} \in \arg\min_x \{L(x) : x \in \mathbb{C}^p\} \tag{4}$$

where L is the negative log-likelihood function (under a constant shift):

$$L(x) := \sum_{i=1}^n \left[-y_i \log\left(|\langle a_i, x\rangle|^2\right) + |\langle a_i, x\rangle|^2\right].$$

The function L, unfortunately, is non-convex, and currently there does not exist a well-guaranteed algorithm for solving the optimization problem.

Motivated by the PhaseLift approach [7, 11], we can reformulate the non-convex optimization problem (4) as follows. Define $A_i := a_i a_i^H$ for all i, and $X^\natural := x^\natural (x^\natural)^H$. Then we have

$$\left|\left\langle a_i, x^\natural\right\rangle\right|^2 = \mathrm{Tr}\left(A_i X^\natural\right) \quad i = 1, \ldots, n$$

where $\mathrm{Tr}(\cdot)$ denotes the trace function, and hence we can rewrite the original optimization problem as

$$\hat{x}_{\mathrm{ML}} \in \arg\min_x \left\{f(X) : X = xx^H, x \in \mathbb{C}^p\right\}$$

where f is given in (2). This is equivalent to the optimization problem

$$\hat{X}_{\mathrm{ML}} \in \arg\min_X \left\{f(X) : X \geq 0, \mathrm{rank}(X) = 1, X \in \mathbb{C}^{p\times p}\right\}.$$

Note that given $\hat{X}_{\mathrm{ML}}$, $\hat{x}_{\mathrm{ML}}$ can be recovered via the relation $\hat{X}_{\mathrm{ML}} = \hat{x}_{\mathrm{ML}}\hat{x}_{\mathrm{ML}}^H$.

As the variable X is always of rank 1, we can consider the convex relaxation given in (1). We then find an eigenvector associated with the largest eigenvalue of $\hat{X}$ as our estimate of $x^\natural$.

It is easy to verify that (1) is a convex optimization problem.

3. A RULE OF THUMB FOR SETTING THE CONSTRAINT

In the convex optimization formulation (1), we leave one parameter c unspecified. The ideal setting should be $c = \left\|X^\natural\right\|_* = \left\|x^\natural\right\|_2^2$. While this setting may not be practically feasible, we need $c > \left\|x^\natural\right\|_2^2$ to ensure that $X^\natural$ is in the constraint set $\mathcal{X}$.

The following theorem shows that choosing $c = (1/n)\sum_{i=1}^n y_i$ suffices, if the sampling scheme satisfies an isometry property with high probability.

Proposition 3.1. *Let $A \in \mathbb{C}^{n\times p}$, whose i-th row is given by a_i^H. Assume that there exists some $\varepsilon > 0$ such that*

$$(1-\varepsilon)\left\|x^\natural\right\|_2^2 \leq \left\|\frac{1}{\sqrt{n}}Ax^\natural\right\|_2^2 \leq (1+\varepsilon)\left\|x^\natural\right\|_2^2 \tag{5}$$

with probability at least $1 - p_\epsilon$. Then we have, for any $t > 0$,

$$\bar{y} := \frac{1}{n}\sum_{i=1}^n y_i > (1+\varepsilon)\left\|x^\natural\right\|_2^2 + t$$

with probability at least $1 - p_\epsilon - p_t$, where

$$p_t := \exp\left[-\frac{nt}{4}\log\left(1 + \frac{t}{2(1+\varepsilon)\left\|x^\natural\right\|_2^2}\right)\right].$$

If $x^\natural$ is sparse, then the isometry condition (5) can be implied by the restricted isometry property (RIP) of A [6, 16, 28]. Even without sparsity, if n is significantly larger than p, a matrix A of independent and identically distributed (i.i.d.) subgaussian random variables can also satisfy (5) with high probability [16].

While the isometry property of the Fourier measurement with a coded diffraction pattern is unclear currently, we show via numerical experiments in Section 6 that this rule of thumb works well on both synthetic and real-world data.

4. REVIEW OF CONVEX OPTIMIZATION TOOLS

We address why several existing convex optimization algorithms are not applicable to (1) in this section.

We note that (1) is a constrained convex minimization problem with a smooth loss function, and there are many well-known algorithms for solving such a problem. State-of-the-art choices for large-scale applications include the proximal gradient-type methods [1, 2, 12, 23, 25, 30], alternating direction method of multipliers (ADMM) [14], and Frank-Wolfe-type algorithms (a.k.a. conditional gradient methods) [17, 18, 19, 21, 24, 31, 32]. There are also well-developed MATLAB packages available on the Internet [3, 30]. Those seemingly ready-to-use convex optimization tools, however, are not desirable for solving our problem (1) for two issues.

The first issue is scalability. When applied to the problem (1), both proximal gradient-type methods and the ADMM require computing the *prox-mapping* given by

$$\mathrm{prox}(X) := \arg\min_S \{\omega(S - X) : S \in \mathcal{X}\}$$

for a given strongly convex "distance generating function" (DGF) ω. A standard choice of DGF for matrix variables is $\omega(X) := (1/2)\|X\|_F^2$, where $\|\cdot\|_F$ denotes the Frobenius norm. For a positive semi-definite matrix $X \in \mathbb{C}^{p\times p}$, whose eigenvalue decomposition is $X = U\mathrm{diag}(v)U^H$, we have $\mathrm{prox}(X) = U\mathrm{diag}(\tilde{v})U^H$, where $\tilde{v}$ is the Euclidean projection of v onto the standard simplex in $\mathbb{R}^p$ scaled by c. While the prox-mapping is simple to describe, the eigenvalue decomposition renders the algorithm slow when the parameter dimension p is large, as its computational complexity is in general $O(p^3)$. Similar issues exist when we choose other DGFs.

Scalability is a major reason why Frank-Wolfe-type algorithms have been attracting attention in recent years. We summarize the standard Frank-Wolfe algorithm (when applied to (1)) in Algorithm 1, where $(\tau_t)_{t=1}^T$ is a sequence of real numbers in the interval $(0, 1]$ to be specified.

Here we have a slight abuse of notations. When applied to our specific problem (1), the variables $x_0, \ldots, x_t$ and $\nabla f(x_t)$ should be understood as their matrix counterparts $X_0, \ldots, X_t$ and $\nabla f(X_t)$, respectively.

Algorithm 1 (The standard Frank-Wolfe algorithm)

Choose an arbitrary $x_0 \in \mathcal{X}$
for $t = 0, \ldots, T$ **do**
  Compute $v_t \in \arg\min_s \{\langle s, \nabla f(x_t)\rangle : s \in \mathcal{X}\}$
  Update $x_{t+1} = (1 - \tau_t)x_t + \tau_t v_t$
end for

The only computational bottleneck is in computing v_t (or its matrix counterpart V_t). For the specific constraint set $\mathcal{X}$ given in (3) and any positive semi-definite matrix X_t, it can be easily verified that V_t is a scaled rank-one approximation of $\nabla f(X_t)$, and hence can be

efficiently computed by the Lanczos method [21]. More precisely, let $u_t \in \mathbb{C}^p$ be an eigenvector of $\nabla f(X_t)$ associated with the largest eigenvalue. We have $V_t = c(u_t u_t^H)$.

Unfortunately, the second issue arises: none of the existing theoretical convergence guarantees for Frank-Wolfe-type algorithms, to the best of our knowledge, is valid for the specific loss function (2). The result in [21] requires a bounded curvature condition; [17, 18, 19] require the gradient of the objective function to be Lipschitz continuous; [24] requires a weaker condition that the gradient is Hölder continuous; the Frank-Wolfe like algorithm in [31, 32] requires the gradient of the conjugate of the objective function to be Hölder continuous. All of the conditions mentioned above implicitly presumes $\mathcal{X} \subseteq \mathrm{dom}(f)$, but this is not the case for (1), since $0 \in \mathcal{X}$ but $0 \notin \mathrm{dom}(f)$.

The second issue also exists for proximal gradient-type methods and the ADMM, as [1, 2, 12, 14, 23, 25] also require the Lipschitz continuity of the gradient. The only exception is the composite self-concordant minimization algorithms proposed in [30]—the logarithmic function is a typical example of self-concordant functions.

There are some works on noiseless phase retrieval by non-convex optimization techniques [7, 26], and provide theoretical convergence guarantees. The convergence guarantees do not extend to the Poisson noise case.

5. CONVERGENCE GUARANTEE

In this section, we provide convergence guarantee of the standard Frank-Wolfe method in Algorithm 1 for the prototype constrained convex optimization optimization problem:

$$g^\star := \min_{X \in \mathcal{C}} \{g(X) : X \in \mathcal{C}\} \tag{6}$$

where $\mathcal{C}$ is a nuclear norm ball in $\mathbb{R}^{p\times p}$, and

$$g(X) := \mathrm{Tr}(\Psi X) - \sum_{i=1}^{n} \eta_i \log \mathrm{Tr}(\Phi_i X) \tag{7}$$

for some $\Psi \in \mathbb{R}^{p\times p}$, non-negative integers $\eta_1, \ldots, \eta_n$, and positive semi-definite matrices $\Phi_1, \ldots, \Phi_n \in \mathbb{R}^p$.

We start with some definitions. Let $\|\cdot\|$ be the spectral norm on $\mathbb{R}^{p\times p}$, and $\|\cdot\|_*$ be the nuclear norm. Define $d_{\mathcal{C}}$ as the diameter of $\mathcal{C}$, i.e.,

$$d_{\mathcal{C}} := \max_{X,Y} \{\|X - Y\| : X, Y \in \mathcal{C}\}.$$

Let $d_\Phi := \max_i \|\Phi_i\|$ and $d_\Psi := \|\Psi\|$. Furthermore, we define

$$\bar{\mu} := \max_{i,x} \{\mathrm{Tr}(\Phi_i X) : 1 \le i \le n, X \in \mathcal{C}\}$$
$$\underline{\mu} := \min_{i} \{\mathrm{Tr}(\Phi_i X_0) : 1 \le i \le n\}.$$

Notice that we need to choose X_0 such that $\underline{\mu} > 0$, due to the presence of logarithmic functions in g.

Our main theoretical result is the following theorem:

Theorem 5.1. *Consider the optimization problem (6). The iterates $(X_t)_{t\ge 0}$ given by Algorithm 1 with*

$$\tau_t := \frac{2}{t+3}$$

satisfies

$$g(X_t) - g^\star < \frac{8\gamma^2 d_\Phi^2 d_{\mathcal{C}}^2}{t+2} + \frac{2 d_{\mathcal{C}} \|\nabla g(X_0)\|}{\underline{\mu}(t+1)(t+2)}$$

The quantity $\gamma := \max\{\gamma_1, \gamma_2, \gamma_3\}$ is a constant independent of t, where

$$\gamma_1 := \frac{2 d_\Psi d_{\mathcal{C}}}{\underline{\mu}}, \quad \gamma_2 := 2\frac{n d_\eta}{\underline{\mu}} \left(\frac{4n\bar{\mu} d_\eta}{\underline{\mu}} + 1\right)^2,$$
$$\gamma_3 := \frac{64 n^2 \bar{\mu}^2 d_\eta^2}{\underline{\mu}^3} \left(\frac{4n\bar{\mu} d_\eta}{\underline{\mu}} + 1\right).$$

Consequently, we have $g(X_t) - g^\star = O(1/t)$.

Theorem 5.1 establishes the validity of using the standard Frank-Wolfe algorithm to solve (1). We note that this theorem is a *worst case* guarantee for all loss functions of the form (7). As we will see in the next section, empirically, both the constant and the convergence rate can be much better.

Our choice of τ_t is slightly different from the standard one in [21, 24], where $\tau_t := 2/(t+2)$. This is due of technical concerns in the proof.

As a short sketch, the key idea is to show the boundedness of $\|\nabla g(X_{t+1}) - \nabla g(X_t)\|$ for all t, where $\|\cdot\|$ denotes the spectral norm. This bound, by the framework in [24], is sufficient to establish the convergence guarantee. This is simple if the gradient is Hölder continuous, since then

$$\|\nabla g(X_{t+1}) - \nabla g(X_t)\| \le L_\nu \|X_{t+1} - X_t\|_*^\nu \le L_\nu d_{\mathcal{C}}^\nu$$

for some $\nu \in (0, 1]$ and $L_\nu > 0$. For the optimization problem (1) we consider, this issue can be reduced to the boundedness of

$$C_t := \sum_{i=1}^{n} \frac{\eta_i}{\mathrm{Tr}(\Phi_i X_t)}$$

for all t. We complete the proof by showing that C_t is bounded above by a constant for all t, if we choose $\tau_t = 2/(t+3)$.

6. NUMERICAL RESULTS

In this section, we present numerical evidence to assess the convergence behaviour and the scalability of the proposed Frank-Wolfe algorithm.

Our numerical experiment is based on coded diffraction pattern measurements with the octonary modulation, which were considered in [9, 31] for the noiseless model. A similar setup was also considered also in [8] for the Poisson noise model.

In [8], the MATLAB package TFOCS [3] was used to solve a convex optimization problem similar to (1). The algorithm, however, is not guaranteed to converge for the problem under our consideration (cf. Section 4). Therefore, we compare the Frank-Wolfe algorithm with the proximal gradient method in the Self-Concordant OPTimization toolbox (SCOPT) [30]. Recall that our loss function is self-concordant, and hence the algorithms in [30] are applicable.

In our first experiment, we consider the random Gaussian signal model: We generate a random complex Gaussian vector $x^\natural \in \mathbb{C}^p$ with i.i.d. entries, where the real and the imaginary parts of the each entry of $x^\natural$ are independent and sampled from the standard Gaussian distribution.

We run both algorithms starting from the same Gaussian initial iterate, sampled from the same distribution as $x^\natural$. We keep track of the objective value and the elapsed time over the iterations, and compute the approximate relative objective residual ($|f - f^*|/|f^*|$) as the performance measure, where the actual optimum value $f^\star$ is approximated by f^*, the minimum objective value obtained by

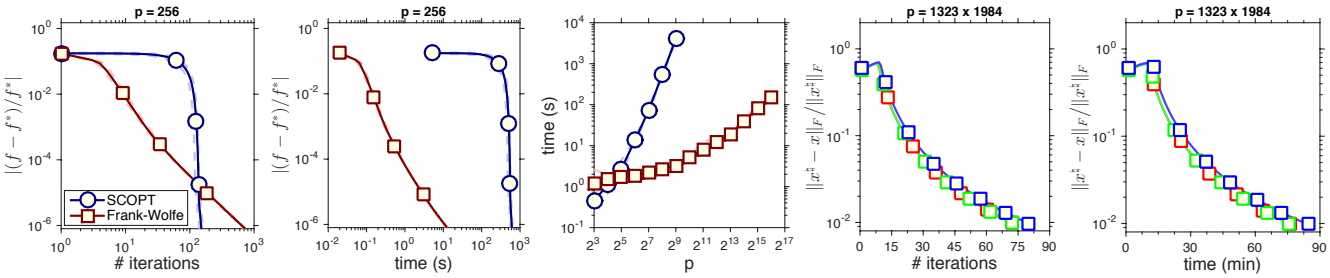

Fig. 1. Convergence behaviour of the algorithms for different data sizes: The three plots on the left correspond to the first experiment. Solid lines show the average performance over 10 random trials, and the two dashed lines show the best and the worst performances, respectively. The two plots on the right correspond to the second experiment. Each color (blue, green, red) represents one color channel.

running 200 iterations of the SCOPT and/or 10000 iterations of the Frank-Wolfe algorithm.

In the second experiment, we test the scalability of the Frank-Wolfe approach, by recovering a real image as in [9, 31]. We choose the EPFL campus image of size 1323×1984 as the signal to be measured, which corresponds to a signal dimension $p = 2624832$. We apply the Frank-Wolfe algorithm to recover three color channels separately, and stop the algorithm when 10^{-2} recovery error ($\|x - x^\natural\|_F / \|x^\natural\|_F$) is reached.

In both experiments, we set the constraint parameter c to the mean of the measurements, following the rule of thumb in Section 3, and we set the number of different modulating waveforms L to 20.

Fig. 2. An EPFL image of size 1323×1984, reconstructed by 75 iterations of the Frank-Wolfe algorithm: PSNR = 44.92 dB.

We implement Algorithm 1 in MATLAB and use the built-in `eigs` function, which is based on the Lanczos algorithm, with 10^{-3} relative error tolerance, to perform the minimization step of the Frank-Wolfe algorithm. In the weighting step, we adapt the efficient thin singular value decomposition updating method of [5] under low rank modifications, as explained in [31], in order to tame the memory growth.

We time our experiments on a computer cluster, and restricting the computational resource to 8 CPU of 2.40 GHz and 32 GB of memory space per simulation.

Figure 1 illustrates the convergence behaviour of the algorithms for different data sizes.

The first three plots on the left correspond to the first experiment. Solid lines show the average performance over 10 random trials, and the two dashed lines show the best and the worst instances, respectively. In the first two plots, we observe that the empirical rate of convergence is about $O(t^{-1.89})$, which is better than the theoretically guaranteed rate $O(t^{-1})$. In the third plot, we show the time required to reach a predefined accuracy level of 10^{-5} in terms of the relative objective residual, for different data sizes.

The last two plots of Figure 1 correspond to the second experiment, which also provides an empirical evidence for the estimation quality using the constraint parameter c. Each color (blue, green, red) represents one color channel.

Finally, Figure 2 shows the estimate x_t, after 75 iterations of the Frank-Wolfe method. The PSNR of the reconstructed image is 44.92dB.

Notice that, considering the lifted dimensions p^2 in the second experiment, even the generation of a simple iterate X_t would require approximately 7 TB of memory space, for a single color channel, when using the prox-mapping-based solver in SCOPT. By avoiding the computation of the prox-mapping, and adapting the efficient low rank updates, the Frank-Wolfe algorithm keeps a low memory footprint, and hence is more scalable compared to the self-concordant optimization method in SCOPT.

7. DISCUSSION

While we focus on the Poisson phase retrieval problem in this paper, our main contribution is in verifying the validity of applying the standard Frank-Wolfe algorithm to optimization problems of the form (1). Therefore, the application of our result is not restricted to Poisson phase retrieval. One interesting application is ML estimation for quantum state tomography [20], where the parameter dimension grows exponentially fast with the number of qubits, and the physical model naturally imposes a nuclear norm constraint.

8. PROOFS

8.1. Proof of Proposition 3.1

Notice that, conditioning on $a_1, \ldots, a_n$, $n\bar{y}$ is a Poisson random variable with mean $\sum_{i=1}^n \lambda_i$. By the tail bound for Poisson random variables [4, 22], conditioning on $a_1, \ldots, a_n$, we have for any $t > 0$,

$$\mathbb{P}\left\{\bar{y} - \mathbb{E}\,\bar{y} > t\right\} \leq \exp\left[-\frac{nt}{4}\log\left(1 + \frac{t}{2\lambda}\right)\right],$$

where $\lambda := (1/n)\sum_{i=1}^n \lambda_i$.

Recall that $\lambda_i := \left|\langle a_i, x^\natural\rangle\right|^2$. By the assumption on A, we have $(1-\delta)\left\|x\right\|_2^2 \le \lambda \le (1+\delta)\left\|x\right\|_2^2$ with probability at least $1-p_\delta$. Moreover, on this event, we have

$$\begin{aligned}&\mathbb{P}\left\{\bar{y} - (1+\delta)\left\|x^\natural\right\|_2^2 > t\right\}\\ &\quad\le \mathbb{P}\left\{\bar{y} - \lambda > t\right\}\\ &\quad\le \exp\left[-\frac{nt}{4}\log\left(1+\frac{t}{2(1+\delta)\left\|x^\natural\right\|_2^2}\right)\right].\end{aligned}$$

This proves the theorem.

8.2. Proof of Theorem 5.1

Let $(\alpha_t)_{t\ge 0}$, $\alpha_0 \neq 0$ be a sequence of non-negative real numbers. We consider step sizes of the form

$$\tau_t = \alpha_{t+1}/S_{t+1}, \tag{8}$$

where $S_t := \sum_{k=0}^t \alpha_t$. Unless otherwise stated, $(X_t)_{t\ge 0}$ refers to the sequence of iterates generated by Algorithm 1, with the step size chosen as in (8). Notice that then the convergence rate of the algorithm can depend on the sequence $(\alpha_t)_{t\ge 0}$.

By the convexity of $\mathcal{C}$, it is obvious that $X_t \in \mathcal{C}$ for all t. Due to the presence of the logarithmic function, we also need to verify that $X_t \in \mathrm{dom}(g)$ for all t.

Proposition 8.1. *The following hold.*

1. $\mathrm{Tr}(\Phi_i V_t) \ge 0$ *for all* i *and* t.
2. *If* $\mathrm{Tr}(\Phi_i X_0) > 0$, *then* $\mathrm{Tr}(\Phi_i X_t) > 0$ *for all* i *and* t.

Proof. See Section 8.3. □

Now we show the boundedness of C_t for all t, as stated in Section 5. Recall that $C_t := \sum_{i=1}^n (\eta_i/\mathrm{Tr}(\Phi_i X_t))$.

Lemma 8.2. *For any* T *such that* $1 - 4n(\bar{\mu}/\underline{\mu})d_\eta\tau_T > 0$, *we have* $C_t \le C$, *where* C *is a constant independent of* t *defined as*

$$C := \max\left\{\frac{2d_\Psi d_{\mathcal{C}}}{\underline{\mu}}, C_0\prod_{i=0}^T\frac{1}{1-\tau_i}, \frac{64n^2\bar{\mu}^2 d_\eta^2}{\underline{\mu}^3\left(1-\frac{4n\bar{\mu}d_\eta}{\underline{\mu}}\tau_T\right)}\right\}.$$

Proof. See Section 8.4. □

The following lemma mimics [24, Lemma 2]. Define

$$\begin{aligned}B_t := \alpha_0\max\left\{\langle\nabla g(X_0), X_0 - X\rangle : X\in\mathcal{X}\right\}\\ + \left(\sum_{k=1}^t \frac{\alpha_k^2}{S_{k-1}}\right)\gamma,\end{aligned}$$

where $\gamma := C^2 d_\Phi^2 d_{\mathcal{C}}^2$.

Lemma 8.3. *For any* $t \ge 0$ *and* $X \in \mathcal{X}$, *we have*

$$S_t g(X_t) \le \sum_{k=0}^t \left\{\alpha_k\left[g(X_k) + \langle\nabla g(X_k), X - X_k\rangle\right]\right\} + B_t$$

Proof. See Section 8.5. □

Set $X = X^\star$, a minimizer, in Lemma 8.3, and notice that

$$g(X_k) + \langle\nabla g(X_k), X^\star - X_k\rangle \le g^\star$$

for all k. We immediately obtain a convergence guarantee for any $(a_t)_{t\ge 0}$.

Corollary 8.4. *We have* $g(X_t) - g^\star \le (B_t/S_t)$.

Now we consider the special case where $\alpha_t = t+1$. As then $S_t = (t+1)(t+2)/2$, this choice corresponds to $\tau_t = 2/(t+3)$ as in Theorem 5.1.

Proposition 8.5. *Choose* $\alpha_t = t+1$. *We have*

$$\frac{B_t}{S_t} < \frac{8\left(\max\left\{\gamma_1,\gamma_2,\gamma_3\right\}\right)^2 d_\Phi^2 d_{\mathcal{C}}^2}{t+2} + \frac{2d_{\mathcal{C}}\left\|\nabla g(X_0)\right\|}{(t+1)(t+2)},$$

where

$$\begin{aligned}\gamma_1 &:= \frac{2d_\Psi d_{\mathcal{C}}}{\underline{\mu}}, \quad \gamma_2 := 2\frac{nd_\eta}{\underline{\mu}}\left(\frac{4n\bar{\mu}d_\eta}{\underline{\mu}}+1\right)^2,\\ \gamma_3 &:= \frac{64n^2\bar{\mu}^2 d_\eta^2}{\underline{\mu}^3}\left(\frac{4n\bar{\mu}d_\eta}{\underline{\mu}}+1\right).\end{aligned}$$

Proof. See Section 8.6. □

8.3. Proof of Proposition 8.1

Recall that V_t is always a positive semi-definite matrix of rank 1, as discussed in Section 4. Since Φ_i is also positive semi-definite, this implies $\mathrm{Tr}(\Phi_i V_t) \ge 0$ for all i and t.

We prove the second claim by induction. The second claim holds true for $t = 0$ by assumption. Suppose $\mathrm{Tr}(\Phi_i X_t) > 0$ for some $t \ge 0$ for all i. Because of the assumption that $\alpha_0 \neq 0$, we always have $\tau_t < 1$ for all t. Then

$$\begin{aligned}\mathrm{Tr}(\Phi_i X_{t+1}) &= (1-\tau)\mathrm{Tr}(\Phi_i X_t) + \tau_t\mathrm{Tr}(\Phi_i V_t)\\ &\ge (1-\tau)\mathrm{Tr}(\Phi_i X_t) > 0,\end{aligned}$$

where the first inequality is by the first claim.

8.4. Proof of Lemma 8.2

Consider the sequence $(C_t)_{t\ge 0}$. Roughly speaking, the idea behind the proof is to show that there exists some $T > 0$, such that $C_{t+1} \le C_t$ for all $t \ge T$; then we can bound C_t from above by C_T for all $t \ge T$, a constant independent of t. Notice that, however, the actual argument in this proof is slightly more delicate (cf. the proof of Proposition 8.8).

A simple bound on C_{t+1} is

$$\begin{aligned}C_{t+1} &= \sum_{i=1}^n \frac{\eta_i}{\mathrm{Tr}(\Phi_i X_{t+1})}\\ &\le \frac{1}{(1-\tau_t)}\sum_{i=1}^n\frac{\eta_i}{\mathrm{Tr}(\Phi_i X_t)} = \frac{1}{1-\tau_t}C_t,\end{aligned} \tag{9}$$

using the fact that $\mathrm{Tr}(\Phi_i X_{t+1}) \ge (1-\tau_t)\mathrm{Tr}(\Phi_i X_t)$. This yields the following simple result.

Proposition 8.6. *We have* $C_t \le C_0\prod_{i=0}^t(1-\tau_t)^{-1}$.

However, as $1-\tau_t < 1$, the upper bound (9) is not sharp enough for our purpose.

Notice that for any k, we have

$$\begin{aligned} C_{t+1} &= \sum_{i\neq k} \frac{\eta_i}{\mathrm{Tr}(\Phi_i X_{t+1})} + \frac{\eta_k}{\mathrm{Tr}(\Phi_k X_{t+1})} \\ &\leq \sum_{i\neq k} \frac{\eta_i}{(1-\tau_t)\mathrm{Tr}(\Phi_i X_t)} + \frac{\eta_k}{\mathrm{Tr}(\Phi_k X_{t+1})} \\ &= \frac{C_t}{1-\tau_t} - \frac{\eta_k}{(1-\tau_t)\mathrm{Tr}(\Phi_k X_t)} + \frac{\eta_k}{\mathrm{Tr}(\Phi_k X_{t+1})} \\ &= \frac{C_t}{1-\tau_t} - \frac{\eta_k \tau_t \mathrm{Tr}(\Phi_k V_t)}{[(1-\tau_t)\mathrm{Tr}(\Phi_k X_t)]\,\mathrm{Tr}(\Phi_k X_{t+1})} \\ &\leq \frac{C_t}{1-\tau_t} - \xi_k \end{aligned} \tag{10}$$

where

$$\xi_k := \frac{\tau_t \mathrm{Tr}(\Phi_k V_t)}{[(1-\tau_t)\mathrm{Tr}(\Phi_k X_t)]\,\mathrm{Tr}(\Phi_k X_{t+1})};$$

the last inequality is due to the fact that either $\eta_k = 0$ or $\eta_k \geq 1$ in the Poisson phase retrieval problem. This bound is sharper than (9), as ξ_k is always non-negative.

Proposition 8.7. *If $C_t > 2\underline{\mu}^{-1} d_\Psi d_{\mathcal{C}}$, then there exists some $k \leq n$ such that*

$$\begin{aligned} \frac{1}{\mathrm{Tr}(\Phi_k X_t)} &\geq \frac{\underline{\mu} C_t}{4n\bar{\mu} d_\eta}, \\ \mathrm{Tr}(\Phi_k V_t) &\geq \frac{\underline{\mu}}{4}. \end{aligned}$$

Proof. We prove by contradiction. By the definition of V_t, we have $\langle V_t, \nabla g(X_t)\rangle \leq \langle X_0, \nabla g X_t\rangle$, and hence

$$\begin{aligned} \sum_{i=1}^n \frac{\eta_i \langle V_t, \Phi_i\rangle}{\langle X_t, \Phi_i\rangle} &\geq \sum_{i=1}^n \left(\frac{\eta_i \langle X_0, \Phi_i\rangle}{\langle X_t, \Phi_i\rangle}\right) + \langle \Psi, X_0 - V_t\rangle \\ &\geq \sum_{i=1}^n \frac{\eta_i \langle X_0, \Phi_i\rangle}{\langle X_t, \Phi_i\rangle} - d_\Psi d_{\mathcal{C}} \\ &\geq \underline{\mu} C_t - d_\Psi d_{\mathcal{C}} \geq \frac{\underline{\mu} C_t}{2}. \end{aligned}$$

Let Ω be the set of i's such that $\langle X_t, \Phi_i\rangle^{-1} \geq \underline{\mu} C_t/(4n\bar{\mu} d_\eta)$. Suppose the claim of the proposition is false, i.e. for all $i \in \Omega$, $\langle V_t, \Phi_i\rangle < \underline{\mu}/4$. Then we have

$$\begin{aligned} \sum_{i=1}^n \frac{\eta_i \langle V_t, \Phi_i\rangle}{\langle X_t, \Phi_i\rangle} &= \sum_{i\in\Omega} \frac{\eta_i \langle V_t, \Phi_i\rangle}{\langle X_t, \Phi_i\rangle} + \sum_{i\notin\Omega} \frac{\eta_i \langle V_t, \Phi_i\rangle}{\langle X_t, \Phi_i\rangle} \\ &< \frac{\underline{\mu}}{4} C_t + n d_\eta \bar{\mu} \frac{\underline{\mu} C_t}{4n\bar{\mu} d_\eta} = \frac{\underline{\mu} C_t}{2}, \end{aligned}$$

a contradiction. This completes the proof. □

Assume $C_t > 2\underline{\mu}^{-1} d_\Psi d_{\mathcal{C}}$. By Proposition 8.7 and (10), we have

$$C_{t+1} \leq C_t \left\{ \frac{1}{1-\tau_t} - \frac{\frac{\tau_t \underline{\mu}}{4}}{(1-\tau_t)\frac{4n\bar{\mu} d_\eta}{\underline{\mu}} \left[(1-\tau_t)\frac{4n\bar{\mu} d_\eta}{\underline{\mu} C_t} + \tau_t \frac{\underline{\mu}}{4}\right]} \right\}.$$

By direct calculation, we obtain $C_{t+1} \leq C_t$, if

$$\begin{aligned} 1 - \frac{4n\bar{\mu} d_\eta}{\underline{\mu}} \tau_t &> 0, \\ C_t \geq \kappa_t &:= \frac{64(1-\tau_t) n^2 \bar{\mu}^2 d_\eta^2}{\underline{\mu}^3 \left(1 - \frac{4n\bar{\mu} d_\eta}{\underline{\mu}} \tau_t\right)}. \end{aligned} \tag{11}$$

Proposition 8.8. *Assume that $C_t > 2\underline{\mu}^{-1} d_\Psi d_{\mathcal{C}}$. Choose T such that* (11) *holds for $t = T$. Then we have*

$$C_t \leq \max\left\{ C_0 \prod_{i=0}^T \frac{1}{1-\tau_i}, \frac{64 n^2 \bar{\mu}^2 d_\eta^2}{\underline{\mu}^3 \left(1 - \frac{4n\bar{\mu} d_\eta}{\underline{\mu}} \tau_T\right)} \right\}.$$

Proof. Since $(\tau_t)_{t\geq 0}$ is a decreasing sequence, the inequality (11) holds for all $t \geq T$.

If $t \leq T$, we can apply Proposition 8.6, and obtain

$$C_t \leq C_0 \prod_{i=0}^t \frac{1}{1-\tau_i} \leq C_0 \prod_{i=0}^T \frac{1}{1-\tau_i}.$$

Consider the case when $t > T$. Suppose $C_T \geq \kappa_T$. We have $C_{t+1} \leq C_t \leq C_T$, which can be bounded using Proposition 8.6, until some t^* such that $C_{t^*} < \kappa_{t^*}$. But then $C_{t+1} \leq (1-\tau_t)^{-1}\kappa_t$ for all $t \geq t^*$. If $C_T < \kappa_T$, similarly, we also obtain $C_{t+1} \leq (1-\tau_t)^{-1}\kappa_t$ for all $t \geq T$. The proposition follows, as

$$\frac{1}{1-\tau_t}\kappa_t = \frac{64 n^2 \bar{\mu}^2 d_\eta^2}{\underline{\mu}^3 \left(1 - \frac{4n\bar{\mu} d_\eta}{\underline{\mu}} \tau_t\right)} \leq \frac{64 n^2 \bar{\mu}^2 d_\eta^2}{\underline{\mu}^3 \left(1 - \frac{4n\bar{\mu} d_\eta}{\underline{\mu}} \tau_T\right)}.$$

□

If $C_t \leq 2\underline{\mu}^{-1} d_\Psi d_{\mathcal{C}}$, then this is already a constant upper bound on C_t. This completes the proof.

8.5. Proof of Lemma 8.3

We prove by induction. The claim is obviously correct for $t = 0$. Suppose the claim holds for some $t \geq 0$. Then we have

$$\begin{aligned} &\sum_{k=0}^{t+1} \alpha_k \left[g(X_k) + \langle \nabla g(X_k), X - X_k\rangle\right] + B_t \\ &\quad \geq S_t g(X_t) + \alpha_{k+1}\left[g(X_{t+1}) + \langle \nabla g(X_{t+1}), X - X_{t+1}\rangle\right] \\ &\quad = S_{t+1} g(X_{t+1}) + S_t\left[g(X_t) - g(X_{t+1})\right] \\ &\qquad + \langle \nabla g(X_{t+1}), \alpha_{t+1}(X - X_{t+1})\rangle. \\ &\quad \geq S_{t+1} g(X_{t+1}) \\ &\qquad + \langle \nabla g(X_{t+1}), \alpha_{t+1}(X - X_{t+1}) + S_t(X_t - X_{t+1})\rangle \\ &\quad = S_{t+1} g(X_{t+1}) + \alpha_{t+1}\langle \nabla g(X_{t+1}), X - V_t\rangle \\ &\quad \geq S_{t+1} g(X_{t+1}) + \alpha_{t+1}\langle \nabla g(X_{t+1}) - \nabla g(X_t), X - V_t\rangle, \end{aligned}$$

where the second inequality is due to convexity of g, and the third inequality is due to the fact that

$$\langle \nabla g(X_t), X - V_t\rangle \geq 0$$

for any $X \in \mathcal{C}$, as V_t minimizes $\langle \nabla g(X_t), \cdot\rangle$ on $\mathcal{C}$.

To complete the proof, we need to show that

$$\alpha_{t+1}\langle \nabla g(X_{t+1}) - \nabla g(X_t), X - V_t\rangle \geq B_t - B_{t+1} = -\frac{\alpha_{t+1}^2}{S_t}\gamma,$$

or

$$\langle \nabla g(X_{t+1}) - \nabla g(X_t), X - V_t \rangle \geq -\frac{\alpha_{t+1}}{S_t}\gamma. \tag{12}$$

By Hölder's inequality, we have

$$\begin{aligned}
&|\langle \nabla g(X_{t+1}) - \nabla g(X_t), X - V_t \rangle| \\
&\quad \leq \|\nabla g(X_{t+1}) - \nabla g(X_t)\| \, \|X - V_t\|_* \\
&\quad \leq \|\nabla g(X_{t+1}) - \nabla g(X_t)\| \, d_{\mathcal{C}},
\end{aligned}$$

where $\|\cdot\|$ denotes the spectral norm.

Now we bound the quantity $\|\nabla g(X_{t+1}) - \nabla g(X_t)\|$. By direct calculation, we obtain

$$\begin{aligned}
&\|\nabla g(X_{t+1}) - \nabla g(X_t)\| \\
&\quad = \left\| \sum_{i=1}^{n} \frac{\eta_i \langle X_t - X_{t+1}, \Phi_i \rangle}{\langle X_t, \Phi_i \rangle \langle X_{t+1}, \Phi_i \rangle} \Phi_i \right\| \\
&\quad \leq d_\Phi \sum_{i=1}^{n} \frac{\eta_i \, |\langle X_t - X_{t+1}, \Phi_i \rangle|}{\langle X_t, \Phi_i \rangle \langle X_{t+1}, \Phi_i \rangle} \\
&\quad = \tau_t d_\Phi \sum_{i=1}^{n} \frac{\eta_i \, |\langle X_t - V_t, \Phi_i \rangle|}{\langle X_t, \Phi_i \rangle \langle X_{t+1}, \Phi_i \rangle} \\
&\quad \leq \tau_t d_\Phi \sum_{i=1}^{n} \frac{\eta_i \, \|X_t - V_t\|_* \, \|\Phi_i\|}{\langle X_t, \Phi_i \rangle \langle X_{t+1}, \Phi_i \rangle} \\
&\quad \leq \tau_t d_\Phi^2 d_{\mathcal{C}} \sum_{i=1}^{n} \frac{\eta_i}{\langle X_t, \Phi_i \rangle \langle X_{t+1}, \Phi_i \rangle}.
\end{aligned}$$

Since either $\eta_i = 0$ or $\eta_i \geq 1$, we have

$$\begin{aligned}
&\|\nabla g(X_{t+1}) - \nabla g(X_t)\| \\
&\quad \leq \tau_t d_\Phi^2 d_{\mathcal{C}} \sum_{i=1}^{n} \frac{\eta_i^2}{\langle X_t, \Phi_i \rangle \langle X_{t+1}, \Phi_i \rangle} \\
&\quad \leq \frac{\tau_t d_\Phi^2 d_{\mathcal{C}}}{1 - \tau_t} \sum_{i=1}^{n} \left(\frac{\eta_i}{\langle X_t, \Phi_i \rangle} \right)^2 \\
&\quad \leq \frac{\tau_t}{1 - \tau_t} d_\Phi^2 d_{\mathcal{C}} \left(\sum_{i=1}^{n} \frac{\eta_i}{\langle X_t, \Phi_i \rangle} \right)^2 \\
&\quad \leq \frac{\alpha_{t+1}}{S_t} d_\Phi^2 d_{\mathcal{C}} \left(\sum_{i=1}^{n} \frac{\eta_i}{\langle X_t, \Phi_i \rangle} \right)^2.
\end{aligned}$$

By Lemma 8.2,

$$\|\nabla g(X_{t+1}) - \nabla g(X_t)\| \leq \frac{\alpha_{t+1}}{S_t} d_\Phi^2 d_{\mathcal{C}} C^2.$$

Hence it suffices to choose $\gamma \geq C^2 d_\Phi^2 d_{\mathcal{C}}^2$.

8.6. Proof of Proposition 8.5

By Hölder's inequality, the first term in the definition of B_t can be bounded above by $\|\nabla g(X_0)\| \, d_{\mathcal{C}}$. The second term can be bounded as

$$\left(\sum_{k=1}^{t} \frac{\alpha_k^2}{S_{k-1}} \right) \gamma = \gamma \sum_{k=1}^{t} \left(2 + \frac{2}{k} \right) \leq 4t\gamma.$$

Then we obtain

$$\begin{aligned}
\frac{B_t}{S_t} &\leq \frac{8t\gamma}{(t+1)(t+2)} + \frac{2d_{\mathcal{C}} \, \|\nabla g(X_0)\|}{(t+1)(t+2)} \\
&< \frac{8\gamma}{t+2} + \frac{2d_{\mathcal{C}} \, \|\nabla g(X_0)\|}{(t+1)(t+2)} \\
&\leq \frac{8C^2 d_\Phi^2 d_{\mathcal{C}}^2}{t+2} + \frac{2d_{\mathcal{C}} \, \|\nabla g(X_0)\|}{(t+1)(t+2)}.
\end{aligned}$$

The definition of C in Lemma 8.2 also involves τ_t. We notice that choosing $T = 8n(\bar{\mu}/\underline{\mu})d_\eta - 1$ suffices to ensure $1 - 4n(\bar{\mu}/\underline{\mu})d_\eta \tau_T \geq 0$. Then we obtain

$$\begin{aligned}
\prod_{k=0}^{T} \frac{1}{1 - \tau_k} &= \frac{(T+2)(T+3)}{2} \\
&< \frac{(T+3)^2}{2} = 2 \left(\frac{4n\bar{\mu} d_\eta}{\underline{\mu}} + 1 \right)^2.
\end{aligned}$$

The quantity C_0 can be easily bounded as $C_0 \leq n\underline{\mu}^{-1} d_\eta$. Finally, we have

$$\frac{64n^2\bar{\mu}^2 d_\eta^2}{\underline{\mu}^3 \left(1 - \frac{4n\bar{\mu} d_\eta}{\underline{\mu}} \tau_T \right)} = \frac{64n^2\bar{\mu}^2 d_\eta^2}{\underline{\mu}^3} \left(\frac{4n\bar{\mu} d_\eta}{\underline{\mu}} + 1 \right).$$

9. REFERENCES

[1] A. Auslender and M. Teboulle, "Interior gradient and proximal methods for convex and conic optimization," *SIAM J. Optim.*, vol. 16, no. 3, pp. 697–725, 2006.

[2] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[3] S. R. Becker, E. J. Candès, and M. C. Grant, "Templates for convex cone problems with applications to sparse signal recovery," *Math. Prog. Comp.*, vol. 3, pp. 165–218, 2011.

[4] S. G. Bobkov and M. Ledoux, "On modified logarithmic Sobolev inequalities for Bernoulli and Poisson meausres," *J. Funct. Anal.*, vol. 156, pp. 347–365, 1998.

[5] M. Brand, "Fast low-rank modifications of the thin singular value decomposition," *Linear Algebra Appl.*, vol. 415, no. 1, pp. 20–30, 2006.

[6] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *C. R. Acad. Sci. Paris, Ser. I*, vol. 346, pp. 589–592, 2008.

[7] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM Rev.*, vol. 57, no. 2, pp. 225–251, 2015.

[8] E. J. Candès, X. Li, and M. Soltanolkotabi, "Phase retrieval from coded diffraction patterns," *Appl. Comput. Harmon. Anal.*, vol. 39, pp. 277–299, 2015.

[9] ——, "Phase retrieval via Wirtinger flow: Theory and algorithms," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1985–2007, 2015.

[10] E. J. Candès and Y. Plan, "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2342–2359, 2011.

[11] E. J. Candès, T. Strohmer, and V. Voroninski, "PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming," *Commun. Pure Appl. Math.*, vol. LXVI, pp. 1241–1274, 2013.

[12] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Model. Simul.*, vol. 4, no. 4, pp. 1168–1200, 2005.

[13] M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters, "1-bit matrix completion," *Inf. Inference*, vol. 3, pp. 189–223, 2014.

[14] J. Eckstein and W. Yao, "Understanding the convergence of the alternating direction method of multipliers: Theoretical and computational perspectives," 2015.

[15] J. R. Fienup, "Phase retrieval algorithms: a comparison," *Appl. Opt.*, vol. 21, no. 15, pp. 2758–2769, 1982.

[16] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*. Basel: Birkhäuser, 2013.

[17] R. M. Freund and P. Grigas, "New analysis and results for the Frank-Wolfe method," *Math. Program., Ser. A*, 2014.

[18] D. Garber and E. Hazan, "Faster rates for the Frank-Wolfe method over strongly-convex sets," in *Proc. 32nd Int. Conf. Machine Learning*, 2015.

[19] Z. Harchaoui, A. Juditsky, and A. Nemirovski, "Conditional gradient algorithms for norm-regularized smooth convex optimization," *Math. Program., Ser. A*, vol. 152, no. 1, pp. 75–112, 2015.

[20] Z. Hradil, "Quantum-state estimation," *Phys. Rev. A*, vol. 55, no. 3, 1997.

[21] M. Jaggi, "Revisiting Frank-Wolfe: Projection-free sparse convex optimization," in *Proc. 30th Int. Conf. Machine Learning*, 2013.

[22] I. Kontoyiannis and M. Madiman, "Measure concentration for compound Poisson distributions," *Electron. Commun. Probab.*, vol. 11, pp. 45–57, 2006.

[23] Y. Nesterov, "Gradient methods for minimizing composite functions," *Math. Program., Ser. B*, vol. 140, pp. 125–161, 2013.

[24] ——, "Complexity bounds for primal-dual methods minimizing the model of objective function," Center for Operations Research and Econometrics, CORE Discussion Paper, 2015.

[25] Y. Nesterov and A. Nemirovski, "On first-order algorithms for ℓ_1/nuclear norm minimization," *Acta Numer.*, pp. 509–575, 2013.

[26] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in *Adv. Neural Information Processing Systems 26*, 2013.

[27] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.

[28] M. Rudelson and R. Vershynin, "On sparse reconstruction from Fourier and Gaussian measurements," *Commun. Pure Appl. Math.*, vol. LXI, pp. 1025–1045, 2008.

[29] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, "Phase retrieval with application to optical imaging," *IEEE Sig. Process. Mag.*, vol. 32, no. 3, pp. 87–109, 2015.

[30] Q. Tran-Dinh, A. Kyrillidis, and V. Cevher, "Composite self-concordant minimization," *J. Mach. Learn. Res.*, vol. 16, pp. 371–416, 2015.

[31] A. Yurtsever, Y.-P. Hsieh, and V. Cevher, "Scalable convex methods for phase retrieval," in *6th IEEE Int. Workshop Computational Advances in Multi-Sensor Adaptive Processing*, 2015.

[32] A. Yurtsever, Q. Tran-Dinh, and V. Cevher, "A universal primal-dual convex optimization framework," in *29th Ann. Conf. Neural Information Processing Systems*, 2015.