

Frank-Wolfe Style Algorithms for Large Scale Optimization

Lijun Ding and Madeleine Udell

Abstract. We introduce a few variants on Frank-Wolfe style algorithms suitable for large scale optimization. We show how to modify the standard Frank-Wolfe algorithm using stochastic gradients, approximate subproblem solutions, and sketched decision variables in order to scale to enormous problems while preserving (up to constants) the optimal convergence rate $\mathcal{O}(\frac{1}{k})$.

Key words. Large scale optimization, Frank-Wolfe algorithm, stochastic gradient, low memory optimization, matrix completion.

AMS subject classifications. 90C06, 90C25.

1 Introduction

This chapter describes variants on Frank-Wolfe style algorithms suitable for large scale optimization. Frank-Wolfe style algorithms enforce constraints by solving a linear optimization problem over the constraint set at each iteration, while competing approaches, such as projected or proximal gradient algorithms, generally require projection onto the constraint set. For important classes of constraints, such as the unit norm ball of the ℓ_1 or nuclear norm, linear optimization over the constraint set is much faster than projection onto the set. This paper provides a gentle introduction to three ideas that can be used to further improve the performance of Frank-Wolfe style algorithms for large scale optimization: stochastic gradients, approximate subproblem solutions, and sketched decision variables. Using these ideas, we show how to

Lijun Ding

Operations Research and Information Engineering, Cornell University e-mail: ld446@cornell.edu

Madeleine Udell

Operations Research and Information Engineering, Cornell University e-mail: udell@cornell.edu

modify the standard Frank-Wolfe algorithm in order to scale to enormous problems while preserving (up to constants) the optimal convergence rate.

To understand the challenges of huge scale optimization, let us start by recalling the original Frank-Wolfe algorithm. The Frank-Wolfe algorithm is designed to solve problems of the form

$$\begin{aligned} & \text{minimize } f(x) \\ & \text{subject to } x \in \Omega, \end{aligned} \tag{1}$$

where f is a real valued convex differentiable function from \mathbf{R}^n to \mathbf{R} , and the set Ω is a nonempty compact convex set in \mathbf{R}^n . Throughout the sequel, we let $x^* \in \arg \min_{x \in \Omega} f(x)$ be an arbitrary solution to (1).

The Frank-Wolfe algorithm is presented as Algorithm 1 below. At each iteration, it computes the gradient of the objective $\nabla f(x)$ at the current iterate x , and finds a feasible point $v \in \Omega$ which maximizes $\nabla f(x)^T v$. The new iterate is taken to be a convex combination of the previous iterate and the point v .

Algorithm 1. Frank-Wolfe Algorithm

```

1: Input: Objective function  $f$  and feasible region  $\Omega$ 
2: Input: A feasible starting point  $x_{-1} \in \Omega$ 
3: Input: Stepsize sequence  $\gamma_k$  and tolerance level  $\varepsilon > 0$ 
4: for  $k = 0, 1, 2, \dots$  do
5:   Compute  $v_k = \arg \min_{v \in \Omega} \nabla f(x_{k-1})^T v$ .
6:   if  $(x_{k-1} - v_k)^T \nabla f(x_{k-1}) \leq \varepsilon$  then
7:     break
8:   end if
9:   Update  $x_k = (1 - \gamma_k)x_{k-1} + \gamma_k v_k$ .
10: end for
11: Output: The last iteration result  $x$ 

```

The Frank-Wolfe algorithm can be used for optimization with matrix variables as well. With some abuse of notation, when $x, \nabla f(x)$, and v are matrices rather than vectors, we use the inner product $\nabla f(x)^T v$ to denote the matrix trace inner product $\text{tr}(\nabla f(x)^T v)$.

Linear Optimization Subproblem. The main bottleneck in implementing Frank-Wolfe is solving the linear optimization subproblem in Line 5 above:

$$\begin{aligned} & \text{minimize } \nabla f(x_{k-1})^T v \\ & \text{subject to } v \in \Omega. \end{aligned} \tag{2}$$

Note that the objective of the subproblem (2) is linear even though the constraint set Ω may not be. Since Ω is compact, the solution to subproblem (2) always exists. Subproblem (2) can easily be solved when the feasible region has atomic structure [1]. We give three examples here.

- *The feasible region is a one norm ball.* For some $\alpha > 0$,

$$\Omega = \{x \in \mathbf{R}^n \mid \|x\|_1 \leq \alpha\}.$$

Let $\{e_i\}_{i=1}^n$ to be the standard basis in \mathbf{R}^n , $S = \operatorname{argmax}_i |\nabla f(x_{k-1})^T e_i|$ and $s_i = \mathbf{sign}(\nabla f(x_{k-1})^T e_i)$. The solution v to subproblem (2) is any vector in the convex hull of $\{-\alpha s_i e_i \mid i \in S\}$:

$$v \in \mathbf{conv}(\{-\alpha s_i e_i \mid i \in S\}).$$

In practice, we generally choose $v = -\alpha s_i e_i$ for some $i \in S$.

- *The feasible region is a nuclear norm ball.* For some $\alpha > 0$,

$$\Omega = \{X \in \mathbf{R}^{m \times n} \mid \|X\|_* \leq \alpha\},$$

where $\|\cdot\|_*$ is the nuclear norm, *i.e.*, the sum of the singular values. Here v , x_{k-1} , and $\nabla f(x_{k-1})$ are matrices in $\mathbf{R}^{m \times n}$, and we recall that the objective in Problem (2), $v^T \nabla f(x_{k-1})$, should be understood as the matrix trace inner product $\mathbf{tr}(\nabla f(x_{k-1})^T v)$. Subproblem (2) in this case is

$$\begin{aligned} & \text{minimize } \mathbf{tr}(\nabla f(x_{k-1})^T v) \\ & \text{subject to } \|v\|_* \leq \alpha. \end{aligned} \quad (3)$$

Denote the singular values of $\nabla f(x_{k-1})$ as $\sigma_1 \geq \dots \geq \sigma_{\min(m,n)}$ and the corresponding singular vectors as $(u_1, v_1), \dots, (u_{\min(m,n)}, v_{\min(m,n)})$. Let $S = \{i \mid \sigma_i = \sigma_1\}$ be the set of indices with maximal singular value. Then the solution to problem (2) is the convex hull of the singular vectors with maximal singular value, appropriately scaled:

$$\mathbf{conv}(\{-\alpha u_i v_i^T \mid i \in S\}).$$

In practice, we often take the solution $-\alpha u_1^T v_1$. This solution is easy to compute compared to the full singular value decomposition. Specifically, suppose $\nabla f(x_{k-1})$ is sparse, and let s be the number of non-zero entries in $\nabla f(x_{k-1})$. For any tolerance level $\varepsilon > 0$, the number of arithmetic operations required to compute the top singular tuple (u_1, v_1) using the Lanczos algorithm such that $u_1^T \nabla f(x_{k-1}) v_1 \geq \sigma_1 - \varepsilon$ is at most $\mathcal{O}(s \frac{\log(m+n)\sqrt{\sigma_1}}{\sqrt{\varepsilon}})$ with high probability [7].

- *The feasible region is a restriction of a nuclear norm ball.* For some $\alpha > 0$,

$$\Omega = \{X \in \mathbf{R}^{n \times n} \mid \|X\|_* \leq \alpha, X \succeq 0\},$$

where $X \succeq 0$ means X is symmetric and positive semidefinite, *i.e.*, every eigenvalue of X is nonnegative. In this case the objective in problem (2) $v^T \nabla f(x_{k-1})$ should be understood as $\mathbf{tr}(\nabla f(x_{k-1})^T v)$, where v , x_{k-1} , and $\nabla f(x_{k-1})$ are matrices in \mathbf{S}^n . The subproblem (2) in this case is just

$$\begin{aligned} & \text{minimize } \mathbf{tr}(\nabla f(x_{k-1})^T v) \\ & \text{subject to } \|v\|_* \leq \alpha \\ & \quad v \succeq 0. \end{aligned} \quad (4)$$

Denote the eigenvalues of $\nabla f(x_{k-1})$ as $\lambda_1 \geq \dots \geq \lambda_n$ and the corresponding eigenvectors as v_1, \dots, v_n . Let $S = \{i \mid \lambda_i = \lambda_n\}$ be the set of indices with smallest eigenvalue. Then the solution to Problem (2) is simply 0 if $\lambda_n \geq 0$, while if $\lambda_n \leq 0$, the solution set consists of the convex hull of the eigenvectors with smallest eigenvalue, appropriately scaled:

$$\mathbf{conv}(\{\alpha v_i v_i^T \mid i \in S\}).$$

In practice, we generally take $\alpha v_n v_n^T$ as a solution (if $\lambda_n \leq 0$). As in the previous case, this solution is easy to compute compared to the full eigenvalue decomposition. Specifically, suppose $\nabla f(x_{k-1})$ is sparse, and let s be the number of non-zero entries in $\nabla f(x_{k-1})$. For any tolerance level $\varepsilon > 0$, the number of arithmetic operations required to compute the eigenvector v_n using the Lanczos algorithm such that $v_1^T \nabla f(x_{k-1}) v_1 \leq \lambda_n + \varepsilon$ is at most $\mathcal{O}(s \frac{\log(2n) \sqrt{\max(|\lambda_1|, |\lambda_n|)}}{\sqrt{\varepsilon}})$ with high probability [3, Lemma 2].

Thus, after k iterations of the Frank-Wolfe algorithm, the sparsity or rank of the iterate x_k in the above three examples is bounded by k . This property has been noted and exploited by many authors [2, 3, 5].

The stopping criterion in Line 6 of Algorithm 1 bounds the suboptimality $f(x_{k-1}) - f(x^*)$, where $x^* \in \arg \min_{x \in \Omega} f(x)$. Indeed,

$$\begin{aligned} f(x_{k-1}) - f(x^*) &\leq (x_{k-1} - x^*)^T \nabla f(x_{k-1}) \\ &\leq (x_{k-1} - v_k)^T \nabla f(x_{k-1}), \end{aligned}$$

where the first inequality is due to convexity and the second line is due to optimality of v_k .

Matrix Completion. To illustrate our previous points, let's consider the example of matrix completion. Keep this example in mind: we will return to this problem again in the coming sections to illustrate our methods.

We consider the optimization problem

$$\begin{aligned} &\text{minimize } f(\mathcal{A}X) \\ &\text{subject to } \|X\|_* \leq \alpha, \\ &\quad X \in \mathcal{S} \end{aligned} \tag{5}$$

with variable $X \in \mathbf{R}^{m \times n}$. Here $\mathcal{A} : \mathbf{R}^{m \times n} \rightarrow \mathbf{R}^d$ is a linear map and $\alpha > 0$ is a positive constant. The set \mathcal{S} represents some additional information of the underlying problem. In this book chapter, the set \mathcal{S} will be either $\mathbf{R}^{m \times n}$ or $\{X \in \mathbf{R}^{m \times n} \mid X \succeq 0\}$. In the first case, the feasible region of Problem (5) is just the nuclear norm ball. In the second case, the feasible region is a restriction of the nuclear norm. In either case, the linear optimization subproblem can be solved efficiently as we just mentioned. The function $f : \mathbf{R}^d \rightarrow \mathbf{R}$ is a loss function that penalizes the misfit between the predictions $\mathcal{A}X$ of our model and our observations from the matrix.

For example, suppose we observe matrix entries c_{ij} with indices i, j in $\mathcal{O} \subset \{1, \dots, m\} \times \{1, \dots, n\}$ from a matrix $X^0 \in \mathbf{R}^{m \times n}$ corrupted by Gaussian noise E :

$$c_{ij} = (X^0)_{ij} + E_{ij}, \quad E_{ij} \stackrel{iid}{\sim} N(0, \sigma^2)$$

for some $\sigma > 0$. A maximum likelihood formulation of problem (5) to recover X^0 would be

$$\begin{aligned} & \text{minimize } \sum_{(i,j) \in \mathcal{O}} (x_{ij} - c_{ij})^2 \\ & \text{subject to } \|X\|_* \leq \alpha. \end{aligned} \quad (6)$$

To rewrite this problem in the form of (5), we choose \mathcal{A} so that $(\mathcal{A}X)_{ij} = x_{ij}$ for $(i, j) \in \mathcal{O}$, so the number of observations d is the cardinality of \mathcal{O} . Since there is no additional information of X^0 , we set $\mathcal{S} = \mathbf{R}^{n \times m}$. The objective f is a sum of quadratic losses in this case.

Since the constraint region Ω is a nuclear norm ball when $\mathcal{S} = \mathbf{R}^{m \times n}$, we can apply Frank-Wolfe to this optimization problem. The resulting algorithm is shown as Algorithm 2. Here Line 5 computes the singular vectors with largest singular value, and Line 9 exploits the fact that at each iteration we can choose a rank one update.

Algorithm 2. Frank-Wolfe Algorithm Applied to Matrix Completion with nuclear ball constraint only

```

1: Input: Objective function  $f$  and  $\alpha > 0$ 
2: Input: A feasible starting point  $\|X_{-1}\| \leq \alpha$ 
3: Input: Stepsize sequence  $\gamma_k$ 
4: for  $k = 0, 1, 2, \dots, K$  do
5:   Compute  $(u_k, v_k)$ , the top singular vectors of  $\nabla f(\mathcal{A}X_k)$ .
6:   if  $\text{tr}((X_{k-1} + \alpha u_k v_k^T)^T \nabla f(X_{k-1})) \leq \varepsilon$  then
7:     break the for loop.
8:   end if
9:   Update  $X_k = (1 - \gamma_k)X_{k-1} - \gamma_k \alpha u_k v_k^T$ .
10: end for
11: Output: The last iteration result  $X_K$ 

```

However, there are three main challenges in the large scale setting that can pose difficulties in applying the Frank-Wolfe algorithm:

1. Solving the linear optimization subproblem (2) exactly,
2. computing the gradient ∇f , and
3. storing the decision variable x .

To understand why each of these steps might present a difficulty, consider again the matrix completion case with nuclear ball constraint only.

1. Due to Galois theory, it is not possible to exactly compute the top singular vector, even in exact arithmetic. Instead, we rely on iterative methods such as the QR algorithm with shifts, or the Lanczos method, which terminate with some approximation error. What error can we allow in an approximate solution of the

linear optimization subproblem (2)? How will this error affect the performance of the Frank-Wolfe algorithm?

2. In many machine learning and statistics problems, the objective $f(X) = \sum_{i=1}^d f_i(X)$ is a sum over d observations, and each f_i measures the error in observation i . As we collect more data, computing ∇f exactly becomes more difficult, but approximating ∇f is generally easy. Can we use an approximate version of ∇f instead of the exact gradient?
3. Storing X , which requires $m \times n$ space in general, can be costly if n and m are large. One way to avoid using $\mathcal{O}(mn)$ memory is to store each updates (u_k, v_k) . But this approach still uses $\mathcal{O}(mn)$ memory when the number of iterations $K \geq \min(m, n)$. Can we exploit structure in the solution X^* to reduce the memory requirements?

We provide a crucial missing piece to address the first challenge, and a gentle introduction to the ideas needed to tackle the second and third challenges. Specifically, we will show the following.

1. Frank-Wolfe type algorithms still converge when we use an approximate oracle to solve the linear optimization subproblem (2). In fact, the convergence rate is preserved up to a multiplicative user-specified constant.
2. Frank-Wolfe type algorithms still converge when the gradient is replaced by an approximate gradient, and the convergence rate is preserved in expectation.
3. Frank-Wolfe type algorithms are amenable to a matrix sketching procedure which can be used to reduce memory requirements, and the convergence rate is not affected.

Based on these ideas, we propose two new Frank-Wolfe Style algorithms which we call SVRF with approximate oracle ($\widetilde{\text{SVRF}}$, pronounced as “tilde SVRF”), and Sketched SVRF ($\widetilde{\text{SSVRF}}$). They can easily scale to extremely large problems.

The rest of this chapter describes how $\widetilde{\text{SSVRF}}$ addresses the three challenges listed above. To address the first challenge, we augment the Frank-Wolfe algorithm with an approximate oracle for the linear optimization subproblem, and prove a convergence rate in this setting. Numerical experiments confirm that using an approximate oracle reduces the time necessary to achieve a given error tolerance. To address the second challenge, we then present a Stochastic Variance Reduced Frank-Wolfe (SVRF) algorithm with approximate oracle, $\widetilde{\text{SVRF}}$. Finally, we show how to use the matrix sketching procedure of [10] to reduce the memory requirements of the algorithm. We call the resulting algorithm $\widetilde{\text{SSVRF}}$.

Notation. We use $\|\cdot\|$ to denote the Euclidean norm when the norm is applied to a vector, and to denote the operator norm (maximum singular value) when applied to a matrix. We use $\|\cdot\|_F$ to denote the Frobenius norm and $\|\cdot\|_*$ to denote the nuclear norm (sum of singular values). The transpose of a matrix A and a vector v is denoted as A^T and v^T . The trace of a matrix $A \in \mathbf{R}^{n \times n}$ is the sum of all its diagonals, *i.e.*, $\text{tr}(A) = \sum_{i=1}^n A_{ii}$. The set of symmetric matrices in $\mathbf{R}^{n \times n}$ is denoted as \mathbf{S}^n . We use $X \succeq 0$ to mean that X is symmetric and positive semidefinite (psd). A

convex function $f : \mathbf{R}^n \rightarrow \mathbf{R}$ is L -smooth if $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ for some finite $L \geq 0$. The diameter D of a set $\Omega \subset \mathbf{R}^n$ is defined as $D = \sup_{x,y \in \Omega} \|x - y\|$. For an arbitrary matrix Z , we define $[Z]_r$ to be the best rank r approximation of Z in Frobenius norm. For a linear operator $\mathcal{A} : \mathbf{R}^{m \times n} \rightarrow \mathbf{R}^l$, where $\mathbf{R}^{m \times n}$ and \mathbf{R}^n are equipped with the trace inner product and the Euclidean inner product, the adjoint of \mathcal{A} is denoted as $\mathcal{A}^* : \mathbf{R}^l \rightarrow \mathbf{R}^{m \times n}$.

2 Frank-Wolfe with Approximate Oracle

In this section, we address the first challenge: the linear optimization subproblem (2) can only be solved approximately. Most of the ideas in this section are drawn from [5]; we include this introduction for the sake of completeness.

We will show that the Frank-Wolfe algorithm with approximate subproblem oracle converges at the same rate as the one with exact subproblem oracle up to a user-specified multiplicative constant.

2.1 Algorithm and convergence

As before, we seek to solve Problem (1),

$$\begin{aligned} & \text{minimize } f(x) \\ & \text{subject to } x \in \Omega. \end{aligned}$$

Let us introduce Algorithm 3, which we call Frank-Wolfe with approximate oracle. The only difference from the original Frank-Wolfe algorithm is the tolerance $\varepsilon_k > 0$: in Line 5, we compute an approximate solution with tolerance ε_k rather than an exact solution.

Algorithm 3. Frank-Wolfe with approximate oracle

- 1: **Input:** Objective function f and feasible region Ω
 - 2: **Input:** A feasible starting point $x_{-1} \in \Omega$
 - 3: **Input:** Stepsize sequence γ_k and tolerance level ε and error sequence $\varepsilon_k > 0$
 - 4: **for** $k = 0, 1, 2, \dots$ **do**
 - 5: Compute v_k such that $\nabla f(x_{k-1})^T v_k \leq \min_{v \in \Omega} \nabla f(x_{k-1})^T v + \varepsilon_k$.
 - 6: **if** $(x_{k-1} - v_k)^T \nabla f(x_{k-1}) \leq \varepsilon$ **then**
 - 7: break
 - 8: **end if**
 - 9: Update $x_k = (1 - \gamma_k)x_{k-1} + \gamma_k v_k$.
 - 10: **end for**
 - 11: **Output:** The last iteration result x_k
-

There are a few variants on this algorithm that use different line search methods. The next iterate might be the point on the line determined by x_{k-1} and v_k with lowest objective value, or the point with best objective value on the polytope with vertices x_{-1}, v_1, \dots, v_k . These variants may reduce the total number of iterations at the cost of

an increased per-iteration complexity. When memory is plentiful and line search or polytope search is easy to implement, these techniques can be employed; otherwise, a predetermined stepsize rule, *i.e.*, γ_k is determined as an input, *e.g.*, $\gamma_k = \frac{2}{k+2}$ or γ_k is a constant, might be preferred. All these techniques enjoy the same complexity bounds as Algorithm 3 since within an iteration, starting from the same iterate x_k , the objective is guaranteed to decrease at least as much under each of these line search rules as using the predetermined stepsize rule in Algorithm 3.

The following theorem gives a guarantee on the primal convergence of the objective value when f is L -smooth.

Theorem 1. *Given an arbitrary $\delta > 0$, if f is L -smooth, Ω has diameter D , $\gamma_k = \frac{2}{k+2}$ and $\varepsilon_k = \frac{LD^2}{2} \gamma_k \delta$, then the iterates x_k of Algorithm 3 satisfy*

$$f(x_k) - f(x^*) \leq \frac{2LD^2}{k+2} (1 + \delta) \quad (7)$$

where $x^* \in \arg \min_{x \in \Omega} f(x)$.

To start, recall an equivalent definition of L -smoothness [9, Theorem 2.1.5]. For completeness, we provide a short proof in the appendix.

Proposition 1. *If the real valued differentiable convex function f with domain \mathbf{R}^n is L -smooth, then for all $x, y \in \mathbf{R}^n$,*

$$f(x) \leq f(y) + \nabla f(y)^T (x - y) + \frac{L}{2} \|x - y\|^2.$$

Proof (Proof of Theorem 1).

Let $v_k^* \in \arg \min_{v \in \Omega} f(x_{k-1})^T v$ in Line 5. Using the update equation $x_k = x_{k-1} + \gamma_k (v_k - x_{k-1})$, we have

$$\begin{aligned} f(x_k) - f(x^*) &\leq f(x_{k-1}) - f(x^*) + \nabla f(x_{k-1})^T (v_k - x_{k-1}) \gamma_k + \frac{L}{2} \gamma_k^2 \|v_k - x_{k-1}\|^2 \\ &\leq f(x_{k-1}) - f(x^*) + \nabla f(x_{k-1})^T (v_k - x_{k-1}) \gamma_k + \frac{LD^2}{2} \gamma_k^2 \\ &\leq f(x_{k-1}) - f(x^*) + \nabla f(x_{k-1})^T (v_k^* - x_{k-1}) \gamma_k + \frac{LD^2}{2} \gamma_k^2 (1 + \delta) \\ &\leq f(x_{k-1}) - f(x^*) + \nabla f(x_{k-1})^T (x^* - x_{k-1}) \gamma_k + \frac{LD^2}{2} \gamma_k^2 (1 + \delta) \\ &\leq (1 - \gamma_k) (f(x_{k-1}) - f(x^*)) + \frac{LD^2}{2} \gamma_k^2 (1 + \delta). \end{aligned} \quad (8)$$

The first inequality is due to Proposition 1. The second inequality uses the diameter D of Ω and the fact that x_k is feasible since $\gamma_k \in (0, 1)$ and x_k is a convex combination of points in the convex set Ω . The third inequality uses the bound on the suboptimality of v_k in Line 5. The fourth inequality uses the optimality of v_k^* for $\min_{v \in \Omega} v^T \nabla f(x_{k-1})$ and fifth uses convexity of f . The conclusion of the above chain of inequalities is

$$f(x_k) - f(x^*) \leq (1 - \gamma_k) (f(x_{k-1}) - f(x^*)) + \frac{LD^2}{2} \gamma_k^2 (1 + \delta). \quad (9)$$

Now we prove inequality (7) by induction. The base case $k = 0$ follows from (8) since $\gamma_0 = 1$. Now suppose inequality (7) is true for $k \leq s$. Then for $k = s + 1$,

$$\begin{aligned}
f(x_{s+1}) - f(x^*) &\leq \left(1 - \frac{2}{s+2+1}\right)(f(x_s) - f(x^*)) + \frac{LD^2}{2} \left(\frac{2}{s+2+1}\right)^2 (1 + \delta) \\
&= \frac{s+1}{s+2+1}(f(x_s) - f(x^*)) + \frac{LD^2}{2} \left(\frac{2}{s+2+1}\right)^2 (1 + \delta) \\
&\leq \left(\frac{s+1}{s+2+1} \frac{2}{s+2} + \frac{2}{(s+2+1)^2}\right) LD^2 (1 + \delta) \\
&= \left(\frac{2s+2}{s+2} + \frac{2}{s+2+1}\right) \frac{LD^2}{s+2+1} (1 + \delta) \\
&\leq \left(\frac{2s+2+2}{s+2}\right) \frac{LD^2}{s+2+1} (1 + \delta) \\
&= \frac{2LD^2}{s+1+2} (1 + \delta).
\end{aligned} \tag{10}$$

We use (9) in the first inequality and the induction hypothesis in the second inequality to bound the term $f(x_s) - f(x^*)$. The last line completes the induction.

2.2 Numerics

In this subsection, we demonstrate that Frank-Wolfe is robust to using an approximate oracle through numerical experiments.

The specific problem we will use as our case study is the following symmetric matrix completion problem which is a special case of Problem (5). The symmetric matrix completion problem seeks to recover an underlying matrix $X^0 \succeq 0$ from a few noisy entries of X^0 . Specifically, let $C = X^0 + E$ be a matrix of noisy observations of X^0 , where E is a symmetric noise matrix. For each $i \geq j$, we observe C_{ij} independently with probability p . The quantity p is called the sample rate.

Let \mathcal{O} be the set of observed entries and m be the number of entries observed. Note that if $(i, j) \in \mathcal{O}$, $(j, i) \in \mathcal{O}$ as well since our matrices are all symmetric.

The optimization problem we solve to recover X^0 is

$$\begin{aligned}
&\text{minimize } f(X) := \frac{1}{2} \|P_{\mathcal{O}}(X) - P_{\mathcal{O}}(C)\|_F^2 \\
&\text{subject to } \|X\|_* \leq \alpha, \\
&\quad X \succeq 0.
\end{aligned} \tag{11}$$

Here the projection operator $P_{\mathcal{O}} : \mathbf{S}^n \rightarrow \mathbf{R}^m$ is

$$[P_{\mathcal{O}}(Y)]_{ij} = \begin{cases} Y_{ij}, & \text{if } (i, j) \in \mathcal{O} \\ 0, & \text{if } (i, j) \notin \mathcal{O}. \end{cases}$$

for any $Y \in \mathbf{S}^n$. By letting $\mathcal{A} = P_{\mathcal{O}}$, the set $\mathcal{S} = \{X \in \mathbf{R}^{n \times n} \mid X \succeq 0\}$ and $f(\cdot) = \|\cdot\|_F^2$, we see it is indeed a special case of Problem (5).

The gradient at X_k is $\nabla f(X_k) = P_{\mathcal{O}}(X_k) - P_{\mathcal{O}}(C)$. As we discussed in the introduction, a solution to the linear optimization subproblem is

$$V_k = \begin{cases} \alpha v_n v_n^T, & \text{if } \lambda_n(\nabla f(X_{k-1})) \leq 0 \\ 0, & \text{if } \lambda_n(\nabla f(X_{k-1})) > 0 \end{cases}$$

where $\lambda_n(\nabla f(X_{k-1}))$ is the smallest eigenvalue of $\nabla f(X_{k-1})$.

When the sample rate $p < 1$ is fixed, *i.e.*, independent of dimension n , the probability we observe all entries on the diagonal of C is very small. Hence the matrix

$\nabla f(X_{k-1})$ is very unlikely to be positive definite, for any k . (Recall that a positive definite matrix has positive diagonal.) Let us suppose that at least one entry on the diagonal is not observed, so that $\lambda_n(\nabla f(X_{k-1})) \leq 0$ for every k . Thus Line 5 of Algorithm 3 reduces to finding an approximate eigenvector v such that

$$\alpha v^T \nabla f(X_{k-1}) v \leq \alpha \lambda_n(\nabla f(X_{k-1})) + \varepsilon_k. \quad (12)$$

However, the solver ARPACK [8], which is the default solver for iterative eigenvalue problems in a variety of languages (*e.g.*, `eigs` in Matlab), does not support specifying the approximation error in the form of (12). Instead, for a given tolerance ξ_k , it finds an approximate vector $v \in \mathbf{R}^n$ with unit two norm, *i.e.*, $\|v\| = 1$, and an approximate eigenvalue $\lambda \in \mathbf{R}$, such that

$$\|\nabla f(X_{k-1})v - \lambda v\| \leq \xi_k \|\nabla f(X_{k-1})\|.$$

For simplicity, we assume that λ returned by our eigenvalue solver is the true smallest eigenvalue $\lambda_n(\nabla f(X_{k-1}))$, for any tolerance ξ_k . We will justify this assumption later through numerical experiments. In this case, the error ε_k is upper bounded by

$$\xi_k \alpha \|\nabla f(X_{k-1})\| \geq \varepsilon_k. \quad (13)$$

This upper bound turns out to be very conservative for large ξ_k : $\xi_k \alpha \|\nabla f(X_{k-1})\|$ might be much larger than the actual error $\varepsilon_k = \alpha v^T \nabla f(X_{k-1}) v - \alpha \lambda_n(\nabla f(X_{k-1}))$, as we will see later.

In the experiments, we set the dimension $n = 1000$ and generated $X^0 = WW^T$, where $W \in \mathbf{R}^{n \times r}$ had independent standard normal distributed entries. We then added symmetric noise $E = \frac{1}{10} \times (L + L^T)$ to X^0 to get $C = X^0 + E$, where $L \in \mathbf{R}^{n \times n}$ had independent standard normal entries. We then sampled uniformly from the upper triangular part of C (including the diagonal) with probability $p = 0.8$.

In each experiment we solved problem (11) with $\alpha = \|X^0\|_*$. In real applications, one usually does not know $\|X^0\|_*$ in advance. In that case, one might solve problem (11) multiple times with different values of α and select the best α according to some criterion.

We ran 9 experiments in total. In each experiment, we chose a rank r of X^0 in $\{10, 50, 100\}$ and ran Frank-Wolfe with approximate oracle with constant tolerance $\xi_k \in \{10^{-15}, 10^{-5}, 1\}$ using the step size rule $\gamma_k = \frac{2}{k+2}$, as required for Theorem 1, and terminated each experiment after 30 seconds; the qualitative performance of the algorithm is similar even after many more iterations. We emphasize that within an experiment, the tolerance ξ_k was the same for each iteration k . See the discussion above Figure 4 for more details about the choice of ξ_k .

Figure 1 shows experimental results on the relationship between the relative objective $\frac{\|P_{\mathcal{O}}(X_k) - P_{\mathcal{O}}(C)\|_F^2}{\|P_{\mathcal{O}}(C)\|_F^2}$ (on a log scale) and the actual clock time under different combinations of rank r and tolerance ξ_k . For a fixed rank, the relative objective $\frac{\|P_{\mathcal{O}}(X_k) - P_{\mathcal{O}}(C)\|_F^2}{\|P_{\mathcal{O}}(C)\|_F^2}$ evolves similarly for any tolerance. When the underlying matrix has relatively high rank, using a lower tolerance allows faster convergence, at least

for the moderate final relative objective achieved in these experiments. The per iteration cost is summarized in Table 1. In fact, these plots show no advantage to using a tighter tolerance in any setting.

Table 1. Average per iteration time (seconds) of Algorithm (3) for problem (11).

	Rank(X^0) = 10	Rank(X^0) = 50	Rank(X^0) = 100
$\xi_k = 10^{-15}$	0.1136	0.1923	0.2400
$\xi_k = 10^{-5}$	0.0997	0.1376	0.1840
$\xi_k = 1$	0.1017	0.1099	0.1220

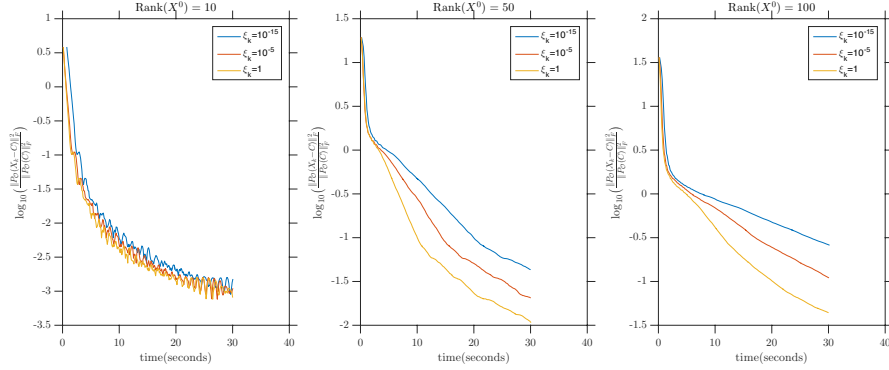


Figure 1. The above plots demonstrate the relation between the relative objective value $\log\left(\frac{\|P_C(X_k) - P_C(C)\|_F^2}{\|P_C(C)\|_F^2}\right)$ and the clock time for different combinations of rank $r = \text{Rank}(X^0)$ and tolerance parameters ξ_k .

One surprising feature of these graphs is the oscillation of relative error that occurs for the model with $r = 10$ once the relative error has reached 10^{-2} or so. This oscillation as the algorithm approaches the optimum is due to the stepsize rule $\gamma_k = \frac{2}{k+2}$. To see how this stepsize leads to oscillation, suppose for simplicity that for some iterate k_0 , $X_{k_0-1} = X^0$. We expect this iterate to have a very low objective value; indeed, in our experiments we found that the relative objective at X^0 is around 5×10^{-4} when $r = 10$. Then in the next iteration, we add V_{k_0} to X_{k_0-1} with step size $\frac{2}{k_0+2}$. Hence X_{k_0} is at least $\frac{2}{k_0+2}\alpha$ away from the true solution. This very likely will increase the relative objective since our p is 0.8. Suppose further that $V_{k_0+1} = -V_{k_0}$. Then we almost return to X^0 in the next iteration and again enjoy a small relative objective. For higher rank X^* , the oscillation begins at later iterations (not shown), as the algorithm approaches the solution.

Using line search eliminates the oscillation, but increases computation time for this problem. We do not consider linesearch further in this paper.

Our goal in this problem is not simply to find the solution of Problem (11) but to produce a matrix X close to X^0 . Hence we also study the numerical convergence of the relative error $\|X - X^0\|_F^2 / \|X^0\|_F^2$. Figure 2 shows experimental results on the relationship between the relative error $\frac{\|X_k - X^0\|_F^2}{\|X^0\|_F^2}$ (on a log scale) and the actual

clock time under different combinations of rank r and tolerance ξ_k . The evolution of $\frac{\|X_k - X^0\|_F^2}{\|X^0\|_F^2}$ is very similar to the evolution of $\frac{\|P_{\mathcal{O}}(X_k) - P_{\mathcal{O}}(C)\|_F^2}{\|P_{\mathcal{O}}(C)\|_F^2}$ in Figure 1.

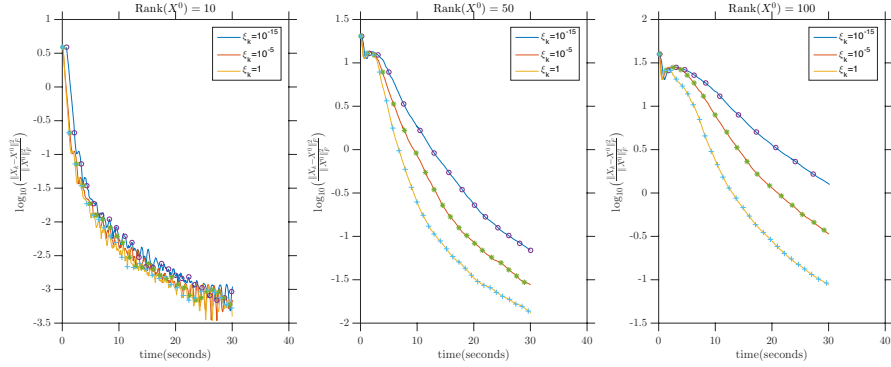


Figure 2. The above plots demonstrate the relation between the relative distance to the solution $\log(\frac{\|X_k - X^0\|_F^2}{\|X^0\|_F^2})$ and the clock time for different combinations of rank $r = \text{Rank}(X^0)$ and tolerance parameters ξ_k . We plot a marker on the line once every ten iterations (in this figure only).

The assumption that the approximate eigenvalue λ returned by the eigenvalue solver is approximately equal to the true smallest eigenvalue λ_n (Equation (12)) is supported by Figure 3. We computed the true eigenvalue λ_n by calling ARPACK with a very tight tolerance. It is interesting that for a low rank model, the estimate λ is very accurate even if ξ_k is large. The relative error in λ is about 10^{-2} on average when $\xi_k = 1$ for high rank models. However, this is not too large: the relative error in our iterate $\frac{\|X_k - X^0\|_F^2}{\|X^0\|_F^2}$ is also about 10^{-2} , hence these two errors are on the same scale.

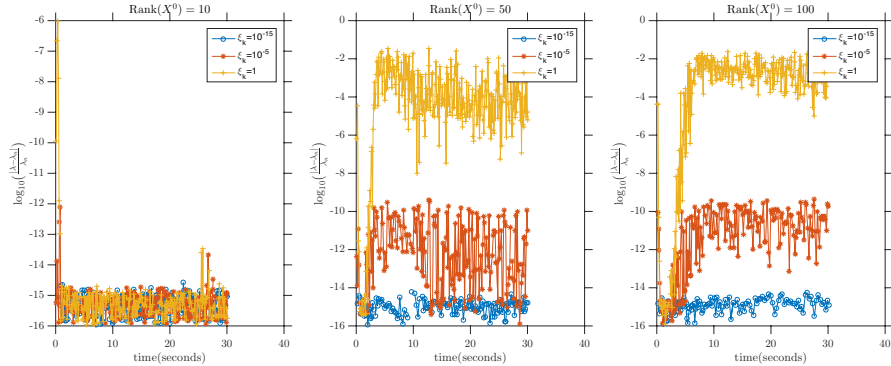


Figure 3. The vertical axis is the relative difference between approximate eigenvalue λ with $\xi_k = \{10^{-15}, 10^{-5}, 1\}$ and the very accurate eigenvalue λ_n of $\lambda_n(\nabla f(X_{k-1}))$, computed with tolerance ξ_k equal to machine precision 10^{-16} .

Figure 4 shows the error $\varepsilon_k = v^T \nabla f(X_{k-1})v - \alpha \lambda_n(\nabla f(X_{k-1}))$ achieved by our linear optimization subproblem solver. It can be seen that for a constant tolerance ξ_k , the error ε_k is also almost constant after some initial transient behavior. Hence controlling ξ_k indeed controls ε_k .

Since our ε_k is approximately constant due to constant choice of ξ_k , rather than decreasing as required by the assumptions of Theorem 1, one might wonder whether the conclusion of Theorem 1 still holds. The answer is yes. In fact, we found that at each iteration throughout our numerical experiments, the inequality $\varepsilon_k = \alpha v^T \nabla f(X_{k-1})v - \alpha \lambda_n(\nabla f(X_{k-1})) \leq \gamma_k L D^2 \delta$ is satisfied, with $\delta = 1$ and $D = 2\|X^0\|_*$. Thus the conclusion of Theorem 1 is still satisfied in our numerical results, although we do not have explicit control over the error ε_k .

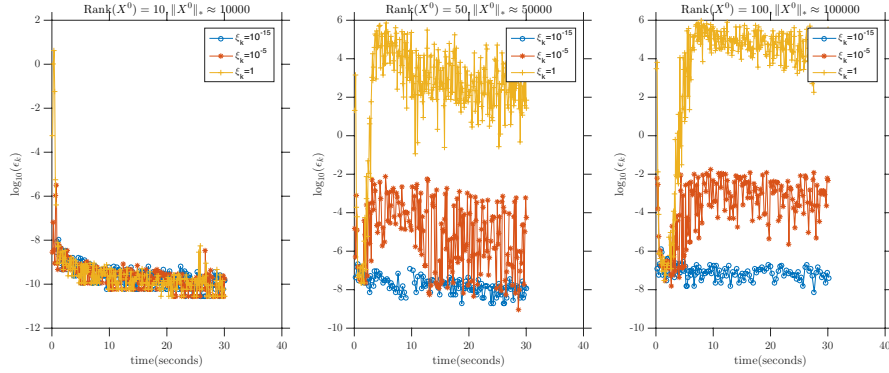


Figure 4. The actual evolution of the error $\varepsilon_k = \alpha v^T \nabla f(X_{k-1})v - \alpha \lambda_n(\nabla f(X_{k-1}))$.

We examine the accuracy of our bound $\xi_k \alpha \|\nabla f(X_{k-1})\|$ on ε_k in Figure 5. It shows that our bound is rather conservative for higher value of ξ_k .

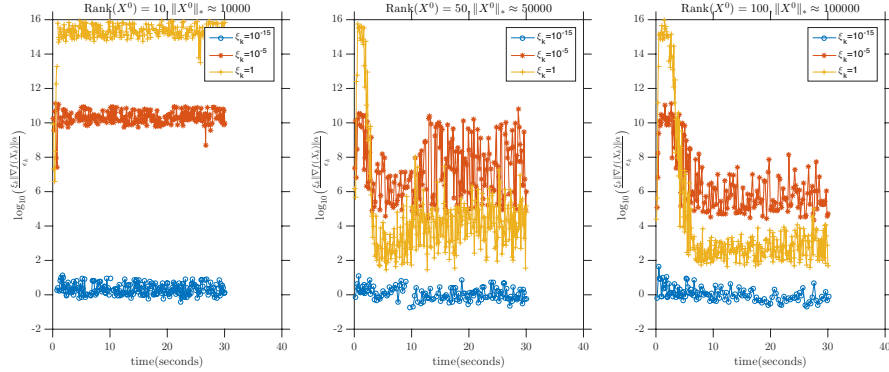


Figure 5. Empirical evolution of the ratio $\frac{\xi_k \alpha \|\nabla f(X_{k-1})\|}{\varepsilon_k}$. Recall that $\xi_k \alpha \|\nabla f(X_{k-1})\|$ is an upper bound of ε_k .

3 Stochastic Variance Reduced Frank-Wolfe (SVRF) algorithm with Approximate Oracle (SVRF)

Having seen that Frank-Wolfe is robust to using an approximate oracle when solving linear optimization subproblem (2), we now turn to our second challenge: computing the gradient ∇f .

To formalize the challenge, we will consider the optimization problem

$$\begin{aligned} & \text{minimize } f(x) := \frac{1}{n} \sum_{i=1}^n f_i(x) \\ & \text{subject to } x \in \Omega, \end{aligned} \quad (14)$$

where $x \in \mathbf{R}^m$. For each $i = 1, \dots, n$, f_i is a convex continuously differentiable real valued function and Ω is a compact convex set in \mathbf{R}^m . This is a particular instance of Problem (1).

Problem (14) is common in statistics and machine learning, where each f_i measures the error in observation i . Computing the gradient ∇f in this setting is a challenge, since the number of observations n can be enormous.

One way to address this challenge is to compute an approximation to the gradient rather than the exact gradient. We sample l elements i_1, \dots, i_l from the set $\{1, \dots, n\}$ with replacement and compute the *stochastic gradient*

$$\tilde{\nabla} f(x) = \frac{1}{l} \sum_{j=1}^l \nabla f_{i_j}(x).$$

The parameter l is called the size of the minibatch $\{i_1, \dots, i_l\}$. The computational benefit here is that we compute only $l \ll n$ derivatives. Intuitively, we expect this method to work since $\mathbf{E}[\tilde{\nabla} f(x)] = \nabla f(x)$.

Of course, the computational benefit does not come for free. This approach suffers one major drawback:

- the stochastic gradient $\tilde{\nabla} f(x)$ may have very large variance $\mathbf{var}(\|\tilde{\nabla} f(x)\|_2)$ even if x is near x^* . Large variance will destabilize any algorithm using $\tilde{\nabla} f(x)$, since even near the solution where $\|\nabla f(x)\|$ is small, $\|\tilde{\nabla} f(x)\|$ may be large.

One simple way to ensure that $\tilde{\nabla} f(x)$ concentrates near $\nabla f(x)$ is to increase the minibatch size l as $\mathbf{var}(\tilde{\nabla} f(x)) = \frac{1}{l} \mathbf{var}(\nabla f_i(x))$, where i is chosen uniformly from $\{1, \dots, n\}$. But using a very large minibatch size l defeats the purpose of using a stochastic gradient.

Variance reduction techniques endeavor to avoid this tradeoff [6]. Instead of using a large minibatch at each iteration, they occasionally compute a full gradient and use it to reduce the variance of $\tilde{\nabla} f(x)$. The modified stochastic gradient is called the variance-reduced stochastic gradient. Johnson and Zhang [6] introduced one way to perform variance reduction. Specifically, they define a variance-reduced stochastic gradient at a point $x \in \Omega$ with respect to some *snapshot* $x_0 \in \Omega$ as

$$\tilde{\nabla} f(x; x_0) = \nabla f_i(x) - (\nabla f_i(x_0) - \nabla f(x_0)),$$

where i is sampled uniformly from $\{1, \dots, n\}$. Notice we require the full gradient $\nabla f(x_0)$ at the snapshot, but only the gradient of the i th function $\nabla f_i(x)$ at the point x . In this case, we still have $\mathbf{E} \tilde{\nabla} f(x; x_0) = \nabla f(x)$, and the variance is

$$\mathbf{var}(\|\tilde{\nabla} f(x; x_0)\|_2) = \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x) - \nabla f_i(x_0) + (\nabla f(x_0) - \nabla f(x))\|_2^2.$$

If x and x_0 are near x^* , the variance will be near zero and so indeed the variance is reduced. We can further reduce the variance using a minibatch by independently sampling l variance-reduced gradients $\tilde{\nabla} f(x; x_0)$ and taking their average.

Hazan and Luo [4, Theorem 1] introduced the stochastic variance reduced Frank-Wolfe (SVRF) algorithm, which augments the Frank-Wolfe algorithm with the variance reduction technique of Johnson and Zhang, and showed that it converges in expectation when an exact oracle is used for the linear optimization subproblem (2). As we will see in Theorem 2, the number of evaluation of full gradient and stochastic gradient is also considerably small.

As we saw in the previous section, Frank-Wolfe with an approximate oracle converges at the same rate as the one using an exact oracle for the linear optimization subproblem (2). One naturally wonders whether an approximate oracle is allowed when we use stochastic gradients. We will show below that the resulting algorithm, which we call SVRF with approximate oracle ($\widetilde{\text{SVRF}}$, pronounced as ‘‘tilde SVRF’’) and present as Algorithm 4, indeed works well. Note that when $\varepsilon_k = 0$ for each k , Algorithm 4 reduces to SVRF.

Algorithm 4. SVRF with approximate oracle ($\widetilde{\text{SVRF}}$)

- 1: **Input:** Objective function $f = \frac{1}{n} \sum_{i=1}^n f_i$
 - 2: **Input:** A feasible starting point $w_{-1} \in \Omega$
 - 3: **Input:** Stepsize γ_k , minibatch size m_k , epoch length N_t and tolerance sequence ε_k
 - 4: **Initialize:** Find x_0 s.t. $\nabla f(w_{-1})^T x_0 \leq \min_{x \in \Omega} \nabla f(w_{-1})^T x + \varepsilon_0$.
 - 5: **for** $t = 1, 2, \dots, T$ **do**
 - 6: Take a snapshot $w_0 = x_{t-1}$ and compute gradient $\nabla f(w_0)$.
 - 7: **for** $k = 1$ to N_t **do**
 - 8: Compute g_k , the average of m_k iid samples of $\tilde{\nabla} f(w_{k-1}, w_0)$.
 - 9: Compute v_k s.t. $g_k^T v_k \leq \min_{v \in \Omega} g_k^T v + \varepsilon_k$.
 - 10: Update $w_k := (1 - \gamma_k)w_{k-1} + \gamma_k v_k$.
 - 11: **end for**
 - 12: Set $x_t = w_{N_t}$.
 - 13: **end for**
 - 14: **Output:** The last iteration result x_T .
-

We give a quantitative description of the objective value convergence $f(x_k) - f(x^*)$ for Algorithm 4 in Theorem 2. Moreover, we show that the convergence rate is the same as the one using the exact subproblem oracle up to a multiplicative user-specified constant.

In Algorithm 4, each time we take a snapshot, we let $k = 1$ again and the algorithm essentially restarts. Another option available is not to restart k . This modification is suggested and implemented in [4]; further, they observe this algorithmic variant is more stable. This modification ensures that the stepsize always decreases, and so intuitively should increase the stability.

We state this modification as Algorithm 5 below. We show it converges in expectation with the same rate as Algorithm 4, and that it converges almost surely. These results are new to the best of our knowledge, and theoretically justify why a diminishing stepsize makes the algorithm more stable: the optimality gap converges almost surely to 0 rather than just in expectation!

Algorithm 5. Stable $\widetilde{\text{SVRF}}$, k increasing in line 7 of Algorithm 4

- 1: ... as Algorithm 4, except replacing the chunk from line 7 to line 12 with the following chunk and start k at $k = 1$ when $t = 1$.
 - 2: **while** $k \leq N_t$ **do**
 - 3: Compute g_k , the average of m_k iid samples of $\tilde{\nabla}f(w_{k-1}, w_0)$.
 - 4: Compute v_k s.t. $g_k^\top v_k \leq \min_{v \in \Omega} g_k^\top v + \varepsilon_k$.
 - 5: Update $w_k := (1 - \gamma_k)w_{k-1} + \gamma_k v_k$ and $k = k + 1$.
 - 6: **end while**
 - 7: Set $x_t = w_{N_t}$.
-

4 Theoretical guarantees for $\widetilde{\text{SVRF}}$

We show below that $\widetilde{\text{SVRF}}$ has the same convergence rate as SVRF, up to constants depending on the error level δ . The proof is analogous to the one in Hazan and Luo [4, Theorem 1], with some additional care in handling the error term.

Theorem 2. *Suppose each f_i is L -smooth and Ω has diameter D . Then for any $\delta > 0$, Algorithms 4 and 5 with parameters*

$$\gamma_k = \frac{2}{k+1}, \quad m_k = 96(k+1), \quad N_t = 2^{t+3} - 2, \quad \varepsilon_k = \frac{LD^2}{2} \gamma_k \delta$$

ensure that for any t ,

$$\mathbf{E}[f(x_t) - f(x^*)] \leq \frac{LD^2(1+\delta)}{2^{t+1}}.$$

Moreover, for any k ,

$$\mathbf{E}[f(w_k) - f(x^*)] \leq \frac{4LD^2(1+\delta)}{k+2}.$$

One might be concerned that $\widetilde{\text{SVRF}}$ is impractical, since the minibatch size required to compute the approximate gradient increases linearly with k . However,

when the number of terms n in the objective is sufficiently large, in fact the complexity of $\widetilde{\text{SVRF}}$ is lower than that of Algorithm 3, Frank-Wolfe with approximate oracle. Under the parameter settings in Theorem 2, with a bit extra work, we see that $\widetilde{\text{SVRF}}$ requires $\mathcal{O}(\ln(\frac{LD^2(1+\delta)}{\varepsilon}))$ full gradient evaluations, $\mathcal{O}(\frac{L^2D^4(1+\delta)^2}{\varepsilon^2})$ stochastic gradient evaluations, and the solution of $\mathcal{O}(\frac{LD^2(1+\delta)}{\varepsilon})$ linear optimization subproblems. As a comparison, Algorithm 3, Frank-Wolfe with approximate oracle, under the parameter settings in Theorem 1, requires $\mathcal{O}(\frac{LD^2(1+\delta)}{\varepsilon})$ full gradient evaluations and the solution of the same number of linear optimization subproblems. Suppose that the cost of computing the full gradient is n times the cost of computing one stochastic gradient. Then $\widetilde{\text{SVRF}}$ enjoys a smaller computational cost than Algorithm 3 if

$$\mathcal{O}\left(\ln\left(\frac{LD^2(1+\delta)}{\varepsilon}\right)\right) + \frac{1}{n}\mathcal{O}\left(\frac{L^2D^4(1+\delta)^2}{\varepsilon^2}\right) < \mathcal{O}\left(\frac{LD^2(1+\delta)}{\varepsilon}\right),$$

which is satisfied for large n .

We begin the proof using the smoothness of f_i [9, Theorem 2.1.5].

Proposition 2. *Suppose a real valued function g is convex and L -smooth over its domain \mathbf{R}^n . Then g satisfies*

$$\|\nabla g(w) - \nabla g(v)\|^2 \leq 2L(g(w) - g(v) - \nabla g(v)^T(w - v))$$

for all $w, v \in \mathbf{R}^n$.

Proof. Consider $h(w) = g(w) - \nabla g(v)^T w$, which is also convex and L -smooth. The minimum of $h(w)$ occurs at $w = v$, since $\nabla h(v) = 0$. Hence

$$\begin{aligned} h(v) - h(w) &\leq h\left(w - \frac{1}{L}\nabla h(w)\right) - h(w) \\ &\leq -(\nabla g(w) - \nabla g(v))^T \left(\frac{1}{L}(\nabla g(w) - \nabla g(v))\right) + \frac{L}{2} \frac{1}{L^2} \|\nabla g(w) - \nabla g(v)\|^2 \\ &\leq -\frac{1}{2L} \|\nabla g(w) - \nabla g(v)\|^2 \end{aligned} \tag{15}$$

where the second inequality is due to the smoothness of h . Substitute $h(w) = g(w) - \nabla g(v)^T w$ back into the above inequality gives Proposition 2.

The second ingredient of the proof is bounding the variance of the reduced variance gradient $\tilde{\nabla}f(x_0, x)$ in terms of the difference between the current value and the optimal function value. Note that $\tilde{\nabla}f(x_0, x)$ is an unbiased estimator of $\nabla f(x)$. The proof relies on Proposition 2 and can be found in Hazan and Luo [4, Lemma 1].

Lemma 1. *For any $x, x_0 \in \Omega$, we have*

$$\mathbf{E}[\|\tilde{\nabla}f(x; x_0) - \nabla f(x)\|^2] \leq 6L(2\mathbf{E}[f(x) - f(x^*)] + \mathbf{E}[f(x_0) - f(x^*)]).$$

Proof.

$$\begin{aligned}
\mathbf{E}[\|\tilde{\nabla}f(x; x_0) - \nabla f(x)\|^2] &= \mathbf{E}[\|\nabla f_i(x) - \nabla f_i(x_0) + \nabla f(x_0) - \nabla f(x)\|^2] \\
&= \mathbf{E}[\|(\nabla f_i(x) - \nabla f_i(x^*)) - (\nabla f_i(x_0) - \nabla f_i(x^*)) \\
&\quad + (\nabla f(x_0) - \nabla f(x^*)) - (\nabla f(x) - \nabla f(x^*))\|^2] \\
&\leq 3\mathbf{E}[\|\nabla f_i(x) - \nabla f_i(x^*)\|^2 + \|(\nabla f_i(x) - \nabla f_i(x^*)) \\
&\quad - (\nabla f(x_0) - \nabla f(x^*))\|^2 + \|\nabla f(x) - \nabla f(x^*)\|^2] \\
&\leq 3\mathbf{E}[\|\nabla f_i(x) - \nabla f_i(x^*)\|^2 + \|\nabla f_i(x_0) - \nabla f_i(x^*)\|^2 \\
&\quad + \|\nabla f(x) - \nabla f(x^*)\|^2]
\end{aligned} \tag{16}$$

where the first inequality is due to Cauchy-Schwarz and the fact that $2ab \leq a^2 + b^2$ for any $a, b \in \mathbf{R}$. The second inequality is the variance $\mathbf{E}[\|(\nabla f_i(x) - \nabla f_i(x^*)) - (\nabla f(x_0) - \nabla f(x^*))\|^2]$ is less than its second moment $\mathbf{E}[\|(\nabla f_i(x) - \nabla f_i(x^*))\|^2]$.

Now we apply Proposition 2 to the three terms above. For example, for the first term, we have

$$\begin{aligned}
\mathbf{E}[\|\nabla f_i(x) - \nabla f_i(x^*)\|^2] &\leq 2L\mathbf{E}[f_i(x) - f_i(x^*) - \nabla f_i(x^*)^T(w - w^*)] \\
&= 2L(f(x) - f(x^*) - \nabla f(x^*)^T(x - x^*)) \\
&\leq 2L(f(x) - f(x^*))
\end{aligned} \tag{17}$$

where the second inequality is due to the optimality of x^* . Applying the proposition similarly to other two terms yields the lemma.

The key to the proof of Theorem 2 is the following lemma.

Lemma 2. *For any t and k in Algorithm 4 and 5, we have*

$$\mathbf{E}[f(w_k) - f(x^*)] \leq \frac{4LD^2(1+\delta)}{k+2}$$

if

$$\mathbf{E}[\|g_s - \nabla f(w_{s-1})\|^2] \leq \frac{L^2D^2(1+\delta)^2}{(s+1)^2}$$

for all $s \leq k$.

Proof. The L -smoothness of f gives that for any $s \leq k$,

$$f(w_s) \leq f(w_{s-1}) + \nabla f(w_{s-1})^T(w_s - w_{s-1}) + \frac{L}{2}\|w_s - w_{s-1}\|^2.$$

Under Algorithm 4 or 5, we have $w_s = (1 - \gamma_s)w_{s-1} + \gamma_s v_s$. Plugging this in the above inequality gives

$$f(w_s) \leq f(w_{s-1}) + \gamma_s \nabla f(w_{s-1})^T(v_s - w_{s-1}) + \frac{L\gamma_s^2}{2}\|v_s - w_{s-1}\|^2.$$

Using the definition of the diameter of Ω , we can rearrange the previous inequality as

$$f(w_s) \leq f(w_{s-1}) + \gamma_s g_s^T(v_s - w_{s-1}) + \gamma_s (\nabla f(w_{s-1}) - g_s)^T(v_s - w_{s-1}) + \frac{LD^2\gamma_s^2}{2}.$$

Since $g_s^T v_s \leq \min_{w \in \Omega} g_s^T w + \frac{\gamma_s \delta L D^2}{2} \leq g_s^T x^* + \frac{\gamma_s \delta L D^2}{2}$, we arrive at

$$\begin{aligned} f(w_s) &\leq f(w_{s-1}) + \gamma_s \nabla f(w_{s-1})^T (x^* - w_{s-1}) \\ &\quad + \gamma_s (\nabla f(w_{s-1}) - g_s)^T (v_s - x^*) + \frac{L D^2 (1 + \delta) \gamma_s^2}{2}. \end{aligned} \quad (18)$$

By convexity, the term $\nabla f(w_{s-1})^T (x^* - w_{s-1})$ is upper bounded by $f(x^*) - f(w_{s-1})$, and Cauchy-Schwarz inequality yields that

$$|(\nabla f(w_{s-1}) - g_s)^T (v_s - x^*)| \leq D \|g_s - \nabla f(w_{s-1})\|.$$

The assumption on $\|g_s - \nabla f(w_{s-1})\|^2$ gives $\mathbf{E}[\|g_s - \nabla f(w_{s-1})\|]$ is at most $\frac{L D (1 + \delta)}{s+1}$ by Jensen's inequality. Recalling $\gamma_s = \frac{2}{s+1}$, we have

$$\begin{aligned} &\mathbf{E}[f(w_s) - f(x^*)] \\ &\leq (1 - \gamma_s) \mathbf{E}[f(w_{s-1}) - f(x^*)] + \frac{L D^2 \gamma_s^2 (1 + \delta)}{2} + \frac{L D^2 \gamma_s^2 (1 + \delta)}{2} \\ &= (1 - \gamma_s) \mathbf{E}[f(w_{s-1}) - f(x^*)] + L D^2 \gamma_s^2 (1 + \delta). \end{aligned}$$

We now prove $\mathbf{E}[f(w_k) - f(x^*)] \leq \frac{4L D^2 (1 + \delta)}{k+2}$ by induction. The base case $k = 1$ is simple by noting $\gamma_1 = 1$ and

$$\mathbf{E}[f(w_1) - f(x^*)] \leq (1 - \gamma_1) \mathbf{E}[f(w_0) - f(x^*)] + \gamma_1 L D^2 (1 + \delta) = L D^2 (1 + \delta).$$

Now suppose for $k = s - 1$, $\mathbf{E}[f(w_{s-1}) - f(x^*)] \leq \frac{4L D^2 (1 + \delta)}{s+1}$. Then with $\gamma_s = \frac{2}{s+1}$, we have for $k = s$

$$\mathbf{E}[f(w_s) - f(x^*)] \leq \frac{4L D^2 (1 + \delta)}{s+1} \left(1 - \frac{2}{s+1} + \frac{1}{s+1}\right) \leq \frac{4L D^2 (1 + \delta)}{s+2},$$

which completes the induction.

With this lemma, we are able to prove Theorem 2

Proof (Proof of Theorem 2). We proceed by induction. In the base case $t = 0$, we have

$$\begin{aligned} f(x_0) &\leq f(w_{-1}) + \nabla f(w_{-1})^T (x_0 - w_{-1}) + \frac{L}{2} \|w_{-1} - x_0\|^2 \\ &\leq f(w_{-1}) + \nabla f(w_{-1})^T (x^* - w_{-1}) + \frac{L D^2}{2} + \frac{L D^2 \delta}{2} \\ &\leq f(x^*) + \frac{L D^2 (1 + \delta)}{2}, \end{aligned}$$

where we use the L -smoothness in the first inequality, the near optimality of x_0 in the second inequality and convexity of f in the last inequality.

Now we assume that $\mathbf{E}[f(x_{t-1}) - f(x^*)] \leq \frac{LD^2(1+\delta)}{2^t}$ and we are in Algorithm 4. We consider iteration of the algorithm and use another induction to show $\mathbf{E}[f(w_k) - f(x^*)] \leq \frac{4LD^2(1+\delta)}{k+1}$ for any $k \leq N_t$. The base case $w_0 = x_{t-1}$ is clearly satisfied because of the induction hypothesis $\mathbf{E}[f(x_{t-1}) - f(x^*)] \leq \frac{LD^2(1+\delta)}{2^t}$. Given the induction hypothesis $\mathbf{E}[f(w_{s-1}) - f(x^*)] \leq \frac{4LD^2(1+\delta)}{s+1}$ for any $s \leq k$, we have

$$\begin{aligned} & \mathbf{E}[\|g_s - \nabla f(w_{s-1})\|^2] \\ & \leq \frac{6L}{m_s} (2\mathbf{E}[f(w_{s-1}) - f(x^*)] + \mathbf{E}[f(w_0) - f(x^*)]) \\ & \leq \frac{6L}{m_s} \left(\frac{8LD^2(1+\delta)}{s+1} + \frac{LD^2(1+\delta)}{2^t} \right) \\ & \leq \frac{6L}{m_s} \left(\frac{8LD^2(1+\delta)}{s+1} + \frac{8LD^2(1+\delta)}{s+1} \right) \\ & = \frac{L^2D^2(1+\delta)}{(s+1)^2} \leq \frac{L^2D^2(1+\delta)^2}{(s+1)^2} \end{aligned}$$

where the first inequality use Lemma 1 and the fact that variance reduced by a factor m_s as g_s is the average of m_s iid samples of $\tilde{\nabla}f(w_{s-1}; w_0)$ and the second and third inequality are due to the two induction hypothesis and $s \leq N_t = 2^{t+3} - 2$. The last equality is due to the choice of m_s . Therefore, we see the condition of Lemma 2 is satisfied and the induction is completed.

Now suppose we are in the situation of Algorithm 5. The only difference here is that we don't restart k at 1. Assuming that $\mathbf{E}[f(x_{s-1}) - f(x^*)] \leq \frac{LD^2(1+\delta)}{2^t}$ for all $s \leq t-1$ and by inspecting previous argument, we only need to show $\mathbf{E}[f(w_k) - f(x^*)] \leq \frac{4LD^2(1+\delta)}{k+1}$ for any $k \leq N_t$. Since our k is always increasing, we cannot directly employ our previous argument. By the structure of our algorithm, we can split the range of k into t cycles $\{1, \dots, N_1\}, \{N_1+1, \dots, N_2\}, \dots, \{N_{t-1}+1, \dots, N_t\}$. Now within each cycle, we can apply the previous argument, and thus we indeed have $\mathbf{E}[f(w_k) - f(x^*)] \leq \frac{4LD^2(1+\delta)}{k+2}$ for any $k \leq N_t$.

By the choice of N_t , we see

$$\begin{aligned} \mathbf{E}[f(w_{N_t}) - f(x^*)] &= \mathbf{E}[f(x_t) - f(x^*)] \\ &\leq \frac{4LD^2(1+\delta)}{N_t+2} \\ &= \frac{LD^2(1+\delta)}{2^{t+1}}. \end{aligned}$$

The authors of [4] mention that Algorithm 5 seems to be more stable than Algorithm 4. We give the following theoretical justification for this empirical observation.

Theorem 3. *Under the same assumption of Theorem 2, we have*

$$\lim_{s \rightarrow \infty} f(w_s) = f(x^*)$$

with probability 1 for Algorithm 5.

The theorem asserts that the objective value will converge to the true optimum under almost any realization while Theorem 2 tells we have convergence in expectation.

The proof relies on the martingale convergence theorem, which we recall here.

Theorem 4 (Martingale convergence theorem). *Let $\{X_t\}_{t=1}^n$ be a sequence of real random variables and \mathbf{E}_s to be the conditional expectation conditional on all $X_i, i \leq s-1$, then if X_t is a supermartingale, i.e.,*

$$\mathbf{E}_s(X_s) \leq X_{s-1}$$

and for all t ,

$$X_t \geq L$$

for some L . Then there is a random variable X that

$$X_s \rightarrow X \quad \text{almost surely.}$$

To make the presentation clear, we first prove a simple lemma in constructing a martingale.

Lemma 3. *Suppose a sequence of random variables $\{X_s\}_{s=1}^\infty$ and a deterministic sequence $\{b_s\}_{s=1}^\infty$ satisfy $\mathbf{E}_s(X_s) \leq X_{s-1} + b_s$ and $X_s \geq L$ for some $L \in \mathbf{R}$ for all s with probability 1. Furthermore, assume that $\sum_{s=1}^\infty b_s = C < \infty$. Then $X_s + a_s$ where $a_s = C - \sum_{i=1}^s b_i$ is a supermartingale.*

Proof. The condition $X_s \geq 0$ is mainly used so that all our expectations make sense. We need to show $\mathbf{E}_s(X_s + a_s) \leq X_{s-1} + a_{s-1}$. Now by moving a_s to the RHS and use the definition of a_s , we see this inequality holds because of the assumption $\mathbf{E}_s(X_s) \leq X_{s-1} + b_s$.

We now prove Theorem 3.

Proof (Proof of Theorem 3). Recall that \mathbf{E}_s denotes the conditional expectation given all the past except the realization of s . Using inequality (18), we see that

$$\begin{aligned} \mathbf{E}_s f(w_s) &\leq f(w_{s-1}) + \gamma_s \nabla f(w_{s-1})^T (x^* - w_{s-1}) \\ &\quad + \gamma_s \mathbf{E}_s [(\nabla f(w_{s-1}) - g_s)^T (v_s - x^*)] + \frac{LD^2(1+\delta)\gamma_s^2}{2}. \end{aligned}$$

By convexity, the term $\nabla f(w_{s-1})^T (x^* - w_{s-1})$ is upper bounded by $f(x^*) - f(w_{s-1})$, and Cauchy-Schwarz inequality yields that $|(\nabla f(w_{s-1}) - g_s)^T (v_s - x^*)| \leq D \|g_s - \nabla f(w_{s-1})\|$. Since $\mathbf{E}_s[\|g_s - \nabla f(w_{s-1})\|] \leq \sqrt{\mathbf{E}_s(\|g_s - \nabla f(w_{s-1})\|^2)}$ by Jensen's inequality. Using lemma 1, we see that

$$\begin{aligned}\sqrt{\mathbf{E}_s[\|g_s - \nabla f(w_{s-1})\|^2]} &\leq \sqrt{\frac{6L}{m_s}(2\mathbf{E}_s[f(w_{s-1}) - f(x^*)] + \mathbf{E}_s[f(w_0) - f(x^*)])} \\ &\leq \sqrt{\frac{18LB}{m_s}}\end{aligned}$$

where $B = \sup_{x \in \Omega} f(x) - f(x^*)$ which is finite as Ω is compact.

Using all previous inequalities, we see that

$$\begin{aligned}\mathbf{E}_s[f(w_s) - f(x^*)] &\leq (1 - \gamma_s)(f(w_{s-1}) - f(x^*)) + \frac{LD^2(1+\delta)\gamma_s^2}{2} + \gamma_s \sqrt{\frac{18LB}{m_s}} \\ &\leq f(w_{s-1}) - f(x^*) + \frac{LD^2(1+\delta)\gamma_s^2}{2} + \gamma_s \sqrt{\frac{18LB}{m_s}}\end{aligned}$$

Since by our choice of γ_s and m_s , we know that by letting $b_s = \frac{LD^2\gamma_s^2}{2} + \gamma_s \sqrt{\frac{18LB}{m_s}}$, $X_s = f(w_s) - f(x^*) \geq 0$, $a_s = \sum_{i=s+1}^{\infty} b_s$, the condition of Lemma 3 is satisfied and thus $X_s + a_s$ is indeed a super martingale.

Now using the martingale convergence theorem, we know that $X_s + a_s$ converges to a certain random variable X . Since $a_s \rightarrow 0$ as $s \rightarrow \infty$, $X_s \rightarrow X$ almost surely. $X_s \geq 0$ then implies $X \geq 0$. But $\mathbf{E}X \leq \mathbf{E}X_s + a_s$ for any s by the supermartingale property. Because $\mathbf{E}X_s \rightarrow 0$ by Theorem 2 and $a_s \rightarrow 0$, $\mathbf{E}X \leq \mathbf{E}X_s + a_s$ implies $\mathbf{E}X \leq 0$. Combine the fact $X \geq 0$ as we just argued, we see $X = 0$. This shows that $X_s \rightarrow 0$ almost surely which is what we need to prove.

The reason that the above argument does not work for Algorithm 4 is that once in a while we restart k and the sequence b_s we used above will be abandoned. More precisely, since the martingale convergence theorem does not tell when the sequence is about to converge, within t th cycle of $k \in \{N_t + 1, \dots, N_{t+1}\}$, we don't know whether the sequence $f(w_s)$ has converged or not. When we enter a new cycle, we start fresh from $k = 1$ with a new b_s . By contrast, for Algorithm 5, we know that k is always increasing and we have only one sequence b_s . This observation explains why Algorithm 5 is likely to be more stable.

5 $\widetilde{\text{SSVRF}}$

In previous sections, we have seen how to augment the standard Frank Wolfe algorithm with

- an approximate oracle for linear optimization subproblem (2),
- stochastic variance reduced gradients.

Now we turn our attention to the third challenge we raised in the introduction, restricting our attention to the case where the decision variable $X \in \mathbf{R}^{m \times n}$ is a matrix: what if storing the decision variable X is also costly?

Of course, if the decision variable at the solution has no structure, there is no hope to store it more cheaply: in general, $m \times n$ space is required simply to output the solution to the problem. However, in many settings X at the solution may enjoy a low rank structure: X at the solution can be well approximated by a low rank matrix.

The idea introduced in [10] is designed to capture this low rank structure. It forms a linear sketch of the column and row spaces of the decision variable X , and then uses the sketched column and row spaces to recover the decision variable. The recovered decision variable approximates the original X well if a low rank structure is present.

The advantage of this procedure in the context of optimization is that the decision variable X may not be low rank at every iteration of the algorithm. However, so long as the *solution* is (approximately) low rank, we can use this procedure to sketch the decision variable and to recover the solution from this sketch, as introduced in [11]. Notably, we need not store the entire decision variable at each iteration, but only the sketch. Hence the memory requirements of the algorithm are substantially reduced.

Specifically, the sketch proposed in [10] is as follows. To sketch a matrix $X \in \mathbf{R}^{m \times n}$, draw two matrices with independent normal entries $\Psi \in \mathbf{R}^{n \times k}$ and $\Phi \in \mathbf{R}^{l \times m}$. We use Y^C and Y^R to capture the column space and the row space of X :

$$Y^C = X\Psi \in \mathbf{R}^{m \times k}, \quad Y^R = \Phi X \in \mathbf{R}^{l \times n}. \quad (19)$$

In the optimization setting of matrix completion with Algorithm 2, we do not observe the matrix X directly. Rather, we observe a stream of rank one updates

$$X \leftarrow \beta_1 X + \beta_2 uv^T,$$

where β_1, β_2 are real scalars. In this setting, Y^C and Y^R can be updated as

$$Y^C \leftarrow \beta_1 Y^C + \beta_2 uv^T \Psi \in \mathbf{R}^{m \times k}, \quad Y^R \leftarrow \beta_1 Y^R + \beta_2 \Phi uv^T \in \mathbf{R}^{l \times n}. \quad (20)$$

This observation allows us to form the sketch Y^C and Y^R from the stream of updates.

We then reconstruct X and get the reconstructed matrix \hat{X} by

$$Y^C = QR, \quad B = (\Phi Q)^\dagger Y^R, \quad \hat{X} = Q[B]_r, \quad (21)$$

where QR is the QR factorization of Y^C and $[\cdot]_r$ returns the best rank r approximation in Frobenius norm. Specifically, the best rank r approximation of a matrix Z is $U\Sigma V^T$, where U and V are right and left singular vectors corresponding to the r largest singular values of Z and Σ is a diagonal matrix with r largest singular values of Z . Note the matrix R is not used.

The following theorem [10, Theorem 5.1] guarantees that the resulting reconstruction approximates X well if X is approximately low rank.

Theorem 5. *Fix a target rank r . Let X be a matrix, and let (Y^C, Y^R) be a sketch as described in equation (19). The procedure (21) yields a rank- r matrix \hat{X} with*

$$\mathbf{E} \|X - \hat{X}\|_F \leq 3\sqrt{2} \|X - [X]_r\|_F.$$

In the paper [11], this matrix sketching procedure is combined with the original Frank-Wolfe (Algorithm 1). We show here that it also works well with SVRF, the stochastic version of Frank-Wolfe and an approximate subproblem oracle.

We use the following matrix completion problem, which is also a particular instance of Problem (5), to illustrate this synthesis:

$$\begin{aligned} & \text{minimize } f(\mathcal{A}W) := \frac{1}{d} \sum_{i \in I} f_i(\mathcal{A}W) \\ & \text{subject to } \|W\|_* \leq \alpha, \end{aligned} \quad (22)$$

where $d = |I|$ is the number of elements in I , $W \in \mathbf{R}^{m \times n}$, $\mathcal{A}: \mathbf{R}^{m \times n} \rightarrow \mathbf{R}^l$ is a linear map, and $\alpha > 0$ is a given constant. By setting $f = \sum_{i \in I} f_i$ and $\mathcal{S} = \mathbf{R}^{m \times n}$, we see it is indeed a special instance of Problem (5). Since SVRF applied to problem (22) updates iterates W_k with a rank-one update at each inner loop iteration, the sketch matrices Y^C and Y^R can be updated using equation (20). In order to compute the gradient $\nabla(f \circ \mathcal{A})(W_k)$ at W_k , we can store the dual variable $z_k = \mathcal{A}W_k$ and compute the gradient from z_k as

$$\nabla(f \circ \mathcal{A})(W_k) = \mathcal{A}^*(\nabla f)(z_k).$$

Using linearity of \mathcal{A} , the dual variable can be updated as

$$z_k := (1 - \gamma_k)z_{k-1} + \gamma_k \mathcal{A}(-\alpha u_k v_k^*).$$

We can store the dual variable efficiently if $l = \mathcal{O}(n)$, and we can update it efficiently if the cost of applying \mathcal{A} to a rank one matrix is $\mathcal{O}(l)$. In many settings we have $l = d$, the number of samples. This means that storing and updating the dual variable z_k could be as costly as computing the full gradient. However, in the over-sampled setting, where $l = \mathcal{O}(n)$ while $d \gg \mathcal{O}(n)$, combining the techniques can be beneficial. In this setting, storing z_k is not too costly, and updating z_k is also efficient so long as applying \mathcal{A} to a matrix costs $\mathcal{O}(l)$.

The combined algorithm, $\widetilde{\text{SSVRF}}$, is shown below as Algorithm 6.

6 Theoretical Guarantees for $\widetilde{\text{SSVRF}}$

The following theorems are analogous to theorems in [11]. In this work, we introduce adaptations to cope with the approximate oracle and stochastic gradient.

Let us first instantiate some definitions. We assume for each i , $f_i \circ \mathcal{A}$ is L -smooth with respect to the Frobenius norm. Note that the diameter of the feasible region is bounded:

$$\sup_{\|X\|_*, \|Y\|_* \leq \alpha} \|X - Y\|_F \leq \sup_{\|X\|_*, \|Y\|_* \leq \alpha} \|X - Y\|_* \leq 2\alpha.$$

Hence the parameter D , the diameter of the feasible set in Theorem 2, can be replaced by 2α . For each t , we denote by \hat{X}_t the matrix reconstructed using $Y_{N_t}^C, Y_{N_t}^R$:

$$Y_{N_t}^C = QR, \quad B = (\Phi Q)^\dagger Y_{N_t}^R, \quad \hat{X}_t = Q[B]_r.$$

Algorithm 6. $\widetilde{\text{SSVRF}}$

-
- 1: **Input:** Objective function $f \circ \mathcal{A} = \frac{1}{d} \sum_{i=1}^d f_i \circ \mathcal{A}$
 - 2: **Input:** Stepsize γ_k , mini-batch size m_k , epoch length N_t and tolerance sequence ε_k
 - 3: **Input:** Target rank r and maximum number of iteration T
 - 4: **Initialize:** Set $x_{-1} = 0, Y^C = 0, Y^R = 0$ and draw $\Phi \in \mathbf{R}^{(4r+3) \times m}, \Psi \in \mathbf{R}^{n \times (2r+1)}$ with standard normal entries.
 - 5: **for** $t = 1, 2, \dots, T$ **do**
 - 6: Take a snapshot $z_0 = x_{t-1}$ and compute gradient $\nabla f(z_0)$
 - 7: **for** $k = 1$ to N_t **do**
 - 8: Compute $\tilde{\nabla}_k$, the average of m_k iid samples of $\tilde{\nabla} f(z_{k-1}, z_0)$
 - 9: Compute u, v such that
 - 10: $-\alpha \mathbf{tr}((\mathcal{A}^* \tilde{\nabla}_k)^T uv^T) \leq \min_{\|X\|_* \leq \alpha} \mathbf{tr}((\mathcal{A}^* \tilde{\nabla}_k)^T X) + \varepsilon_k$
 - 11: Compute $h_k = \mathcal{A}(-\alpha uv^T)$
 - 12: Update $z_k := (1 - \gamma_k)z_{k-1} + \gamma_k h_k$
 - 13: Update $Y_k^C = (1 - \gamma_k)Y_{k-1}^C + \gamma_k(-\alpha uv^T)\Psi$
 - 14: Update $Y_k^R = (1 - \gamma_k)Y_{k-1}^R + \gamma_k \Phi(-\alpha uv^T)$
 - 15: **end for**
 - 16: Set $x_t = z_{N_t}$
 - 17: **end for**
 - 18: Compute QR factorization of the $Y_{N_t}^C = QR$ and compute $B = (\Phi Q)^\dagger Y_{N_t}^R$
 - 19: Compute the top r many left and right singular vectors U, V of B and the diagonal matrix Σ with top r singular values.
 - 20: **Output:** (U, Σ, V) .
-

The matrix \hat{X}_t can be considered as the reconstruction of X_t (the snapshot, not the inner loop iterate) in $\widetilde{\text{SSVRF}}$. We use the same parameters as in Theorem 2 with D replaced by 2α to achieve the following theoretical guarantee:

Theorem 6. *Suppose we apply Algorithm 4 or 5 to the optimization problem (22) and that for a particular realization of the stochastic gradients, the iterates X_t converge to a matrix X_∞ . Further suppose that in Algorithm 6, we use the same stochastic gradients.*

Then

$$\lim_{t \rightarrow \infty} \mathbf{E}_{\Psi, \Phi} \|\hat{X}_t - X_\infty\|_F \leq 3\sqrt{2} \|X_\infty - [X_\infty]_r\|_F.$$

Proof. The proof exactly follows the proof of [10, Theorem 6].

When the solution set of optimization problem (22) contains only matrices with rank $\leq r$, we can prove a stronger guarantee for Algorithm 6:

Theorem 7. *Suppose that the solution set S_* of the optimization problem (22) contains only matrices with rank $\leq r$. Then Algorithm 6 attains*

$$\lim_{t \rightarrow \infty} \mathbf{E} \mathbf{dist}_F(\hat{X}_t, S_*) = 0,$$

where $\mathbf{dist}_F(X, S_*) = \inf_{Y \in S_*} \|X - Y\|_F$.

Proof. The triangle inequality implies that

$$\mathbf{E} \mathbf{dist}_F(\hat{X}_t, S_*) \leq \mathbf{E} \|\hat{X}_t - X_t\|_F + \mathbf{E} \mathbf{dist}_F(X_t, S_*).$$

We claim that the second term, $\mathbf{E} \mathbf{dist}_F(X_t, S_*)$, converges to 0. If so, we may conclude that the first term converges to zero by the following inequality.

$$\begin{aligned} \mathbf{E} \|\hat{X}_t - X_t\|_F &\leq 3\sqrt{2} \mathbf{E} \|X_t - [X_t]_r\|_F \\ &\leq 3\sqrt{2} \mathbf{E}(\mathbf{dist}_F(X_t, S_*)) \rightarrow 0. \end{aligned}$$

The first inequality is Theorem 5, and the second bound is due to the optimality of $[X_t]_r$.

It remains only to prove the claim $\mathbf{E} \mathbf{dist}_F(X_t, S_*) \rightarrow 0$. Let $g = f \circ \mathcal{A}$ and g_* to be the optimal value of g in program (22). Now fix a number $\varepsilon > 0$. Define

$$E = \{X \in \mathbf{R}^{m \times n} : \|X\|_* \leq \alpha \text{ and } \mathbf{dist}(X, S_*) \geq \varepsilon\},$$

and $v = \inf\{g(X), X \in E\}$. If E is empty, then $v = +\infty$. Otherwise, the continuous function g attains the value v on the compact set E . In either case, $v > g_*$ because E contains no optimal point of (22). Thus

$$\mathbf{Prob}(X_t \in E) \leq \mathbf{Prob}(g(X_t) - g_* > v - g_*) \leq \frac{\mathbf{E}(g(X_t) - g_*)}{v - g_*},$$

where the first inequality is due to the optimality of v , and the second is just the Markov inequality. Notice

$$\begin{aligned} \mathbf{E} \mathbf{dist}_F(X_t, S_*) &= \mathbf{E} \mathbf{dist}_F(X_t, S_*) \mathbf{1}_{\{X_t \in E\}} + \mathbf{E} \mathbf{dist}_F(X_t, S_*) \mathbf{1}_{\{X_t \notin E\}} \\ &\leq 2\alpha \mathbf{Prob}(X_t \in E) + \varepsilon \\ &\leq 2\alpha \frac{\mathbf{E}(g(X_t) - g_*)}{v - g_*} + \varepsilon, \end{aligned}$$

where the inequality is due to the definition of E , and the feasible region is $\|X\|_* \leq \alpha$. Since $\mathbf{E}(g(X_t)) \rightarrow g_*$ by Theorem 2, we know $\lim_{t \rightarrow \infty} \mathbf{E} \mathbf{dist}_F(X_t, S_*) \leq \varepsilon$ for any $\varepsilon > 0$. Thus the claim is proved.

When the solution to the optimization problem (22) is unique and the function f has a strong curvature property, we can also bound the distance to the optimal solution in expectation.

Theorem 8. Fix $\kappa > 0$ and $v \geq 1$. Suppose the unique solution X^* of (22) has rank less than or equal to r and

$$f(\mathcal{A}X) - f(\mathcal{A}X^*) \geq \kappa \|X - X^*\|_F^v \quad (23)$$

for all $\|X\|_* \leq \alpha$. Then we have the error bound

$$\mathbf{E} \|\hat{X}_t - X^*\|_F \leq 6 \left(\frac{4\kappa^{-1}LD^2(1+\delta)}{2^{t+1}} \right)^{\frac{1}{v}}$$

for all t .

Proof. Let $g = f \circ \mathcal{A}$. The proof of Theorem 2 tells us that

$$\mathbf{E}(g(X_t) - g(X^*)) \leq \frac{LD^2(1+\delta)}{2^{t+1}}.$$

Since the iterate X_t is feasible, the assumption in (23) gives us

$$\mathbf{E}(g(X_t) - g(X^*)) \geq \kappa \mathbf{E} \|X_t - X^*\|_F^v \quad (24)$$

$$\begin{aligned} &\geq \kappa \mathbf{E} \|X_t - [X_t]_r\|_F^v \\ &\geq \kappa [\mathbf{E}(\|X_t - [X_t]_r\|_F)]^v \\ &\geq \frac{\kappa}{(3\sqrt{2})^v} (\mathbf{E}(\|X_t - \hat{X}_t\|_F))^v. \end{aligned} \quad (25)$$

The second inequality is due to the optimality of $[X_t]_r$ and X^* has rank less than r . The third is because of Jensen's inequality and the last is from Theorem 5. We now conclude that

$$\begin{aligned} \mathbf{E} \|\hat{X}_t - X^*\|_F &\leq \mathbf{E} \|\hat{X}_t - X_t\| + \mathbf{E} \|X_t - X^*\| \\ &\leq 3\sqrt{2} \left(\frac{\kappa^{-1}LD^2(1+\delta)}{2^{t+1}} \right)^{1/v} + \left(\frac{\kappa^{-1}LD^2(1+\delta)}{2^{t+1}} \right)^{1/v}. \end{aligned}$$

The last bound follows from inequality (24) and (25). To reach the final conclusion shown in the theorem, simplify the numerical constant, use the assumption that $v \geq 1$ and note that $D \leq 2\alpha$.

Acknowledgements. This work was supported by DARPA Award FA8750-17-2-0101. The authors are grateful for helpful discussions with Joel Tropp, Volkan Cevher, and Alp Yurtsever.

Appendix

We prove the following simple proposition about L -smooth functions used in Section 2.

Proposition 3. *If f is a real valued differentiable convex function with domain \mathbf{R}^n and satisfies $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$, then for all $x, y \in \mathbf{R}^n$,*

$$f(x) \leq f(y) + \nabla f(y)^T(x - y) + \frac{L}{2}\|x - y\|^2.$$

Proof. The inequality follows from the following computation:

$$\begin{aligned}
f(x) - f(y) - \nabla f(y)^T(x - y) &= \int_0^1 (\nabla f(y + t(x - y)) - \nabla f(y))^T(x - y) dt \\
&\leq \int_0^1 \|(\nabla f(y + t(x - y)) - \nabla f(y))^T(x - y)\| dt \\
&\leq \int_0^1 \|(\nabla f(y + t(x - y)) - \nabla f(y))\| \|x - y\| dt \\
&\leq \int_0^1 Lt \|x - y\|^2 dt \\
&= \frac{L}{2} \|x - y\|^2.
\end{aligned}
\tag{26}$$

References

- [1] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849, 2012.
- [2] R. M. Freund, P. Grigas, and R. Mazumder. An Extended Frank-Wolfe Method with "In-Face" Directions, and Its Application to Low-Rank Matrix Completion. *SIAM Journal on Optimization*, 27(1):319–346, 2017.
- [3] E. Hazan. Sparse approximate solutions to semidefinite programs. *Lecture Notes in Computer Science*, 4957:306–316, 2008.
- [4] E. Hazan and H. Luo. Variance-reduced and projection-free stochastic optimization. In *International Conference on Machine Learning*, pages 1263–1271, 2016.
- [5] M. Jaggi. Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization. In *ICML (1)*, pages 427–435, 2013.
- [6] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.
- [7] J. Kuczyński and H. Woźniakowski. Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start. *SIAM Journal on matrix analysis and applications*, 13(4):1094–1122, 1992.
- [8] R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK users' guide: solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods*. SIAM, 1998.
- [9] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.
- [10] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher. Randomized single-view algorithms for low-rank matrix approximation. *arXiv preprint arXiv:1609.00048*, 2016.
- [11] A. Yurtsever, M. Udell, J. A. Tropp, and V. Cevher. Sketchy Decisions: Convex Low-Rank Matrix Optimization with Optimal Storage. *arXiv preprint arXiv:1702.06838*, 2017.