Convergence Analysis of a Stochastic Projection-free Algorithm

Jean Lafond[*], Hoi-To Wai[†‡], Eric Moulines[§]

October 6, 2015

Abstract

This paper presents and analyzes a stochastic version of the Frank-Wolfe algorithm (a.k.a. conditional gradient method or projection-free algorithm) for constrained convex optimization. We first prove that when the quality of gradient estimate improves as $\mathcal{O}(\sqrt{\eta_t^\Delta/t})$, where t is the iteration index and η_t^Δ is an increasing sequence, then the objective value of the stochastic Frank-Wolfe algorithm converges in at least the same order. When the optimal solution lies in the interior of the constraint set, the convergence rate is accelerated to $\mathcal{O}(\eta_t^\Delta/t)$. Secondly, we study how the stochastic Frank-Wolfe algorithm can be applied to a few practical machine learning problems. Tight bounds on the gradient estimate errors for these examples are established. Numerical simulations support our findings.

1 Introduction

We are interested in solving the convex optimization problem:

$$\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \text{ s.t. } \boldsymbol{\theta} \in \mathcal{C} \,, \tag{1}$$

where $\mathcal{C}$ is a bounded convex set included in a real euclidian space $\mathbf{E}$ and $f(\cdot)$ is a differentiable convex function. Notice that $f(\boldsymbol{\theta})$ may take the form as $f(\boldsymbol{\theta}) = \mathbb{E}_\omega[f(\boldsymbol{\theta};\omega)]$, i.e., an expected risk.

For large-scale problems (e.g., when $\mathbf{E}$ is high-dimensional or when $\mathcal{C}$ is defined by many equalities/inequalities), a popular algorithm for solving (1) is the projected gradient method. While the latter algorithm has been extensively studied in the literature (Beck and Teboulle, 2009; Juditsky and Nemirovski, 2012a; Juditsky and Nemirovski, 2012b), in many applications, efficient implementations are not available for computing the projection onto $\mathcal{C}$. An alternative approach that is gaining popularity is to solve (1) by the Frank-Wolfe algorithm (Frank and Wolfe, 1956) (a.k.a. conditional gradient method and projection free method); see the recent advances in (Jaggi, 2013; Freund and Grigas, 2013; Lacoste-Julien and Jaggi, 2013; Garber and Hazan, 2015; Ghosh and Lam, 2015).

[*]Institut Mines-Telecom, Telecom ParisTech, CNRS LTCI, Paris, France. Email: jean.lafond@telecom-paristech.fr

[†]School of Electrical, Computer and Energy Engineering, Arizona State University, AZ, USA. Email: htwai@asu.edu

[‡]J. Lafond and H.-T. Wai have contributed equally.

[§]CMAP, Ecole Polytechnique, Palaiseau, France. Email: eric.moulines@polytechnique.edu

The aim of this paper is to study a stochastic version of the Frank-Wolfe algorithm. At iteration $t \in \mathbb{N}$, we have access to a noisy estimate of the gradient:

$$\hat{\nabla} f(\boldsymbol{\theta}_t) = \nabla f(\boldsymbol{\theta}_t) + \boldsymbol{\epsilon}_t \ , \tag{2}$$

where $\boldsymbol{\epsilon}_t$ is the error term. We assume in the sequel that $\boldsymbol{\epsilon}_t$ decays to zero at a given rate with high probability as $t \to \infty$. This setting has been considered in (Jaggi, 2013; Freund and Grigas, 2013; Lacoste-Julien et al., 2013) and is motivated by online learning and incremental learning; see section 3 for some interesting examples. The stochastic Frank-Wolfe (sFW) algorithm is summarized below.

Algorithm 1 Stochastic Frank-Wolfe Algorithm.

1: **Initialize:** $\boldsymbol{\theta}_0 \in \mathcal{C}$, $t = 1$;
2: **while** *convergence is not reached* **do**
3: Solve the linear optimization with noisy gradient:

$$\boldsymbol{a}_t \leftarrow \arg\min_{\boldsymbol{a} \in \mathcal{C}} \ \left\langle \boldsymbol{a}, \hat{\nabla} f(\boldsymbol{\theta}_t) \right\rangle \ . \tag{3}$$

4: Update $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \gamma_t(\boldsymbol{a}_t - \boldsymbol{\theta}_t)$, where $\gamma_t = K/(t + K - 1)$ and $K \in \mathbb{Z}_+^\star$ is a step-size parameter.
5: $t \leftarrow t + 1$.
6: **end while**
7: **Return:** $\boldsymbol{\theta}_t$.

Compared to the classical formulation due to (Frank and Wolfe, 1956), Algorithm 1 takes a noisy version of the gradient $\hat{\nabla} f(\boldsymbol{\theta}_t)$ in lieu of the exact gradient $\nabla f(\boldsymbol{\theta}_t)$. For simplicity we have considered here only step-sizes of the form $K/(K + t - 1)$.

The run-time complexity of the sFW algorithm depends on the existence of efficient solution to the linear optimization (3). When $\mathcal{C}$ is an atomic set (Chandrasekaran et al., 2012), i.e., $\mathcal{C} = \text{conv}\{\boldsymbol{a}_1, ..., \boldsymbol{a}_m\}$, solving (3) amounts to finding the most uncorrelated atom with the gradient $\hat{\nabla} f(\boldsymbol{\theta}_t)$. A few examples of $\mathcal{C}$ with efficient solution to (3) are in order.

- The ℓ_1 ball — $\mathcal{C} = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_1 \leq r\} \subseteq \mathbb{R}^n$ — In this case, the linear optimization in (3) can be solved by setting $\boldsymbol{a}_t = -r \cdot \text{sign}([\hat{\nabla} f(\boldsymbol{\theta}_t)]_i) \cdot \boldsymbol{e}_i$, where $i = \arg\max_{j \in [n]} |[\hat{\nabla} f(\boldsymbol{\theta}_t)]_j|$ and $\boldsymbol{e}_i$ is the canonical basis vector.

- The matrix trace-norm ball — $\mathcal{C} = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_{\sigma,1} \leq r\} \subseteq \mathbb{R}^{m \times n}$ — In this case, the linear optimization (3) can be solved by $\boldsymbol{a}_t = -r\boldsymbol{u}_1\boldsymbol{v}_1^T$, where $\boldsymbol{u}_1, \boldsymbol{v}_1$ are the top singular vectors of $\hat{\nabla} f(\boldsymbol{\theta}_t)$. Notice that $\boldsymbol{u}_1, \boldsymbol{v}_1$ can be computed in $\mathcal{O}(\max\{m, n\})$.

See (Jaggi, 2013) for a comprehensive overview on constraint sets $\mathcal{C}$ with efficient implementation.

Organization. section 2 describes the main analytical result on the sFW algorithm. section 3 presents the statistical analysis of a few application examples of the sFW algorithm. Lastly, section 4 presents numerical results to support our findings on sFW.

Notation. For any $n \in \mathbb{N}$, let $[n]$ denote the set $\{1, \cdots, n\}$. The inner product on $\mathbf{E}$ is denoted by $\langle \cdot, \cdot \rangle$ and the associated Euclidian norm by $\| \cdot \|_2$. The space $\mathbf{E}$ is also equipped with a

norm $\|\cdot\|$ and its dual norm $\|\cdot\|_\star$. The diameter of the constrained $\mathcal{C}$ *w.r.t.* $\|\cdot\|_\star$ is denoted by ρ, that is

$$\rho := \sup_{\theta,\theta' \in \mathcal{C}} \|\theta - \theta'\|_\star \ . \tag{4}$$

The ith element in a vector $\boldsymbol{x}$ is denoted by $[\boldsymbol{x}]_i$.

1.1 Contributions & Related Works

This paper presents several new results pertaining to the stochastic Frank-Wolfe (sFW) algorithm. Our main contributions can be summarized as follows.

On the theoretical ground, we have analyzed the convergence rate of the sFW algorithm under a variety of standard assumptions on the function f (see H2, H3 and H4) and the following on the gradient estimate.

H1. *With probability at least* $1 - \Delta$*,*

$$\|\boldsymbol{\epsilon}_t\| \leq \sigma \sqrt{\eta_t^\Delta / (K + t - 1)}, \ \forall \ t \geq 1 \ , \tag{5}$$

for some $\sigma > 0$*,* η_t^Δ *is a non-decreasing sequence satisfying* $\eta_t^\Delta \geq 1$ *and* $\Delta \in [0, 1]$*. Furthermore,* $\eta_t^\Delta/(K + t - 1)$ *is non-increasing and* $\lim_{t\to\infty} \eta_t^\Delta / t = 0$*.*

The $\mathcal{O}(\sqrt{\eta_t^\Delta / t})$ rate of the noise $\|\boldsymbol{\epsilon}_t\|$ is motivated by the stochastic approximation step in many machine learning applications. Notice that (Jaggi, 2013; Lacoste-Julien et al., 2013) have only considered the case when $\|\boldsymbol{\epsilon}_t\|$ is decaying at a rate of $\mathcal{O}(1/t)$.

Under H1, if f has a finite curvature (see H2), then with high probability the objective value of the sFW algorithm converges as $\mathcal{O}(\sqrt{\eta_t^\Delta / t})$. If in addition, the function is strongly convex and the optimal solution lies in the interior of the constraint set (see H3 and H4), then the convergence rate can be further improved to $\mathcal{O}(\eta_t^\Delta / t)$. This is the first improved convergence rate obtained for sFW algorithm. Furthermore, the obtained rate is comparable to that of the popular stochastic gradient descent method (Juditsky and Nemirovski, 2012a; Juditsky and Nemirovski, 2012b; Duchi et al., 2011).

Our first convergence rate of $\mathcal{O}(\sqrt{\eta_t^\Delta / t})$ can be derived from (Freund and Grigas, 2013, Proposition 5.1); however the improved rate is a new result to our best knowledge. (Hazan and Kale, 2012) is another relevant work, who have analyzed an Online Frank-Wolfe (OFW) algorithm for empirical risk minimization. We show that when applied to large scale learning problem, the sFW algorithm is equivalent to a randomized version of the OFW algorithm. A faster convergence rate can be guaranteed ($\mathcal{O}(\log(t)/t)$ versus $\mathcal{O}(1/t^{2/3})$) under the interior optimal point assumption (H4). The improved rate stems from our analysis that focuses on controlling the gradient estimate error directly (see section 2). A related result can also be found in (Ghosh and Lam, 2015), yet the analysis therein is restricted to expected risk minimization.

On the application side, we have analyzed the sFW algorithm applied to several machine learning examples. These examples include an online sparse learning, an online matrix completion and a large-scale learning problems. For the considered problems, we apply stochastic approximation (SA) to obtain the desired online gradient estimate and derive bounds on the convergence of primal optimality.

While online algorithms to sparse learning and matrix completion have been studied (Garrigues and El Ghaoui, 2008; Langford et al., 2009; Recht and Re, 2013), most of the prior works have only

analyzed the convergence in terms of the empirical risk that depends on the set of data collected so far. The convergence of online algorithms in terms of the expected risk has only been considered in a few papers (Shalev-Shwartz et al., 2009; Bottou, 1998; Ghosh and Lam, 2015). Using our Theorem 1 for sFW algorithm, we are able to analyze the convergence of expected risk for these problems.

Lastly, we have applied the sFW algorithm on the machine learning examples using both synthetic and real world data. The numerical experiments are consistent with our analysis on the convergence rate.

2 Main Results

Consider the following assumptions.

H2. *The curvature constant C_f, defined as:*

$$C_f := \sup_{\substack{\boldsymbol{x},\boldsymbol{s}\in\mathcal{C},\\ \gamma\in[0,1]}} \frac{2}{\gamma^2}\Big(f(\boldsymbol{x}+\gamma(\boldsymbol{s}-\boldsymbol{x})) - f(\boldsymbol{x}) - \gamma\langle \boldsymbol{s}-\boldsymbol{x}, \nabla f(\boldsymbol{x})\rangle\Big), \tag{6}$$

is finite.

Note that if the gradient is L-Lipschitz, then H2 is satisfied for $C_f \leq L \sup_{\boldsymbol{\theta},\boldsymbol{\theta}'\in\mathcal{C}} \|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_2^2 < \infty$ (see *e.g.,* (Jaggi, 2013)).

H3. *The function f is μ-strongly convex, i.e., such that*

$$f(\boldsymbol{x}) \leq f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{x}-\boldsymbol{y}\rangle - \frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2, \tag{7}$$

for all $\boldsymbol{x},\boldsymbol{y}\in\mathcal{C}$.

H4. *The optimal solution $\boldsymbol{\theta}^\star$ to (1) lies in the interior of $\mathcal{C}$, i.e.,*

$$\delta = \inf_{\boldsymbol{s}\in\partial\mathcal{C}} \|\boldsymbol{s}-\boldsymbol{\theta}^\star\| > 0, \tag{8}$$

where $\partial\mathcal{C}$ denotes the boundary set of $\mathcal{C}$.

Define $D' = (1+(2K-1)^{-1})(\rho\sigma + KC_f/2)$ and $D = 3(1+K^{-1})(\rho\sigma + KC_f/2)^2/(\delta^2\mu)$, where $\rho, \sigma, \eta_t^\Delta, C_f, \mu$ and δ are constants defined in (4), (5), (6), (7) and (8), respectively.

Theorem 1. *Consider Algorithm 1 with the gradient estimate given by (2). Assume H1 and H2. Then, the following holds with probability at least $1-\Delta$:*

$$f(\boldsymbol{\theta}_t) - f(\boldsymbol{\theta}^\star) \leq D'\sqrt{\frac{\eta_t^\Delta}{t+K-1}}\,, \tag{9}$$

for all $t \geq 2$. Assume in addition H3 and H4 are satisfied. Then, the following also holds with probability at least $1-\Delta$:

$$f(\boldsymbol{\theta}_t) - f(\boldsymbol{\theta}^\star) \leq D\frac{\eta_t^\Delta}{t+K-1}\,, \tag{10}$$

for all $t \geq 2$.

Remark 1. *To illustrate the tightness of the bounds in Theorem 1, let us consider problem* (1) *with* $\mathcal{C} = [a, b]$ *and the following objective function*

$$f(\boldsymbol{\theta}) := \mathbb{E}_{y_t}[(y_t - \boldsymbol{\theta})^2] \,, \tag{11}$$

where $y_t \sim \mathcal{N}(\bar{\boldsymbol{\theta}}, 1)$ *is a standard Gaussian random variable with unknown mean* $\bar{\boldsymbol{\theta}}$*. Consider an online algorithm for the above problem. We are sequentially given the i.i.d. observations* y_t *and the gradient can be estimated online as* $\hat{\nabla} f_t(\boldsymbol{\theta}) = \boldsymbol{\theta} - \bar{y}_t$*, where* $\bar{y}_t = (1 - t^{-1})\bar{y}_{t-1} + t^{-1} y_t$ *is updated in an online fashion. We can prove that* $|\hat{\nabla} f_t(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta})| = \mathcal{O}(\sqrt{\log t / t})$ *with high probability for any* $\boldsymbol{\theta} \in \mathcal{C}$ *(see subsection 3.1). Let the mean parameter* $\bar{\boldsymbol{\theta}}$ *be in the interior of* $\mathcal{C}$*, then it is also the optimal solution of problem* (1)*. Notice that H 1, H 2, H 3 and H 4 are satisfied, Theorem 1 shows that the convergence rate of sFW is* $\mathcal{O}(\log t / t)$*. On the other hand, the rate of* $\mathcal{O}(1/t)$ *is a well known statistical (achievable) lower bound (see* e.g.*, (Tsybakov, 2009)) for Gaussian estimation. This implies in particular that* $\mathcal{O}(1/t)$ *is a lower bound for the convergence rate of Algorithm 1.*

Proof. For simplicity, we assume that $K = 1$. The general case is considered in the supplementary material.

$\mathcal{O}(\sqrt{\eta_t^\Delta / t})$ **bound:** Let us define $h_t = f(\boldsymbol{\theta}_t) - f(\boldsymbol{\theta}^\star)$, then by H 2 we get

$$h_{t+1} \leq h_t + \frac{1}{t}\langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{a}_t - \boldsymbol{\theta}_t \rangle + \frac{C_f}{2t^2} \,. \tag{12}$$

On the other hand, the following also holds:

$$\begin{aligned}
\langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{a}_t - \boldsymbol{\theta}_t \rangle &= \langle \hat{\nabla} f(\boldsymbol{\theta}_t), \boldsymbol{a}_t - \boldsymbol{\theta}_t \rangle - \langle \boldsymbol{\epsilon}_t, \boldsymbol{a}_t - \boldsymbol{\theta}_t \rangle, \\
&\leq \langle \hat{\nabla} f(\boldsymbol{\theta}_t), \boldsymbol{\theta}^\star - \boldsymbol{\theta}_t \rangle - \langle \boldsymbol{\epsilon}_t, \boldsymbol{a}_t - \boldsymbol{\theta}_t \rangle \\
&= \langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{\theta}^\star - \boldsymbol{\theta}_t \rangle + \langle \boldsymbol{\epsilon}_t, \boldsymbol{\theta}^\star - \boldsymbol{a}_t \rangle \\
&\leq -h_t + \rho \|\boldsymbol{\epsilon}_t\| \,.
\end{aligned} \tag{13}$$

where the second line follows from the definition of $\boldsymbol{a}_t$ and the last inequality from the convexity of f and the definition of the diameter (4). Plugging (13) into (12) and using H 1 shows that with probability at least $1 - \Delta$ and for all $t \geq 1$

$$h_{t+1} \leq (1 - \frac{1}{t}) h_t + \rho\sigma \sqrt{\frac{\eta_t^\Delta}{t^3}} + \frac{C_f}{2t^2} \,. \tag{14}$$

We now proceed by induction to prove the first bound of the Theorem. Recall $D' = 2(\rho\sigma + C_f/2)$. The initialization is done by applying (14) with $t = 1$. Assume that $h_t \leq D' \sqrt{\eta_t^\Delta / t}$ for some $t \geq 1$. From (14) we get:

$$\begin{aligned}
&h_{t+1} - D' \sqrt{\frac{\eta_{t+1}^\Delta}{t+1}} \\
&\leq D' \left(\sqrt{\frac{\eta_t^\Delta}{t}} - \sqrt{\frac{\eta_{t+1}^\Delta}{t+1}} \right) + \left(\rho\sigma \sqrt{\eta_t^\Delta} + \frac{C_f}{2\sqrt{t}} - D' \sqrt{\eta_t^\Delta} \right) / t^{3/2} \\
&\leq D' \sqrt{\eta_t^\Delta} \left(\frac{1}{\sqrt{t}} - \frac{1}{\sqrt{t+1}} \right) + \left(\rho\sigma + \frac{C_f}{2} - D' \right) \frac{\sqrt{\eta_t^\Delta}}{t^{3/2}} \\
&\leq (D'/2 + D'/2 - D') \sqrt{\eta_t^\Delta} / t^{3/2} = 0 \,,
\end{aligned}$$

where we used the fact that η_t^Δ is increasing and greater than 1 for the second inequality and the definition of D' for the last line.

$\mathcal{O}(\eta_t^\Delta/t)$ **bound:** Define

$$g_t = \max_{\boldsymbol{s}\in\mathcal{C}} \langle \boldsymbol{\theta}_t - \boldsymbol{s}, \nabla f(\boldsymbol{\theta}_t)\rangle \ , \tag{15}$$

as the duality gap at $\boldsymbol{\theta}_t$. We use the following Lemma which is borrowed from (Lacoste-Julien and Jaggi, 2013).

Lemma 1. *Under H2, H3 and H4, we have*

$$g_t^2 \geq 2\mu\delta^2 \cdot h_t \quad \textit{and} \quad C_f \geq (\mu\delta^2) \ , \tag{16}$$

where ρ, C_f μ and δ are defined in (4), (6) (7) *and* (8), *respectively.*

Proof. See (Lacoste-Julien and Jaggi, 2013, Proof of Theorem 3). □

Define $\boldsymbol{s}_t \in \arg\max_{\boldsymbol{s}\in\mathcal{C}} \langle \boldsymbol{\theta}_t - \boldsymbol{s}, \nabla f(\boldsymbol{\theta}_t)\rangle$. In the same line as (13) we have

$$\begin{aligned}
&\langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{a}_t - \boldsymbol{\theta}_t\rangle \leq \langle \hat{\nabla} f(\boldsymbol{\theta}_t), \boldsymbol{s}_t - \boldsymbol{\theta}_t\rangle - \langle \boldsymbol{\epsilon}_t, \boldsymbol{a}_t - \boldsymbol{\theta}_t\rangle, \\
&= \langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{s}_t - \boldsymbol{\theta}_t\rangle + \langle \boldsymbol{\epsilon}_t, \boldsymbol{s}_t - \boldsymbol{a}_t\rangle, \\
&= -g_t + \rho\|\boldsymbol{\epsilon}_t\| \leq -\delta\sqrt{2\mu h_t} + \rho\|\boldsymbol{\epsilon}_t\| \ ,
\end{aligned} \tag{17}$$

where g_t is defined in (15) and the last equation follows from Lemma 1. Plugging (17) into (12) shows that with probability at least $1-\Delta$ and for all $t \geq 1$,

$$h_{t+1} \leq h_t - \frac{\delta}{t}\sqrt{2\mu h_t} + \rho\sigma\sqrt{\frac{\eta_t^\Delta}{t^3}} + \frac{C_f}{2t^2} \ . \tag{18}$$

Recall $D = 6(\rho\sigma + C_f/2)^2/(\delta^2\mu)$. Again, we proceed by induction and suppose that $h_t \leq D\eta_t^\Delta/t$ for some $t \geq 1$. There are two cases. If $h_t - \gamma_t\delta\sqrt{2\mu h_t} \leq 0$, then (18) yields

$$\begin{aligned}
h_{t+1} \leq \rho\sigma\sqrt{\eta_t^\Delta}/t^{3/2} + C_f/2t^2 &\leq (\rho\sigma + \frac{C_f}{2})\eta_{t+1}^\Delta/t^{3/2} \ , \\
&= (\delta\sqrt{\mu D/6})\eta_{t+1}^\Delta/t^{3/2}
\end{aligned}$$

where we used that η_t^Δ is increasing and larger than 1. By Lemma 1 $\mu\delta^2 \leq C_f$ holds, this implies in particular $\delta\sqrt{\mu} \leq \sqrt{D}$. Hence

$$h_{t+1} \leq D\eta_{t+1}^\Delta/(6t^{3/2}) \leq D\eta_{t+1}^\Delta/(t+1) \ .$$

Otherwise, if $h_t - \gamma_t\delta\sqrt{2\mu h_t} > 0$, then by induction we get

$$\begin{aligned}
h_t - \frac{\delta}{t}\sqrt{2\mu h_t} &= \sqrt{h_t}(\sqrt{h_t} - \frac{\delta}{t}\sqrt{2\mu}) \\
&\leq \sqrt{\frac{\eta_t^\Delta}{t}}(\sqrt{\frac{\eta_t^\Delta}{t}} - \frac{\delta}{t}\sqrt{2\mu}) \ ,
\end{aligned}$$

which yields with (18)

$$
\begin{aligned}
& h_{t+1} - D\frac{\eta_{t+1}^{\Delta}}{t+1} \\
& \leq D(\frac{\eta_t^{\Delta}}{t} - \frac{\eta_{t+1}^{\Delta}}{t+1}) + (\rho\sigma\sqrt{\eta_t^{\Delta}} + \frac{C_f}{2\sqrt{t}} - \delta\sqrt{2\mu D\eta_t^{\Delta}})/t^{3/2} \\
& \leq \frac{D\eta_t^{\Delta}}{t^2} + (\rho\sigma\sqrt{\eta_t^{\Delta}} + \frac{C_f\sqrt{\eta_t^{\Delta}}}{2} - \delta\sqrt{2\mu D\eta_t^{\Delta}})/t^{3/2} \ , \qquad (19)
\end{aligned}
$$

where we used the fact that η_t^{Δ} is increasing and greater than 1 and $t \geq 1$. Define $t_0 := \inf\{t \geq 1 | t \geq D\eta_t^{\Delta}/(\delta^2\mu)\}$. Since η_t^{Δ}/t is decreasing to 0, t_0 exists and for any $t \geq t_0$ we have

$$
h_{t+1} - D\frac{\eta_{t+1}^{\Delta}}{t+1} \leq (\rho\sigma + \frac{C_f}{2} - (\sqrt{2}-1)\delta\sqrt{\mu D})\frac{\sqrt{\eta_t^{\Delta}}}{t^{3/2}} \leq 0 \ .
$$

We now consider the case when $t \leq t_0$. By Lemma 1, $C_f \geq \delta^2\mu$, therefore applying (14) for $t = 1$ yields $h_2 \leq D/2 \leq D\eta_2^{\Delta}/2$. Finally, for $t \leq t_0$ we have by monotony of η_t^{Δ}/t, $t \leq D\eta_t^{\Delta}/(\delta^2\mu)$ and therefore

$$
\delta\sqrt{\mu}\sqrt{\frac{D\eta_t^{\Delta}}{t}} \leq \frac{D\eta_t^{\Delta}}{t} \ . \qquad (20)
$$

Since $D' \leq \delta\sqrt{\mu D}$, the left hand side in the above is lower bounded by $D'\sqrt{\eta_t^{\Delta}/t}$ and we have $D'\sqrt{\eta_t^{\Delta}/t} \leq D\eta_t^{\Delta}/t$. Hence, the first bound obtained in the proof combinend with (20) yields

$$
h_t \leq D'\sqrt{\eta_t^{\Delta}/t} \leq D\eta_t^{\Delta}/t \ .
$$

□

3 Applications

We discuss several machine learning applications of the sFW algorithm. We demonstrate that H1 is satisfied in the examples below. The proofs can be found in the supplementary material.

3.1 Online sparse learning

In the online sparse learning problem, we are sequentially given i.i.d. observations $(\boldsymbol{Y}_t, \boldsymbol{A}_t)$ such that $\boldsymbol{Y}_t \in \mathbb{R}^m$ is the response, $\boldsymbol{A}_t \in \mathbb{R}^{m\times n}$ is the random design and

$$
\boldsymbol{Y}_t = \boldsymbol{A}_t\bar{\boldsymbol{\theta}} + \boldsymbol{w}_t \ , \qquad (21)
$$

where the vector $\boldsymbol{w}_t$ is i.i.d., $[\boldsymbol{w}_t]_i$ is independent of $[\boldsymbol{w}_t]_j$ for $i \neq j$ and $[\boldsymbol{w}_t]_i$ is zero-mean and sub-Gaussian with parameter σ_w. Furthermore, we suppose that $\bar{\boldsymbol{\theta}}$ is sparse. To learn $\bar{\boldsymbol{\theta}}$, we solve the following optimization problem:

$$
\min_{\boldsymbol{\theta}\in\mathbb{R}^n} f(\boldsymbol{\theta}) := \frac{1}{2}\mathbb{E}_{\bar{\boldsymbol{\theta}}}[\|\boldsymbol{Y}_t - \boldsymbol{A}_t\boldsymbol{\theta}\|_2^2] \text{ s.t. } \|\boldsymbol{\theta}\|_1 \leq r, \qquad (22)
$$

where $r > 0$ is a regularization constant.

Similar setup has been studied in (Garrigues and El Ghaoui, 2008; Langford et al., 2009), which considered the case of empirical squared loss, i.e., when $f(\boldsymbol{\theta}) = (1/2T)\sum_{s=1}^{T} \|\boldsymbol{Y}_s - \boldsymbol{A}_s\boldsymbol{\theta}\|_2^2$ with $T \gg 1$.

We propose the following online algorithm to (22).

Algorithm. At time $t \geq 1$, we perform:

1. Select $\boldsymbol{A}_t$ and observe $\boldsymbol{Y}_t$ according to (21).

2. Evaluate the sufficient statistics:

$$\overline{(\boldsymbol{A}^\top \boldsymbol{Y})}_t := (1 - t^{-1})\overline{(\boldsymbol{A}^\top \boldsymbol{Y})}_{t-1} + t^{-1}\boldsymbol{A}_s^\top \boldsymbol{Y}_s,$$
$$\overline{(\boldsymbol{A}^\top \boldsymbol{A})}_t := (1 - t^{-1})\overline{(\boldsymbol{A}^\top \boldsymbol{A})}_{t-1} + t^{-1}\boldsymbol{A}_s^\top \boldsymbol{Y}_s,$$

   and form the gradient estimate:

$$\hat{\nabla} f_t(\boldsymbol{\theta}) = \overline{(\boldsymbol{A}^\top \boldsymbol{A})}_t \boldsymbol{\theta} - \overline{(\boldsymbol{A}^\top \boldsymbol{Y})}_t \,. \tag{23}$$

3. Set $\boldsymbol{a}_t = -r \cdot \operatorname{sign}([\hat{\nabla} f_t(\boldsymbol{\theta}_t)]_i) \cdot \boldsymbol{e}_i$, where $i = \arg\max_{j \in [n]} |[\hat{\nabla} f_t(\boldsymbol{\theta}_t)]_j|$.

4. Update $\boldsymbol{\theta}_{t+1} = (1 - \gamma_t)\boldsymbol{\theta}_t + \gamma_t \boldsymbol{a}_t$, where γ_t is defined in Algorithm 1.

Notice that (23) is an SA step for the stochastic gradient function. We can derive:

Proposition 1. *Assume that* $\|\boldsymbol{A}_t^\top \boldsymbol{A}_t - \mathbb{E}[\boldsymbol{A}^\top \boldsymbol{A}]\|_{max} \leq B_1$ *and* $\|\boldsymbol{A}_t\|_{max} \leq B_2$ *almost surely, with* $\|\cdot\|_{max}$ *being the matrix max norm. Define* $\boldsymbol{\epsilon}_t(\boldsymbol{\theta}) := \hat{\nabla} f_t(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta})$ *and* $c := \max_{\boldsymbol{\theta} \in \mathcal{C}} \|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|_1$. *With probability at least* $1 - (1 + 1/n)(\pi^2 \Delta/6)$, *the following holds for all* $\boldsymbol{\theta} \in \mathcal{C}$ *and all* $t \geq 1$*:*

$$\|\boldsymbol{\epsilon}_t(\boldsymbol{\theta})\|_\infty \leq (cB_1 + \sqrt{mB_2\sigma_w^2})\sqrt{\frac{2(\log(2n^2t^2) - \log \Delta)}{t}} \,, \tag{24}$$

where $\|\cdot\|_\infty$ *is the infinity norm and the dual norm of* $\|\cdot\|_1$.

Consequently, we observe that H 1 is satisfied with η_t^Δ asymptotically equivalent to $4\sqrt{\log(t)}$. The convergence analysis from Theorem 1 applies for online LASSO.

Remark 2. *(Langford et al., 2009) has applied a stochastic proximal gradient (sPG) method to solve an* ℓ_1 *regularized variant of* (22) *for empirical risk minimization. From their result, if the number of iterations run* T *is known a-priori, then the convergence rate is* $\mathcal{O}(1/\sqrt{T})$. *This is similar to the sFW algorithm, yet the sFW algorithm has a better sparsity-accuracy tradeoff; see (Jaggi, 2013).*

3.2 Online Matrix Completion

We consider the matrix completion problem when the observation distribution belongs to the natural exponential family. This distribution family is general enough to encompass many distributions encountered in practice. For example, Gaussian matrix completion can be used for recommender systems (Candès and Plan, 2010; Negahban and Wainwright, 2012; Klopp, 2014), Poisson matrix completion for image recovery (Cao and Xie, 2015) and logistic for one-bit completion (Davenport et al., 2012). We are sequentially given i.i.d. observations in the form (k_t, l_t, Y_t),

with $(k_t, l_t) \in [m_1] \times [m_2]$ and $Y_t \in \mathbb{R}$. The conditional distribution of Y_t *w.r.t.* the sampling is parametrized by an unknown matrix $\bar{\boldsymbol{\theta}} \in \mathbb{R}^{m_1 \times m_2}$ and supposed to belong to the exponential family. More precisely, the conditional density of Y_t is given by

$$p_{\bar{\boldsymbol{\theta}}}(Y_t | k_t, l_t) := \mathrm{m}(Y_t) \exp\left(Y_t \bar{\boldsymbol{\theta}}_{k_t, l_t} - \mathrm{A}(\bar{\boldsymbol{\theta}}_{k_t, l_t})\right) , \tag{25}$$

where $\mathrm{m}(\cdot)$ and $\mathrm{A}(\cdot)$ are the base measure and log-partition functions, respectively.

For any $\boldsymbol{\theta} \in \mathbb{R}^{m_1 \times m_2}$, the Kullback-Leibler divergence between the distribution associated to $\boldsymbol{\theta}$ and the true distribution is (up to an irrelevant additive constant):

$$f(\boldsymbol{\theta}) := \mathbb{E}_{\bar{\boldsymbol{\theta}}}\left[\mathrm{A}(\boldsymbol{\theta}_{k_1, l_1}) - Y_1 \boldsymbol{\theta}_{k_1, l_1}\right] . \tag{26}$$

Note that in the example of Gaussian noise, $f(.)$ is simply the square loss function. We are interested in minimizing the Kullback-Leibler divergence (26) under a penalty favoring low rank solutions. In the following we consider a trace norm penalty $\|\cdot\|_{\sigma,1}$ defined as the sum of the singular values. In such case the problem boils down to

$$\min_{\boldsymbol{\theta} \in \mathcal{C}_R} f(\boldsymbol{\theta}) \quad \text{with} \quad \mathcal{C}_R := \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_{\sigma,1} \leq R\} , \tag{27}$$

We propose the following online algorithm to (27):
Algorithm. At time $t \geq 1$, we perform:

1. Receive a new observation (k_t, l_t, Y_t) and update the sufficient statistics as follows:

$$S_t^{(1)} = (1 - t^{-1}) S_{t-1}^{(1)} + t^{-1} Y_t e_{k_t} e_{l_t}^{\prime \top}$$
$$S_t^{(2)} = (1 - t^{-1}) S_{t-1}^{(2)} + t^{-1} e_{k_t} e_{l_t}^{\prime \top}$$

   with $(e_k)_{k=1}^{m_1}$ (*resp.* $(e'_l)_{l=1}^{m_2}$) the canonical basis of of $\mathbb{R}^{m_1}$ (*resp.* $\mathbb{R}^{m_2}$).

2. Evaluate the gradient estimate as:

$$[\hat{\nabla} f_t(\boldsymbol{\theta})]_{k,l} = \mathrm{A}'(\boldsymbol{\theta}_{k,l}) [S_t^{(2)}]_{k,l} - [S_t^{(1)}]_{k,l} ,$$

   for all $k, l \in [m_1] \times [m_2]$.

3. Set $\boldsymbol{a}_t = -R \boldsymbol{u}_1 \boldsymbol{v}_1^\top$, where $\boldsymbol{u}_1, \boldsymbol{v}_1$ are the top singular vectors of $\hat{\nabla} f_t(\boldsymbol{\theta}_t)$.

4. Update $\boldsymbol{\theta}_{t+1} = (1 - \gamma_t) \boldsymbol{\theta}_t + \gamma_t \boldsymbol{a}_t$.

In order to control the gradient estimation error, we assume the following on the noise.

A1. *The noise variance is finite, that is there exists a constant* $\bar{\sigma} > 0$ *such that for all* $\vartheta \in \mathbb{R}$, $0 \leq \mathrm{A}''(\vartheta) \leq \bar{\sigma}^2$, *and the noise is sub-exponential* i.e., *there exist a constant* $\lambda \geq 1$ *such that for all* $(k, l) \in [m_1] \times [m_2]$*:*

$$\int \exp\left(\lambda^{-1} \left|y - \mathrm{A}'(\bar{\boldsymbol{\theta}}_{k,l})\right|\right) p_{\bar{\boldsymbol{\theta}}}(y | k, l) dy \leq \mathrm{e} , \tag{28}$$

where $p_{\bar{\boldsymbol{\theta}}}(\cdot)$ *is define in* (25) *and* e *is the natural number.*

A2. *There exists a finite constant $\kappa > 0$ such that for all $\boldsymbol{\theta} \in \mathcal{C}$, $k \in [m_1]$, $l \in [m_2]$*

$$\kappa \geq \max \left(\sqrt{\sum_{l=1}^{m_2} \mathrm{A}'(\boldsymbol{\theta}_{k,l})^2}, \sqrt{\sum_{k=1}^{m_1} \mathrm{A}'(\boldsymbol{\theta}_{k,l})^2} \right) . \tag{29}$$

Note that A 1 and A 2 are satisfied by all the above mentioned exponential family distributions. Define $d := m_1 + m_2$, the approximation error can be controlled as follows.

Proposition 2. *Assume A 1, A 2 and that the sampling distribution is uniform. Define the approximation error $\boldsymbol{\epsilon}_t(\boldsymbol{\theta}) := \hat{\nabla} f_t(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta})$. With probability at least $1 - \Delta$, for any $t \geq T_\Delta := (\lambda/\bar{\sigma})^2 \log^2(\lambda/\bar{\sigma}) \log(d + 2d/\Delta)$, and any $\boldsymbol{\theta} \in \mathcal{C}_R$:*

$$\|\boldsymbol{\epsilon}_t(\boldsymbol{\theta})\|_{\sigma,\infty} = \mathcal{O}\left(c_\lambda \, (\kappa + \bar{\sigma}) \sqrt{\frac{\log(d(1 + t^2/\Delta))}{t(m_1 \wedge m_2)}} \right) ,$$

with $\|\cdot\|_{\sigma,\infty}$ the operator norm, c_λ a constant which depends only on λ and where λ, $\bar{\sigma}$ and κ are defined in A 1 and A 2.

Remark 3. *Here we assume that the sampling is uniform only for simplicity. More general sampling scheme can be controlled similarly under additional assumptions (see* e.g., *(Klopp, 2014)).*

In light of Proposition 2, if the sFW algorithm is used after T_Δ samples are observed, then H 1 is satisfied and the results from Theorem 1 applies.

3.3 Large-scale learning with empirical risk

In this example, our aim is to minimize the following empirical risk:

$$\min_{\boldsymbol{\theta} \in \mathcal{C}} F_T(\boldsymbol{\theta}) := \frac{1}{T} \sum_{s=1}^{T} f_s(\boldsymbol{\theta}), \tag{30}$$

where $\mathcal{C} \subseteq \mathbb{R}^n$. We assume that the function $F_T(\cdot)$ is convex and the gradient $\nabla f_t(\cdot)$ is L-Lipschitz continuous with respect to $\|\cdot\|_2$. Problem (30) has been considered by (Zinkevich, 2003; Hazan and Kale, 2012).

We propose the following sFW algorithm for (30), which can be seen as a randomized version of the online Frank-Wolfe (OFW) algorithm in (Hazan and Kale, 2012).

Algorithm. At time $t \geq 1$, we perform

1. Randomly draw a new integers $i(t)$ uniformly from $[T]$ without replacement. Evaluate the gradient estimate:
$$\hat{\nabla} F_t(\boldsymbol{\theta}_t) = \frac{1}{t} \sum_{s=1}^{t} \nabla f_{i(s)}(\boldsymbol{\theta}_t), \tag{31}$$
2. Set $\boldsymbol{a}_t = \arg\min_{\boldsymbol{a} \in \mathcal{C}} \langle \boldsymbol{a}, \hat{\nabla} F_t(\boldsymbol{\theta}_t) \rangle$.
3. Update $\boldsymbol{\theta}_{t+1} = (1 - \gamma_t)\boldsymbol{\theta}_t + \gamma_t \boldsymbol{a}_t$.

By applying a result from (Shalev-Shwartz et al., 2009), the error in the gradient estimate can be bounded as:

Proposition 3. *Define* $\boldsymbol{\epsilon}_t(\boldsymbol{\theta}) := \hat{\nabla} F_t(\boldsymbol{\theta}) - \nabla F_T(\boldsymbol{\theta})$. *With probability at least* $1 - \Delta$, *the following holds for all* $\boldsymbol{\theta} \in \mathcal{C} \subset \mathbb{R}^n$ *and for all* $1 \leq t \leq T$,

$$\|\boldsymbol{\epsilon}_t(\boldsymbol{\theta})\|_\infty = \mathcal{O}\Big(\rho L \sqrt{\frac{n \log(1+t) \log(nt/\Delta)}{t}}\Big) \tag{32}$$

where ρ *is the diameter of* $\mathcal{C}$; *and* $\|\boldsymbol{\epsilon}_t(\boldsymbol{\theta})\|_\infty = 0$ *if* $t > T$.

Following Proposition 3, the error $\boldsymbol{\epsilon}_t(\boldsymbol{\theta})$ satisfies H 1 and since the gradient is L Lipschitz, H 2 is also satisfied. The convergence results from Theorem 1 can be applied. Notice that the error bound is independent of T and depends only on $\sqrt{n \log(n)}$.

Remark 4. *A deterministic algorithm for* (30) *with Lipschitz continuous function* $f_t(\cdot)$ *was studied by (Hazan and Kale, 2012). We have shown that a randomized version of their algorithm converges at a rate of* $\mathcal{O}(\sqrt{\log(t)/t})$ *(compared to* $\mathcal{O}(1/\sqrt{t})$*); moreover, if H3, H4 are satisfied by* $F_T(\boldsymbol{\theta})$, *then the algorithm converges at* $\mathcal{O}(\log(t)/t)$ *(compared to* $\mathcal{O}(1/t^{2/3})$*).*

4 Numerical Experiments

We conduct numerical experiments to demonstrate the practical performance of the sFW algorithm. We focus on the three examples given in section 3.

4.1 Online sparse learning

Data model. We consider a randomly generated data model for online sparse learning. In particular, $\boldsymbol{A}_t = \boldsymbol{A}$ is fixed for all t with dimension 60×100 and the parameter $\bar{\boldsymbol{\theta}} \in \mathbb{R}^{100}$ is a sparse vector with 15% sparsity and independent $\mathcal{N}(0,1)$ elements. We also set $\sigma_w = 10$. We consider two models for $\boldsymbol{A}$ — when $\boldsymbol{A}$ is generated as (i) a random Gaussian matrix with independent $\mathcal{N}(0,1)$ elements; (ii) $\mathbf{1}\mathbf{1}^\top + \boldsymbol{G}$, where $\boldsymbol{G}$ is a random Gaussian matrix with independent $\mathcal{N}(0, 10^{-4})$ elements. Notice that although $\boldsymbol{A}$ is not full rank in the both cases, but the matrix generated in model (i) satisfies the restricted isometry property (RIP) (Candès and Tao, 2005) of order greater than $2\|\bar{\boldsymbol{\theta}}\|_0$ with high probability. Therefore, model (i) roughly corresponds to the situation when f is strongly convex. On the other hand, the $\boldsymbol{A}$ in model (ii) do not satisfy the RIP of desired order and the model corresponds to a non-strongly convex f.

Implementation. For benchmarking purpose, we have compared the sFW's performance with the stochastic projected gradient (sPG) method (Rosasco et al., 2014). The sPG algorithm was implemented with a fixed step size $\gamma_t = 1/L$ using the same gradient estimate in (23). Both algorithms were programmed in `Python 2.7` with `numpy`.

Results. Figure 1 plots the primal optimality $h_t := f(\boldsymbol{\theta}_t) - f(\boldsymbol{\theta}^\star)$ with the iteration number t when $\boldsymbol{A}$ is generated with model (i) and model (ii).

Convergence rate. We focus on the primal convergence rate between the two settings of r in Figure 1. For the case when $r = 0.15\|\bar{\boldsymbol{\theta}}\|_1$, we notice that the primal convergence rate is slower than $\mathcal{O}(1/t)$; while for the case with $r = 1.1\|\bar{\boldsymbol{\theta}}\|_1$, the convergence rate is clearly $\mathcal{O}(1/t)$. The latter case corresponds to the scenario under H 4 as $\boldsymbol{\theta}^\star = \bar{\boldsymbol{\theta}} \in \mathcal{C}$ and $\boldsymbol{\theta}^\star$ belongs to the interior of $\mathcal{C}$.

The simulation result corroborates with our analysis in Theorem 1, which indicates an accelerated convergence rate. For Figure 1 (Bottom), we see that H 3 is not satisfied by the problem yet H 4 holds. In this case, we observe that the convergence rate is roughly $\mathcal{O}(\sqrt{1/t})$, indicating that the accelerated convergence is sensitive to strong convexity.

Step size rule. We compare the primal convergence with $K = 1, 2$ in Algorithm 1. While the asymptotic convergence rate is not sensitive to the choice of K, Figure 1 indicates that choosing $K = 2$ can better track the noise in the data. This can be explained by noticing that a higher weight was put on the direction $\boldsymbol{a}_t - \boldsymbol{\theta}_t$ generated from the current observation.

Comparison with sPG. The primal convergence rate of sPG is similar to sFW. However, the per-iteration complexity of sPG is $\mathcal{O}(n \log n)$, while it is $\mathcal{O}(n)$ for the sFW.

4.2 Online matrix completion

Data model. We consider a rank-30, 100×100 matrix of the form $\bar{\boldsymbol{\theta}} := \frac{1}{\sqrt{30}} \sum_{i=1}^{30} u_i v_i^\top$, where u_i and v_i entries are independent $\mathcal{N}(0, 1)$ elements. In both cases, the observation indexes (k_s, l_s) were sampled uniformly over $[100] \times [100]$ and the values Y_s were generated according to a Gaussian distribution of mean $\bar{\boldsymbol{\theta}}_{k_s,l_s}$ and standard deviation 3 *i.e.,* $Y_s \sim \mathcal{N}(\bar{\boldsymbol{\theta}}_{k_s,l_s}, 3)$. In this scenario, the objective function is given by: $f(\boldsymbol{\theta}) := \mathbb{E}[(Y_s - \boldsymbol{\theta}_{k_s,l_s})^2]$. The constraint set radius R has been chosen as $R = 1.1\|\bar{\boldsymbol{\theta}}\|_{\sigma,1}$, so that the optmimal solution lies in the interior.

Step size rule. The stepsize was chosen with $K = 2$.

Comparison with sPG. As for the online Lasso experiments, the results are benchmarked against the stochastic projected gradient. Note that a crucial difference between the two is the iteration cost. Indeed, sPG requires to perform a full SVD at each step of complexity $\mathcal{O}((m_1+m_2)^3)$, whereas the sFW only requires to compute the top singular eigen vectors of the gradient. Since the gradient at step t is at most t-sparse, this can be performed in $\mathcal{O}(t(m_1 + m_2))$ complexity using Lanczos iteration algorithm.

Results. Figure 2 plots the primal optimality $h_t := f(\boldsymbol{\theta}_t) - f(\boldsymbol{\theta}^\star)$ with the iteration number (Top) and the execution time (Bottom).

Rate of convergence In this simulation, the optimum lies in the interior of the constraint set and the objective is strongly convex. As expected by the theoretical results (up to a logarithmic factor), the rate of convergence of the sFW algorithm is $\mathcal{O}(1/t)$. The same rate is achieved by the sPG algorithm (see Figure 2, Top). When the primal optimality is plotted against the time of execution, the sFW algorithm converges faster than the sPG gradient algorithm (see Figure 2, Bottom).

4.3 Large scale learning

Dataset. We consider learning a sparse parameter $\boldsymbol{\theta}$ from the dataset `R64.mat` available from (Duarte et al., 2008). The dataset consists of $T_{tot} = 4319$ one-bit measurements of an image of size 64×64. The squared loss function is chosen such that $f_t(\boldsymbol{\theta}) = (y_t - \boldsymbol{a}_t^\top \boldsymbol{\theta})^2$, where $\boldsymbol{a}_t \in \mathbb{R}^n$ is a binary measurement vector and $n = 4096$ is the vectorized image. The constraint set is chosen as $\mathcal{C} = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_1 \leq r\}$ similar to the sparse learning example.

Implementation. We have compared the OFW algorithm (Hazan and Kale, 2012) with our sFW algorithm. Both algorithms were programmed using `MATLAB`. For the algorithm in subsection 3.3, we have (i) used batch processing by drawing 5 new integers (without replacement) in step 1 and (ii) introduced an inner loop to the algorithm by repeating step 2 to 3 for 100 times

within each iteration. Similar batch processing and inner loop acceleration are applied to OFW for fair comparison.

Results. Figure 3 compares the primal objective value $F_T(\boldsymbol{\theta}_t)$ against the iteration number. We also show the reconstructed image after $t_f = 500$ iterations of the sFW and OFW algorithm.

Comparison with OFW. Figure 3 shows that the convergence of primal objective of sFW is similar to that of OFW. Both algorithms converge at a rate of $\mathcal{O}(1/t)$. While our analysis for sFW has successfully predicted the convergence rate, we remark that the analysis for OFW in (Hazan and Kale, 2012) can only guarantee a convergence rate of $\mathcal{O}(1/t^{2/3})$.

References

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202.

Bottou, L. (1998). Online learning and stochastic approximations. In Saad, D., editor, *Online Learning and Neural Networks*. Cambridge University Press.

Candès, E. J. and Plan, Y. (2010). Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936.

Candès, E. J. and Tao, T. (2005). Decoding by linear programming. *IEEE Trans. Inf. Theory*, 51(12):4203–4215.

Cao, Y. and Xie, Y. (2015). Poisson matrix recovery and completion. *CoRR*, abs/1504.05229.

Chandrasekaran, V., Recht, B., Parrilo, P. A., , and Willsky, A. S. (2012). The convex geometry of linear inverse problems. *Found. Comp. Math.*, 12(6):805–849.

Davenport, M. A., Plan, Y., van den Berg, E., and Wootters, M. (2012). 1-bit matrix completion. *CoRR*, abs/1209.3672.

Duarte, M., Davenport, M., Takhar, D., Laska, J., Sun, T., Kelly, K., and Baraniuk, R. (2008). Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine*, 25(2):83–91.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *JMLR*, 12:2121–2159.

Frank, M. and Wolfe, P. (1956). An algorithm for quadratic programming. *Naval Res. Logis. Quart.*

Freund, R. M. and Grigas, P. (2013). New analysis and results for the frank-wolfe method. *CoRR*, abs/1307.0873v2.

Garber, D. and Hazan, E. (2015). Faster rates for the frank-wolfe method over strongly-convex sets. *ICML*.

Garrigues, P. J. and El Ghaoui, L. (2008). An homotopy algorithm for the lasso with online observations. *NIPS*.

Ghosh, S. and Lam, H. (2015). Computing worst-case input models in stochastic simulation. *CoRR*, abs/1507.05609.

Hazan, E. and Kale, S. (2012). Projection-free online learning. *ICML*.

Horn, R. A. and Johnson, C. R. (1994). *Topics in matrix analysis*. Cambridge University Press, Cambridge. Corrected reprint of the 1991 original.

Jaggi, M. (2013). Revisiting frank-wolfe: Projection-free sparse convex optimization. *ICML*.

Juditsky, A. B. and Nemirovski, A. S. (2012a). *First-Order Methods for Nonsmooth Convex Large-Scale Optimization, I: General Purpose Methods*.

Juditsky, A. B. and Nemirovski, A. S. (2012b). *First-Order Methods for Nonsmooth Convex Large-Scale Optimization, II: Utilizing Problem's Structure*.

Klopp, O. (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 2(1):282–303.

Koltchinskii, V. (2013). *A remark on low rank matrix recovery and noncommutative Bernstein type inequalities*, volume Volume 9 of *Collections*, pages 213–226. Institute of Mathematical Statistics.

Lacoste-Julien, S. and Jaggi, M. (2013). An affine invariant linear convergence analysis for frank-wolfe algorithms. *NIPS*.

Lacoste-Julien, S., Jaggi, M., Schmidt, M., and Pletscher, P. (2013). Block-coordinate frank-wolfe optimization for structural svms. *ICML*.

Langford, J., Li, L., and Zhang, T. (2009). Sparse online learning via truncated gradient. *NIPS*.

Negahban, S. and Wainwright, M. J. (2012). Restricted strong convexity and weighted matrix completion: optimal bounds with noise. *J. Mach. Learn. Res.*, 13.

Recht, B. and Re, C. (2013). Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226.

Rosasco, L., Villa, S., and Vu, B. C. (2014). Convergence of Stochastic Proximal Gradient Algorithm. *CoRR*, abs/1403.5074v3.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2009). Stochastic convex optimization. *COLT*.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York.

Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. *ICML*.

A Additional Results

Proof of Theorem 1

$\mathcal{O}(\sqrt{\eta_t^\Delta/t})$ bound: Let us define $h_t = f(\boldsymbol{\theta}_t) - f(\boldsymbol{\theta}^\star)$, then by H 2 we get

$$h_{t+1} \leq h_t + \gamma_t \langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{a}_t - \boldsymbol{\theta}_t \rangle + \frac{1}{2}\gamma_t^2 C_f \ . \tag{33}$$

On the other hand, the following also hods:

$$\begin{aligned}
\langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{a}_t - \boldsymbol{\theta}_t \rangle &= \langle \hat{\nabla} f(\boldsymbol{\theta}_t), \boldsymbol{a}_t - \boldsymbol{\theta}_t \rangle - \langle \boldsymbol{\epsilon}_t, \boldsymbol{a}_t - \boldsymbol{\theta}_t \rangle, \\
&\leq \langle \hat{\nabla} f(\boldsymbol{\theta}_t), \boldsymbol{\theta}^\star - \boldsymbol{\theta}_t \rangle - \langle \boldsymbol{\epsilon}_t, \boldsymbol{a}_t - \boldsymbol{\theta}_t \rangle \\
&= \langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{\theta}^\star - \boldsymbol{\theta}_t \rangle + \langle \boldsymbol{\epsilon}_t, \boldsymbol{\theta}^\star - \boldsymbol{a}_t \rangle \\
&\leq -h_t + \rho \|\boldsymbol{\epsilon}_t\| \ .
\end{aligned} \tag{34}$$

where the second line follows from the definition of $\boldsymbol{a}_t$ and the last inequality is due to the convexity of f and the definition of the diameter (4). Plugging (34) into (12) and using H1 yields the following with probability at least $1 - \Delta$ and for all $t \geq 1$

$$h_{t+1} \leq (1 - \gamma_t) h_t + \gamma_t \rho \sigma \sqrt{\frac{\eta_t^\Delta}{t}} + \frac{1}{2}\gamma_t^2 C_f \ . \tag{35}$$

We now proceed by induction to prove the first bound of the Theorem. Define

$$D' = 2(K\rho\sigma + K^2 C_f/2)/(2K - 1) \ .$$

The initialization is done by applying (35) with $t = 1$. Assume that $h_t \leq D'\sqrt{\eta_t^\Delta/(K + t - 1)}$ for some $t \geq 1$. Since by definition $\gamma_t = K/(t + K - 1)$, from (35) we get:

$$\begin{aligned}
&h_{t+1} - D'\sqrt{\frac{\eta_{t+1}^\Delta}{K + t}} \\
&\leq D'\Big(\sqrt{\frac{\eta_t^\Delta}{K + t - 1}} - \sqrt{\frac{\eta_{t+1}^\Delta}{K + t}}\Big) + \Big(-D'K\sqrt{\eta_t^\Delta} + K\rho\sigma + \frac{K^2 C_f}{2\sqrt{K + t - 1}}\Big)/(K + t - 1)^{3/2} \\
&\leq \Big(\frac{D'}{2} - D'K + K\rho\sigma + \frac{K^2 C_f}{2}\Big)\sqrt{\eta_t^\Delta}/(K + t - 1)^{3/2} \leq 0
\end{aligned}$$

where we used the fact that η_t^Δ is increasing and larger that 1 for the second indequality.

$\mathcal{O}(\eta_t^\Delta/t)$ bound: Define $g_t = \max_{\boldsymbol{s} \in \mathcal{C}} \langle \boldsymbol{\theta}_t - \boldsymbol{s}, \nabla f(\boldsymbol{\theta}_t) \rangle$ as the duality gap at $\boldsymbol{\theta}_t$. Define $\boldsymbol{s}_t \in \arg\max_{\boldsymbol{s} \in \mathcal{C}} \langle \boldsymbol{\theta}_t - \boldsymbol{s}, \nabla f(\boldsymbol{\theta}_t) \rangle$. By the same arguments used to derive (34) we have

$$\begin{aligned}
\langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{a}_t - \boldsymbol{\theta}_t \rangle &\leq \langle \hat{\nabla} f(\boldsymbol{\theta}_t), \boldsymbol{s}_t - \boldsymbol{\theta}_t \rangle - \langle \boldsymbol{\epsilon}_t, \boldsymbol{a}_t - \boldsymbol{\theta}_t \rangle, \\
&= \langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{s}_t - \boldsymbol{\theta}_t \rangle + \langle \boldsymbol{\epsilon}_t, \boldsymbol{s}_t - \boldsymbol{a}_t \rangle, \\
&= -g_t + \rho \|\boldsymbol{\epsilon}_t\| \leq -\delta\sqrt{2\mu h_t} + \rho \|\boldsymbol{\epsilon}_t\| \ .
\end{aligned} \tag{36}$$

where the last line follows from Lemma 1. Plugging (36) into (33) yields the following with probability at least $1 - \Delta$ and for all $t \geq 1$,

$$h_{t+1} \leq h_t - \gamma_t \delta \sqrt{2\mu h_t} + \gamma_t \rho \sigma \sqrt{\frac{\eta_t^\Delta}{t}} + \frac{1}{2}\gamma_t^2 C_f \ . \tag{37}$$

Let us define

$$D = 3(K+1)/K^3(\rho\sigma K + C_f K^2/2)^2/(\delta^2\mu) \ ,$$

and proceed by induction. Suppose that $h_t \leq D\eta_t^\Delta/t$ for some $t \geq 1$. There are two cases. If $h_t - \gamma_t\delta\sqrt{2\mu h_t} \leq 0$, then since $\gamma_t = K/(K+t-1)$, (37) yields

$$\begin{aligned} h_{t+1} &\leq \rho\sigma K\frac{\sqrt{\eta_t^\Delta}}{(K+t-1)^{3/2}} + \frac{C_f K^2}{2(K+t-1)^2} \\ &\leq (\rho\sigma K + C_f K^2/2)\frac{\eta_{t+1}^\Delta}{(K+t-1)^{3/2}} \ , \\ &\leq (\rho\sigma K + C_f K^2/2)\frac{K+1}{K}\frac{\eta_{t+1}^\Delta}{(K+t)} \end{aligned}$$

where we used that η_t^Δ is increasing and larger than 1. To conclude, one just needs to check that

$$(\rho\sigma K + C_f K^2/2)\frac{K+1}{K} \leq D \ ,$$

or equivalently

$$3/(K^2\delta^2\mu)(\rho\sigma K + C_f K^2/2) \geq 1 \ . \tag{38}$$

But by Lemma 1, since $C_f \geq \delta^2\mu$ this is always satifsfied. Hence

$$h_{t+1} \leq D\eta_{t+1}^\Delta/(K+t) \ .$$

Otherwise, if $h_t - \gamma_t\delta\sqrt{2\mu h_t} > 0$, then by induction (37) gives

$$\begin{aligned} &h_{t+1} - D\frac{\eta_{t+1}^\Delta}{K+t} \\ &\leq D(\frac{\eta_t^\Delta}{K+t-1} - \frac{\eta_{t+1}^\Delta}{K+t}) + (\rho\sigma K\sqrt{\eta_t^\Delta} + \frac{K^2 C_f}{2\sqrt{K+t-1}} - \delta K\sqrt{2\mu D\eta_t^\Delta})/(K+t-1)^{3/2} \\ &\leq \frac{D\eta_t^\Delta}{(K+t-1)^2} + (\rho\sigma K\sqrt{\eta_t^\Delta} + \frac{K^2 C_f\sqrt{\eta_t^\Delta}}{2} - \delta K\sqrt{2\mu D\eta_t^\Delta})/(K+t-1)^{3/2} \ , \end{aligned} \tag{39}$$

where we used the fact that η_t^Δ is increasing and larger than 1 and $t \geq 1$. Define $t_0 := \inf\{t \geq 1 | K+t-1 \geq D\eta_t^\Delta/(\delta^2\mu)\}$. Since $\eta_t^\Delta/(K+t-1)$ is decreasing to 0, t_0 exists and for any $t \geq t_0$ we have

$$h_{t+1} - D\frac{\eta_{t+1}^\Delta}{K+t} \leq (K\rho\sigma + \frac{K^2 C_f}{2} - (\sqrt{2}K-1)\delta\sqrt{\mu D})\frac{\sqrt{\eta_t^\Delta}}{(K+t-1)^{3/2}} \ . \tag{40}$$

We conclude by noticing that

$$3\frac{K+1}{K^3} \geq \frac{1}{(\sqrt{2}K-1)^2}$$

holds and therefore the RHS of (40) is nonpositive. We now consider the case when $t \leq t_0$. By Lemma 1, $C_f \geq \delta^2\mu$, therefore applying (35) for $t=1$ yields $h_2 \leq D/2 \leq D\eta_2^\Delta/2$. Finally, for $t \leq t_0$ we have $t \leq D\eta_t^\Delta/(\delta^2\mu)$ and therefore

$$\delta\sqrt{\mu}\sqrt{\frac{D\eta_t^\Delta}{K+t-1}} \leq \frac{D\eta_t^\Delta}{K+t-1} \ . \tag{41}$$

Since $D' \leq \delta\sqrt{\mu D}$, the left hand side in the above is lower bounded by $D'\sqrt{\eta_t^\Delta/(K+t-1)}$ and we have $D'\sqrt{\eta_t^\Delta/(K+t-1)} \leq D\eta_t^\Delta/t$.

Proof of Proposition 1

Notice that the gradient vector is given by:

$$\nabla f(\boldsymbol{\theta}) = \mathbb{E}[\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{Y})] = \mathbb{E}[\boldsymbol{A}^\top\boldsymbol{A}]\boldsymbol{\theta} - \mathbb{E}[\boldsymbol{A}^\top\boldsymbol{Y}]. \tag{42}$$

By noting that $\overline{(\boldsymbol{A}^\top\boldsymbol{Y})}_t = \overline{(\boldsymbol{A}^\top\boldsymbol{A})}_t\bar{\boldsymbol{\theta}} + (1/t)\sum_{s=1}^t \boldsymbol{A}_s^\top\boldsymbol{w}_s$ and $\mathbb{E}[\boldsymbol{A}^\top\boldsymbol{Y}] = \mathbb{E}[\boldsymbol{A}^\top\boldsymbol{A}]\bar{\boldsymbol{\theta}}$, we can bound the gradient estimation error as:

$$\|\boldsymbol{\epsilon}_t\|_\infty \leq \Big\|\frac{1}{t}\sum_{s=1}^t \boldsymbol{A}_s^\top\boldsymbol{w}_s\Big\|_\infty + \Big\|\frac{1}{t}\sum_{s=1}^t \big(\boldsymbol{A}_s^\top\boldsymbol{A}_s - \mathbb{E}[\boldsymbol{A}^\top\boldsymbol{A}]\big)(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})\Big\|_\infty \tag{43}$$

To bound the second term in (43), we define $\boldsymbol{Z}_s := \boldsymbol{A}_s^\top\boldsymbol{A}_s - \mathbb{E}[\boldsymbol{A}^\top\boldsymbol{A}]$. Observe that

$$\Big\|\frac{1}{t}\sum_{s=1}^t \boldsymbol{Z}_s(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})\Big\|_\infty = \max_{i\in[n]}\Big|\frac{1}{t}\sum_{s=1}^t \boldsymbol{z}_{s,i}(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})\Big|, \tag{44}$$

where $\boldsymbol{z}_{s,i}$ denotes the ith row vector in $\boldsymbol{Z}_s$. Furthermore, by the Holder's inequality,

$$\big|\frac{1}{t}\sum_{s=1}^t \boldsymbol{z}_{s,i}(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})\big| \leq \|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|_1\big\|\frac{1}{t}\sum_{s=1}^t \boldsymbol{z}_{s,i}\big\|_\infty,$$

Now that $\boldsymbol{z}_{s,i}$ is a zero-mean, independent random vector with elements bounded in $[-B_1, B_1]$, applying the union bound and the Hoefding's inequality gives:

$$\mathbb{P}\big(\big\|(1/t)\textstyle\sum_{s=1}^t \boldsymbol{z}_{s,i}\big\|_\infty \geq x,\ \forall\ i\big) \leq 2n^2\mathrm{e}^{-\frac{x^2t}{2B_1^2}}. \tag{45}$$

Setting $x = B_1\sqrt{2(\log(2n^2t^2) - \log\Delta)/t}$ gives Δ/t^2 on the right hand side. With probability at least $1 - \Delta/t^2$, we have

$$\Big\|\frac{1}{t}\sum_{s=1}^t \boldsymbol{Z}_s\boldsymbol{\theta}\Big\|_\infty \leq cB_1\sqrt{2(\log(2n^2t^2) - \log\Delta)/t}, \tag{46}$$

To bound the first term in (43), we find that the ith element of the vector $\boldsymbol{A}_s^\top\boldsymbol{w}_s$ is zero-mean. Furthermore, it can be verified that

$$\mathbb{E}\big[\mathrm{e}^{\left(\lambda\sum_{j=1}^m A_{s,i,j}w_{s,j}\right)}\big] \leq \mathrm{e}^{\lambda^2\cdot m\sigma_w^2B_2/2}, \tag{47}$$

for all $\lambda \in \mathbb{R}$, where $A_{s,i,j}$ is the (i,j)th element of $\boldsymbol{A}_s$ and $w_{s,j}$ is the jth element of $\boldsymbol{w}_s$. In other words, the ith element of $\boldsymbol{A}_s^\top\boldsymbol{w}_s$ is sub-Gaussian with parameter $m\sigma_w^2B_2$. It follows by the Hoefding's inequality that

$$\mathbb{P}\big(\big\|\frac{1}{t}\sum_{s=1}^t \boldsymbol{A}_s^\top\boldsymbol{w}_s\big\|_\infty \geq x\big) \leq 2n\mathrm{e}^{-\frac{x^2t}{2mB_2\sigma_w^2}}. \tag{48}$$

Setting $x = \sigma_w\sqrt{2mB_2(\log(2n^2t^2) - \log\Delta)/t}$ yields $\Delta/(nt^2)$ on the right hand side. Combining (45), (48) and using a union bound argument yields the desired result.

Proof of Proposition 2

For a fixed $\boldsymbol{\theta}$, by the triangle inequality

$$\|\boldsymbol{\epsilon}_t(\boldsymbol{\theta})\|_{\sigma,\infty} \leq \|\frac{1}{t}\sum_{s=1}^{t} Y_s e_{k_s} e_{l_s}^{'\top} - \mathbb{E}[Y_s e_{k_s} e_{l_s}^{'\top}]\|_{\sigma,\infty} + \|\frac{1}{t}\sum_{s=1}^{t} \mathrm{A}'(\boldsymbol{\theta}_{k_s,l_s}) e_{k_s} e_{l_s}^{'\top} - \mathbb{E}[\mathrm{A}'(\boldsymbol{\theta}_{k_s,l_s}) e_{k_s} e_{l_s}^{'\top}]\|_{\sigma,\infty}$$

Define $Z_s := Y_s e_{k_s} e_{l_s}^{'\top} - \mathbb{E}[Y_s e_{k_s} e_{l_s}^{'\top}]$, then

$$\begin{aligned}
\|\mathbb{E}[Z_s Z_s^\top]\|_{\sigma,\infty} &\leq \|\mathbb{E}[Y_s^2 e_{k_s} e_{l_s}^{'\top} e_{l_s}^{'} e_{k_s}^\top]\|_{\sigma,\infty} \,, \\
&= \|\frac{1}{m_1 m_2} \operatorname{diag}\left(\left(\sum_{l=1}^{m_2} \mathbb{E}[Y_s^2|k,l]\right)_{k=1}^{m_1}\right)\|_{\sigma,\infty} \,, \\
&= \frac{1}{m_1 m_2} \max_{k\in[m_1]} \left(\sum_{l=1}^{m_2} \mathrm{A}''(\bar{\boldsymbol{\theta}}_{k,l}) + (\mathrm{A}'(\bar{\boldsymbol{\theta}}_{k,l}))^2\right) \,, \\
&\leq \frac{\bar{\sigma}^2}{m_1 \wedge m_2} + \frac{\kappa^2}{m_1 m_2} \leq \frac{\bar{\sigma}^2 + \kappa^2}{m_1 \wedge m_2} \,,
\end{aligned}$$

where we used the fact that the distribution belongs to the exponential family for the second equality. Similarly one shows that $\|\mathbb{E}[Z_s^\top Z_s]\|_{\sigma,\infty}$ satisfies the same upper bound. Hence by Proposition 4 and A 1, with probability at least $1 - \mathrm{e}^{-\nu}$, it holds

$$\|\frac{1}{t}\sum_{s=1}^{t} Z_s\|_{\sigma,\infty} \leq c_\lambda \sqrt{\frac{(\bar{\sigma}^2 + \kappa^2)(\nu + \log(d))}{t(m_1 \wedge m_2)}} \,,$$

for t larger than the threshold given in the proposition statement. For the second term, define $P_t := 1/t \sum_{s=1}^{t} e_{k_s} e_{l_s}^{'\top} - (m_1 m_2)^{-1} \mathbf{1}\mathbf{1}^\top$, we get

$$\|\frac{1}{t}\sum_{s=1}^{t} \mathrm{A}'(\boldsymbol{\theta}_{k_s,l_s}) e_{k_s} e_{l_s}^{'\top} - \mathbb{E}[\mathrm{A}'(\boldsymbol{\theta}_{k_s,l_s}) e_{k_s} e_{l_s}^{'\top}]\|_{\sigma,\infty} = \|P_t \odot (\mathrm{A}'(\boldsymbol{\theta}_{k,l}))_{k,l}\|_{\sigma,\infty} \leq \kappa \|P_t\|_{\sigma,\infty} \,,$$

where $\odot$ denotes the Hardamard product and we have used (Horn and Johnson, 1994, Theorem 5.5.3) for the last inequality. Define $Z'_s := e_{k_s} e_{l_s}^{'\top} - (m_1 m_2)^{-1} \mathbf{1}\mathbf{1}^\top$. Since by definition, $\lambda \geq 1$, one can again apply Proposition 4 for $U = \lambda$ and get with probability at least $1 - \mathrm{e}^{-\nu}$,

$$\|P_t\|_{\sigma,\infty} \leq c_\lambda \sqrt{\frac{\nu + \log(d)}{t(m_1 \wedge m_2)}} \,.$$

Hence, by a union bound argument we find that with probability at least $1 - 2\mathrm{e}^{-\nu}$

$$\|\boldsymbol{\epsilon}_t\|_{\sigma,\infty} \leq c_\lambda \left(2\kappa + \bar{\sigma}\right) \sqrt{\frac{\nu + \log(d)}{t(m_1 \wedge m_2)}} \,.$$

Taking $\nu = \log(1 + 2t^2/\Delta)$ and applying a union bound argument yields the result.

Proof of Proposition 3

Observe that $\mathbb{E}[\nabla f_{i(s)}(\boldsymbol{\theta})] = \nabla F_T(\boldsymbol{\theta})$, we can write

$$\boldsymbol{\epsilon}_t(\boldsymbol{\theta}) = \frac{1}{t}\sum_{s=1}^{t} \nabla f_{i(s)}(\boldsymbol{\theta}) - \mathbb{E}[\nabla f_{i(s)}(\boldsymbol{\theta})] \tag{49}$$

Let $\nabla f_{i(s),j}(\boldsymbol{\theta})$ be the jth coordinate of $\nabla f_{i(s)}$. By the assumption, $\nabla f_{i(s),j}(\boldsymbol{\theta})$ is Lipschitz-continuous with parameter L. Applying (Shalev-Shwartz et al., 2009, Theorem 5) shows that with probability at least $1 - x$, we have

$$\sup_{\boldsymbol{\theta}\in\mathcal{C}} |\boldsymbol{\epsilon}_{t,j}(\boldsymbol{\theta})| = \mathcal{O}\Big(\rho L\sqrt{\frac{n\log(t)\log(nt/x)}{t}}\Big). \tag{50}$$

Setting $x = \Delta/(nt^2)$ and applying a union bound argument (for all j and t) yields the desired result.

Proposition 4. *Consider a finite sequence of independent random matrices $(Z_s)_{1\leq s\leq t} \in \mathbb{R}^{m_1\times m_2}$ satisfying $\mathbb{E}[Z_i] = 0$. For some $U > 0$, assume*

$$\inf\{\lambda > 0 : \mathbb{E}[\exp(\|Z_i\|_{\sigma,\infty}/\lambda)] \leq \mathrm{e}\} \leq U \quad \forall i \in [n],$$

and there exists σ_Z s.t.

$$\sigma_Z^2 \geq \max\left\{ \left\| \frac{1}{t}\sum_{s=1}^{t} \mathbb{E}[Z_s Z_s^\top] \right\|_{\sigma,\infty}, \left\| \frac{1}{t}\sum_{s=1}^{t} \mathbb{E}[Z_s^\top Z_s] \right\|_{\sigma,\infty} \right\}.$$

Then for any $\nu > 0$, with probability at least $1 - \mathrm{e}^{-\nu}$

$$\left\| \frac{1}{t}\sum_{i=1}^{t} Z_i \right\|_{\sigma,\infty} \leq c_U \max\left\{ \sigma_Z\sqrt{\frac{\nu+\log(d)}{t}}, U\log(\frac{U}{\sigma_Z})\frac{\nu+\log(d)}{t} \right\},$$

with c_U an increasing constant with U.

Proof. This result is proved in (Koltchinskii, 2013, Theorem 4) for symetric matrices. Here we state a slightly different result because σ_Z^2 is an upper bound of the variance and not the variance itself. However, it does not the alter the proof and the result stays valid. This concentration is extended to rectangular matrices by dilation, see (Klopp, 2014, Proposition 11) for details. □

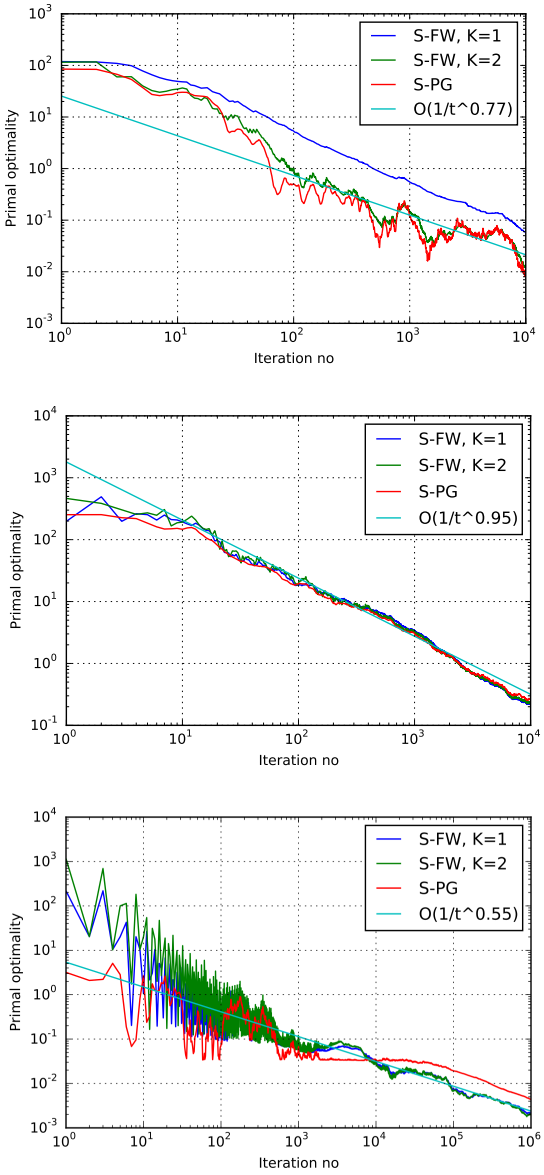

Figure 1: Convergence of the primal optimality for online sparse learning (Top) model (i) and $r = 0.15\|\bar{\boldsymbol{\theta}}\|_1 = \|\boldsymbol{\theta}^\star\|_1$; (Middle) model (i) and $r = 1.1\|\bar{\boldsymbol{\theta}}\|_1 > \|\boldsymbol{\theta}^\star\|_1$; (Bottom) model (ii) and $r = 1.1\|\bar{\boldsymbol{\theta}}\|_1 > \|\boldsymbol{\theta}^\star\|_1$.

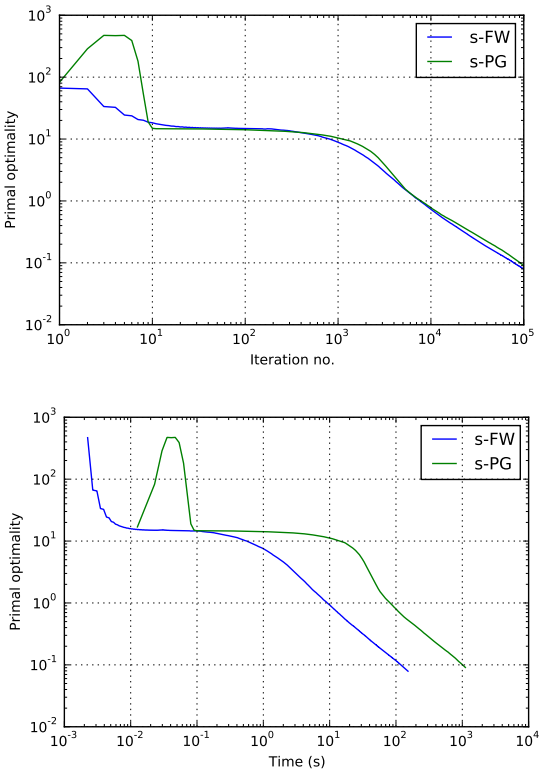

Figure 2: Convergence of the primal optimality for online matrix completion with number of iterations (Top) and time (Bottom).

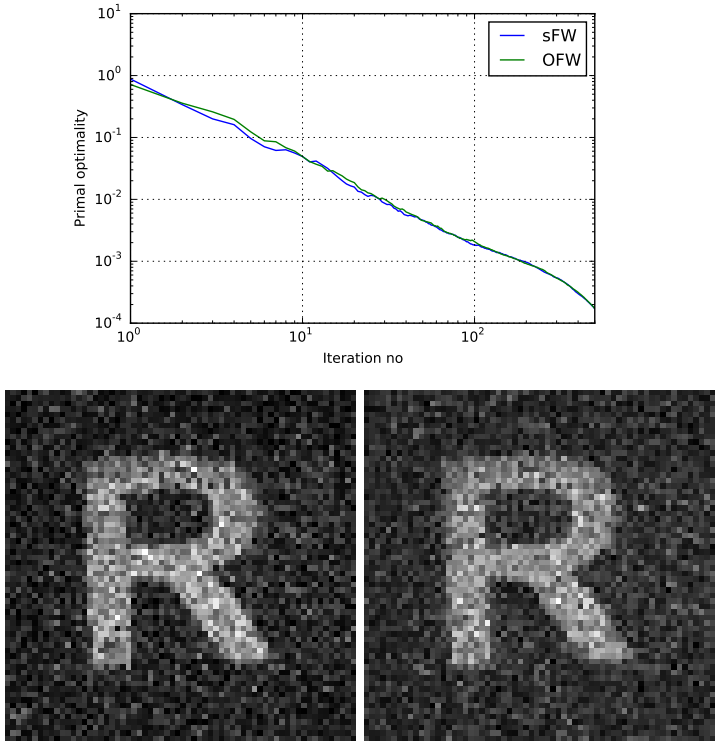

Figure 3: (Top) Convergence of the primal optimality for large-scale learning problem. (Bottom) Reconstructed image after 500 iterations of sFW / OFW (only 2500 measurements are used). (Left) sFW, (Right) OFW