Supplementary Material: Linear Convergence of Stochastic Frank Wolfe Variants

Donald Goldfarb
Columbia University

Garud Iyengar
Columbia University

Chaoxu Zhou
Columbia University

1 Proof of Lemma 2

Proof. To prove the result, we use brackets of the type $[f_\theta - \epsilon g/2, f_\theta + \epsilon g/2]$ for θ that ranging over a suitably chosen subset of Θ and these brackets have L_1-size $\epsilon\|g\|_1$. If $\|\theta_1 - \theta_2\| \le \epsilon/2$, then by the Lipschitz condition that

$$|f_{\theta_1}(\xi) - f_{\theta_2}(\xi)| \le g(\xi)\|\theta_1 - \theta_2\|, \tag{1}$$

we have $f_{\theta_1} - \epsilon g/2 \le f_{\theta_2} \le f_{\theta_1} + \epsilon g/2$. Therefore, the brackets cover $\mathcal{F}$ if θ ranges over a grid of meshwidth $\epsilon/\sqrt{p}$ over Θ. This grid has at most $(\sqrt{p}D_\Theta/\epsilon)^p$ grid points. Therefore the bracketing number $N_{[]}(\epsilon\|g\|_1, \mathcal{F}, L_1)$ can be bounded by $(\sqrt{p}D_\Theta/\epsilon)^p$. □

2 Proof of Lemma 3

Proof. Consider the function class $\mathcal{F} = \{f(\cdot, \mathbf{x}) \mid \mathbf{x} \in \mathcal{P}\}$ as defined in (SP1), that is $f(i, \mathbf{x}) = f_i(\mathbf{x})$. Since $f_i(\cdot)$ each is assumed to be Lipschitz continuous with Lipschitz constant L_i, we must have $|f_i(\mathbf{x}) - f_i(\mathbf{y})| \le L_F\|\mathbf{x} - \mathbf{y}\|$, where $L_F \equiv \max\{L_1, \dots, L_n\}$. Moreover, the index set $\mathcal{P} \in \mathbb{R}^p$ for the function class $\mathcal{F}$ is assume to be bounded. Therefore all conditions for Lemma 2 are satisfied and hence the number of brackets of the type $[f(\cdot, \mathbf{x}) - \epsilon L_F, f(\cdot, \mathbf{x}) + \epsilon L_F]$ satisfies

$$N_{[]}(\epsilon L_F, \mathcal{F}, L_1) \le K_{\mathcal{P}}(\frac{D}{\epsilon})^p,$$

for every $0 < \epsilon < D$, where $D = \sup\{\|\mathbf{x} - \mathbf{y}\| \mid \mathbf{x}, \mathbf{y} \in \mathcal{P}\}$ and $K_{\mathcal{P}} = (\sqrt{p})^p$. Let $\Gamma \subset \mathcal{P}$ denote the set of indices of the centers of these brackets and $\xi_1, \dots \xi_{m^{(k)}}$ be the i.i.d. samples drawn at the k-th iteration of the algorithm. Since the brackets centered at Γ cover $\mathcal{F}$, we must have

$$\begin{aligned}
&\sup_{\mathbf{x}\in\mathcal{P}} |\frac{1}{m^{(k)}} \sum_{i=1}^{m^{(k)}} f(\xi_i, \mathbf{x}) - \mathbb{E}f(\xi_i, \mathbf{x})| \\
&\le \max\{|\frac{1}{m^{(k)}} \sum_{i=1}^{m^{(k)}} f(\xi_i, \mathbf{y}) - \mathbb{E}f(\xi_i, \mathbf{y})| \mid \mathbf{y} \in \Gamma\} + 2\epsilon L_F.
\end{aligned}$$

Consequently, for every $\delta \ge 0$ and $\epsilon < \min\{\delta/(2L_F), D\}$,

$$\begin{aligned}
&\mathbb{P}\{\sup_{\mathbf{x}\in\mathcal{P}} |\frac{1}{m^{(k)}} \sum_{i=1}^{m^{(k)}} f(\xi_i, \mathbf{x}) - \mathbb{E}f(\xi_i, \mathbf{x})| \ge \delta\} \\
&\le \mathbb{P}\{\max\{|\frac{1}{m^{(k)}} \sum_{i=1}^{m^{(k)}} f(\xi_i, \mathbf{y}) - \mathbb{E}f(\xi_i, \mathbf{y})| \mid \mathbf{y} \in \Gamma\} \\
&\quad + 2\epsilon L_F \ge \delta\} \\
&\le \sum_{\mathbf{y}\in\Gamma} \mathbb{P}\{|\frac{1}{m^{(k)}} \sum_{i=1}^{m^{(k)}} f(\xi_i, \mathbf{y}) - \mathbb{E}f(\xi_1, \mathbf{y})| \ge \delta - 2\epsilon L_F\} \\
&\hfill \text{(union bound)} \\
&\le \sum_{\mathbf{y}\in\Gamma} 2\exp\{-\frac{2m^{(k)}(\delta - 2L_F\epsilon)^2}{(u_F - l_F)^2}\} \\
&\hfill \text{(Hoeffding inequality)} \\
&\le 2K_{\mathcal{P}}(\frac{D}{\epsilon})^p \exp\{-\frac{2m^{(k)}(\delta - 2L_F\epsilon)^2}{(u_F - l_F)^2}\}. \\
&\hfill (|\Gamma| \le K_{\mathcal{P}}(\tfrac{D}{\epsilon})^p)
\end{aligned}$$

Since by definition, $F^{(k)}(\mathbf{x}) = \frac{1}{m^{(k)}} \sum_{i=1}^{m^{(k)}} f(\xi_i, \mathbf{x})$ and $F(\mathbf{x}) = \mathbb{E}f(\xi_i, \mathbf{x})$, the desired result follows. □

3 Proof of Corollary 1

Proof. First note that both $F^{(k)}(\cdot)$ and $F(\cdot)$ are bounded by l_F and u_F; hence, $\sup_{\mathbf{x}\in\mathcal{P}} |F^{(k)}(\mathbf{x}) - F(\mathbf{x})| \le 2(|u_F| + |l_F|)$. Then for every $\delta \ge 0$, we have

$$\begin{aligned}
&\mathbb{E}\sup_{\mathbf{x}\in\mathcal{P}} |F^{(k)}(\mathbf{x}) - F(\mathbf{x})| \\
&\le 2(|u_F| + |l_F|)\mathbb{P}\{\sup_{\mathbf{x}\in\mathcal{P}} |F^{(k)}(\mathbf{x}) - F(\mathbf{x})| \ge \delta\} \\
&\quad + \delta\, \mathbb{P}\{\sup_{\mathbf{x}\in\mathcal{P}} |F^{(k)}(\mathbf{x}) - F(\mathbf{x})| < \delta\} \\
&\le 4(|u_F| + |l_F|)K_{\mathcal{P}}(\frac{D}{\epsilon})^p \exp\{-\frac{2m^{(k)}(\delta - 2L_F\epsilon)^2}{(u_F - l_F)^2}\} + \delta \\
&\le 4(|u_F| + |l_F|)K_{\mathcal{P}}D^p \exp\{-\frac{2m^{(k)}(\delta - 2L_F\epsilon)^2}{(u_F - l_F)^2} + p\log\frac{1}{\epsilon}\} + \delta.
\end{aligned}$$

Now let $\delta = \frac{(u_F - l_F)\sqrt{4(p+1)\log\sqrt{m^{(k)}}}}{\sqrt{m^{(k)}}\sqrt{2}}$, $\epsilon = \frac{(u_F - l_F)}{2L_F\sqrt{m^{(k)}}\sqrt{2}}$. Then

$$\begin{aligned}
&\mathbb{E}\sup_{\mathbf{x}\in\mathcal{P}} |F^{(k)}(\mathbf{x}) - F(\mathbf{x})| \\
&\le 4(|u_F| + |l_F|)K_{\mathcal{P}} D^p \exp\{-(\sqrt{4(p+1)\log\sqrt{m^{(k)}}} - 1)^2 \\
&\quad - p(\log\frac{u_F - l_F}{2\sqrt{2}L_F}) + p\log\sqrt{m^{(k)}}\} \\
&\quad + \frac{(u_F - l_F)\sqrt{4(p+1)\log\sqrt{m^{(k)}}}}{\sqrt{m^{(k)}}\sqrt{2}}.
\end{aligned}$$

Note that $(x-1)^2 \ge x^2/4$ when $x \ge 2$. Thus, for $m^{(k)} \ge 3$ and $p \ge 1$, $\sqrt{4(p+1)\log\sqrt{m^{(k)}}} \ge 2$. Therefore

$$\begin{aligned}
&\mathbb{E}\sup_{\mathbf{x}\in\mathcal{P}} |F^{(k)}(\mathbf{x}) - F(\mathbf{x})| \\
&\le 4(|u_F| + |l_F|)K_{\mathcal{P}} D^p \exp\{-(p+1)\log(\sqrt{m^{(k)}}) \\
&\quad + p\log\sqrt{m^{(k)}} - p(\log\frac{u_F - l_F}{2\sqrt{2}L_F})\} \\
&\quad + \frac{(u_F - l_F)\sqrt{4(p+1)\log\sqrt{m^{(k)}}}}{\sqrt{m^{(k)}}\sqrt{2}} \\
&\le C_1\sqrt{\frac{\log m^{(k)}}{m^{(k)}}},
\end{aligned}$$

where $C_1 = 4(|u_F| + |l_F|)K_{\mathcal{P}} D^p \exp\{-p(\log\frac{u_F - l_F}{2\sqrt{2}L_F})\} + (u_F - l_F)\sqrt{p+1}$.

Next, we will obtain a bound for $\mathbb{E}|F^{(k)}(\mathbf{x}_*^{(k)}) - F(\mathbf{x}^*)|$. Lemma 3 implies both

$$F(\mathbf{x}_*^{(k)}) - \delta \le F^{(k)}(\mathbf{x}_*^{(k)}) \le F(\mathbf{x}_*^{(k)}) + \delta \tag{2}$$

and

$$F(\mathbf{x}^*) - \delta \le F^{(k)}(\mathbf{x}^*) \le F(\mathbf{x}^*) + \delta \tag{3}$$

happen with probability at least $1 - 2K_{\mathcal{P}}(\frac{D}{\epsilon})^p \exp\{-\frac{m^{(k)}(\delta - 2L_F\epsilon)^2}{2(u_F - l_F)^2}\}$. Consequently, on one hand

$$\begin{aligned}
F^{(k)}(\mathbf{x}_*^{(k)}) &\ge F(\mathbf{x}_*^{(k)}) - \delta && \text{(by (2))} \\
&\ge F(\mathbf{x}^*) - \delta && \text{(optimality of } \mathbf{x}^* \text{ for } F(\cdot))
\end{aligned}$$

On the other hand,

$$\begin{aligned}
F^{(k)}(\mathbf{x}_*^{(k)}) &\le F^{(k)}(\mathbf{x}^*) && \text{(optimiality of } \mathbf{x}_*^{(k)} \text{ for } F^{(k)}(\cdot)) \\
&\le F(\mathbf{x}^*) + \delta && \text{(by (3))}
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\mathbb{P}\{|F^{(k)}(\mathbf{x}_*^{(k)}) - F(\mathbf{x}^*)| \ge \delta\} \\
&\le 2K_{\mathcal{P}}(\frac{D}{\epsilon})^p \exp\{-\frac{m^{(k)}(\delta - 2L_F\epsilon)^2}{2(u_F - l_F)^2}\},
\end{aligned}$$

and hence $\mathbb{E}|F^{(k)}(\mathbf{x}_*^{(k)}) - F(\mathbf{x}^*)| = C_1\sqrt{\frac{\log m^{(k)}}{m^{(k)}}}$. □

4 Proof of Lemma 4

Proof. The right hand side of the stated result in Lemma 4 is obtained by setting $b_i = 1$ for $i \le m$ and $b_i = 0$ for $i > m$. We will show that this choice of $\{b_i\}$ maximizes $\sum_{k=1}^n a^{\sum_{j=k}^n b_j} c_k$. Consider an assignment of b_i that there is a $b_r = 0$ for $r \le m$ and $b_s = 1$ for $s > m$. Define a new assignment b_i' such that there is $b_i' = b_i$ for $i \ne r, s$, $b_r' = 1$ and $b_s' = 0$. Then

$$\begin{aligned}
&\sum_{k=1}^n a^{\sum_{j=k}^n b_j} c_k \\
&= \sum_{k=s+1}^n a^{\sum_{j=k}^n b_j} c_k + \sum_{k=r}^s a^{\sum_{j=k}^n b_j} c_k + \sum_{k=1}^{r-1} a^{\sum_{j=k}^n b_j} c_k \\
&= \sum_{k=s+1}^n a^{\sum_{j=k}^n b_j'} c_k + \sum_{k=r+1}^s a^{\sum_{j=k}^n b_j} c_k + \sum_{k=1}^{r} a^{\sum_{j=k}^n b_j'} c_k \\
&= \sum_{k=s+1}^n a^{\sum_{j=k}^n b_j'} c_k + a\sum_{k=r+1}^s a^{\sum_{j=k}^n b_j'} c_k + \sum_{k=1}^{r} a^{\sum_{j=k}^n b_j'} c_k \\
&\le \sum_{k=s+1}^n a^{\sum_{j=k}^n b_j'} c_k + \sum_{k=r+1}^s a^{\sum_{j=k}^n b_j'} c_k + \sum_{k=1}^{r} a^{\sum_{j=k}^n b_j'} c_k \\
&= \sum_{k=1}^n a^{\sum_{j=k}^n b_j'} c_k.
\end{aligned}$$

Therefore, such interchanges will always increase the value of $\sum_{k=1}^n a^{\sum_{j=k}^n b_j} c_k$ and hence setting $b_i = 1$ for $i \le m$ and $b_i = 0$ for $i > m$ maximizes it. □

5 Proof of Theorem 1

Proof. At iteration k, let $\mathbf{x}^{(k)}$ denote the current solution, $\xi_1, \ldots, \xi_{m^{(k)}}$ denote the samples obtained in the algorithm, $\mathbf{d}^{(k)}$ denote the direction that the algorithm will take at this step and $\gamma^{(k)}$ denote the step length. Define $F^{(k)}(\mathbf{x}) = \frac{1}{m^{(k)}}\sum_{i=1}^{m^{(k)}} f(\xi_i, \mathbf{x})$, $\mathbf{x}_*^{(k)} = \arg\min_{\mathbf{x}\in\mathcal{P}} F^{(k)}(\mathbf{x})$ and $F_*^{(k)} = F^{(k)}(\mathbf{x}_*^{(k)})$. Note that $F^{(k)}$ is Lipschitz continuous with Lipschitz constant $L^{(k)} = \frac{1}{m^{(k)}}\sum_{i=1}^{m^{(k)}} L_{\xi_i}$ and strongly convex with constant $\sigma^{(k)} = \frac{1}{m^{(k)}}\sum_{i=1}^{m^{(k)}} \sigma_{\xi_i}$. In addition, the stochastic gradient $\mathbf{g}^{(k)} = \nabla F^{(k)}(\mathbf{x})$. From the choice of $\mathbf{d}^{(k)}$ in the algorithm,

$$\begin{aligned}
\langle \mathbf{g}^{(k)}, \mathbf{d}^{(k)}\rangle &\le \frac{1}{2}(\langle \mathbf{g}^{(k)}, \mathbf{p}^{(k)} - \mathbf{x}^{(k)}\rangle + \langle \mathbf{g}^{(k)}, \mathbf{x}^{(k)} - \mathbf{u}^{(k)}\rangle) \\
&= \frac{1}{2}\langle \mathbf{g}^{(k)}, \mathbf{p}^{(k)} - \mathbf{u}^{(k)}\rangle \le 0.
\end{aligned}$$

Hence, we can lower bound $\langle \mathbf{g}^{(k)}, \mathbf{d}^{(k)} \rangle^2$ by

$$\begin{aligned}
\langle \mathbf{g}^{(k)}, \mathbf{d}^{(k)} \rangle^2 &\geq \frac{1}{4} \langle \mathbf{g}^{(k)}, \mathbf{u}^{(k)} - \mathbf{p}^{(k)} \rangle^2 \\
&\geq \frac{1}{4} \max_{\mathbf{p} \in V, \mathbf{u} \in U^{(k)}} \langle \mathbf{g}^{(k)}, \mathbf{u} - \mathbf{p} \rangle^2 \\
&\qquad \text{(definition of } \mathbf{p}^{(k)} \text{ and } \mathbf{u}^{(k)}) \\
&= \frac{1}{4} \max_{\mathbf{p} \in V, \mathbf{u} \in U^{(k)}} \langle \nabla F^{(k)}(\mathbf{x}^{(k)}), \mathbf{u} - \mathbf{p} \rangle^2 \\
&\qquad (\mathbf{g}^{(k)} = \nabla F^{(k)}(\mathbf{x}^{(k)})) \\
&\geq \frac{1}{4} \frac{\Omega_{\mathcal{P}}^2}{|U^{(k)}|^2} \frac{\langle \nabla F^{(k)}(\mathbf{x}^{(k)}), \mathbf{x}^{(k)} - \mathbf{x}_*^{(k)} \rangle^2}{\|\mathbf{x}^{(k)} - \mathbf{x}_*^{(k)}\|^2} \\
&\qquad \text{(by Lemma 1)} \\
&\geq \frac{\Omega_{\mathcal{P}}^2}{4N^2} \frac{\{F^{(k)}(\mathbf{x}^{(k)}) - F_*^{(k)}\}^2}{\|\mathbf{x}^{(k)} - \mathbf{x}_*^{(k)}\|^2} \\
&\qquad \text{(Convexity of } F^{(k)}(\cdot)) \\
&\geq \frac{\Omega_{\mathcal{P}}^2 \sigma^{(k)}}{8N^2} \{F^{(k)}(\mathbf{x}^{(k)}) - F_*^{(k)}\} \\
&\qquad \text{(by strong convexity of } F^{(k)}(\cdot)) \\
&\geq \frac{\Omega_{\mathcal{P}}^2 \sigma_F}{8N^2} \{F^{(k)}(\mathbf{x}^{(k)}) - F_*^{(k)}\}.
\end{aligned}$$

Similarly, we can upper bound $\langle \mathbf{g}^{(k)}, \mathbf{d}^{(k)} \rangle$ by

$$\begin{aligned}
\langle \mathbf{g}^{(k)}, \mathbf{d}^{(k)} \rangle &\leq \frac{1}{2} \langle \mathbf{g}^{(k)}, \mathbf{p}^{(k)} - \mathbf{u}^{(k)} \rangle \\
&\leq \frac{1}{2} \langle \mathbf{g}^{(k)}, \mathbf{x}_*^{(k)} - \mathbf{x}^{(k)} \rangle \\
&\qquad \text{(definition of } \mathbf{p}^{(k)} \text{ and } \mathbf{u}^{(k)}) \\
&= \frac{1}{2} \langle \nabla F^{(k)}(\mathbf{x}^{(k)}), \mathbf{x}_*^{(k)} - \mathbf{x}^{(k)} \rangle \\
&\qquad (\mathbf{g}^{(k)} = \nabla F^{(k)}(\mathbf{x}^{(k)})) \\
&\leq \frac{1}{2} \{F_*^{(k)} - F^{(k)}(\mathbf{x}^{(k)})\}. \\
&\qquad \text{(Convexity of } F(\cdot))
\end{aligned}$$

With the above bounds, we can separate our analysis into the following four cases at iteration k

($A^{(k)}$) $\gamma_{\max}^{(k)} \geq 1$ and $\gamma^{(k)} \leq 1$.

($B^{(k)}$) $\gamma_{\max}^{(k)} \geq 1$ and $\gamma^{(k)} \geq 1$.

($C^{(k)}$) $\gamma_{\max}^{(k)} < 1$ and $\gamma^{(k)} < \gamma_{\max}^{(k)}$.

($D^{(k)}$) $\gamma_{\max}^{(k)} < 1$ and $\gamma^{(k)} = \gamma_{\max}^{(k)}$.

By the descent lemma, we have

$$F^{(k)}(\mathbf{x}^{(k+1)}) = F^{(k)}(\mathbf{x}^{(k)} + \gamma^{(k)} \mathbf{d}^{(k)}) \tag{4}$$

$$\leq F^{(k)}(\mathbf{x}^{(k)}) + \gamma^{(k)} \langle \nabla F^{(k)}(\mathbf{x}^{(\mathbf{k})}), \mathbf{d}^{(k)} \rangle + \frac{L^{(k)}(\gamma^{(k)})^2}{2} \|\mathbf{d}^{(k)}\|^2$$

$$= F^{(k)}(\mathbf{x}^{(k)}) + \gamma^{(k)} \langle \mathbf{g}^{(k)}, \mathbf{d}^{(k)} \rangle + \frac{L^{(k)}(\gamma^{(k)})^2}{2} \|\mathbf{d}^{(k)}\|^2. \tag{5}$$

In case $(A^{(k)})$, let $\delta_{A^{(k)}}$ denote the indicator function for this case. Then

$$\begin{aligned}
&\delta_{A^{(k)}} \{F^{(k)}(\mathbf{x}^{(k+1)}) - F_*^{(k)}\} \\
&\leq \delta_{A^{(k)}} \{F^{(k)}(\mathbf{x}^{(k)}) - F_*^{(k)} + \gamma^{(k)} \langle \mathbf{g}^{(k)}, \mathbf{d}^{(k)} \rangle + \\
&\quad \frac{L^{(k)}(\gamma^{(k)})^2}{2} \|\mathbf{d}^{(k)}\|^2\} \\
&= \delta_{A^{(k)}} \{F^{(k)}(\mathbf{x}^{(k)}) - F_*^{(k)} - \frac{\langle \mathbf{g}^{(k)}, \mathbf{d}^{(k)} \rangle^2}{2L^{(k)} \|\mathbf{d}^{(k)}\|^2}\} \\
&\qquad \text{(definition of } \gamma^{(k)} \text{ in case } A^{(k)}) \\
&\leq \delta_{A^{(k)}} \{(1 - \frac{\Omega_{\mathcal{P}}^2 \sigma_F}{16N^2 L^{(k)} D^2})(F^{(k)}(\mathbf{x}^{(k)}) - F_*^{(k)})\} \\
&\leq \delta_{A^{(k)}} \{(1 - \frac{\Omega_{\mathcal{P}}^2 \sigma_F}{16N^2 L_F D^2})(F^{(k)}(\mathbf{x}^{(k)}) - F_*^{(k)})\}
\end{aligned}$$

In case $(B^{(k)})$, since $\gamma^{(k)} > 1$, we have

$$-\langle \mathbf{g}^{(k)}, \mathbf{d}^{(k)} \rangle > L^{(k)} \|\mathbf{d}^{(k)}\|^2 \qquad \text{and} \tag{6}$$

$$\gamma^{(k)} \langle \mathbf{g}^{(k)}, \mathbf{d}^{(k)} \rangle + \frac{L^{(k)}(\gamma^{(k)})^2}{2} \|\mathbf{d}^{(k)}\|^2 \tag{7}$$

$$\leq \langle \mathbf{g}^{(k)}, \mathbf{d}^{(k)} \rangle + \frac{L^{(k)}}{2} \|\mathbf{d}^{(k)}\|^2. \tag{8}$$

Use $\delta_{B^{(k)}}$ to denote the indicator function for this case. Then,

$$\begin{aligned}
&\delta_{B^{(k)}} \{F^{(k)}(\mathbf{x}^{(k+1)}) - F_*^{(k)}\} \\
&\leq \delta_{B^{(k)}} \{F^{(k)}(\mathbf{x}^{(k)}) - F_*^{(k)} + \\
&\quad \gamma^{(k)} \langle \nabla F^{(k)}(\mathbf{x}^{(k)}), \mathbf{d}^{(k)} \rangle + \frac{L^{(k)}(\gamma^{(k)})^2}{2} \|\mathbf{d}^{(k)}\|^2\} \\
&= \delta_{B^{(k)}} \{F^{(k)}(\mathbf{x}^{(k)}) - F_*^{(k)} + \gamma^{(k)} \langle \mathbf{g}^{(k)}, \mathbf{d}^{(k)} \rangle \\
&\quad + \frac{L^{(k)}(\gamma^{(k)})^2}{2} \|\mathbf{d}^{(k)}\|^2 \\
&\leq \delta_{B^{(k)}} \{F^{(k)}(\mathbf{x}^{(k)}) - F_*^{(k)} + \langle \mathbf{g}^{(k)}, \mathbf{d}^{(k)} \rangle + \frac{L^{(k)}}{2} \|\mathbf{d}^{(k)}\|^2\} \\
&\qquad \text{(by (8))} \\
&\leq \delta_{B^{(k)}} \{F^{(k)}(\mathbf{x}^{(k)}) - F_*^{(k)} + \frac{1}{2} \langle \mathbf{g}^{(k)}, \mathbf{d}^{(k)} \rangle\} \quad \text{(by (6))} \\
&\leq \delta_{B^{(k)}} \{\frac{1}{2}(F^{(k)}(\mathbf{x}^{(k)}) - F_*^{(k)})\}
\end{aligned}$$

In case $(C^{(k)})$, let $\delta_{C^{(k)}}$ be the indicator function for this case and we can use exactly the same argument as in case (A) to obtain the following inequality

$$\begin{aligned}
&\delta_{C^{(k)}} \{F^{(k)}(\mathbf{x}^{(k+1)}) - F_*^{(k)}\} \\
&\leq \delta_{C^{(k)}} \{F^{(k)}(\mathbf{x}^{(k)}) - F_*^{(k)} - \frac{\langle \mathbf{g}^{(k)}, \mathbf{d}^{(k)} \rangle^2}{2L^{(k)} \|\mathbf{d}^{(k)}\|^2}\} \\
&\leq \delta_{C^{(k)}} \{(1 - \frac{\Omega_{\mathcal{P}}^2 \sigma_F}{16N^2 L_F D^2})(F^{(k)}(\mathbf{x}^{(k)}) - F_*^{(k)})\}
\end{aligned}$$

Case $(D^{(k)})$ is the so called "drop step" in the conditional gradient algorithm with away-steps. Use $\delta_{D^{(k)}}$ to denote the indicator function for this case. Note that $\gamma^{(k)} = \gamma_{\max}^{(k)} \le -\langle \mathbf{g}^{(k)}, \mathbf{d}^{(k)}\rangle/(L^{(k)}\|\mathbf{d}^{(k)}\|^2)$ in this case. Hence, we have

$$
\begin{aligned}
&\delta_{D^{(k)}}\{(F^{(k)}(\mathbf{x}^{(k+1)}) - F_*^{(k)})\} \\
&\le \delta_{D^{(k)}}\{F^{(k)}(\mathbf{x}^{(k)}) - F_*^{(k)} + \gamma^{(k)}\langle \nabla F^{(k)}(\mathbf{x}^{(\mathbf{k})}), \mathbf{d}^{(k)}\rangle \\
&\quad + \frac{L^{(k)}(\gamma^{(k)})^2}{2}\|\mathbf{d}^{(\mathbf{k})}\|^2\} \\
&= \delta_{D^{(k)}}\{F^{(k)}(\mathbf{x}^{(k)}) - F_*^{(k)} + \gamma^{(k)}\langle \mathbf{g}^{(k)}, \mathbf{d}^{(k)}\rangle + \\
&\quad \frac{L^{(k)}(\gamma^{(k)})^2}{2}\|\mathbf{d}^{(\mathbf{k})}\|^2\} \\
&\le \delta_{D^{(k)}}\{F^{(k)}(\mathbf{x}^{(k)}) - F_*^{(k)} + \frac{\gamma^{(k)}}{2}\langle \mathbf{g}^{(k)}, \mathbf{d}^{(k)}\rangle\} \\
&\le \delta_{D^{(k)}}\{F^{(k)}(\mathbf{x}^{(k)}) - F_*^{(k)}\}.
\end{aligned}
$$

Define $\rho = \min\{\frac{1}{2}, \frac{\Omega_{\mathcal{P}}^2\sigma_F}{16N^2L_FD^2}\}$. Note that ρ is a deterministic constant between 0 and 1. Therefore we have

$$
\begin{aligned}
&F^{(k)}(\mathbf{x}^{(k+1)}) - F_*^{(k)} \\
&\le (\{1-\rho)^{\{1-\delta_{D^{(k)}}\}}(F^{(k)}(\mathbf{x}^{(k)}) - F_*^{(k)}) \\
&= (1-\rho)^{\{1-\delta_{D^{(k)}}\}}(F^{(k-1)}(\mathbf{x}^{(k)}) - F_*^{(k-1)}) \\
&\quad + (1-\rho)^{\{1-\delta_{D^{(k)}}\}}\{F^{(k)}(\mathbf{x}^{(k)}) - F_*^{(k)} \\
&\quad - F^{(k-1)}(\mathbf{x}^{(k)}) + F_*^{(k-1)}\} \\
&= (1-\rho)^{\{1-\delta_{D^{(k)}}\}}(F^{(k-1)}(\mathbf{x}^{(k)}) - F_*^{(k-1)}) \\
&\quad + (1-\rho)^{\{1-\delta_{D^{(k)}}\}}\{F^{(k)}(\mathbf{x}^{(k)}) - F(\mathbf{x}^{(k)}) + F(\mathbf{x}^{(k)}) \\
&\quad - F^{(k-1)}(\mathbf{x}^{(k)}) + F^* - F_*^{(k)} + F_*^{(k-1)} - F^*\} \\
&\le (1-\rho)^{\{1-\delta_{D^{(k)}}\}}(F^{(k-1)}(\mathbf{x}^{(k)}) - F_*^{(k-1)}) \\
&\quad + (1-\rho)^{\{1-\delta_{D^{(k)}}\}}\{|F^{(k)}(\mathbf{x}^{(k)}) - F(\mathbf{x}^{(k)})| \\
&\quad + |F^{(k-1)}(\mathbf{x}^{(k)}) - F(\mathbf{x}^{(k)})| + |F_*^{(k)} - F^*| \\
&\quad + |F_*^{(k-1)} - F^*|\} \\
&\le (1-\rho)^{\sum_{i=1}^k\{1-\delta_{D^{(i)}}\}}(F^{(0)}(\mathbf{x}^{(1)}) - F_*^{(0)}) + \\
&\quad \sum_{i=1}^k (1-\rho)^{\sum_{j=i}^k\{1-\delta_{D^{(j)}}\}}\{|F^{(i)}(\mathbf{x}^{(i)}) - F(\mathbf{x}^{(i)})| \\
&\quad + |F^{(i-1)}(\mathbf{x}^{(i)}) - F(\mathbf{x}^{(i)})| + |F_*^{(i)} - F^*| \\
&\quad + |F_*^{(i-1)} - F^*|\}.
\end{aligned}
$$

At iteration k, there are at most $(k+1)/2$ drop steps, i.e., at most $(k+1)/2$ $\delta_{D^{(i)}}$'s equal to 1. Then by Lemma **??**, we have

$$
\begin{aligned}
&\sum_{i=1}^k (1-\rho)^{\sum_{j=i}^k\{1-\delta_{D^{(j)}}\}}\{|F^{(i)}(\mathbf{x}^{(i)}) - F(\mathbf{x}^{(i)})| \\
&\quad + |F^{(i-1)}(\mathbf{x}^{(i)}) - F(\mathbf{x}^{(i)})| + |F_*^{(i)} - F^*| + |F_*^{(i-1)} - F^*|\} \\
&\le \sum_{i=k/2}^k \{|F^{(i)}(\mathbf{x}^{(i)}) - F(\mathbf{x}^{(i)})| + |F^{(i-1)}(\mathbf{x}^{(i)}) - F(\mathbf{x}^{(i)})| \\
&\quad + |F_*^{(i)} - F^*| + |F_*^{(i-1)} - F^*|\} \\
&\quad + \sum_{i=1}^{k/2-1} (1-\rho)^{k/2-i}\{|F^{(i)}(\mathbf{x}^{(i)}) - F(\mathbf{x}^{(i)})| \\
&\quad + |F^{(i-1)}(\mathbf{x}^{(i)}) - F(\mathbf{x}^{(i)})| + |F_*^{(i)} - F^*| + |F_*^{(i-1)} - F^*|\}.
\end{aligned}
$$

Therefore

$$
\begin{aligned}
&F^{(k)}(\mathbf{x}^{(k+1)}) - F_*^{(k)} \\
&\le (1-\rho)^{\frac{k-1}{2}}(u_F - l_F) + \sum_{i=k/2}^k \{|F^{(i)}(\mathbf{x}^{(i)}) - F(\mathbf{x}^{(i)})| \\
&+ |F^{(i-1)}(\mathbf{x}^{(i)}) - F(\mathbf{x}^{(i)})| + |F_*^{(i)} - F^*| + |F_*^{(i-1)} - F^*|\} \\
&+ \sum_{i=1}^{k/2-1} (1-\rho)^{k/2-i}\{|F^{(i)}(\mathbf{x}^{(i)}) - F(\mathbf{x}^{(i)})| \\
&+ |F^{(i-1)}(\mathbf{x}^{(i)}) - F(\mathbf{x}^{(i)})| + |F_*^{(i)} - F^*| + |F_*^{(i-1)} - F^*|\}.
\end{aligned}
$$

In addition, $F^{(k)}(\mathbf{x}^{(k+1)}) - F_*^{(k)} = F(\mathbf{x}^{(k+1)}) - F^* + (F^{(k)}(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^{(k+1)})) + (F^* - F_*^{(k)})$. Thus

$$
\begin{aligned}
&F(\mathbf{x}^{(k+1)}) - F^* \\
&\le (1-\rho)^{\frac{k-1}{2}}(u_F - l_F) + \sum_{i=k/2}^{k+1} \{|F^{(i)}(\mathbf{x}^{(i)}) - F(\mathbf{x}^{(i)})| \\
&\quad + |F^{(i-1)}(\mathbf{x}^{(i)}) - F(\mathbf{x}^{(i)})| + |F_*^{(i)} - F^*| + |F_*^{(i-1)} - F^*|\} \\
&\quad + \sum_{i=1}^{k/2-1} (1-\rho)^{k/2-i}\{|F^{(i)}(\mathbf{x}^{(i)}) - F(\mathbf{x}^{(i)})| \\
&\quad + |F^{(i-1)}(\mathbf{x}^{(i)}) - F(\mathbf{x}^{(i)})| + |F_*^{(i)} - F^*| + |F_*^{(i-1)} - F^*|\}.
\end{aligned}
$$

Note that for any deterministic $\mathbf{x} \in \mathcal{P}$, we have $\mathbb{E}F^{(k)}(\mathbf{x}) = F(\mathbf{x})$. In addition, by Corollary **??**, the following bound holds for every iteration k

$$
\begin{aligned}
&\mathbb{E}|F^{(k)}(\mathbf{x}^{(k)}) - F(\mathbf{x}^{(k)})| \\
&\le \mathbb{E}\sup_{\mathbf{x}\in\mathcal{P}}|F^{(k)}(\mathbf{x}) - F(\mathbf{x})| \le C_1\sqrt{\frac{\log m^{(k)}}{m^{(k)}}}
\end{aligned}
$$

and

$$
\mathbb{E}|F_*^{(k)} - F^*| \le C_1\sqrt{\frac{\log m^{(k)}}{m^{(k)}}}.
$$

Combining all above bounds and use $m^{(i)} = \lceil 1/(1-\rho)^{2i+2} \rceil$, we have

$$\begin{aligned}
&\mathbb{E}\{F(\mathbf{x}^{(k+1)}) - F^*\} \\
&\le (1-\rho)^{\frac{k-1}{2}}(u_F - l_F) \\
&\quad + 2C_1\{\sum_{i=k/2}^{k+1}(\sqrt{\frac{\log m^{(i)}}{m^{(i)}}} + \sqrt{\frac{\log m^{(i-1)}}{m^{(i-1)}}}) \\
&\quad + \sum_{i=1}^{k/2-1}(1-\rho)^{k/2-i}(\sqrt{\frac{\log m^{(i)}}{m^{(i)}}} + \sqrt{\frac{\log m^{(i-1)}}{m^{(i-1)}}})\} \\
&\le (1-\rho)^{\frac{k-1}{2}}(u_F - l_F) + 4C_1\{\sum_{i=k/2}^{k+1}\sqrt{\frac{\log m^{(i-1)}}{m^{(i-1)}}} \\
&\quad + \sum_{i=1}^{k/2-1}(1-\rho)^{k/2-i}\sqrt{\frac{\log m^{(i-1)}}{m^{(i-1)}}}\} \\
&\qquad\qquad (\tfrac{\log x}{x} \text{ decreases for } x > e) \\
&\le (1-\rho)^{\frac{k-1}{2}}(u_F - l_F) \\
&\quad + 4C_1\sqrt{2\log\frac{1}{1-\rho}}\{\sum_{i=k/2}^{k+1}(1-\rho)^i\sqrt{i} \\
&\quad + \sum_{i=1}^{k/2-1}(1-\rho)^{k/2}\sqrt{i}\} \\
&\le C_2(1-\beta)^{\frac{k-1}{2}}
\end{aligned}$$

for some constant C_2 and $0 < \beta < \rho < 1$. □

6 Proof of Corollary 3

Proof. Let k be the total number of iterations performed by Algorithm 2 so that an ϵ-accurate solution is obtained for the first time. Theorem 1 implies $C_2(1-\beta)^{\frac{k-1}{2}} < \epsilon$ and hence $k \ge 1 + 2\log\epsilon/\log(1-\beta)$. In iteration i of Algorithm 2, $m^{(i)} = 1/(1-\rho)^{2i+2}$ of stochastic gradient evaluations are performed. Thus, the total number of stochastic gradient evaluations until iteration k is

$$\begin{aligned}
\sum_{i=1}^{k} m^{(i)} &= \sum_{i=1}^{k}\frac{1}{(1-\rho)^{(2i+2)}} \\
&= \frac{1}{(1-\rho)^2}\frac{1/(1-\rho)^2 - 1/(1-\rho)^{2k+2}}{1 - 1/(1-\rho)^2} \\
&\le \frac{2}{(1-\rho)^{2k+4}} \le \frac{2}{(1-\rho)^4}\exp\{-2k\log(1-\rho)\} \\
&\le \frac{2}{(1-\rho)^4}\exp\{-2\log(1-\rho) - 4\frac{\log\epsilon\log(1-\rho)}{\log(1-\beta)}\} \\
&= O((\frac{1}{\epsilon})^{\frac{4\log(1-\rho)}{\log(1-\beta)}}) \\
&= O((\frac{1}{\epsilon})^{4\eta}).
\end{aligned}$$

□

7 Proof of Theorem 2

Proof. Since $\mathbf{d}^{(k)} = \mathbf{p}^{(k)} - \mathbf{u}^{(k)}$, similar to the proof of Theorem 1, we have

$$\begin{aligned}
\langle \mathbf{g}^{(k)}, \mathbf{d}^{(k)} \rangle^2 &\ge \frac{\Omega_{\mathcal{P}}^2\sigma_F}{4N^2}\{F^{(k)}(\mathbf{x}^{(k)}) - F_*^{(k)}\} \\
\langle \mathbf{g}^{(k)}, \mathbf{d}^{(k)} \rangle &\le \frac{1}{2}(F_*^{(k)} - F^{(k)}(\mathbf{x}^{(k)})).
\end{aligned}$$

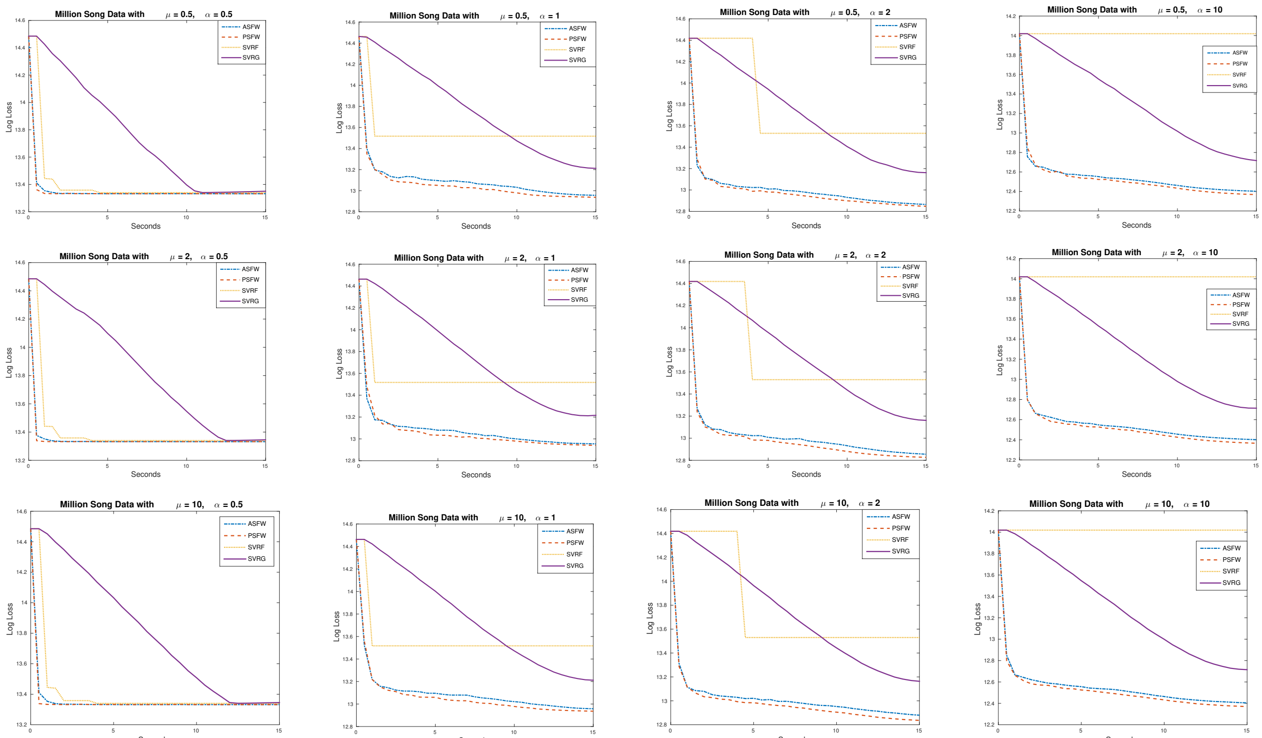
The remaining proof for Theorem 1 could also apply here except that the case $D^{(k)}$ can be either a 'drop step' or a so-called 'swap step'. A swap step moves the weight of a active vertex to another active vertex. There are at most $(1 - \frac{1}{3|V|!+1})k$ drop steps and swap steps after k iteration. The same argument as in Theorem 1 implies

$$\mathbb{E}\{F(\mathbf{x}^{(k+1)}) - F^*\} \le C_3(1-\phi)^{k/(3|V|!+1)}$$

for a deterministic constant C_3 and $0 < \phi < \kappa \le 1/2$. □

8 More Figures for Million Song Dataset Experiment

We tested the algorithms on the Million Song Dataset for different choices of μ and α. The performances of the algorithms follow the same pattern as we described in the paper.


Million Song Data with μ = 0.5, α = 0.5
Million Song Data with μ = 0.5, α = 1
Million Song Data with μ = 0.5, α = 2
Million Song Data with μ = 0.5, α = 10
Million Song Data with μ = 2, α = 0.5
Million Song Data with μ = 2, α = 1
Million Song Data with μ = 2, α = 2
Million Song Data with μ = 2, α = 10
Million Song Data with μ = 10, α = 0.5
Million Song Data with μ = 10, α = 1
Million Song Data with μ = 10, α = 2
Million Song Data with μ = 10, α = 10
ASFW
PSFW
SVRF
SVRG
Log Loss
Seconds