1 Proof of Lemma 2

Proof. To prove the result, we use brackets of the type $[f_0 - \epsilon g/2, f_0 + \epsilon g/2]$ for $\theta$ that ranging over a suitably chosen subset of $\Theta$ and these brackets have $L_1$-size $\epsilon \|g\|_1$. If $\|\theta_1 - \theta_2\| \leq \epsilon/2$, then by the Lipschitz condition that

$$|f_{\theta_1}(\xi) - f_{\theta_2}(\xi)| \leq g(\xi)\|\theta_1 - \theta_2\|,$$

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

2 Proof of Lemma 3

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

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

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

$$\sup_{x \in \mathcal{P}} \left| \frac{1}{m(k)} \sum_{i=1}^{m(k)} f(\xi_i, x) - E f(\xi_i, x) \right| \leq \max \left\{ \frac{1}{m(k)} \sum_{i=1}^{m(k)} f(\xi_i, y) - E f(\xi_i, y) \mid y \in \Gamma \right\} + 2\epsilon L_F.$$

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

$$P\left\{ \sup_{x \in \mathcal{P}} \left| \frac{1}{m(k)} \sum_{i=1}^{m(k)} f(\xi_i, x) - E f(\xi_i, x) \right| \geq \delta \right\}$$

$$\leq P\left\{ \max \left\{ \left| \frac{1}{m(k)} \sum_{i=1}^{m(k)} f(\xi_i, y) - E f(\xi_i, y) \right| \mid y \in \Gamma \right\} + 2\epsilon L_F \geq \delta \right\}$$

$$\leq \sum_{y \in \Gamma} P\left\{ \left| \frac{1}{m(k)} \sum_{i=1}^{m(k)} f(\xi_i, y) - E f(\xi_i, y) \right| \geq \delta - 2\epsilon L_F \right\}$$

$$\leq \sum_{y \in \Gamma} 2 \exp \left\{ - \frac{2m(k)(\delta - 2\epsilon L_F)^2}{(u_F - l_F)^2} \right\}$$

(union bound)

$$\leq 2K_{\mathcal{F}} \frac{D}{\epsilon}^p \exp \left\{ - \frac{2m(k)(\delta - 2\epsilon L_F)^2}{(u_F - l_F)^2} \right\},$$

($|\Gamma| \leq K_{\mathcal{F}} \frac{D}{\epsilon}^p$)

Since by definition, $F^{(k)}(x) = \frac{1}{m(k)} \sum_{i=1}^{m(k)} f(\xi_i, x)$ and $F(x) = E f(\xi_i, x)$, 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, $\sup_{x \in \mathcal{P}} |F^{(k)}(x) - F(x)| \leq 2(|u_F| + |l_F|)$. Then for every $\delta \geq 0$, we have

$$E \sup_{x \in \mathcal{P}} |F^{(k)}(x) - F(x)|$$

$$\leq 2(|u_F| + |l_F|) P\left\{ \sup_{x \in \mathcal{P}} |F^{(k)}(x) - F(x)| \geq \delta \right\} + \delta P\left\{ \sup_{x \in \mathcal{P}} |F^{(k)}(x) - F(x)| < \delta \right\}$$

$$\leq 4(|u_F| + |l_F|) K_{\mathcal{F}} \frac{D}{\epsilon}^p \exp \left\{ - \frac{2m(k)(\delta - 2\epsilon L_F)^2}{(u_F - l_F)^2} \right\} + \delta$$

$$\leq 4(|u_F| + |l_F|) K_{\mathcal{F}} D^p \exp \left\{ - \frac{2m(k)(\delta - 2\epsilon L_F)^2}{(u_F - l_F)^2} + p \log \frac{1}{\epsilon} \right\} + \delta.$$
Now let $\delta = \frac{(u_F - l_F)\sqrt{4(p+1)\log m(k)}}{\sqrt{m(k)}}, \epsilon = \frac{(u_F - l_F)}{2L_F\sqrt{m(k)}},$

Then

$$\mathbb{E} \sup_{x \in P} |F^{(k)}(x) - F(x)|$$

$$\leq 4(|u_F| + |l_F|)K_P D^p \exp\{-\sqrt{4(p+1)\log m(k)} - 1\}^2$$

$$- p\log \frac{u_F - l_F}{2\sqrt{2L_F}} + p\log m(k)$$

$$+ \frac{(u_F - l_F)\sqrt{4(p+1)\log m(k)}}{\sqrt{m(k)}},$$

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

$$\mathbb{E} \sup_{x \in P} |F^{(k)}(x) - F(x)|$$

$$\leq 4(|u_F| + |l_F|)K_P D^p \exp\{-\sqrt{4(p+1)\log m(k)}\}$$

$$+ \frac{p\log m(k)}{2\sqrt{2L_F}} + p\log m(k)$$

$$+ \frac{(u_F - l_F)\sqrt{4(p+1)\log m(k)}}{\sqrt{m(k)}},$$

where $C_1 = 4(|u_F| + |l_F|)K_P D^p \exp\{-\sqrt{4(p+1)\log m(k)}\}. \quad \Box$

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

$$F(x^{(k)}) - \delta \leq F(k)(x^{(k)}) \leq F(x^{(k)}) + \delta \quad (2)$$

and

$$F(x^*) - \delta \leq F^{(k)}(x^*) \leq F(x^*) + \delta \quad (3)$$

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

$$F^{(k)}(x^{(k)}) \geq F(x^{(k)}) - \delta \quad \text{(by (2))}$$

$$\geq F(x^*) - \delta \quad \text{(optimality of } x^{(k)} \text{ for } F^{(k)})$$

On the other hand,

$$F^{(k)}(x^{(k)}) \leq F(k)(x^*) \quad \text{(optimality of } x^{(k)} \text{ for } F^{(k)})$$

$$\leq F(x^*) + \delta \quad \text{(by (3))}$$

Therefore, we have

$$\mathbb{P}\{F^{(k)}(x^{(k)}) - F(x^*) \geq \delta\} \leq 2K_P \left(\frac{D}{2}\right)^p \exp\{-\frac{m(k)(\delta - 2L_F\epsilon)^2}{2(u_F - l_F)^2}\},$$

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

\section{Proof of Lemma 4}

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

$$\sum_{k=1}^{n} a^k \sum_{j=k}^{b_i} c_k$$

$$= \sum_{k=s+1}^{n} a^k \sum_{j=k}^{b_i} c_k + \sum_{k=1}^{s} a^k \sum_{j=k}^{b_i} c_k + \sum_{k=1}^{r-1} a^k \sum_{j=k}^{b_i} c_k$$

$$= \sum_{k=s+1}^{n} a^k \sum_{j=k}^{b_i} c_k + \sum_{k=1}^{r-1} a^k \sum_{j=k}^{b_i} c_k + \sum_{k=1}^{r-1} a^k \sum_{j=k}^{b_i} c_k$$

$$= \sum_{k=1}^{n} a^k \sum_{j=k}^{b_i} c_k.$$ \quad \Box

\section{Proof of Theorem 1}

\textbf{Proof.} At iteration $k,$ let $x^{(k)}$ denote the current solution, $\xi_1, \ldots, \xi_{m(k)}$ denote the samples obtained in the algorithm, $d^{(k)}$ denote the direction that the algorithm will take at this step and $\gamma^{(k)}$ denote the step length. Define $F^{(k)}(x) = \frac{1}{m(k)} \sum_{i=1}^{m(k)} f(\xi_i, x)$ and $F^{(k)}(x) = \nabla F^{(k)}(x).$ 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 $g^{(k)} = \nabla F^{(k)}(x).$ From the choice of $d^{(k)}$ in the algorithm,

$$\langle g^{(k)}, d^{(k)} \rangle \leq \frac{1}{2} \langle (g^{(k)}, p^{(k)} - x^{(k)}) + (g^{(k)}, x^{(k)} - u^{(k)}) \rangle \leq \frac{1}{2} \langle g^{(k)}, p^{(k)} - u^{(k)} \rangle \leq 0.$$ \Box
Hence, we can lower bound $\langle g^{(k)}, d^{(k)} \rangle^2$ by
\[
\langle g^{(k)}, d^{(k)} \rangle^2 \geq \frac{1}{4} \langle g^{(k)}, u^{(k)} - p^{(k)} \rangle^2 \\
\geq \frac{1}{4} \max_{p \in V, u \in U^{(k)}} \langle g^{(k)}, u - p \rangle^2 \\
= \frac{1}{4} \max_{p \in V, u \in U^{(k)}} \langle \nabla F^{(k)}(x^{(k)}), u - p \rangle^2 \\
= \frac{\Omega_2}{4} \frac{\langle \nabla F^{(k)}(x^{(k)}), x^{(k)} - x^*_k \rangle}{\|x^{(k)} - x^*_k\|^2}^2 \\
(\text{by Lemma 1}) \\
\geq \frac{\Omega_2}{8N^2} \left\{ F^{(k)}(x^{(k)}) - F^*_k \right\}^2 \\
(\text{Convexity of } F^{(k)}) \\
\geq \frac{\Omega_2^{2} \sigma F}{8N^2} \left\{ F^{(k)}(x^{(k)}) - F^*_k \right\} \\
(\text{by strong convexity of } F^{(k)}) \\
\geq \frac{\Omega_2^{2} \sigma F}{8N^2} \left\{ F^{(k)}(x^{(k)}) - F^*_k \right\}.
\]

Similarly, we can upper bound $\langle g^{(k)}, d^{(k)} \rangle$ by
\[
\langle g^{(k)}, d^{(k)} \rangle \leq \frac{1}{2} \langle g^{(k)}, p^{(k)} - u^{(k)} \rangle \\
\leq \frac{1}{2} \langle g^{(k)}, x^*_k - x^{(k)} \rangle \\
(\text{definition of } p^{(k)} \text{ and } u^{(k)}) \\
= \frac{1}{2} \langle \nabla F^{(k)}(x^{(k)}), x^*_k - x^{(k)} \rangle \\
(\text{by strong convexity of } F^{(k)}) \\
\leq \frac{1}{2} \left\{ F^*_k - F^{(k)}(x^{(k)}) \right\}.
\]

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

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

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

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

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

By the descent lemma, we have
\[
F^{(k)}(x^{(k+1)}) = F^{(k)}(x^{(k)}) + \gamma^{(k)} d^{(k)} \\
\leq F^{(k)}(x^{(k)}) + \gamma^{(k)} \langle \nabla F^{(k)}(x^{(k)}), d^{(k)} \rangle + \frac{L^{(k)}(\gamma^{(k)})^2}{2} \|d^{(k)}\|^2 \\
= F^{(k)}(x^{(k)}) + \gamma^{(k)} \langle g^{(k)}, d^{(k)} \rangle + \frac{L^{(k)}(\gamma^{(k)})^2}{2} \|d^{(k)}\|^2.
\]

In case (A), let $\delta_{A^{(k)}}$ denote the indicator function for this case. Then
\[
\delta_{A^{(k)}} \left\{ F^{(k)}(x^{(k+1)}) - F^*_k \right\} \\
\leq \delta_{A^{(k)}} \left\{ F^{(k)}(x^{(k)}) - F^*_k + \gamma^{(k)} \langle g^{(k)}, d^{(k)} \rangle + \frac{L^{(k)}(\gamma^{(k)})^2}{2} \|d^{(k)}\|^2 \right\} \\
= \delta_{A^{(k)}} \left\{ F^{(k)}(x^{(k)}) - F^*_k - \frac{\langle g^{(k)}, d^{(k)} \rangle^2}{2L^{(k)} \|d^{(k)}\|^2} \right\} \\
(\text{definition of } \gamma^{(k)} \text{ in case } A^{(k)}) \\
\leq \delta_{A^{(k)}} \left\{ 1 - \frac{\Omega_2^{2} \sigma F}{16N^2L^{(k)} D^2} (F^{(k)}(x^{(k)}) - F^*_k) \right\} \\
\leq \delta_{A^{(k)}} \left\{ 1 - \frac{\Omega_2^{2} \sigma F}{16N^2L^{(k)} D^2} (F^{(k)}(x^{(k)}) - F^*_k) \right\}.
\]

In case (B), since $\gamma^{(k)} > 1$, we have
\[
- \langle g^{(k)}, d^{(k)} \rangle > L^{(k)} \|d^{(k)}\|^2 \quad \text{and} \quad \frac{\Omega_2^{2} \sigma F}{8N^2} \left\{ F^{(k)}(x^{(k)}) - F^*_k \right\} \\
\leq \langle g^{(k)}, d^{(k)} \rangle + \frac{L^{(k)}(\gamma^{(k)})^2}{2} \|d^{(k)}\|^2. \quad \text{(6)}
\]

Use $\delta_{B^{(k)}}$ to denote the indicator function for this case. Then,
\[
\delta_{B^{(k)}} \left\{ F^{(k)}(x^{(k+1)}) - F^*_k \right\} \\
\leq \delta_{B^{(k)}} \left\{ F^{(k)}(x^{(k)}) - F^*_k + \gamma^{(k)} \langle g^{(k)}, d^{(k)} \rangle + \frac{L^{(k)}(\gamma^{(k)})^2}{2} \|d^{(k)}\|^2 \right\} \\
= \delta_{B^{(k)}} \left\{ F^{(k)}(x^{(k)}) - F^*_k + \gamma^{(k)} \langle g^{(k)}, d^{(k)} \rangle \right. \\
\left. + \frac{L^{(k)}(\gamma^{(k)})^2}{2} \|d^{(k)}\|^2 \right\} \\
\leq \delta_{B^{(k)}} \left\{ F^{(k)}(x^{(k)}) - F^*_k + \langle g^{(k)}, d^{(k)} \rangle + \frac{L^{(k)}(\gamma^{(k)})^2}{2} \|d^{(k)}\|^2 \right\} \quad \text{(by (8))} \\
\leq \delta_{B^{(k)}} \left\{ F^{(k)}(x^{(k)}) - F^*_k + \frac{1}{2} \|g^{(k)}\|^2 \right\} \quad \text{(by (6))} \\
\leq \delta_{B^{(k)}} \frac{1}{2} \left\{ F^{(k)}(x^{(k)}) - F^*_k \right\} \quad \text{(by (6))}.
\]

In case (C), 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
\[
\delta_{C^{(k)}} \left\{ F^{(k)}(x^{(k+1)}) - F^*_k \right\} \\
\leq \delta_{C^{(k)}} \left\{ F^{(k)}(x^{(k)}) - F^*_k - \frac{\langle g^{(k)}, d^{(k)} \rangle^2}{2L^{(k)} \|d^{(k)}\|^2} \right\} \\
\leq \delta_{C^{(k)}} \left\{ 1 - \frac{\Omega_2^{2} \sigma F}{16N^2L^{(k)} D^2} (F^{(k)}(x^{(k)}) - F^*_k) \right\}.
\]

Case (D) 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 $\gamma^{(k)}\leq \gamma^{(k)} = \frac{\delta^{(k)}}{\max\{\langle \rho, g^{(k)} \rangle, \|d^{(k)}\|^2\}}$ in this case. Hence, we have

$$
\delta_{D^{(k)}}\{(F^{(k)}(x^{(k+1)}) - F^{(k)}(x^{(k+1)})
\leq \delta_{D^{(k)}}\{F^{(k)}(x^{(k)}) - F^{(k)}(x^{(k)}) + \gamma^{(k)}\langle \nabla F^{(k)}(x^{(k)}), d^{(k)} \rangle +
\frac{L^{(k)}(\gamma^{(k)})^2}{2}\|d^{(k)}\|^2\}
\leq \delta_{D^{(k)}}\{F^{(k)}(x^{(k)}) - F^{(k)}(x^{(k)}) + \gamma^{(k)}\langle g^{(k)}, d^{(k)} \rangle +
\frac{L^{(k)}(\gamma^{(k)})^2}{2}\|d^{(k)}\|^2\}
\leq \delta_{D^{(k)}}\{F^{(k)}(x^{(k)}) - F^{(k)}(x^{(k)}) + \gamma^{(k)}\langle g^{(k)}, d^{(k)} \rangle +
\frac{L^{(k)}(\gamma^{(k)})^2}{2}\|d^{(k)}\|^2\}
\leq \delta_{D^{(k)}}\{F^{(k)}(x^{(k)}) - F^{(k)}(x^{(k)}) + \gamma^{(k)}\langle g^{(k)}, d^{(k)} \rangle +
\frac{L^{(k)}(\gamma^{(k)})^2}{2}\|d^{(k)}\|^2\}
\leq \delta_{D^{(k)}}\{F^{(k)}(x^{(k)}) - F^{(k)}(x^{(k)}) + \gamma^{(k)}\langle g^{(k)}, d^{(k)} \rangle +
\frac{L^{(k)}(\gamma^{(k)})^2}{2}\|d^{(k)}\|^2\}
$$

Define $\rho = \min\{\frac{1}{2}, \frac{\Omega^{2}L_{F}}{2\sigma_{F}}\}$. Note that $\rho$ is a deterministic constant between 0 and 1. Therefore we have

$$
F^{(k)}(x^{(k+1)}) - F^{(k)}(x^{(k)})
\leq (1 - \rho)^{\delta_{D^{(k)}}}\{F^{(k)}(x^{(k)}) - F^{(k)}(x^{(k)})
\leq (1 - \rho)^{\delta_{D^{(k)}}}\{F^{(k)}(x^{(k)}) - F^{(k)}(x^{(k)})
\leq (1 - \rho)^{\delta_{D^{(k)}}}\{F^{(k)}(x^{(k)}) - F^{(k)}(x^{(k)})
\leq (1 - \rho)^{\delta_{D^{(k)}}}\{F^{(k)}(x^{(k)}) - F^{(k)}(x^{(k)})
\leq (1 - \rho)^{\delta_{D^{(k)}}}\{F^{(k)}(x^{(k)}) - F^{(k)}(x^{(k)})
\leq (1 - \rho)^{\delta_{D^{(k)}}}\{F^{(k)}(x^{(k)}) - F^{(k)}(x^{(k)})
$$

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

$$
\sum_{i=1}^{k}(1 - \rho)\sum_{j=1}^{k-h}(1 - \delta_{D^{(j)}})\{F^{(i)}(x^{(i)}) - F^{(i)}(x^{(i)})
\leq \sum_{i=1}^{k}(1 - \rho)\sum_{j=1}^{k-h}(1 - \delta_{D^{(j)}})\{F^{(i)}(x^{(i)}) - F^{(i)}(x^{(i)})
\leq \sum_{i=1}^{k}(1 - \rho)\sum_{j=1}^{k-h}(1 - \delta_{D^{(j)}})\{F^{(i)}(x^{(i)}) - F^{(i)}(x^{(i)})
\leq \sum_{i=1}^{k}(1 - \rho)\sum_{j=1}^{k-h}(1 - \delta_{D^{(j)}})\{F^{(i)}(x^{(i)}) - F^{(i)}(x^{(i)})
\leq \sum_{i=1}^{k}(1 - \rho)\sum_{j=1}^{k-h}(1 - \delta_{D^{(j)}})\{F^{(i)}(x^{(i)}) - F^{(i)}(x^{(i)})
\leq \sum_{i=1}^{k}(1 - \rho)\sum_{j=1}^{k-h}(1 - \delta_{D^{(j)}})\{F^{(i)}(x^{(i)}) - F^{(i)}(x^{(i)})
$$

Supplementary Material: Linear Convergence of Stochastic Frank Wolfe Variants
Combining all above bounds and use $m^{(i)} = \lceil 1/(1 - \rho)^{2i+2} \rceil$, we have

$$
\mathbb{E}\{F(x^{(k+1)}) - F^*\} \\
\leq (1 - \rho)^{k+1} (u_F - l_F) \\
+ 2C_1 \left\{ \sum_{i=k/2}^{k+1} \left( \frac{\log m^{(i)}}{m^{(i)}} + \sqrt{\frac{\log m^{(i-1)}}{m^{(i-1)}}} \right) \right\} \\
+ \sum_{i=1}^{k/2-1} (1 - \rho)^{k/2-i} \left( \frac{\log m^{(i)}}{m^{(i)}} + \sqrt{\frac{\log m^{(i-1)}}{m^{(i-1)}}} \right) \\
\leq (1 - \rho)^{k+1} (u_F - l_F) \\
+ 4C_1 \sum_{i=k/2}^{k+1} (1 - \rho)^i \sqrt{i} \\
+ \sum_{i=1}^{k/2-1} (1 - \rho)^{k/2-i} \\
\leq C_2 (1 - \rho)^{k/2} \\
$$

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 $\epsilon$-accurate solution is obtained for the first time. Theorem 1 implies $C_2 (1 - \beta)^{k+1} < \epsilon$ and hence $k \geq 1 + 2 \log \epsilon / \log(1 - \beta)$. 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

$$
\sum_{i=1}^{k} m^{(i)} = \sum_{i=1}^{k} \frac{1}{(1 - \rho)^{2i+2}} \\
= \frac{1}{(1 - \rho)^2} \sum_{i=1}^{k+1} \frac{1 - \rho^{2i+2}}{1 - (1 - \rho)^{2i+2}} \\
\leq \frac{2}{(1 - \rho)^{2k+4}} \leq \frac{2}{(1 - \rho)^4} \exp\{-2k \log(1 - \rho)\} \\
\leq \frac{2}{(1 - \rho)^4} \exp\{-2 \log(1 - \rho) - 4 \log \epsilon \log(1 - \rho)\} \\
= O\left( \frac{1}{\epsilon^{4\log(1 - \rho)}} \right) \\
= O\left( \frac{1}{\epsilon^4} \right).
$$

7 Proof of Theorem 2

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

$$
\langle g^{(k)}, d^{(k)} \rangle^2 \geq \frac{\Omega_p^2 \sigma_F}{4N^2} \{ F^{(k)}(x^{(k)}) - F^* \} \\
\langle g^{(k)}, d^{(k)} \rangle \leq \frac{1}{2} \{ F^* - F^{(k)}(x^{(k)}) \}.
$$

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 $(1 - \frac{1}{\sqrt{V'}})^k$ drop steps and swap steps after $k$ iteration. The same argument as in Theorem 1 implies

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

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

8 More Figures for Million Song Dataset Experiment

We tested the algorithms on the Million Song Dataset for different choices of $\mu$ and $\alpha$. The performances of the algorithms follow the same pattern as we described in the paper.
Million Song Data with $\mu = 0.5$, $\alpha = 0.5$

Million Song Data with $\mu = 0.5$, $\alpha = 1$

Million Song Data with $\mu = 0.5$, $\alpha = 2$

Million Song Data with $\mu = 0.5$, $\alpha = 10$

Million Song Data with $\mu = 2$, $\alpha = 0.5$

Million Song Data with $\mu = 2$, $\alpha = 1$

Million Song Data with $\mu = 2$, $\alpha = 2$

Million Song Data with $\mu = 2$, $\alpha = 10$

Million Song Data with $\mu = 10$, $\alpha = 0.5$

Million Song Data with $\mu = 10$, $\alpha = 1$

Million Song Data with $\mu = 10$, $\alpha = 2$

Million Song Data with $\mu = 10$, $\alpha = 10$