Supplementary Material: $O(\log T)$ Projections for Stochastic Optimization of Smooth and Strongly Convex Functions

Lijun Zhang* ZHANGLIJ@MSU.EDU

Tianbao Yang† TYANG@GE.COM

Rong Jin* RONGJIN@CSE.MSU.EDU

Xiaofei He‡ XIAOFEIHE@CAD.ZJU.EDU.CN

*Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA
†GE Global Research, San Ramon, CA 94583, USA
‡State Key Laboratory of CAD&CG, College of Computer Science, Zhejiang University, Hangzhou 310027, China

A. Proof of Lemma 1

We need the following lemma that characterizes the property of the extra-gradient descent.

Lemma 8 (Lemma 3.1 in (Nemirovski, 2005))**.** *Let $\mathcal{Z}$ be a convex compact set in Euclidean space $\mathcal{E}$ with inner product $\langle\cdot,\cdot\rangle$, let $\|\cdot\|$ be a norm on $\mathcal{E}$ and $\|\cdot\|_*$ be its dual norm, and let $\omega(\mathbf{z}): \mathcal{Z} \mapsto \mathbb{R}$ be a α-strongly convex function with respect to $\|\cdot\|$. The Bregman distance associated with ω for points $\mathbf{z}, \mathbf{w} \in \mathcal{Z}$ is defined as*

$$B_\omega(\mathbf{z}, \mathbf{w}) = \omega(\mathbf{z}) - \omega(\mathbf{w}) - \langle \mathbf{z} - \mathbf{w}, \nabla\omega(\mathbf{w})\rangle.$$

Let $\mathcal{U}$ be a convex and closed subset of $\mathcal{Z}$, and let $\mathbf{z}_- \in \mathcal{Z}$, let $\boldsymbol{\xi}, \boldsymbol{\eta} \in \mathcal{E}$, and let $\gamma > 0$. Consider the points

$$\mathbf{w} = \operatorname*{argmin}_{\mathbf{y}\in\mathcal{U}} \{\langle \gamma\boldsymbol{\xi} - \nabla\omega(\mathbf{z}_-), \mathbf{y}\rangle + \omega(\mathbf{y})\},$$
$$\mathbf{z}_+ = \operatorname*{argmin}_{\mathbf{y}\in\mathcal{U}} \{\langle \gamma\boldsymbol{\eta} - \nabla\omega(\mathbf{z}_-), \mathbf{y}\rangle + \omega(\mathbf{y})\}.$$

Then for all $\mathbf{z} \in \mathcal{U}$ one has

$$\langle \mathbf{w} - \mathbf{z}, \gamma\boldsymbol{\eta}\rangle \leq B_\omega(\mathbf{z}, \mathbf{z}_-) - B_\omega(\mathbf{z}, \mathbf{z}_+) + \frac{\gamma^2}{\alpha}\|\boldsymbol{\eta} - \boldsymbol{\xi}\|_*^2 - \frac{\alpha}{2}\{\|\mathbf{w} - \mathbf{z}_-\|^2 + \|\mathbf{z}_+ - \mathbf{w}\|^2\}.$$

Proof of Lemma 1. We first state the inner loop in Algorithm 1 below.

for $t = 1$ to M **do**

Compute the average gradient at $\mathbf{w}_t^k$ over B^k calls to the gradient oracle

$$\bar{\mathbf{g}}_t^k = \frac{1}{B^k}\sum_{i=1}^{B^k} \hat{\mathbf{g}}(\mathbf{w}_t^k, i)$$

Update

$$\mathbf{z}_t^k = \Pi_{\mathcal{D}}\left(\mathbf{w}_t^k - \eta\bar{\mathbf{g}}_t^k\right)$$

Compute the average gradient at $\mathbf{z}_t^k$ over B^k calls to the gradient oracle

$$\bar{\mathbf{f}}_t^k = \frac{1}{B^k}\sum_{i=1}^{B^k} \hat{\mathbf{g}}(\mathbf{z}_t^k, i)$$

Update

$$\mathbf{w}_{t+1}^k = \Pi_{\mathcal{D}}\left(\mathbf{w}_t^k - \eta\bar{\mathbf{f}}_t^k\right)$$

end for

To simplify the notation, we define

$$\mathbf{g}_t^k = \nabla F(\mathbf{w}_t^k) \text{ and } \mathbf{f}_t^k = \nabla F(\mathbf{z}_t^k).$$

Let the two norms $\|\cdot\|$ and $\|\cdot\|_*$ in Lemma 8 be the vector ℓ_2 norm. Each iteration in the inner loop satisfies the conditions in Lemma 8 by doing the mappings below:

$$\mathcal{U} = \mathcal{Z} = \mathcal{E} \leftarrow \mathcal{D},\ \omega(\mathbf{z}) \leftarrow \frac{1}{2}\|\mathbf{z}\|^2,\ \alpha \leftarrow 1,\ \gamma \leftarrow \eta,\ \mathbf{z}_- \leftarrow \mathbf{w}_t^k,\ \boldsymbol{\xi} \leftarrow \bar{\mathbf{g}}_t^k,\ \boldsymbol{\eta} \leftarrow \bar{\mathbf{f}}_t^k,\ \mathbf{w} \leftarrow \mathbf{z}_t^k,\ \mathbf{z}_+ \leftarrow \mathbf{w}_{t+1}^k,\ \mathbf{z} \leftarrow \mathbf{w}_*.$$

Following Lemma 8, we have

$$\begin{aligned}
&\langle \mathbf{z}_t^k - \mathbf{w}_*, \eta\bar{\mathbf{f}}_t^k\rangle \\
\leq&\frac{\|\mathbf{w}_t^k - \mathbf{w}_*\|^2}{2} - \frac{\|\mathbf{w}_{t+1}^k - \mathbf{w}_*\|^2}{2} + \eta^2\|\bar{\mathbf{g}}_t^k - \bar{\mathbf{f}}_t^k\|^2 - \frac{1}{2}\|\mathbf{w}_t^k - \mathbf{z}_t^k\|^2 \\
\leq&\frac{\|\mathbf{w}_t^k - \mathbf{w}_*\|^2}{2} - \frac{\|\mathbf{w}_{t+1}^k - \mathbf{w}_*\|^2}{2} + 3\eta^2\left(\|\bar{\mathbf{g}}_t^k - \mathbf{g}_t^k\|^2 + \|\bar{\mathbf{f}}_t^k - \mathbf{f}_t^k\|^2 + \|\mathbf{g}_t^k - \mathbf{f}_t^k\|^2\right) - \frac{1}{2}\|\mathbf{w}_t^k - \mathbf{z}_t^k\|^2 \\
\leq&\frac{\|\mathbf{w}_t^k - \mathbf{w}_*\|^2}{2} - \frac{\|\mathbf{w}_{t+1}^k - \mathbf{w}_*\|^2}{2} + 3\eta^2\left(\|\bar{\mathbf{g}}_t^k - \mathbf{g}_t^k\|^2 + \|\bar{\mathbf{f}}_t^k - \mathbf{f}_t^k\|^2\right) + 3\eta^2\|\mathbf{g}_t^k - \mathbf{f}_t^k\|^2 - \frac{1}{2}\|\mathbf{w}_t^k - \mathbf{z}_t^k\|^2 \\
\leq&\frac{\|\mathbf{w}_t^k - \mathbf{w}_*\|^2}{2} - \frac{\|\mathbf{w}_{t+1}^k - \mathbf{w}_*\|^2}{2} + 3\eta^2\left(\|\bar{\mathbf{g}}_t^k - \mathbf{g}_t^k\|^2 + \|\bar{\mathbf{f}}_t^k - \mathbf{f}_t^k\|^2\right) + 3\eta^2L^2\|\mathbf{w}_t^k - \mathbf{z}_t^k\|^2 - \frac{1}{2}\|\mathbf{w}_t^k - \mathbf{z}_t^k\|^2 \\
\leq&\frac{\|\mathbf{w}_t^k - \mathbf{w}_*\|^2}{2} - \frac{\|\mathbf{w}_{t+1}^k - \mathbf{w}_*\|^2}{2} + 3\eta^2\left(\|\bar{\mathbf{g}}_t^k - \mathbf{g}_t^k\|^2 + \|\bar{\mathbf{f}}_t^k - \mathbf{f}_t^k\|^2\right),
\end{aligned} \tag{11}$$

where in the fifth line we use the smoothness assumption

$$\|\mathbf{g}_t^k - \mathbf{f}_t^k\| = \|\nabla F(\mathbf{w}_t^k) - \nabla F(\mathbf{z}_t^k)\| \leq L\|\mathbf{w}_t^k - \mathbf{z}_t^k\|.$$

From the property of λ-strongly convex function and (11), we obtain

$$\begin{aligned}
&F(\mathbf{z}_t^k) - F(\mathbf{w}_*) \\
\leq&\langle \mathbf{f}_t^k, \mathbf{z}_t^k - \mathbf{w}_*\rangle - \frac{\lambda}{2}\|\mathbf{z}_t^k - \mathbf{w}_*\|^2 \\
=&\langle \bar{\mathbf{f}}_t^k, \mathbf{z}_t^k - \mathbf{w}_*\rangle + \langle \mathbf{f}_t^k - \bar{\mathbf{f}}_t^k, \mathbf{z}_t^k - \mathbf{w}_*\rangle - \frac{\lambda}{2}\|\mathbf{z}_t^k - \mathbf{w}_*\|^2 \\
\leq&\frac{\|\mathbf{w}_t^k - \mathbf{w}_*\|^2}{2\eta} - \frac{\|\mathbf{w}_{t+1}^k - \mathbf{w}_*\|^2}{2\eta} + 3\eta\left(\|\bar{\mathbf{g}}_t^k - \mathbf{g}_t^k\|^2 + \|\bar{\mathbf{f}}_t^k - \mathbf{f}_t^k\|^2\right) + \langle \mathbf{f}_t^k - \bar{\mathbf{f}}_t^k, \mathbf{z}_t^k - \mathbf{w}_*\rangle - \frac{\lambda}{2}\|\mathbf{z}_t^k - \mathbf{w}_*\|^2.
\end{aligned}$$

Summing up over all $t = 1, 2, \ldots, M$, we have

$$\begin{aligned}
&\sum_{t=1}^M F(\mathbf{z}_t^k) - MF(\mathbf{w}_*) \\
\leq&\frac{\|\mathbf{w}_1^k - \mathbf{w}_*\|^2}{2\eta} + 3\eta\left(\sum_{t=1}^M\|\bar{\mathbf{g}}_t^k - \mathbf{g}_t^k\|^2 + \sum_{t=1}^M\|\bar{\mathbf{f}}_t^k - \mathbf{f}_t^k\|^2\right) + \sum_{t=1}^M\langle \mathbf{f}_t^k - \bar{\mathbf{f}}_t^k, \mathbf{z}_t^k - \mathbf{w}_*\rangle - \frac{\lambda}{2}\sum_{t=1}^M\|\mathbf{z}_t^k - \mathbf{w}_*\|^2.
\end{aligned}$$

Dividing both sides by M and following Jensen's inequality, we have

$$
\begin{aligned}
& F\left(\frac{1}{M}\sum_{t=1}^{M}\mathbf{z}_t^k\right) - F(\mathbf{w}_*) \\
\leq & \frac{1}{M}\sum_{t=1}^{M} F(\mathbf{z}_t^k) - F(\mathbf{w}_*) \\
\leq & \frac{\|\mathbf{w}_1^k - \mathbf{w}_*\|^2}{2M\eta} + \frac{3\eta}{M}\left(\sum_{t=1}^{M}\|\bar{\mathbf{g}}_t^k - \mathbf{g}_t^k\|^2 + \sum_{t=1}^{M}\|\bar{\mathbf{f}}_t^k - \mathbf{f}_t^k\|^2\right) + \frac{1}{M}\sum_{t=1}^{M}\langle \mathbf{f}_t^k - \bar{\mathbf{f}}_t^k, \mathbf{z}_t^k - \mathbf{w}_*\rangle - \frac{\lambda}{2M}\sum_{t=1}^{M}\|\mathbf{z}_t^k - \mathbf{w}_*\|^2.
\end{aligned} \tag{12}
$$

which gives the first inequality in Lemma 1.

Let $\mathrm{E}_{k-1}[\cdot]$ denote the expectation conditioned on all the randomness up to epoch $k-1$ and $\mathrm{E}_k^{t-1}[\cdot]$ denote the expectation conditioned on all the randomness up to the $t-1$-th iteration in the k-th epoch. Taking the conditional expectation of (12), we have

$$
\begin{aligned}
& \mathrm{E}_{k-1}\left[F\left(\frac{1}{M}\sum_{t=1}^{M}\mathbf{z}_t^k\right)\right] - F(\mathbf{w}_*) \\
\leq & \frac{\|\mathbf{w}_1^k - \mathbf{w}_*\|^2}{2M\eta} + \frac{3\eta}{M}\left(\sum_{t=1}^{M}\mathrm{E}_{k-1}\left[\|\bar{\mathbf{g}}_t^k - \mathbf{g}_t^k\|^2\right] + \sum_{t=1}^{M}\mathrm{E}_{k-1}\left[\|\bar{\mathbf{f}}_t^k - \mathbf{f}_t^k\|^2\right]\right) + \frac{1}{M}\sum_{t=1}^{M}\mathrm{E}_{k-1}\left[\langle \mathbf{f}_t^k - \bar{\mathbf{f}}_t^k, \mathbf{z}_t^k - \mathbf{w}_*\rangle\right],
\end{aligned} \tag{13}
$$

where we drop the last term, since it is negative. To bound $\mathrm{E}_{k-1}\left[\|\bar{\mathbf{g}}_t^k - \mathbf{g}_t^k\|^2\right]$, we have

$$
\begin{aligned}
& \mathrm{E}_{k-1}\left[\|\bar{\mathbf{g}}_t^k - \mathbf{g}_t^k\|^2\right] = \mathrm{E}_{k-1}\left[\left\|\frac{1}{B^k}\sum_{i=1}^{B^k}\hat{\mathbf{g}}(\mathbf{w}_t^k, i) - \mathbf{g}_t^k\right\|^2\right] = \mathrm{E}_{k-1}\left[\left\|\frac{1}{B^k}\sum_{i=1}^{B^k}\left(\hat{\mathbf{g}}(\mathbf{w}_t^k, i) - \mathbf{g}_t^k\right)\right\|^2\right] \\
= & \frac{1}{[B^k]^2}\left(\sum_{i=1}^{B^k}\mathrm{E}_{k-1}\left[\left\|\hat{\mathbf{g}}(\mathbf{w}_t^k, i) - \mathbf{g}_t^k\right\|^2\right] + \mathrm{E}_{k-1}\left[\sum_{i\neq j}\left\langle \mathrm{E}_k^{t-1}\left[\hat{\mathbf{g}}(\mathbf{w}_t^k, i) - \mathbf{g}_t^k\right], \mathrm{E}_k^{t-1}\left[\hat{\mathbf{g}}(\mathbf{w}_t^k, j) - \mathbf{g}_t^k\right]\right\rangle\right]\right) \\
= & \frac{1}{[B^k]^2}\left(\sum_{i=1}^{B^k}\mathrm{E}_{k-1}\left[\left\|\hat{\mathbf{g}}(\mathbf{w}_t^k, i) - \mathbf{g}_t^k\right\|^2\right]\right) \leq \frac{G^2}{B^k},
\end{aligned} \tag{14}
$$

where we make use of the facts $\hat{\mathbf{g}}(\mathbf{w}_t^k, i)$ and $\hat{\mathbf{g}}(\mathbf{w}_t^k, j)$ are independent when $i \neq j$, and

$$
\mathrm{E}_k^{t-1}\left[\hat{\mathbf{g}}(\mathbf{w}_t^k, i) - \mathbf{g}_t^k\right] = 0,\ \mathrm{E}_k^{t-1}\left[\|\hat{\mathbf{g}}(\mathbf{w}_t^k, i) - \mathbf{g}_t^k\|^2\right] \leq \mathrm{E}_k^{t-1}\left[\|\hat{\mathbf{g}}(\mathbf{w}_t^k, i)\|^2\right] \leq G^2,\ \forall i = 1, \ldots, B^k.
$$

Similarly, we also have

$$
\mathrm{E}_{k-1}\left[\|\bar{\mathbf{f}}_t^k - \mathbf{f}_t^k\|^2\right] \leq \frac{G^2}{B^k}. \tag{15}
$$

Notice that $\bar{\mathbf{f}}_t^k$ is an unbiased estimate of $\mathbf{f}_t^k$, thus

$$
\mathrm{E}_{k-1}\left[\langle \mathbf{f}_t^k - \bar{\mathbf{f}}_t^k, \mathbf{z}_t^k - \mathbf{w}_*\rangle\right] = \mathrm{E}_{k-1}\left[\langle \mathrm{E}_k^{t-1}\left[\mathbf{f}_t^k - \bar{\mathbf{f}}_t^k\right], \mathbf{z}_t^k - \mathbf{w}_*\rangle\right] = 0. \tag{16}
$$

Substituting (14), (15), and (16) into (13), we get the second inequality in Lemma 1. □

B. Proof of Lemma 4

Recall that $\bar{\mathbf{g}}_t^k = \frac{1}{B^k}\sum_{i=1}^{B^k}\hat{\mathbf{g}}(\mathbf{w}_t^k, i)$, thus

$$
\|\bar{\mathbf{g}}_t^k - \mathbf{g}_t^k\| = \left\|\frac{1}{B^k}\sum_{i=1}^{B^k}\hat{\mathbf{g}}(\mathbf{w}_t^k, i) - \mathbf{g}_t^k\right\|.
$$

Since $\|\hat{\mathbf{g}}(\mathbf{w}_t^k, i)\| \leq G$, and $\mathrm{E}[\hat{\mathbf{g}}(\mathbf{w}_t^k, i)] = \mathbf{g}_t^k$, we have with a probability at least $1-\delta$

$$\|\bar{\mathbf{g}}_t^k - \mathbf{g}_t^k\| \leq \frac{4G}{\sqrt{B^k}} \log \frac{2}{\delta}.$$

We obtain (8) by the union bound and setting $\tilde{\delta}/2 = M\delta$. The inequality in (9) can be proved in the same way.

C. Proof of Lemma 5

We first state the Bernstein's inequality for martingales (Cesa-Bianchi & Lugosi, 2006), which is used in the proof below.

Theorem 3. *(Bernstein's inequality for martingales). Let $X_1, \ldots, X_n$ be a bounded martingale difference sequence with respect to the filtration $\mathcal{F} = (\mathcal{F}_i)_{1 \leq i \leq n}$ and with $|X_i| \leq K$. Let*

$$S_i = \sum_{j=1}^{i} X_j$$

be the associated martingale. Denote the sum of the conditional variances by

$$\Sigma_n^2 = \sum_{t=1}^{n} \mathrm{E}\left[X_t^2 | \mathcal{F}_{t-1}\right].$$

Then for all constants t, $\nu > 0$,

$$\Pr\left[\max_{i=1,\ldots,n} S_i > t \text{ and } \Sigma_n^2 \leq \nu\right] \leq \exp\left(-\frac{t^2}{2(\nu + Kt/3)}\right),$$

and therefore,

$$\Pr\left[\max_{i=1,\ldots,n} S_i > \sqrt{2\nu t} + \frac{2}{3}Kt \text{ and } \Sigma_n^2 \leq \nu\right] \leq e^{-t}.$$

To simplify the notation, we define

$$\begin{aligned} A &= \sum_{i=1}^{M} \|\mathbf{z}_t^k - \mathbf{w}_*\|^2 \leq \frac{4MG^2}{\lambda^2}, \\ C &= \frac{4G}{\sqrt{B^k}} \log \frac{8M}{\tilde{\delta}}. \end{aligned}$$

In the analysis below, we consider two different scenarios, i.e., $A \leq \eta G^2/[\lambda B^k]$ and $A > \eta G^2/[\lambda B^k]$.

C.1. $A \leq \eta G^2/[\lambda B^k]$

On event E_1, we can bound

$$Z_t^k \leq \|\mathbf{f}_t^k - \bar{\mathbf{f}}_t^k\| \|\mathbf{z}_t^k - \mathbf{w}_*\| \leq \frac{\eta}{4}\|\mathbf{f}_t^k - \bar{\mathbf{f}}_t^k\|^2 + \frac{1}{\eta}\|\mathbf{z}_t^k - \mathbf{w}_*\|^2 \leq \frac{\eta}{4}C^2 + \frac{1}{\eta}\|\mathbf{z}_t^k - \mathbf{w}_*\|^2.$$

Summing up over all $t = 1, 2, \ldots, M$,

$$\sum_{t=1}^{M} Z_t^k \leq \frac{\eta M C^2}{4} + \frac{1}{\eta}\sum_{t=1}^{M} \|\mathbf{z}_t^k - \mathbf{w}_*\|^2 \leq \frac{\eta M C^2}{4} + \frac{G^2}{\lambda B^k}. \tag{17}$$

C.2. $A > \eta G^2/[\lambda B^k]$

Similar to the above proof, on event E_1, we bound

$$|Z_t^k| \leq \|\mathbf{f}_t^k - \bar{\mathbf{f}}_t^k\|\|\mathbf{z}_t^k - \mathbf{w}_*\| \leq \frac{1}{\theta}\|\mathbf{f}_t^k - \bar{\mathbf{f}}_t^k\|^2 + \frac{\theta}{4}\|\mathbf{z}_t^k - \mathbf{w}_*\|^2 \leq \frac{C^2}{\theta} + \frac{\theta A}{4},$$

where θ can be any nonnegative real number. Denote the sum of conditional variances by

$$\Sigma_M^2 = \sum_{t=1}^{M} \mathrm{E}_k^{t-1}\left[[Z_t^k]^2\right] \leq C^2 \sum_{t=1}^{M} \|\mathbf{z}_t - \mathbf{w}_*\|^2 = C^2 A,$$

where $\mathrm{E}_k^{t-1}[\cdot]$ denote the expectation conditioned on all the randomness up to the $t-1$-th iteration in the k-th epoch.

Notice that A in the upper bound for $|Z_t^k|$ and Σ_M^2 is a random variable, thus we cannot directly apply Theorem 3. To address this challenge, we make use of the peeling technique described in (Bartlett et al., 2005), and have

$$\begin{aligned}
&\Pr\left(\sum_{t=1}^{M} Z_t^k \geq 2\sqrt{C^2 A\tau} + \frac{4}{3}\left(\frac{C^2}{\theta} + \frac{\theta A}{4}\right)\tau\right) \\
=&\Pr\left(\sum_{t=1}^{M} Z_t^k \geq 2\sqrt{C^2 A\tau} + \frac{4}{3}\left(\frac{C^2}{\theta} + \frac{\theta A}{4}\right)\tau, \frac{\eta G^2}{\lambda B^k} < A \leq \frac{4MG^2}{\lambda^2}\right) \\
=&\Pr\left(\sum_{t=1}^{M} Z_t^k \geq 2\sqrt{C^2 A\tau} + \frac{4}{3}\left(\frac{C^2}{\theta} + \frac{\theta A}{4}\right)\tau, \max_t |Z_t^k| \leq \frac{C^2}{\theta} + \frac{\theta A}{4}, \Sigma_M^2 \leq C^2 A, \frac{\eta G^2}{\lambda B^k} < A \leq \frac{4MG^2}{\lambda^2}\right) \\
\leq&\sum_{i=1}^{n}\Pr\left(\sum_{t=1}^{M} Z_t^k \geq 2\sqrt{C^2 A\tau} + \frac{4}{3}\left(\frac{C^2}{\theta} + \frac{\theta A}{4}\right)\tau, \max_t |Z_t^k| \leq \frac{C^2}{\theta} + \frac{\theta A}{4}, \Sigma_M^2 \leq C^2 A, \frac{\eta G^2}{\lambda B^k}2^{i-1} < A \leq \frac{\eta G^2}{\lambda B^k}2^i\right) \\
\leq&\sum_{i=1}^{n}\Pr\left(\sum_{t=1}^{M} Z_t^k \geq 2\sqrt{\left(C^2\frac{\eta G^2}{\lambda B^k}2^{i-1}\right)\tau} + \frac{4}{3}\left(\frac{C^2}{\theta} + \frac{\theta}{4}\frac{\eta G^2}{\lambda B^k}2^{i-1}\right)\tau, \max_t |Z_t^k| \leq \frac{C^2}{\theta} + \frac{\theta}{4}\frac{\eta G^2}{\lambda B^k}2^i, \Sigma_M^2 \leq C^2\frac{\eta G^2}{\lambda B^k}2^i\right) \\
\leq&\sum_{i=1}^{n}\Pr\left(\sum_{t=1}^{M} Z_t^k \geq \sqrt{2\left(C^2\frac{\eta G^2}{\lambda B^k}2^{i}\right)\tau} + \frac{2}{3}\left(\frac{C^2}{\theta} + \frac{\theta}{4}\frac{\eta G^2}{\lambda B^k}2^{i}\right)\tau, \max_t |Z_t^k| \leq \frac{C^2}{\theta} + \frac{\theta}{4}\frac{\eta G^2}{\lambda B^k}2^i, \Sigma_M^2 \leq C^2\frac{\eta G^2}{\lambda B^k}2^i\right) \\
\leq& n e^{-\tau},
\end{aligned}$$

where

$$n = \left\lceil \log_2 \frac{4MB^k}{\eta\lambda} \right\rceil,$$

and the last step follows the Bernstein's inequality for martingales in Theorem 3. Setting

$$\begin{aligned}
\theta &= \frac{3\lambda}{4\tau}, \\
\tau &= \log\frac{4n}{\tilde{\delta}},
\end{aligned}$$

with a probability at least $1 - \tilde{\delta}/4$ we have

$$\begin{aligned}
&\sum_{t=1}^{M} Z_t^k \\
\leq& 2\sqrt{C^2 A\tau} + \frac{4}{3}\left(\frac{C^2}{\theta} + \frac{\theta A}{4}\right)\tau = 2\sqrt{C^2 A\tau} + \frac{16C^2}{9\lambda}\tau^2 + \frac{\lambda A}{4} \\
\leq& \frac{4}{\lambda}C^2\tau + \frac{\lambda A}{4} + \frac{16C^2}{9\lambda}\tau^2 + \frac{\lambda A}{4} = \frac{4C^2}{\lambda}\left(\log\frac{4n}{\tilde{\delta}} + \frac{4}{9}\log^2\frac{4n}{\tilde{\delta}}\right) + \frac{\lambda A}{2}.
\end{aligned} \tag{18}$$

We complete the proof by combining (17) and (18).

D. Proof of Lemma 7

We follow the logic used in the proof of Lemma 2.

It is straightforward to check that

$$B^k = \alpha\eta\lambda 2^{k-1} = \frac{2\alpha\eta G^2}{V_k}.$$

When $k = 1$, with a probability $(1 - \tilde{\delta})^{1-1} = 1$, we have

$$\Delta_1 = F(\mathbf{w}_1^1) - F(\mathbf{w}_*) \overset{(1)}{\leq} \frac{2G^2}{\lambda} = \frac{G^2}{\lambda 2^{1-2}} = V_1.$$

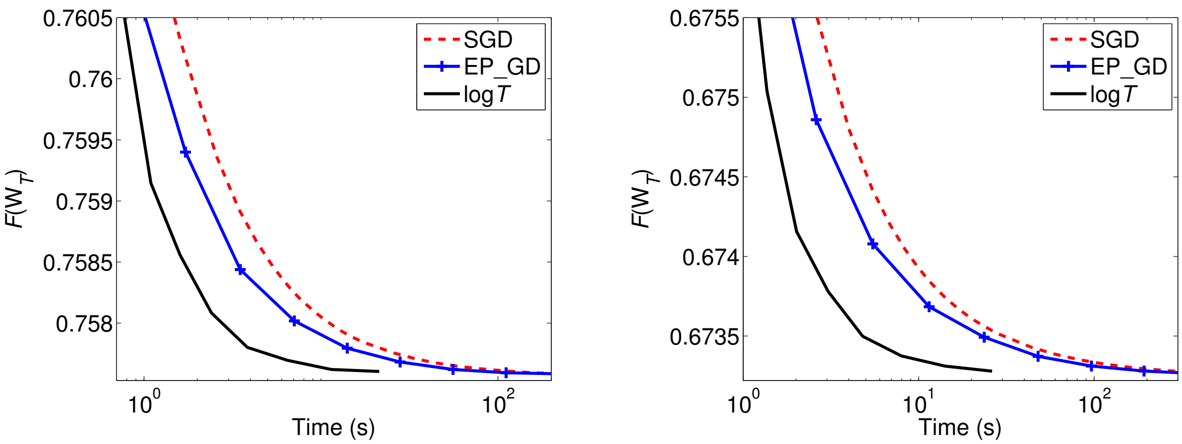
Assume that with a probability at least $(1 - \tilde{\delta})^{k-1}$, $\Delta_k \leq V_k$ for some $k \geq 1$. We now prove the case for $k + 1$. Notice that N defined in (4) is larger than n defined in (10). From Lemma 6, with a probability at least $1 - \tilde{\delta}$, we have

$$\begin{aligned}
&\Delta_{k+1} = F(\mathbf{w}_1^{k+1}) - F(\mathbf{w}_*) \\
\leq& \frac{\|\mathbf{w}_1^k - \mathbf{w}_*\|^2}{2M\eta} + \frac{100G^2\eta}{B^k} \log^2 \frac{8M}{\tilde{\delta}} + \frac{G^2}{\lambda B^k M} \left[1 + 64 \log^2 \frac{8M}{\tilde{\delta}} \left(\log \frac{4N}{\tilde{\delta}} + \frac{4}{9} \log^2 \frac{4N}{\tilde{\delta}}\right)\right] \\
\leq& \frac{\Delta_k}{4} + \frac{400}{\alpha} \log^2 \frac{8M}{\tilde{\delta}} \frac{V_k}{8} + \frac{1}{\alpha} \left[1 + 64 \log^2 \frac{8M}{\tilde{\delta}} \left(\log \frac{4N}{\tilde{\delta}} + \frac{4}{9} \log^2 \frac{4N}{\tilde{\delta}}\right)\right] \frac{V_k}{8}.
\end{aligned}$$

Using the definition of α in (3), with a probability at least $(1 - \tilde{\delta})^k$ we have,

$$\Delta_{k+1} \leq \frac{1}{4} V_k + \frac{1}{8} V_k + \frac{1}{8} V_k = \frac{1}{2} V_k = V_{k+1}.$$

E. More Results for the Regularized Distance Metric Learning



(a) Mushrooms (b) Adult

Figure 3. Results for the regularized distance metric learning on the Mushrooms and Adult data sets. $F(W_T)$ is measured on 10^4 testing pairs and the horizontal axis measures the training time. The experiments are repeated 10 times and the averages are reported.