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, USA
†GE Global Research, San Ramon, CA, USA
‡State Key Laboratory of CAD&CG, College of Computer Science, Zhejiang University, Hangzhou, 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

$$\begin{aligned} \mathbf{w} &= \mathop{\arg\min}_{\mathbf{y}\in\mathcal{U}} \{\langle \gamma\boldsymbol{\xi} - \nabla\omega(\mathbf{z}_-), \mathbf{y}\rangle + \omega(\mathbf{y})\}, \\ \mathbf{z}_+ &= \mathop{\arg\min}_{\mathbf{y}\in\mathcal{U}} \{\langle \gamma\boldsymbol{\eta} - \nabla\omega(\mathbf{z}_-), \mathbf{y}\rangle + \omega(\mathbf{y})\}. \end{aligned}$$

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

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

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
        $\mathbf{g}_t^k = \frac{1}{B^k}\sum_{i=1}^{B^k} \mathbf{g}(\mathbf{w}_t^k, i)$
    Update
        $\mathbf{z}_t^k = \Pi_{\mathcal{D}}\left(\mathbf{w}_t^k - \eta\mathbf{g}_t^k\right)$
    Compute the average gradient at $\mathbf{z}_t^k$ over $B^k$ calls to the gradient oracle
        $\mathbf{f}_t^k = \frac{1}{B^k}\sum_{i=1}^{B^k} \mathbf{g}(\mathbf{z}_t^k, i)$
```

pdate

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

end for

o s p fy the notat on we de ne

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

Let the two nor s $\|\cdot\|$ and $\|\cdot\|$ n Le a be the vector ℓ_2 nor Each terat on n the nner oop sat s es the cond t ons n Le a by do ng the app ngs be ow

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

$$\|\mathbf{z}_t^k$$

Fo ow ng Le a we have

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

wher d f t h e hfth n wes the s oothness

$$\|\mathbf{g}_t^k - \mathbf{f}_t^k\| \quad \|\nabla \mathbf{w}_t^k \ F \ \nabla \mathbf{z}_t^k \ L\|\mathbf{w}_t^k - \mathbf{z}_t^k\|.$$

Dividing both sides by M and fo owing Jensen's inequa ity we have

$$F\left(\frac{}{M}\sum_{t=1}^{M}\mathbf{z}_t^k\right) - F\ \mathbf{w}$$
$$\leq \frac{}{M}\sum_{t=1}^{M} F\ \mathbf{z}_t^k\ - F\ \mathbf{w}$$
$$\leq \frac{\|\mathbf{w}_1^k - \mathbf{w}\ \|^2}{M\eta} + \frac{\eta}{M}\left(\sum_{t=1}^M \|\mathbf{g}_t^k - \mathbf{g}_t^k\|^2 + \sum_{t=1}^M \|\mathbf{f}_t^k - \mathbf{f}_t^k\|^2\right) + \frac{}{M}\sum_{t=1}^M\langle \mathbf{f}_t^k - \mathbf{f}_t^k, \mathbf{z}_t^k - \mathbf{w}\ \rangle - \frac{\lambda}{M}\sum_{t=1}^M \|\mathbf{z}_t^k - \mathbf{w}\ \|^2.$$

which gives the first inequa ity in Le a

Let $\mathrm{E}_{k\ \ 1}\ \cdot$ denote the expectation conditioned on a the rando ness up to epoch $k-$ and $\mathrm{E}_k^{t\ \ 1}\ \cdot$ denote the expectation conditioned on a the rando ness up to the $t-$ th iteration in the k th epoch a ing the conditiona expectation of we have

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

where we drop the ast ter since it is negative o bound $\mathrm{E}_{k\ \ 1}\left[\|\mathbf{g}_t^k-\mathbf{g}_t^k\|^2\right]$ we have

$$\mathrm{E}_{k\ \ 1}\left[\|\mathbf{g}_t^k-\mathbf{g}_t^k\|^2\right]\quad \mathrm{E}_{k\ \ 1}\left[\left\|\frac{}{B^k}\sum_{i=1}^{B^k}\mathbf{g}\ \mathbf{w}_t^k, i\ -\mathbf{g}_t^k\right\|^2\right]\quad \mathrm{E}_{k\ \ 1}\left[\left\|\frac{}{B^k}\sum_{i=1}^{B^k}\left(\mathbf{g}\ \mathbf{w}_t^k,i\ -\mathbf{g}_t^k\right)\right\|^2\right]$$
$$\frac{}{B^{k\ 2}}\left(\sum_{i=1}^{B^k}\mathrm{E}_{k\ \ 1}\left[\left\|\mathbf{g}\ \mathbf{w}_t^k,i\ -\mathbf{g}_t^k\right\|^2\right] + \mathrm{E}_{k\ \ 1}\left[\sum_{i=j}\left\langle \mathrm{E}_k^{t\ \ 1}\left[\mathbf{g}\ \mathbf{w}_t^k,i\ -\mathbf{g}_t^k\right], \mathrm{E}_k^{t\ \ 1}\left[\mathbf{g}\ \mathbf{w}_t^k,j\ -\mathbf{g}_t^k\right]\right\rangle\right]\right)$$
$$\frac{}{B^{k\ 2}}\left(\sum_{i=1}^{B^k}\mathrm{E}_{k\ \ 1}\left[\left\|\mathbf{g}\ \mathbf{w}_t^k,i\ -\mathbf{g}_t^k\right\|^2\right]\right)\leq\frac{G^2}{B^k},$$

where we a e use of the facts $\mathbf{g}\ \mathbf{w}_t^k, i$ and $\mathbf{g}\ \mathbf{w}_t^k, j$ are independent when $i\ /\ j$ and

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

i i ar y we a so have

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

Notice that $\mathbf{f}_t^k$ is an unbiased esti ate of $\mathbf{f}_t^k$ thus

$$\mathrm{E}_{k\ \ 1}\left[\langle\mathbf{f}_t^k-\mathbf{f}_t^k,\mathbf{z}_t^k-\mathbf{w}\ \rangle\right]\quad \mathrm{E}_{k\ \ 1}\left[\left\langle \mathrm{E}_k^{t\ \ 1}\left[\mathbf{f}_t^k-\mathbf{f}_t^k\right],\mathbf{z}_t^k-\mathbf{w}\ \right\rangle\right]\qquad .$$

ubstituting and into we get the second inequa ity in Le a □

B. Proof of Lemma

Reca that $\mathbf{g}_t^k\quad \frac{1}{B^k}\sum_{i=1}^{B^k}\mathbf{g}\ \mathbf{w}_t^k, i$ thus

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

nce $\|\mathbf{g}\ \mathbf{w}_t^k, i\ \| \leq G$ and $\mathrm{E}\,\mathbf{g}\ \mathbf{w}_t^k, i\qquad \mathbf{g}_t^k$ we have w th a probab ty at east $-\delta$

$$\|\mathbf{g}_t^k -$$

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

ı ı ar to the above proof on event E_1 we bound

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

where θ can be any nonnegatıve rea nu ber Denote the su of condıtıona varıances by

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

where $\mathrm{E}_k^{t \quad 1} \cdot$ denote the expectatıon condıtıoned on a the rando ness up to the $t -$ th ıteratıon ın the k th epoch

Notıce that A ın the upper bound for $|Z_t^k|$ and ^2_M ıs a rando varıab e thus we cannot dırect y app y heore o address thıs cha enge we a e use of the pee ıng technıque descrıbed ın Bart ett et a and have

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

where

$$n \quad \left\lceil \text{og}_2 \frac{MB^k}{\eta\lambda} \right\rceil,$$

and the ast step fo ows the Bernsteın s ınequa ıty for artınga es ın heore ettıng

$$
\begin{aligned}
\theta &\quad \frac{\lambda}{\tau},\\
\tau &\quad \text{og} \frac{n}{\delta},
\end{aligned}
$$

wıth a probabı ıty at east $- \delta/$ we have

$$
\begin{aligned}
&\sum_{t=1}^M Z_t^k\\
\leq& \sqrt{C^2 A\tau} + \; - \left(\frac{C^2}{\theta} + \frac{\theta A}{}\right)\tau \quad \sqrt{C^2 A\tau} + \frac{C^2}{\lambda}\tau^2 + \frac{\lambda A}{}\\
\leq& \frac{}{\lambda} C^2\tau + \frac{\lambda A}{} + \frac{C^2}{\lambda}\tau^2 + \frac{\lambda A}{} \quad \frac{C^2}{\lambda}\left(\text{og} \frac{n}{\delta} + \; - \text{og}^2 \frac{n}{\delta}\right) + \frac{\lambda A}{}.
\end{aligned}
$$

e co p ete the proof by co bınıng and

D. Proof of Lemma 7

e fo ow the ogic used in the proof of Le a

It is straightforward to chec that

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

hen k with a probability $-\delta \; ^{1 \quad 1}$ we have

$$_1 \quad F \;\; \mathbf{w}_1^1 \;\; - F \;\; \mathbf{w} \quad \overset{(1)}{\leq} \frac{G^2}{\lambda} \quad \frac{G^2}{\lambda \;\; ^{1 \quad 2}} \quad V_1.$$

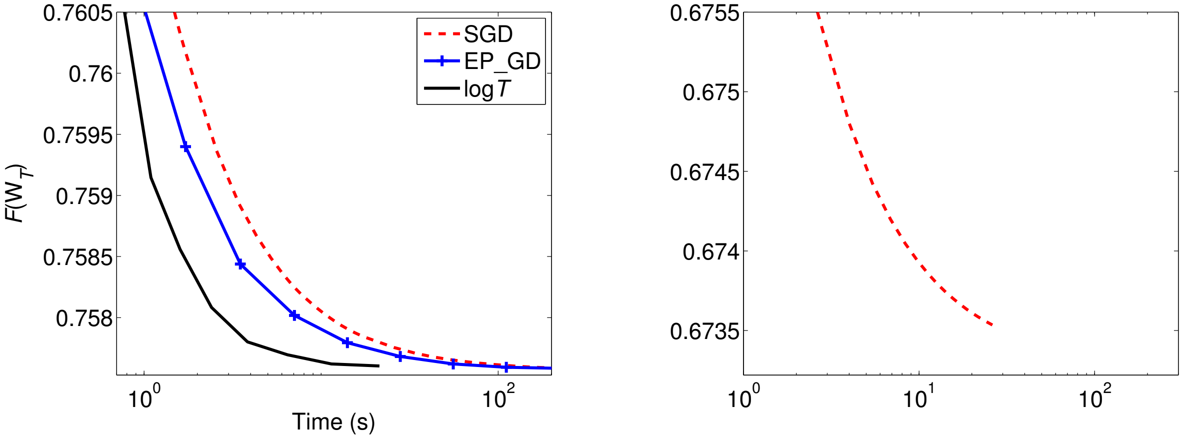
Assu e that with a probability at east $-\delta \; ^{k \quad 1}$ $_k \leq V_k$ for so e $k \geq$ e now prove the case for $k+$ Notice that N de ned in is arger than n de ned in Fro Le a with a probability at east $-\delta$ we have

$$\begin{aligned} & _{k+1} \quad F \;\; \mathbf{w}_1^{k+1} \;\; - F \;\; \mathbf{w} \\ \leq & \frac{\|\mathbf{w}_1^k - \mathbf{w} \;\; \|^2}{M\eta} + \quad \frac{G^2\eta}{B^k} \text{ og}^2 \frac{M}{\delta} + \frac{G^2}{\lambda B^k M}\left[+ \quad \text{ og}^2 \frac{M}{\delta} \left(\text{og} \frac{N}{\delta} + \; - \; \text{og}^2 \frac{N}{\delta} \right) \right] \\ \leq & \frac{\;\; k}{} + \quad \frac{}{\alpha} \text{ og}^2 \frac{M}{\delta} \frac{V_k}{} + \; \frac{}{\alpha} \left[+ \quad \text{ og}^2 \frac{M}{\delta} \left(\text{og} \frac{N}{\delta} + \; - \; \text{og}^2 \frac{N}{\delta} \right) \right] \frac{V_k}{}. \end{aligned}$$

sing the de nition of α in with a probability at east $-\delta \; ^{k}$ we have

$$_{k+1} \leq -V_k + \;\; -V_k + \;\; -V_k \quad -V_k \quad V_{k+1}.$$

E. More Results for the Regularized Distance Metric Learning



(a) Mushrooms