Supplementary Material: Stochastic Optimization for Kernel PCA

Lijun Zhang[1,2] and **Tianbao Yang**[3] and **Jinfeng Yi**[4] and **Rong Jin**[5] and **Zhi-Hua Zhou**[1,2]
[1]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[2]Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, China
[3]Department of Computer Science, the University of Iowa, Iowa City, USA
[4]IBM Thomas J. Watson Research Center, Yorktown Heights, USA
[5]Alibaba Group, Seattle, USA
{zhanglj, zhouzh}@lamda.nju.edu.cn, tianbao-yang@uiowa.edu, jinfengy@us.ibm.com, jinrong.jr@alibaba-inc.com

Proof of Theorem 1

The proof is similar to that of Proposition 1 in (Rakhlin, Shamir, and Sridharan 2012), which shows the $O(1/T)$ rate of SGD. The differences are (i) we need to consider the effect of the non-smooth regularizer $\lambda\|Z\|_*$, and (ii) we use a different step size which leads to a dramatic simplification of the analysis involving martingales.

First, we present one lemma that bounds the difference between Z_{t+1} and $\widehat{K}$ by the difference between Z_t and $\widehat{K}$.

Lemma 1 *Define*

$$\delta_t = \left\langle \xi_t - K, Z_t - \widehat{K} \right\rangle, \textit{ and } S^2 = \max_{t\in[T]} \|Z_t - \xi_t\|_F^2.$$

We have

$$\|Z_{t+1} - \widehat{K}\|_F^2 \leq \eta_t^2 S^2 + 2\eta_t\delta_t + 2\eta_t\lambda\left(\|Z_t\|_* - \|Z_{t+1}\|_*\right) + (1-2\eta_t)\|Z_t - \widehat{K}\|_F^2.$$

The above lemma is proved by exploiting the optimality of Z_{t+1}.

Based on Lemma 1, we obtain the following upper bound for $\|Z_{T+1} - \widehat{K}\|_F$ by choosing an appropriate step size.

Lemma 2 *Define* $\gamma = \max_{t\in[T]} \|Z_t\|_*$. *By setting* $\eta_t = 2/t$, *we have*

$$\|Z_{T+1} - \widehat{K}\|_F^2 \leq \frac{4(S^2+\lambda\gamma)}{T} + \frac{2}{T(T-1)}\left[2\sum_{t=2}^{T}(t-1)\delta_t - \sum_{t=2}^{T}(t-1)\|Z_t - \widehat{K}\|_F^2\right].$$

Note that by taking expectation over both sides, we immediately get an $O(1/T)$ upper bound that holds in expectation. Since our goal is to prove a stronger high probability bound, we proceed by developing an upper bound for the summation of martingale difference sequence $\sum_{t=2}^{T}(t-1)\delta_t$.

Lemma 3 *Assume*

$$\|\xi_t - K\|_F \leq G, \textit{ and } \|Z_t - \widehat{K}\|_F \leq D, \ \forall t > 2.$$

With a probability at least $1-\delta$, *we have*

$$\sum_{t=2}^{T}(t-1)\delta_t \leq \frac{1}{2}\sum_{t=2}^{T}(t-1)\|Z_t - \widehat{K}\|_F^2 + 2G^2\tau(T-1) + \frac{2}{3}GD(T-1)\tau + GD(T-1)$$

where $\tau = \log\frac{\lceil 2\log_2 T\rceil}{\delta}$.

The above lemma is built up the Bernstein's inequality for martingales (Cesa-Bianchi and Lugosi 2006) and the peeling technique (Bartlett, Bousquet, and Mendelson 2005).

Combining Lemmas 2 and 3, with a probability at least $1-\delta$, we have

$$\|Z_{T+1} - \widehat{K}\|_F^2 \leq \frac{4}{T}\left(S^2 + \lambda\gamma + 2G^2\tau + \frac{2}{3}GD\tau + GD\right). \tag{4}$$

As can be seen, there is a nice cancellation of $\sum_{t=2}^{T}(t-1)\|Z_t - \widehat{K}\|_F^2$, which is due to the special setting $\eta_t = 2/t$.

Then, we provide upper bounds for the constants S, γ, G, and D in the above inequality.

Lemma 4 *Assume* $\|\xi\|_F \leq C$. *By setting* $\eta_t = 2/t$, *we have*

$$S^2 \leq 10C^2, \ \gamma \leq 2C\max_{t\in[T]}\sqrt{r_t}, \ G = 2C, \textit{ and } D = 3C$$

where r_t *is the rank of* Z_t.

We complete the proof by substituting those upper bounds into (4).

Proof of Lemma 1

To simplify notations, we define

$$F(Z) = \frac{1}{2}\mathrm{E}\left[\|Z - \xi\|_F^2\right], \text{ and } f_t(Z) = \frac{1}{2}\|Z - \xi_t\|_F^2.$$

From the property of strongly convex, i.e., (2) of (Hazan and Kale 2011), the updating rule implies

$$\begin{aligned}
&\frac{1}{2}\|Z_{t+1} - Z_t\|_F^2 + \eta_t\langle Z_{t+1} - Z_t, \nabla f_t(Z_t)\rangle + \eta_t\lambda\|Z_{t+1}\|_* \\
\leq&\frac{1}{2}\|\widehat{K} - Z_t\|_F^2 + \eta_t\langle \widehat{K} - Z_t, \nabla f_t(Z_t)\rangle + \eta_t\lambda\|\widehat{K}\|_* - \frac{1}{2}\|\widehat{K} - Z_{t+1}\|_F^2.
\end{aligned} \tag{5}$$

Since $F(Z)$ is 1-strongly convex, we also have

$$\begin{aligned}
&\frac{1}{2}\|Z_t - \widehat{K}\|_F^2 \\
\leq&F(Z_t) + \lambda\|Z_t\|_* - F(\widehat{K}) - \lambda\|\widehat{K}\|_* \\
\leq&\langle Z_t - \widehat{K}, \nabla F(Z_t)\rangle - \frac{1}{2}\|Z_t - \widehat{K}\|_F^2 + \lambda\|Z_t\|_* - \lambda\|\widehat{K}\|_* \\
=&\langle Z_t - \widehat{K}, \nabla f_t(Z_t)\rangle - \lambda\|\widehat{K}\|_* - \frac{1}{2\eta_t}\|Z_t - \widehat{K}\|_F^2 \\
&+ \lambda\|Z_t\|_* - \frac{1}{2}\|Z_t - \widehat{K}\|_F^2 + \frac{1}{2\eta_t}\|Z_t - \widehat{K}\|_F^2 + \langle \nabla F(Z_t) - \nabla f_t(Z_t), Z_t - \widehat{K}\rangle \\
\overset{(5)}{\leq}&\langle Z_t - Z_{t+1}, \nabla f_t(Z_t)\rangle - \lambda\|Z_{t+1}\|_* - \frac{1}{2\eta_t}\|Z_t - Z_{t+1}\|_F^2 - \frac{1}{2\eta_t}\|Z_{t+1} - \widehat{K}\|_F^2 + \\
&+ \lambda\|Z_t\|_* + \frac{1}{2}\left(\frac{1}{\eta_t} - 1\right)\|Z_t - \widehat{K}\|_F^2 + \langle \nabla F(Z_t) - \nabla f_t(Z_t), Z_t - \widehat{K}\rangle \\
\leq&\max_W\left(\langle W, \nabla f_t(Z_t)\rangle - \frac{1}{2\eta_t}\|W\|_F^2\right) - \lambda\|Z_{t+1}\|_* - \frac{1}{2\eta_t}\|Z_{t+1} - \widehat{K}\|_F^2 + \\
&+ \lambda\|Z_t\|_* + \frac{1}{2}\left(\frac{1}{\eta_t} - 1\right)\|Z_t - \widehat{K}\|_F^2 + \langle \nabla F(Z_t) - \nabla f_t(Z_t), Z_t - \widehat{K}\rangle \\
=&\frac{\eta_t}{2}\|\nabla f_t(Z_t)\|_F^2 + \lambda\|Z_t\|_* - \lambda\|Z_{t+1}\|_* + \frac{1}{2}\left(\frac{1}{\eta_t} - 1\right)\|Z_t - \widehat{K}\|_F^2 - \frac{1}{2\eta_t}\|Z_{t+1} - \widehat{K}\|_F^2 \\
&+ \langle \nabla F(Z_t) - \nabla f_t(Z_t), Z_t - \widehat{K}\rangle.
\end{aligned} \tag{6}$$

We complete the proof by noticing

$$S^2 = \max_{t\in[T]}\|\nabla f_t(Z_t)\|_F^2, \text{ and } \langle \nabla F(Z_t) - \nabla f_t(Z_t), Z_t - \widehat{K}\rangle = \langle \xi_t - K, Z_t - \widehat{K}\rangle.$$

Proof of Lemma 2

From Lemma 1 and the definition of η_t, we have

$$\begin{aligned}
&\|Z_{t+1} - \widehat{K}\|_F^2 \\
\leq&\frac{4S^2}{t^2} + \frac{4\lambda}{t}\left(\|Z_t\|_* - \|Z_{t+1}\|_*\right) + \left(1 - \frac{4}{t}\right)\|Z_t - \widehat{K}\|_F^2 + \frac{4\delta_t}{t} \\
=&\frac{4S^2}{t^2} + \frac{4\lambda}{t}\left(\|Z_t\|_* - \|Z_{t+1}\|_*\right) + \frac{t-2}{t}\|Z_t - \widehat{K}\|_F^2 + \frac{2}{t}\left(2\delta_t - \|Z_t - \widehat{K}\|_F^2\right).
\end{aligned}$$

Following the strategy in (Rakhlin, Shamir, and Sridharan 2012), we unwind the above recursive inequality from $t = T$ till

$t = 2$ and obtain

$$
\begin{aligned}
&\|Z_{T+1} - \widehat{K}\|_F^2 \\
\leq &\frac{4S^2}{T^2} + \frac{4\lambda}{T}\left(\|Z_T\|_* - \|Z_{T+1}\|_*\right) + \frac{2}{T}\left(2\delta_T - \|Z_T - \widehat{K}\|_F^2\right) \\
&+ \frac{T-2}{T}\left[\frac{4S^2}{(T-1)^2} + \frac{4\lambda}{T-1}\left(\|Z_{T-1}\|_* - \|Z_T\|_*\right) + \frac{2}{T-1}\left(2\delta_{T-1} - \|Z_{T-1} - \widehat{K}\|_F^2\right)\right] \\
&+ \frac{T-2}{T}\frac{T-3}{T-1}\|Z_{T-1} - \widehat{K}\|_F^2 \\
\leq &\cdots \\
\leq &4S^2\sum_{t=2}^{T}\frac{1}{t^2}\prod_{i=t+1}^{T}\frac{i-2}{i} + 4\lambda\sum_{t=2}^{T}\frac{1}{t}\prod_{i=t+1}^{T}\frac{i-2}{i}\left(\|Z_t\|_* - \|Z_{t+1}\|_*\right) \\
&+ 2\sum_{t=2}^{T}\frac{1}{t}\prod_{i=t+1}^{T}\frac{i-2}{i}\left(2\delta_t - \|Z_t - \widehat{K}\|_F^2\right)
\end{aligned} \tag{7}
$$

Since $\prod_{i=t+1}^{T}\frac{i-2}{i} = \frac{t(t-1)}{T(T-1)}$, we have

$$
\begin{aligned}
&\sum_{t=2}^{T}\frac{1}{t^2}\prod_{i=t+1}^{T}\frac{i-2}{i} = \sum_{t=2}^{T}\frac{1}{t^2}\frac{t(t-1)}{T(T-1)} \leq \frac{1}{T}, \\
&\sum_{t=2}^{T}\frac{1}{t}\prod_{i=t+1}^{T}\frac{i-2}{i}\left(\|Z_t\|_* - \|Z_{t+1}\|_*\right) \\
=&\frac{1}{T(T-1)}\sum_{t=2}^{T}(t-1)\left(\|Z_t\|_* - \|Z_{t+1}\|_*\right) \leq \frac{1}{T(T-1)}\sum_{t=2}^{T}\|Z_t\|_* \leq \frac{\gamma}{T}, \\
&\sum_{t=2}^{T}\frac{1}{t}\prod_{i=t+1}^{T}\frac{i-2}{i}\left(2\delta_t - \|Z_t - \widehat{K}\|_F^2\right) = \frac{1}{T(T-1)}\sum_{t=2}^{T}(t-1)\left(2\delta_t - \|Z_t - \widehat{K}\|_F^2\right)
\end{aligned}
$$

We complete the proof by substituting the above inequalities into (7).

Proof of Lemma 3

We need the Bernstein's inequality for martingales (Cesa-Bianchi and Lugosi 2006), which is stated below.

Theorem 2 *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,

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

From the assumption about the random matrix, it is easy to check that $b_t = (t-1)\delta_t, t = 2, \dots, T$ is a martingale difference sequence. Furthermore,

$$|b_t| \leq (t-1)\|\xi_t - K\|_F \|Z_t - \widehat{K}\|_F \leq GD(T-1).$$

Define the martingale $B_T = \sum_{t=2}^T b_t$. Define the conditional variance Σ_T^2 as

$$\Sigma_T^2 = \sum_{t=2}^T \mathrm{E}_{t-1}\left[b_t^2\right] \leq (T-1)G^2 \underbrace{\sum_{t=2}^T (t-1)\|Z_t - \widehat{K}\|_F^2}_{A_T}$$

where $\mathrm{E}_{t-1}[\cdot]$ denotes the expectation conditioned on all the randomness up to the $t-1$-th iteration.

In the following, we consider two different scenarios: $A_T \leq D^2$ and $A_T > D^2$.

$A_T \leq D^2$ In this case, we have

$$B_T = \sum_{t=2}^T b_t \leq G \sum_{t=2}^T (t-1)\|Z_t - \widehat{K}\|_F \leq G\sqrt{\sum_{t=2}^T (t-1)}\sqrt{A_T} \leq GD(T-1). \tag{8}$$

$A_T > D^2$ Since A_T in the upper bound for Σ_T^2 is a random variable, we cannot apply Bernstein's inequality directly. To address this issue, we make use of the peeling process described in (Bartlett, Bousquet, and Mendelson 2005). Notice that we have both a lower bound and an upper bound for A_T, i.e.,

$$D^2 < A_T \leq T^2 D^2.$$

We have

$$\begin{aligned}
&\Pr\left[B_T \geq 2\sqrt{(T-1)G^2 A_T \tau} + \frac{2}{3}GD(T-1)\tau\right] \\
=&\Pr\left[B_T \geq 2\sqrt{(T-1)G^2 A_T \tau} + \frac{2}{3}GD(T-1)\tau, D^2 < A_T \leq T^2 D^2\right] \\
=&\Pr\left[B_T \geq 2\sqrt{(T-1)G^2 A_T \tau} + \frac{2}{3}GD(T-1)\tau, \Sigma_T^2 \leq (T-1)G^2 A_T, D^2 < A_T \leq T^2 D^2\right] \\
\leq&\sum_{i=1}^m \Pr\left[B_T \geq 2\sqrt{(T-1)G^2 A_T \tau} + \frac{2}{3}GD(T-1)\tau, \Sigma_T^2 \leq (T-1)G^2 A_T, D^2 2^{i-1} < A_T \leq D^2 2^i\right] \\
\leq&\sum_{i=1}^m \Pr\left[B_T \geq \sqrt{2(T-1)G^2 D^2 2^i \tau} + \frac{2}{3}GD(T-1)\tau, \Sigma_T^2 \leq (T-1)G^2 D^2 2^i\right] \\
\leq& m e^{-\tau},
\end{aligned}$$

where $m = \lceil 2\log_2 T \rceil$, and the last step follows from Theorem 2. By setting

$$\tau = \log \frac{m}{\delta},$$

with a probability at least $1-\delta$, we have

$$B_T \leq 2\sqrt{(T-1)G^2 A_T \tau} + \frac{2}{3}GD(T-1)\tau. \tag{9}$$

Combining (8) and (9), with a probability at least $1-\delta$, we have

$$\begin{aligned}
B_T \leq& 2\sqrt{(T-1)G^2 A_T \tau} + \frac{2}{3}GD(T-1)\tau + GD(T-1) \\
\leq& \frac{1}{2}A_T + 2G^2\tau(T-1) + \frac{2}{3}GD(T-1)\tau + GD(T-1).
\end{aligned}$$

Proof of Lemma 4

We first prove $\|Z_t\|_F \leq 2C, \forall t \geq 1$ by induction. Since $\|Z_1\|_F = 0$, we have

$$\|Z_2\|_F = \|\mathcal{D}_{\eta_1 \lambda}\left[2\xi_1\right]\|_F \leq \|2\xi_1\|_F \leq 2C.$$

Suppose $\|Z_t\|_F \leq 2C$ for some $t \geq 2$. We will show that it leads to $\|Z_{t+1}\|_F \leq 2C$. To see this, we have

$$\begin{aligned}\|Z_{t+1}\|_F =& \|\mathcal{D}_{\eta_t \lambda}\left[(1-\eta_t)Z_t + \eta_t \xi_t\right]\|_F \leq \|(1-\eta_t)Z_t + \eta_t \xi_t\|_F \\ \leq& \|(1-\eta_t)Z_t\|_F + \|\eta_t \xi_t\|_F = \frac{t-2}{t}\|Z_t\|_F + \frac{2}{t}\|\xi_t\|_F \leq 2C.\end{aligned}$$

Then, for the constant S in Lemma 1, we have

$$S^2 \leq \max_{t \in [T]} 2\|Z_t\|_F^2 + 2\|\xi_t\|_F^2 \leq 10C^2.$$

For the constant γ in Lemma 2, we have

$$\gamma = \max_{t \in [T]} \|Z_t\|_* \leq \max_{t \in [T]} \sqrt{r_t}\|Z_t\|_F \leq 2C \max_{t \in [T]} \sqrt{r_t}$$

where r_t is the rank of Z_t. Since

$$\begin{aligned}\|\xi_t - K\|_F \leq \|\xi_t\|_F + \|K\|_F \leq \|\xi_t\|_F + \mathrm{E}\left[\|\xi\|_F\right] \leq 2C, \\ \|Z_t - \widehat{K}\|_F \leq \|Z_t\|_F + \|\widehat{K}\|_F \leq 2C + \|K\|_F \leq 3C,\end{aligned}$$

we can set G and D in Lemma 3 as

$$G = 2C, \text{ and } D = 3C.$$