Supplementary Material:
A Single-Pass Algorithm for Efficiently Recovering Sparse Cluster Centers of High-dimensional Data

Jinfeng Yi JINFENGY@US.IBM.COM

IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA

Lijun Zhang ZHANGLJ@LAMDA.NJU.EDU.CN

National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

Jun Wang WANGJUN@US.IBM.COM

IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA

Rong Jin RONGJIN@CSE.MSU.EDU

Anil K. Jain JAIN@CSE.MSU.EDU

Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824 USA

Theorem 1. *Let $\epsilon \leq 1/(6m)$ be a parameter to control the success probability. Assume*

$$\Delta_* \leq \Delta^1 \leq \Delta_{\max}, \tag{1}$$

$$\frac{\Delta^1}{2\sqrt{2s}} \leq \lambda^1 \leq c\frac{\Delta^1}{2\sqrt{2s}}, \tag{2}$$

$$T \geq \max\left(\frac{18}{\mu_0}\ln\frac{2K}{\epsilon}, \frac{3c_2\eta_0}{\lambda^1}, \left(\frac{6c_3\sigma}{\lambda^1}\right)^2(\ln n + \ln d)\right) \tag{3}$$

where c, c_2 and c_3 are some universal constants. Then, with a probability at least $1 - 6m\epsilon$, we have

$$\Delta^{m+1} = \max_{1\leq i\leq K}\|\widehat{\mathbf{c}}_i^{m+1} - \mathbf{c}_i\| \leq \max\left(\Delta_*, \frac{c\Delta^1}{\sqrt{2^m}}\right).$$

Corollary 1. *The convergence rate for Δ, the maximum difference between the optimal cluster centers and the estimated ones, is $O(\sqrt{(s\log d)/n})$ before reaching the optimal difference Δ_*.*

1. Proof of Corollary 1

According to the assumption of λ^1 in (2), we know that $\frac{1}{\lambda^1} \propto \frac{\sqrt{s}}{\Delta^1}$. Since the value of T is dominated by the last term in the right side of (3), we have $T \propto \frac{s\log d}{\Delta^1 \cdot \Delta^1}$, which implies

$$n \propto 2^m T \propto 2^m \frac{s\log d}{\Delta^1 \cdot \Delta^1}.$$

Combining with the conclusion $\Delta_{m+1} \propto \frac{\Delta^1}{\sqrt{2^m}}$, we have

$$\Delta_{m+1} \propto \sqrt{\frac{s \log d}{n}}.$$

Lemma 1. *Let Δ^t be the maximum difference between the optimal cluster centers and the ones estimated from iteration t, and $\epsilon \in (0,1)$ be the failure probability. Assume*

$$\Delta^t \leq \frac{1-\rho}{2} - \sigma\sqrt{5\ln(3K)} \triangleq \Delta_{\max}, \tag{4}$$

$$|\mathcal{S}^t| \geq \frac{18}{\mu_0} \ln \frac{2K}{\epsilon}, \tag{5}$$

$$\lambda^t \geq c_1 \exp\left(-\frac{(1-2\Delta^t-\rho)^2}{8(1+\Delta^t)^2\sigma^2}\right)(\eta_0 + \sigma\sqrt{\ln|\mathcal{S}^t|}) + \frac{c_2\eta_0}{|\mathcal{S}^t|} + c_3\sigma\frac{\sqrt{\ln|\mathcal{S}^t|}+\sqrt{\ln d}}{\sqrt{|\mathcal{S}^t|}}, \tag{6}$$

for some constants c_1, c_2 and c_3. Then with a probability $1-6\epsilon$, we have

$$\Delta^{t+1} \leq 2\sqrt{s}\lambda^t.$$

2. Proof of Lemma 1

For the simplicity of analysis, we will drop the superscript t through this analysis.

2.1. Preliminaries

We denote by $\mathcal{C}_k$ the support of $\mathbf{c}_k$ and $\overline{\mathcal{C}}_k = [d] \setminus \mathcal{C}_k$. For any vector $\mathbf{z}$, $\mathbf{z}(\mathcal{C})$ is defined as $[\mathbf{z}(\mathcal{C})]_i = z_i$ if $i \in \mathcal{C}$ and zero, otherwise.

For any $\mathbf{x}_i \in \mathcal{S}$, we use k_i to denote the index of the true cluster, and $\widehat{k}_i$ to denote index of the cluster assigned by the nearest neighbor search, i.e.,

$$\begin{aligned} \mathbf{x}_i =& \mathbf{c}_{k_i} + \mathbf{g}_i \text{ and } \mathbf{g}_i \sim N(0, \sigma^2 I), \\ \widehat{k}_i =& \arg\max_{j \in [K]} \widehat{\mathbf{c}}_j^\top \mathbf{x}_i. \end{aligned}$$

Then, we can partition data points in $\mathcal{S}$ based on either the ground truth or the assigned cluster. Let $\mathcal{S}_k$ be the subset of data points in $\mathcal{S}$ that belong to the k-th cluster, i.e.,

$$\mathcal{S}_k = \{\mathbf{x}_i \in \mathcal{S} : \mathbf{x}_i = \mathbf{c}_k + \mathbf{g}_i \text{ and } \mathbf{g}_i \sim N(0, \sigma^2 I)\} \tag{7}$$

Let $\widehat{\mathcal{S}}_k$ be the subset of data points that are assigned to the k-th cluster based on the nearest neighbor search, i.e.,

$$\widehat{\mathcal{S}}_k = \{\mathbf{x}_i \in \mathcal{S} : k = \arg\max_{j \in [K]} \widehat{\mathbf{c}}_j^\top \mathbf{x}_i\} \tag{8}$$

2.2. The Main Analysis

Let $\mathcal{L}_k(\mathbf{c})$ be the objective function in Step 11 of Algorithm 1. We expand $\mathcal{L}_k(\mathbf{c})$ as

$$\begin{aligned} &\mathcal{L}_k(\mathbf{c}) \\ =&\lambda\|\mathbf{c}\|_1 + \|\mathbf{c}-\mathbf{c}_k\|^2 + \frac{1}{|\widehat{\mathcal{S}}_k|}\sum_{\mathbf{x}_i \in \widehat{\mathcal{S}}_k} \|\mathbf{x}_i - \mathbf{c}_k\|^2 - \frac{2}{|\widehat{\mathcal{S}}_k|}\sum_{\mathbf{x}_i \in \widehat{\mathcal{S}}_k} (\mathbf{c}-\mathbf{c}_k)^\top(\mathbf{x}_i - \mathbf{c}_k) \\ =&\lambda\|\mathbf{c}\|_1 + \|\mathbf{c}-\mathbf{c}_k\|^2 + \frac{1}{|\widehat{\mathcal{S}}_k|}\sum_{\mathbf{x}_i \in \widehat{\mathcal{S}}_k} \|\mathbf{x}_i - \mathbf{c}_k\|^2 \\ &- \underbrace{2(\mathbf{c}-\mathbf{c}_k)^\top \frac{1}{|\widehat{\mathcal{S}}_k|}\sum_{\mathbf{x}_i \in \widehat{\mathcal{S}}_k \setminus \mathcal{S}_k} (\mathbf{c}_{k_i} - \mathbf{c}_k)}_{A_k} - \underbrace{2(\mathbf{c}-\mathbf{c}_k)^\top \frac{1}{|\widehat{\mathcal{S}}_k|}\sum_{\mathbf{x}_i \in \widehat{\mathcal{S}}_k} \mathbf{g}_i}_{B_k}. \end{aligned} \tag{9}$$

Let $\mathbf{c}_k^*$ be the optimal solution that minimizes $\mathcal{L}_k(\mathbf{c})$, and define $\mathbf{f}_k = \mathbf{c}_k^* - \mathbf{c}_k$. We have

$$\begin{aligned}
&\mathcal{L}_k(\mathbf{c}_k^*) - \mathcal{L}_k(\mathbf{c}_k)\\
=&\lambda\|\mathbf{f}_k + \mathbf{c}_k\|_1 + \|\mathbf{f}_k\|^2 - 2\mathbf{f}_k^\top A_k - 2\mathbf{f}_k^\top B_k - \lambda\|\mathbf{c}_k\|_1\\
\geq&\lambda\|\mathbf{c}_k\|_1 - \lambda\|\mathbf{f}_k(\mathcal{C}_k)\|_1 + \lambda\|\mathbf{f}_k(\overline{\mathcal{C}}_k)\|_1 + \|\mathbf{f}_k\|^2 - 2\mathbf{f}_k^\top A_k - 2\mathbf{f}_k^\top B_k - \lambda\|\mathbf{c}_k\|_1\\
\geq& - \lambda\|\mathbf{f}_k(\mathcal{C}_k)\|_1 + \lambda\|\mathbf{f}_k(\overline{\mathcal{C}}_k)\|_1 + \|\mathbf{f}_k\|^2 - 2\|\mathbf{f}_k\|_1\|A_k\|_\infty - 2\|\mathbf{f}_k\|_1\|B_k\|_\infty\\
=& - (\lambda + 2\|A_k\|_\infty + 2\|B_k\|_\infty)\|\mathbf{f}_k(\mathcal{C}_k)\|_1 + (\lambda - 2\|A_k\|_\infty - 2\|B_k\|_\infty)\|\mathbf{f}_k(\overline{\mathcal{C}}_k)\|_1 + \|\mathbf{f}_k\|^2\\
\geq& - \sqrt{|\mathcal{C}_k|}(\lambda + 2\|A_k\|_\infty + 2\|B_k\|_\infty)\|\mathbf{f}_k(\mathcal{C}_k)\| + (\lambda - 2\|A_k\|_\infty - 2\|B_k\|_\infty)\|\mathbf{f}_k(\overline{\mathcal{C}}_k)\|_1 + \|\mathbf{f}_k\|^2.
\end{aligned}$$

Thus, if

$$\lambda \geq 2\|A_k\|_\infty + 2\|B_k\|_\infty,$$

we have

$$\|\mathbf{f}_k(\mathcal{C}_k)\|^2 \leq \|\mathbf{f}_k\|^2 \leq (\lambda + 2\|A_k\|_\infty + 2\|B_k\|_\infty)\sqrt{|\mathcal{C}_k|}\|\mathbf{f}_k(\mathcal{C}_k)\| \leq 2\lambda\sqrt{|\mathcal{C}_k|}\|\mathbf{f}_k(\mathcal{C}_k)\| \Rightarrow \|\mathbf{f}_k(\mathcal{C}_k)\| \leq 2\lambda\sqrt{|\mathcal{C}_k|},$$

and thus

$$\|\mathbf{f}_k\|^2 \leq 2\lambda\sqrt{|\mathcal{C}_k|}\|\mathbf{f}_k(\mathcal{C}_k)\| \leq 4\lambda^2|\mathcal{C}_k| \Rightarrow \|\mathbf{f}_k\| \leq 2\lambda\sqrt{|\mathcal{C}_k|}.$$

In summary, if

$$\lambda \geq 2\|A_k\|_\infty + 2\|B_k\|_\infty, \forall k \in [K]$$

we have

$$\max_{1\leq k\leq K} \|\mathbf{c}_k^* - \mathbf{c}_k\| \leq 2\sqrt{s}\lambda.$$

In the following, we discuss how to bound $\|A_k\|_\infty$ and $\|B_k\|_\infty$.

2.3. Bound for $\|A_k\|_\infty$

From the definition of A_k in (9), we have

$$\|A_k\|_\infty \leq 2\eta_0\frac{|\widehat{\mathcal{S}}_k \setminus \mathcal{S}_k|}{|\widehat{\mathcal{S}}_k|}.$$

2.3.1. Lower bound of $|\widehat{\mathcal{S}}_k|$

First, we show that the size of $\mathcal{S}_k$ is lower-bounded, which means a significant amount of data points in S belong to the k-th cluster. Recall that $\mu_1, \ldots, \mu_K$ are the weight of the Gaussian mixtures, and $\mu_0 = \min_{1\leq i\leq K} \mu_i$. According to the Chernoff bound (Angluin & Valiant, 1979) provided in Appendix A, we have, with a probability at least $1-\epsilon$

$$|\mathcal{S}_k| \geq \mu_k|\mathcal{S}|\left(1 - \sqrt{\frac{2}{\mu_k|\mathcal{S}|}\ln\frac{K}{\epsilon}}\right) \overset{(5)}{\geq} \frac{2}{3}\mu_k|\mathcal{S}|,\ \forall k \in [K]. \tag{10}$$

Next, we prove that a larger amount of data points in $\mathcal{S}_k$ belong to $\widehat{\mathcal{S}}_k$. We begin by analyzing the probability that the assigned cluster $\widehat{k}_i$ of $\mathbf{x}_i$ is the true cluster k_i. The similarity between $\mathbf{x}_i$ and the estimated cluster centers can be bounded by

$$\begin{aligned}
\widehat{\mathbf{c}}_{k_i}^\top\mathbf{x}_i =&\widehat{\mathbf{c}}_{k_i}^\top(\mathbf{c}_{k_i} + \mathbf{g}_i) = \|\mathbf{c}_{k_i}\|^2 + [\widehat{\mathbf{c}}_{k_i} - \mathbf{c}_{k_i}]^\top\mathbf{c}_{k_i} + \widehat{\mathbf{c}}_{k_i}^\top\mathbf{g}_i\\
&\geq 1 - \|\widehat{\mathbf{c}}_{k_i} - \mathbf{c}_{k_i}\| - |\widehat{\mathbf{c}}_{k_i}^\top\mathbf{g}_i| \geq 1 - \Delta - (1+\Delta)\left|\mathbf{g}_i^\top\frac{\widehat{\mathbf{c}}_{k_i}}{\|\widehat{\mathbf{c}}_{k_i}\|}\right|,\\
\widehat{\mathbf{c}}_j^\top\mathbf{x}_i =&\widehat{\mathbf{c}}_j^\top(\mathbf{c}_{k_i} + \mathbf{g}_i) = \mathbf{c}_j^\top\mathbf{c}_{k_i} + [\widehat{\mathbf{c}}_j - \mathbf{c}_j]^\top\mathbf{c}_{k_i} + \widehat{\mathbf{c}}_j^\top\mathbf{g}_i\\
&\leq \rho + \|\widehat{\mathbf{c}}_j - \mathbf{c}_j\| + |\widehat{\mathbf{c}}_j^\top\mathbf{g}_i| \leq \rho + \Delta + (1+\Delta)\left|\mathbf{g}_i^\top\frac{\widehat{\mathbf{c}}_j}{\|\widehat{\mathbf{c}}_j\|}\right|,\ j \neq k_i.
\end{aligned}$$

Hence, $\mathbf{x}_i$ will be assigned to cluster k_i if

$$1 - \Delta - (1+\Delta)\left|\mathbf{g}_i^\top \frac{\widehat{\mathbf{c}}_{k_i}}{\|\widehat{\mathbf{c}}_{k_i}\|}\right| \geq \rho + \Delta + (1+\Delta)\left|\mathbf{g}_i^\top \frac{\widehat{\mathbf{c}}_j}{\|\widehat{\mathbf{c}}_j\|}\right|, \ \forall j \neq k_i,$$

which leads to the following sufficient condition

$$\max_{1\leq j\leq K}\left|\mathbf{g}_i^\top \frac{\widehat{\mathbf{c}}_j}{\|\widehat{\mathbf{c}}_j\|}\right| \leq \frac{1-2\Delta-\rho}{2(1+\Delta)} \triangleq g_0 \overset{(4)}{\geq} \frac{2\sigma\sqrt{5\ln(3K)}}{3} \geq \sigma\sqrt{2\ln(3K)}. \tag{11}$$

It is easy to verify that for any fixed direction $\widehat{\mathbf{c}}$ with $\|\widehat{\mathbf{c}}\| = 1$, $\mathbf{g}_i^\top \mathbf{c}$ is a Gaussian random variable with mean 0 and variance σ^2. Based on the tail bound for the Gaussian distribution (Chang et al., 2011) provided in Appendix B, we have

$$\Pr\left[\max_{1\leq j\leq K}\left|\mathbf{g}_i^\top \frac{\widehat{\mathbf{c}}_j}{\|\widehat{\mathbf{c}}_j\|}\right| \leq g_0\right] \geq 1 - K\exp\left(-\frac{g_0^2}{2\sigma^2}\right).$$

Define

$$\delta = K\exp\left(-\frac{g_0^2}{2\sigma^2}\right) \overset{(11)}{\leq} \frac{1}{3}. \tag{12}$$

In summary, we have proved the following lemma.

Lemma 2. *Under the condition in (4), with a probability at least* $1-\delta$, $\mathbf{x}_i = \mathbf{c}_{k_i} + \mathbf{g}_i \in \mathcal{S}_{k_i} \subset S$ *satisfies*

$$\max_{1\leq j\leq K}\left|\mathbf{g}_i^\top \frac{\widehat{\mathbf{c}}_j}{\|\widehat{\mathbf{c}}_j\|}\right| \leq g_0,$$

and is assigned to the correct cluster k_i *based on the nearest neighbor search (i.e.,* $\widehat{k}_i = k_i$*).*

Define

$$\mathcal{S}_k^1 = \left\{\mathbf{x}_i \in \mathcal{S}_k : \max_{1\leq j\leq K}\left|\mathbf{g}_i^\top \frac{\widehat{\mathbf{c}}_j}{\|\widehat{\mathbf{c}}_j\|}\right| \leq g_0\right\} \subset \widehat{\mathcal{S}}_k \cap \mathcal{S}_k. \tag{13}$$

Since each data point in $\mathcal{S}_k$ has a probability at least $1-\delta$ to be assigned to set $\mathcal{S}_k^1$, using the Chernoff bound again, we have, with a probability at least $1-\epsilon$,

$$\begin{aligned} |\widehat{\mathcal{S}}_k| \geq |\widehat{\mathcal{S}}_k \cap \mathcal{S}_k| \geq |\mathcal{S}_k^1| &\geq \mathrm{E}\left[|\mathcal{S}_k^1|\right]\left(1 - \sqrt{\frac{2}{\mathrm{E}\left[|\mathcal{S}_k^1|\right]}\ln\frac{K}{\epsilon}}\right) \\ &\geq (1-\delta)\,|\mathcal{S}_k|\left(1 - \sqrt{\frac{2}{(1-\delta)\,|\mathcal{S}_k|}\ln\frac{K}{\epsilon}}\right) \\ &\overset{(12)}{\geq} \frac{2}{3}|\mathcal{S}_k|\left(1 - \sqrt{\frac{3}{|\mathcal{S}_k|}\ln\frac{K}{\epsilon}}\right) \overset{(5),(10)}{\geq} \frac{1}{3}|\mathcal{S}_k|, \forall k \in [K]. \end{aligned} \tag{14}$$

2.3.2. UPPER BOUND OF $|\widehat{\mathcal{S}}_k \setminus \mathcal{S}_k|$

Define

$$\mathcal{O} = \cup_{k=1}^K \mathcal{S}_k^1 \subset \mathcal{S} \text{ and } \overline{\mathcal{O}} = \cup_{k=1}^K \left(\widehat{\mathcal{S}}_k \setminus \mathcal{S}_k^1\right) = \mathcal{S} \setminus \mathcal{O} \subset \mathcal{S}.$$

From Lemma 2, we know that with a probability at least $1-\delta$, each $\mathbf{x}_i \in \mathcal{S}_k$ belongs to the set $\mathcal{S}_k^1 \subset \mathcal{O}$. Thus, with probability at least $1-\delta$, each $\mathbf{x}_i \in \mathcal{S}$ belongs to $\mathcal{O}$. In other words, with probability *at most* δ, each $\mathbf{x}_i \in \mathcal{S}$ belongs to $\overline{\mathcal{O}}$. Based on the Chernoff bound, we have, with a probability at least $1-\epsilon$,

$$|\overline{\mathcal{O}}| \leq 2\mathrm{E}\left[|\overline{\mathcal{O}}|\right] + 2\ln\frac{1}{\epsilon} \leq 2\delta|\mathcal{S}| + 2\ln\frac{1}{\epsilon}. \tag{15}$$

Since $\mathcal{S}_k^1 \subset \mathcal{S}_k$, we have $\widehat{\mathcal{S}}_k \setminus \mathcal{S}_k \subset \widehat{\mathcal{S}}_k \setminus \mathcal{S}_k^1 \subset \overline{\mathcal{O}}$. Therefore, with a probability at least $1-\epsilon$, we have

$$|\widehat{\mathcal{S}}_k \setminus \mathcal{S}_k| \leq 2\delta|\mathcal{S}| + 2\ln\frac{1}{\epsilon}, \forall k \in [K]. \tag{16}$$

Combining (10), (14) and (16), we have, with probability at least $1-3\epsilon$

$$\|A_k\|_\infty \leq 2\eta_0 \frac{2\delta|\mathcal{S}| + 2\ln\frac{1}{\epsilon}}{\frac{2}{9}\mu_k|\mathcal{S}|} = \frac{18\eta_0}{\mu_k}\left(\delta + \frac{1}{|\mathcal{S}|}\ln\frac{1}{\epsilon}\right) = O(\delta\eta_0) + O\left(\frac{\eta_0}{|\mathcal{S}|}\right), \forall k \in [K]. \tag{17}$$

2.4. Bound for $\|B_k\|_\infty$

Notice that $\{\mathbf{g}_i : \mathbf{x}_i \in \widehat{\mathcal{S}}_k\}$, determined by the estimated centers $\widehat{\mathbf{c}}_1, \ldots, \widehat{\mathbf{c}}_K$, is a specific subset of $\{\mathbf{g}_i : \mathbf{x}_i \in \mathcal{S}\}$. Although $\mathbf{g}_i$ is drawn from the Gaussian distribution $N(0, \sigma^2 I)$, the distribution of elements in $\{\mathbf{g}_i : \mathbf{x}_i \in \widehat{\mathcal{S}}_k\}$ is unknown. As a result, we cannot direct apply concentration inequality of Gaussian random vectors to bound $\|B_k\|_\infty$. Let $U_1 \in \mathbb{R}^{d\times K}$ be a matrix whose columns are basis vectors of the subspace spanned by $\widehat{\mathbf{c}}_1, \ldots, \widehat{\mathbf{c}}_K$, and $U_2 \in \mathbb{R}^{d\times(d-K)}$ be a matrix whose columns are basis vectors of the complementary subspace. We then divide each $\mathbf{g}_i$ as

$$\mathbf{g}_i = \mathbf{g}_i^{\|} + \mathbf{g}_i^{\perp},$$

where $\mathbf{g}_i^{\|} = U_1U_1^\top \mathbf{g}_i$, and $\mathbf{g}_i^{\perp} = U_2U_2^\top \mathbf{g}_i$.

First, we upper bound $\|B_k\|_\infty$ as

$$\|B_k\|_\infty \leq \underbrace{\left\|\frac{1}{|\widehat{\mathcal{S}}_k|}\sum_{\mathbf{x}_i\in\widehat{\mathcal{S}}_k}\mathbf{g}_i^{\perp}\right\|_\infty}_{\widehat{B}_k^1} + \underbrace{\frac{|\widehat{\mathcal{S}}_k\backslash\mathcal{S}_k^1|}{|\widehat{\mathcal{S}}_k|}\left\|\frac{1}{|\widehat{\mathcal{S}}_k\backslash\mathcal{S}_k^1|}\sum_{\mathbf{x}_i\in\widehat{\mathcal{S}}_k\backslash\mathcal{S}_k^1}\mathbf{g}_i^{\|}\right\|_\infty}_{\widehat{B}_k^2} + \underbrace{\frac{|\mathcal{S}_k^1|}{|\widehat{\mathcal{S}}_k|}\left\|\frac{1}{|\mathcal{S}_k^1|}\sum_{\mathbf{x}_i\in\mathcal{S}_k^1}\mathbf{g}_i^{\|}\right\|_\infty}_{\widehat{B}_k^3}. \tag{18}$$

In the following, we discuss how to bound each term in the right hand side of (18).

2.4.1. Upper bound of $\widehat{B}_k^1$

Following the property of Gaussian random vector, $\sum_{\mathbf{x}_i\in\widehat{\mathcal{S}}_k} U_2^\top\mathbf{g}_i/\left(\sigma\sqrt{|\widehat{\mathcal{S}}_k|}\right)$ can be treated as a $(d-K)$-dimensional Gaussian random vector. As a result, each element of $U_2\sum_{\mathbf{x}_i\in\widehat{\mathcal{S}}_k} U_2^\top\mathbf{g}_i/\left(\sigma\sqrt{|\widehat{\mathcal{S}}_k|}\right)$ is a Gaussian random variable with variance smaller than 1. Based on the tail bound for the Gaussian distribution (Chang et al., 2011) provided in Appendix B and the union bound, with a probability at least $1-\epsilon$, we have

$$\left\|\sum_{\mathbf{x}_i\in\widehat{\mathcal{S}}_k}\mathbf{g}_i^{\perp}/\left(\sigma\sqrt{|\widehat{\mathcal{S}}_k|}\right)\right\|_\infty = \left\|U_2\sum_{\mathbf{x}_i\in\widehat{\mathcal{S}}_k}U_2^\top\mathbf{g}_i/\left(\sigma\sqrt{|\widehat{\mathcal{S}}_k|}\right)\right\|_\infty \leq \sqrt{2\ln\frac{Kd}{\epsilon}}, \forall k\in[K],$$

which implies

$$\widehat{B}_k^1 \leq \sigma\sqrt{\frac{2\ln\frac{Kd}{\epsilon}}{|\widehat{\mathcal{S}}_k|}} \overset{(10),(14)}{\leq} \sigma\sqrt{\frac{2\ln\frac{Kd}{\epsilon}}{2\mu_k|\mathcal{S}|/9}} = O\left(\sigma\sqrt{\frac{\ln d}{|\mathcal{S}|}}\right), \forall k\in[K]. \tag{19}$$

2.4.2. Upper bound of $\widehat{B}_k^2$

First, we have

$$\left\|\frac{1}{|\widehat{\mathcal{S}}_k\backslash\mathcal{S}_k^1|}\sum_{\mathbf{x}_i\in\widehat{\mathcal{S}}_k\backslash\mathcal{S}_k^1}\mathbf{g}_i^{\|}\right\|_\infty = \left\|\frac{1}{|\widehat{\mathcal{S}}_k\backslash\mathcal{S}_k^1|}\sum_{\mathbf{x}_i\in\widehat{\mathcal{S}}_k\backslash\mathcal{S}_k^1}U_1U_1^\top\mathbf{g}_i\right\|_\infty \leq \left\|\frac{1}{|\widehat{\mathcal{S}}_k\backslash\mathcal{S}_k^1|}\sum_{\mathbf{x}_i\in\widehat{\mathcal{S}}_k\backslash\mathcal{S}_k^1}U_1^\top\mathbf{g}_i\right\| \tag{20}$$

Since $U_1^\top\mathbf{g}_i/\sigma$ can be treated as a K-dimensional Gaussian random vector, based on the tail bound for the χ^2 distribution (Laurent & Massart, 2000), we have with a probability at least $1-\epsilon$,

$$\|U_1^\top\mathbf{g}_i\| \leq \sigma\left(\sqrt{K} + \sqrt{2\log\frac{1}{\epsilon}}\right)$$

Applying the union bound again, with a probability at least $1-\epsilon$, we have

$$\max_{1\leq i\leq |\mathcal{S}|}\left\|U_1^\top \mathbf{g}_i\right\| \leq \sigma\left(\sqrt{K}+\sqrt{2\log\frac{|\mathcal{S}|}{\epsilon}}\right) \tag{21}$$

Combining (20) and (21), we have

$$\widehat{B}_k^2 \leq \frac{9\sigma}{\mu_k}\left(\delta+\frac{1}{|\mathcal{S}|}\ln\frac{1}{\epsilon}\right)\left(\sqrt{K}+\sqrt{2\log\frac{|\mathcal{S}|}{\epsilon}}\right) = O(\delta\sigma\sqrt{\ln|\mathcal{S}|})+O\left(\sigma\frac{\sqrt{\ln|\mathcal{S}|}}{|\mathcal{S}|}\right), \forall k\in[K]. \tag{22}$$

2.4.3. Upper bound of $\widehat{B}_k^3$

First, we have

$$\left\|\frac{1}{|\mathcal{S}_k^1|}\sum_{\mathbf{x}_i\in\mathcal{S}_k^1}\mathbf{g}_i^{\|}\right\|_\infty = \left\|U_1\frac{1}{|\mathcal{S}_k^1|}\sum_{\mathbf{x}_i\in\mathcal{S}_k^1}U_1^\top\mathbf{g}_i\right\|_\infty \leq \left\|\frac{1}{|\mathcal{S}_k^1|}\sum_{\mathbf{x}_i\in\mathcal{S}_k^1}U_1^\top\mathbf{g}_i\right\| := u_k \tag{23}$$

Recall the definition of $\mathcal{S}_k^1$ in (13). Due to the fact that the domain is symmetric, we have $\mathrm{E}\left[U_1^\top\mathbf{g}_i\right]=0$. Under the condition in (21), we can invoke the following lemma to bound u_k.

Lemma 3. *(Lemma 2 from (Smale & Zhou, 2007)) Let $\mathcal{H}$ be a Hilbert space and ξ be a random variable on (Z,ρ) with values in $\mathcal{H}$. Assume $\|\xi\|\leq M<\infty$ almost surely. Denote $\sigma^2(\xi)=\mathrm{E}(\|\xi\|^2)$. Let $\{z_i\}_{i=1}^m$ be independent random drawers of ρ. For any $0<\delta<1$, with confidence $1-\delta$,*

$$\left\|\frac{1}{m}\sum_{i=1}^m(\xi_i-\mathrm{E}[\xi_i])\right\| \leq \frac{2M\ln(2/\delta)}{m}+\sqrt{\frac{2\sigma^2(\xi)\ln(2/\delta)}{m}}$$

From Lemma 3 and the union bound, with a probability at least $1-\epsilon$, we have

$$u_k \leq \sigma\left(\sqrt{K}+\sqrt{2\log\frac{|\mathcal{S}|}{\epsilon}}\right)\left(\frac{2\ln(2K/\epsilon)}{|\mathcal{S}_k^1|}+\sqrt{\frac{2\ln(2K/\epsilon)}{|\mathcal{S}_k^1|}}\right),\ \forall k\in[K]. \tag{24}$$

Combining (23) and (24), we have

$$\begin{aligned}\widehat{B}_k^3 &\leq \sigma\left(\sqrt{K}+\sqrt{2\log\frac{|\mathcal{S}|}{\epsilon}}\right)\left(\frac{2}{|\mathcal{S}_k^1|}\ln\frac{2K}{\epsilon}+\sqrt{\frac{2}{|\mathcal{S}_k^1|}\ln\frac{2K^2}{\epsilon}}\right)\\ &\overset{(10),(14),(5)}{\leq} \sigma\left(\sqrt{K}+\sqrt{2\log\frac{|\mathcal{S}|}{\epsilon}}\right)2\sqrt{\frac{9}{\mu_k|\mathcal{S}|}\ln\frac{2K}{\epsilon}} = O\left(\sigma\sqrt{\frac{\ln|\mathcal{S}|}{|\mathcal{S}|}}\right), \forall k\in[K].\end{aligned} \tag{25}$$

In summary, under the condition that (10), (14) and (15) are true, with a probability at least $1-3\epsilon$,

$$\|B_k\|_\infty \leq O(\delta\sigma\sqrt{\ln|\mathcal{S}|})+O\left(\sigma\frac{\sqrt{\ln|\mathcal{S}|}+\sqrt{\ln d}}{\sqrt{|\mathcal{S}|}}\right), \forall k\in[K]. \tag{26}$$

A. Chernoff Bound

Theorem 2 (Multiplicative Chernoff Bound (Angluin & Valiant, 1979))**.** *Let $X_1, X_2,\dots,X_n$ be independent binary random variables with $\Pr[X_i=1]=p_i$. Denote $S=\sum_{i=1}^n X_i$ and $\mu=\mathrm{E}[S]=\sum_{i=1}^n p_i$. We have*

$$\Pr\left[S\leq(1-\epsilon)\mu\right]\leq\exp\left(-\frac{\epsilon^2}{2}\mu\right),\ \textit{for } 0<\epsilon<1,$$

$$\Pr\left[S\geq(1+\epsilon)\mu\right]\leq\exp\left(-\frac{\epsilon^2}{2+\epsilon}\mu\right),\ \textit{for } \epsilon>0.$$

Therefore,

$$\Pr\left[S \leq \left(1-\sqrt{\frac{2}{\mu}\ln\frac{1}{\delta}}\right)\mu\right] \leq \delta, \text{ for } \exp\left(-\frac{2}{\mu}\right) < \delta < 1,$$
$$\Pr\left[S \geq 2\mu + 2\ln\frac{1}{\delta} \geq \left(1+\frac{\ln\frac{1}{\delta}+\sqrt{2\mu\ln\frac{1}{\delta}}}{\mu}\right)\mu\right] \leq \delta, \text{ for } 0 < \delta < 1.$$

B. Tail bounds for the Gaussian distribution

Theorem 3 (Chernoff-type upper bound for the Q-function (Chang et al., 2011))**.** *The Q-function defined as*

$$Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty \exp\left(-\frac{t^2}{2}\right)dt$$

is the tail probability of the standard Gaussian distribution. When $x > 0$, we have

$$Q(x) \leq \frac{1}{2}\exp\left(-\frac{x^2}{2}\right).$$

Let $X \sim \mathcal{N}(0,1)$ be a Gaussian random variable. According to Theorem 3, we have

$$\Pr\left[|X| \geq \epsilon\right] \leq \exp\left(-\frac{\epsilon^2}{2}\right), \text{or}$$
$$\Pr\left[|X| \geq \sqrt{2\ln\frac{1}{\delta}}\right] \leq \delta.$$

References

Angluin, D. and Valiant, L.G. Fast probabilistic algorithms for hamiltonian circuits and matchings. *Journal of Computer and System Sciences*, 18(2):155–193, 1979.

Chang, Seok-Ho, Cosman, Pamela C., and Milstein, Laurence B. Chernoff-type bounds for the gaussian error function. *IEEE Transactions on Communications*, 59(11):2939–2944, 2011.

Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28 (5):1302–1338, 2000.

Smale, Steve and Zhou, Ding-Xuan. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007.