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* $1/(6m)$ *be a parameter to control the success probability. Assume*

$$ {}_{*} \qquad ; \tag{1} $$

$$ \frac{p}{2\sqrt{2s}} \qquad c\frac{p}{2\sqrt{2s}}; \tag{2} $$

$$ T \quad \max\left(\frac{18}{} \ln \frac{2K}{}; \frac{3c}{}; \left(\frac{6c}{}\right) (\ln n + \ln d) \right) \tag{3} $$

where c, c *and* c *are some universal constants. Then, with a probability at least* $1 \quad 6m$ *, we have*

$$ {}^{m} = \max_{\leq i \leq K} \|\hat{c}_i^m \quad c_i\| \quad \max\left({}_{*}; \frac{c}{p\sqrt{2^m}} \right): $$

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

1. Proof of Corollary 1

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

$$ n \;/\; 2^m T \;/\; 2^m \frac{s \log d}{}: $$

Combining with the conclusion $_m \propto \frac{1}{\sqrt{m}}$, we have

$$_m \propto \sqrt{\frac{s \log d}{n}}.$$

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

$$^t \quad \frac{1}{2} \quad \sqrt{5 \ln(3K)}\,, \quad ; \tag{4}$$

$$|S^t| \quad \frac{18}{} \ln \frac{2K}{}; \tag{5}$$

$$^t \quad c \exp\left(\frac{(1 \quad 2 \quad ^t \quad)}{8(1 + \quad ^t)} \right)(\quad + \quad \sqrt{\ln |S^t|}) + \frac{c}{|S^t|} + c \quad \frac{\sqrt{\ln |S^t|} + \sqrt{\ln d}}{\sqrt{|S^t|}}; \tag{6}$$

for some constants c, c *and* c. *Then with a probability* $1 \quad 6$, *we have*

$$^t \quad 2\sqrt{s} \quad ^t.$$

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

For any $\mathbf{x}_i \in S$, 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.,

$$\mathbf{x}_i = \mathbf{c}_{k_i} + \mathbf{g}_i \text{ and } \mathbf{g}_i \quad N(0, \quad I);$$
$$\widehat{k}_i = \arg\max_{j \in K} \widehat{\mathbf{c}}_j^\top \mathbf{x}_i.$$

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

$$S_k = \{\mathbf{x}_i \in S : \mathbf{x}_i = \mathbf{c}_k + \mathbf{g}_i \text{ and } \mathbf{g}_i \quad N(0, \quad I)\} \tag{7}$$

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

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

2.2. The Main Analysis

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

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

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

$$
\begin{aligned}
& L_k(\mathbf{c}_k^*) \quad L_k(\mathbf{c}_k) \\
= \; & k\mathbf{f}_k + \mathbf{c}_k k \; + k\mathbf{f}_k k \quad 2\mathbf{f}_k^\top A_k \quad 2\mathbf{f}_k^\top B_k \quad k\mathbf{c}_k k \\
& k\mathbf{c}_k k \quad k\mathbf{f}_k(C_k)k \; + \; k\mathbf{f}_k(\overline{C}_k)k \; + k\mathbf{f}_k k \quad 2\mathbf{f}_k^\top A_k \quad 2\mathbf{f}_k^\top B_k \quad k\mathbf{c}_k k \\
& k\mathbf{f}_k(C_k)k \; + \; k\mathbf{f}_k(\overline{C}_k)k \; + k\mathbf{f}_k k \quad 2k\mathbf{f}_k k \; kA_k k_\infty \quad 2k\mathbf{f}_k k \; kB_k k_\infty \\
= \; & (\; + 2kA_k k_\infty + 2kB_k k_\infty)k\mathbf{f}_k(C_k)k \; + (\quad 2kA_k k_\infty \quad 2kB_k k_\infty)k\mathbf{f}_k(\overline{C}_k)k \; + k\mathbf{f}_k k \\
& \sqrt{jC_k j}(\; + 2kA_k k_\infty + 2kB_k k_\infty)k\mathbf{f}_k(C_k)k + (\quad 2kA_k k_\infty \quad 2kB_k k_\infty)k\mathbf{f}_k(\overline{C}_k)k \; + k\mathbf{f}_k k :
\end{aligned}
$$

Thus, if

$$2kA_k k_\infty + 2kB_k k_\infty;$$

we have

$$k\mathbf{f}_k(C_k)k \quad k\mathbf{f}_k k \quad (\; + 2kA_k k_\infty + 2kB_k k_\infty)\sqrt{jC_k j}k\mathbf{f}_k(C_k)k \quad 2 \; \sqrt{jC_k j}k\mathbf{f}_k(C_k)k \;) \; k\mathbf{f}_k(C_k)k \quad 2 \; \sqrt{jC_k j};$$

and thus

$$k\mathbf{f}_k k \quad 2 \; \sqrt{jC_k j}k\mathbf{f}_k(C_k)k \quad 4 \quad jC_k j \;) \; k\mathbf{f}_k k \quad 2 \; \sqrt{jC_k j}:$$

In summary, if

$$2kA_k k_\infty + 2kB_k k_\infty; 8k \; 2 \; [K]$$

we have

$$\max_{\le k \le K} k\mathbf{c}_k^* \quad \mathbf{c}_k k \quad 2^p \overline{s} \; :$$

In the following, we discuss how to bound $kA_k k_\infty$ and $kB_k k_\infty$.

2.3. Bound for $kA_k k_\infty$

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

$$kA_k k_\infty \quad 2 \quad \frac{j\widehat{S}_k \; n \; S_k j}{j\widehat{S}_k j}:$$

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

First, we show that the size of S_k is lower-bounded, which means a significant amount of data points in S belong to the k-th cluster. Recall that $\; ; : : : ; \; _K$ are the weight of the Gaussian mixtures, and $\; = \min_{\le i \le K} \; _i$. According to the Chernoff bound (Angluin & Valiant, 1979) provided in Appendix A, we have, with a probability at least 1

$$jS_k j \quad _k jSj \left(1 \quad \sqrt{\frac{2}{_k jSj} \ln \frac{K}{}} \right) \stackrel{(5)}{} \frac{2}{3} \; _k jSj; \; 8k \; 2 \; [K]: \tag{10}$$

Next, we prove that a larger amount of data points in S_k belong to $\widehat{S}_k$. 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

$$
\begin{aligned}
\widehat{\mathbf{c}}_{k_i}^\top \mathbf{x}_i = & \widehat{\mathbf{c}}_{k_i}^\top (\mathbf{c}_{k_i} + \mathbf{g}_i) = k\mathbf{c}_{k_i} k \; + [\widehat{\mathbf{c}}_{k_i} \quad \mathbf{c}_{k_i}]^\top \mathbf{c}_{k_i} + \widehat{\mathbf{c}}_{k_i}^\top \mathbf{g}_i \\
& 1 \quad k\widehat{\mathbf{c}}_{k_i} \quad \mathbf{c}_{k_i} k \quad j\widehat{\mathbf{c}}_{k_i}^\top \mathbf{g}_i j \quad 1 \quad (1 + \quad) \left| \mathbf{g}_i^\top \frac{\widehat{\mathbf{c}}_{k_i}}{k\widehat{\mathbf{c}}_{k_i} k} \right| ; \\
\widehat{\mathbf{c}}_j^\top \mathbf{x}_i = & \widehat{\mathbf{c}}_j^\top (\mathbf{c}_{k_i} + \mathbf{g}_i) = \mathbf{c}_j^\top \mathbf{c}_{k_i} + [\widehat{\mathbf{c}}_j \quad \mathbf{c}_j]^\top \mathbf{c}_{k_i} + \widehat{\mathbf{c}}_j^\top \mathbf{g}_i \\
& + k\widehat{\mathbf{c}}_j \quad \mathbf{c}_j k + j\widehat{\mathbf{c}}_j^\top \mathbf{g}_i j \quad + \quad + (1 + \quad) \left| \mathbf{g}_i^\top \frac{\widehat{\mathbf{c}}_j}{k\widehat{\mathbf{c}}_j k} \right| ; \; j \neq k_i :
\end{aligned}
$$

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

$$1 \quad (1 + \quad) \left| \mathbf{g}_i^\top \frac{\widehat{\mathbf{c}}_{k_i}}{\|\widehat{\mathbf{c}}_{k_i}\|} \right| \quad + \quad + (1 + \quad) \left| \mathbf{g}_i^\top \frac{\widehat{\mathbf{c}}_j}{\|\widehat{\mathbf{c}}_j\|} \right|; 8j \not 6 k_i;$$

which leads to the following sufficient condition

$$\max_{\leq j \leq K} \left| \mathbf{g}_i^\top \frac{\widehat{\mathbf{c}}_j}{\|\widehat{\mathbf{c}}_j\|} \right| \quad \frac{1 \quad 2}{2(1 + \quad)} \, , \, g \quad \stackrel{(4)}{} \frac{2 \quad \sqrt{5 \ln(3K)}}{3} \quad \sqrt{2 \ln(3K)}: \tag{11}$$

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 . Based on the tail bound for the Gaussian distribution (Chang et al., 2011) provided in Appendix B, we have

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

Define

$$= K \exp \left(\quad \frac{g}{2} \right) \stackrel{(11)}{} \frac{1}{3}: \tag{12}$$

In summary, we have proved the following lemma.

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

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

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

Define

$$S_k = \left\{ \mathbf{x}_i \; 2 \; S_k : \max_{\leq j \leq K} \left| \mathbf{g}_i^\top \frac{\widehat{\mathbf{c}}_j}{\|\widehat{\mathbf{c}}_j\|} \right| \quad g \right\} \quad \widehat{S}_k \setminus S_k: \tag{13}$$

Since each data point in S_k has a probability at least $1 \quad$ to be assigned to set S_k, using the Chernoff bound again, we have, with a probability at least $1 \quad$,

$$\begin{aligned} j\widehat{S}_k j \quad j\widehat{S}_k \setminus S_k j \quad jS_k j \quad \mathrm{E}\left[jS_k j\right] &\left(1 \quad \sqrt{\frac{2}{\mathrm{E}\left[jS_k j\right]} \ln \frac{K}{}}\right) \\ (1 \quad) \, jS_k j &\left(1 \quad \sqrt{\frac{2}{(1 \quad) \, jS_k j} \ln \frac{K}{}}\right) \\ \stackrel{(12)}{} \quad \frac{2}{3} jS_k j &\left(1 \quad \sqrt{\frac{3}{jS_k j} \ln \frac{K}{}}\right) \stackrel{(5), (10)}{} \frac{1}{3} jS_k j; 8k \; 2 \; [K]: \end{aligned} \tag{14}$$

2.3.2. Upper bound of $j\widehat{S}_k \; n \; S_k j$

Define

$$O = [_{k}^{K} \quad S_k \quad S \text{ and } \overline{O} = [_{k}^{K} \quad \left(\widehat{S}_k \; n \; S_k\right) = S \; n \; O \quad S:$$

From Lemma 2, we know that with a probability at least $1 \quad$, each $\mathbf{x}_i \; 2 \; S_k$ belongs to the set $S_k \quad O$. Thus, with probability at least $1 \quad$, each $\mathbf{x}_i \; 2 \; S$ belongs to O. In other words, with probability *at most* , each $\mathbf{x}_i \; 2 \; S$ belongs to $\overline{O}$. Based on the Chernoff bound, we have, with a probability at least $1 \quad$,

$$j\overline{O}j \quad 2\mathrm{E}\left[j\overline{O}j\right] + 2 \ln \frac{1}{} \quad 2 \;\; jSj + 2 \ln \frac{1}{}: \tag{15}$$

Since $S_k \quad S_k$, we have $\widehat{S}_k \; n \; S_k \quad \widehat{S}_k \; n \; S_k \quad \overline{O}$. Therefore, with a probability at least $1 \quad$, we have

$$j\widehat{S}_k \; n \; S_k j \quad 2 \;\; jSj + 2 \ln \frac{1}{}; 8k \; 2 \; [K]: \tag{16}$$

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

$$\|A_k\|_\infty \quad 2 \frac{2 \; |S| + 2\ln \frac{}{\epsilon}}{- \;_k|S|} = \frac{18}{_k}\left(\; + \frac{1}{|S|}\ln\frac{1}{}\right) = O(\quad) + O\left(\frac{}{|S|}\right); \forall k \in [K]. \tag{17}$$

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

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

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

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

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

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

Following the property of Gaussian random vector, $\sum_{\mathbf{x}_i\in\widehat{S}_k} U^\top\mathbf{g}_i/\left(\;\sqrt{|\widehat{S}_k|}\right)$ can be treated as a $(d \quad K)$-dimensional Gaussian random vector. As a result, each element of $U \;\sum_{\mathbf{x}_i\in\widehat{S}_k} U^\top\mathbf{g}_i/\left(\;\sqrt{|\widehat{S}_k|}\right)$ 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 , we have

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

which implies

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

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

First, we have

$$\left\|\frac{1}{|\widehat{S}_k\setminus S_k|}\sum_{\mathbf{x}_i\in\widehat{S}_k\setminus S_k^1}\mathbf{g}_i^{\|}\right\|_\infty = \left\|\frac{1}{|\widehat{S}_k\setminus S_k|}\sum_{\mathbf{x}_i\in\widehat{S}_k\setminus S_k^1}U\,U^\top\mathbf{g}_i\right\|_\infty \quad \left\|\frac{1}{|\widehat{S}_k\setminus S_k|}\sum_{\mathbf{x}_i\in\widehat{S}_k\setminus S_k^1}U^\top\mathbf{g}_i\right\| \tag{20}$$

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

$$\|U^\top\mathbf{g}_i\| \quad \left(\sqrt{K} + \sqrt{2\log\frac{1}{}}\right)$$

Applying the union bound again, with a probability at least 1 , we have

$$\max_{\leq i \leq |\mathcal{S}|} \left\| U^\top \mathbf{g}_i \right\| \quad \left(\sqrt{K} + \sqrt{2 \log \frac{jSj}{}} \right) \tag{21}$$

Combining (20) and (21), we have

$$\widehat{B}_k \quad \frac{9}{\ _k} \left(\ + \frac{1}{jSj} \ln \frac{1}{} \right) \left(\sqrt{K} + \sqrt{2 \log \frac{jSj}{}} \right) = O(\quad \sqrt{\ln jSj}) + O \left(\frac{\sqrt{\ln jSj}}{jSj} \right); 8k\ 2\ [K]: \tag{22}$$

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

First, we have

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

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

Lemma 3. *(Lemma 2 from (Smale & Zhou, 2007)) Let H be a Hilbert space and be a random variable on $(Z; \ )$ with values in H. Assume $k\ k \quad M < 1$ almost surely. Denote $(\ ) = \mathrm{E}(k\ k\ )$. Let $fz_i g_i^m$ be independent random drawers of . For any $0 < \ < 1$, with confidence 1 ,*

$$\left\| \frac{1}{m} \sum_i^m (\ _i \quad \mathrm{E}[\ _i]) \right\| \quad \frac{2M \ln(2= \ )}{m} + \sqrt{\frac{2 \quad (\ ) \ln(2= \ )}{m}}$$

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

$$u_k \quad \left(\sqrt{K} + \sqrt{2 \log \frac{jSj}{}} \right) \left(\frac{2 \ln(2K= \ )}{jS_kj} + \sqrt{\frac{2 \ln(2K= \ )}{jS_kj}} \right); \ 8k\ 2\ [K]: \tag{24}$$

Combining (23) and (24), we have

$$\widehat{B}_k \quad \left(\sqrt{K} + \sqrt{2 \log \frac{jSj}{}} \right) \left(\frac{2}{jS_kj} \ln \frac{2K}{} + \sqrt{\frac{2}{jS_kj} \ln \frac{2K}{}} \right)$$

$$\stackrel{(10), (14), (5)}{} \quad \left(\sqrt{K} + \sqrt{2 \log \frac{jSj}{}} \right) 2 \sqrt{\frac{9}{\ _k jSj} \ln \frac{2K}{}} = O \left(\sqrt{\frac{\ln jSj}{jSj}} \right); 8k\ 2\ [K]: \tag{25}$$

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

$$kB_k k_\infty \quad O(\quad \sqrt{\ln jSj}) + O \left(\frac{\sqrt{\ln jSj} + \sqrt{\ln d}}{\sqrt{jSj}} \right); 8k\ 2\ [K]: \tag{26}$$

A. Chernoff Bound

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

$$\Pr[S \quad (1 \quad) \ ] \quad \exp\left(\quad \frac{}{2} \right); \text{ for } 0 < \ < 1;$$

$$\Pr[S \quad (1 + \ ) \ ] \quad \exp\left(\quad \frac{}{2 + } \right); \text{ for } \ > 0:$$

Therefore,

$$\Pr\left[S \quad \left(1 \quad \sqrt{\frac{2}{\ }\ln\frac{1}{\ }}\right) \quad \right] \quad ; \text{ for } \exp\left(\ \frac{2}{\ }\right) < \quad < 1;$$

$$\Pr\left[S \quad 2 \quad + 2\ln\frac{1}{\ } \quad \left(1 + \frac{\ln\frac{\ }{\delta} + \sqrt{2 \quad \ln\frac{\ }{\delta}}}{\ }\right) \quad \right] \quad ; \text{ for } 0 < \quad < 1:$$

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*

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

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

$$Q(x) \quad \frac{1}{2}\exp\left(\ \frac{x}{2}\right):$$

Let $X \quad N(0; 1)$ be a Gaussian random variable. According to Theorem 3, we have

$$\Pr[jXj \quad] \quad \exp\left(\ \frac{\ }{2}\right); \text{or}$$

$$\Pr\left[jXj \quad \sqrt{2\ln\frac{1}{\ }}\right] \quad :$$

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.