Appendix A presents several basic probability tools. Appendix B states some applications of concentration inequalities. Appendix C states some anti-contraction result and its generalization. Appendix D discusses about sensitivity. Appendix E finally proves our main result. Appendix F show several more experimental results.

A Probability tools

In this section we present a number of classical probability tools used in the proof. Lemma A.1 (Chernoff), A.2 (Hoeffding) and A.3 (Bernstein) are about tail bounds for random scalar variables. Lemma A.5 and Lemma A.4 state two standard results for random Gaussian variable. Lemma A.6 is a probability for Chi-square distribution. Finally, Lemma A.7 is a concentration result on random matrices.

We state the classical Chernoff bound which is named after Herman Chernoff but due to Herman Rubin. It gives exponentially decreasing bounds on tail distributions of sums of independent random variables.

Lemma A.1 (Chernoff bound Chernoff (1952))**.** *Let $X = \sum_{i=1}^n X_i$, where $X_i = 1$ with probability p_i and $X_i = 0$ with probability $1 - p_i$, and all X_i are independent. Let $\mu = \mathbb{E}[X] = \sum_{i=1}^n p_i$. Then*
1. $\Pr[X \geq (1+\delta)\mu] \leq \exp(-\delta^2\mu/3)$, $\forall \delta > 0$;
2. $\Pr[X \leq (1-\delta)\mu] \leq \exp(-\delta^2\mu/2)$, $\forall 0 < \delta < 1$.

We state the Hoeffding bound:

Lemma A.2 (Hoeffding bound Hoeffding (1963))**.** *Let $X_1, \cdots, X_n$ denote n independent bounded variables in $[a_i, b_i]$. Let $X = \sum_{i=1}^n X_i$, then we have*

$$\Pr[|X - \mathbb{E}[X]| \geq t] \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

We state the Bernstein inequality:

Lemma A.3 (Bernstein inequality Bernstein (1924))**.** *Let $X_1, \cdots, X_n$ be independent zero-mean random variables. Suppose that $|X_i| \leq M$ almost surely, for all i. Then, for all positive t,*

$$\Pr\left[\sum_{i=1}^n X_i > t\right] \leq \exp\left(-\frac{t^2/2}{\sum_{j=1}^n \mathbb{E}[X_j^2] + Mt/3}\right).$$

We state two bounds for Gaussian random variable:

Lemma A.4 (folklore)**.** *Let $X \sim \mathcal{N}(0, \sigma^2)$, then for all $t \geq 0$, we have*

$$\Pr[X \geq t] \leq \exp(-t^2/2\sigma^2).$$

Lemma A.5 (folklore)**.** *Let $X \sim \mathcal{N}(0, \sigma^2)$, that is, the probability density function of X is given by $\phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$. Then*

$$\Pr[|X| \leq t] \leq \frac{4}{5}\frac{t}{\sigma}.$$

We state a tool for Chi-square distribution:

Lemma A.6 (Lemma 1 on page 1325 of Laurent and Massart Laurent & Massart (2000))**.** *Let $X \sim \mathcal{X}_k^2$ be a chi-squared distributed random variable with k degrees of freedom. Each one has zero mean and σ^2 variance. Then*

$$\Pr[X - k\sigma^2 \geq (2\sqrt{kt} + 2t)\sigma^2] \leq \exp(-t),$$
$$\Pr[k\sigma^2 - X \geq 2\sqrt{kt}\sigma^2] \leq \exp(-t).$$

Matrix concentration inequalities have a large number of applications, for more details, we refer the readers to a survey by Tropp Tropp (2015). Recently, there are several non-trivial generalizations, e.g., Expander walk Garg et al. (2018); Naor et al. (2019), Strongly Rayleigh distributions Kyng & Song (2018), and matrix Poincare inequality Aoun et al. (2019). Here, we state matrix Bernstein inequality, which can be thought of as a matrix generalization of Lemma A.3.

Lemma A.7 (Matrix Bernstein, Theorem 6.1.1 in Tropp (2015))**.** *Consider a set of m i.i.d. matrices $\{X_1, \cdots, X_m\} \subset \mathbb{R}^{n_1 \times n_2}$. Assume that*

$$\mathbb{E}[X_i] = 0, \forall i \in [m] \quad \text{and} \quad \|X_i\| \leq M, \forall i \in [m].$$

Let $X = \sum_{i=1}^m X_i$. Let $\mathrm{Var}[X]$ be the matrix variance statistic of sum:

$$\mathrm{Var}[X] = \max \left\{ \Big\| \sum_{i=1}^m \mathbb{E}[X_i X_i^\top] \Big\|, \Big\| \sum_{i=1}^m \mathbb{E}[X_i^\top X_i] \Big\| \right\}.$$

Then

$$\mathbb{E}[\|X\|] \leq (2\mathrm{Var}[X] \cdot \log(n_1 + n_2))^{1/2} + M \cdot \log(n_1 + n_2)/3.$$

Furthermore, for all $t \geq 0$,

$$\Pr[\|X\| \geq t] \leq (n_1 + n_2) \cdot \exp\left(- \frac{t^2/2}{\mathrm{Var}[X] + Mt/3} \right).$$

B Application of concentration inequality

B.1 Application of concentration inequality, truncated Gaussian

Lemma B.1 (Inner product between two vectors)**.** *Let $a > 0$. Let $u_1, \cdots, u_d$ denote i.i.d. random variables satisfying $\forall i \in [d]$ $u_i = y_i \cdot z_i$ where $y_i \sim \mathcal{N}(0, \sigma^2)$ and*

$$z_i = \begin{cases} 1, & |y_i| \leq a; \\ 0, & |y_i| > a. \end{cases}$$

Then, for any fixed vector $x \in \mathbb{R}^d$, for any failure probability $\delta \in (0, 1/10)$, we have

$$\Pr_u[|\langle u, x \rangle| \geq 10\|x\|_2 a(\sqrt{a/\sigma} + 1)\log(1/\delta)] \leq \delta.$$

Proof. First, we can compute can $\mathbb{E}[u_i]$

$$\mathbb{E}[u_i] = \mathbb{E}[u_i] = 0.$$

Second, we can upper bound $\mathbb{E}[(u_i)^2]$ using Lemma A.5

$$\begin{aligned} \mathbb{E}[(u_i)^2] = & \ \mathbb{E}[u_i^2] \\ \leq & \ a^2 \cdot \Pr[|u_i| \leq a] \\ \leq & \ a^2 \cdot \frac{4}{5} \frac{a}{\sigma} \\ \leq & \ a^3/\sigma. \end{aligned}$$

Third, we can upper bound $|u_i x_i|$ by $a \cdot \|x\|_\infty$.

Using Bernstein inequality, we have

$$\begin{aligned} \Pr[|\langle u, x \rangle| \geq t] \leq & \ \exp\Big(-\frac{t^2/2}{\|x\|_2^2 \, \mathbb{E}[u_i^2] + a\|x\|_\infty t/3}\Big) \\ \leq & \ \exp\Big(-\frac{t^2/2}{\|x\|_2^2 a^3/\sigma + a\|x\|_\infty t/3}\Big). \end{aligned}$$

Choosing

$$t = 5\|x\|_2 a^{1.5} \sigma^{-0.5} \sqrt{\log(1/\delta)} + 5\|x\|_\infty a \log(1/\delta)$$

gives us

$$\Pr[|\langle u, x \rangle| \geq 10\|x\|_2 a(\sqrt{a/\sigma} + 1)\log(1/\delta)] \leq \delta.$$

□

Lemma B.2 (Matrix vector multiplication)**.** *Let $a > 0$. Let $A_{i,j}$ denote i.i.d. random variables satisfying $\forall i \in [m], j \in [d]$. $A_{i,j} = y_{i,j} \cdot z_{i,j}$ where $y_{i,j} \sim \mathcal{N}(0, \sigma^2)$ and*

$$z_{i,j} = \begin{cases} 1, & |y_{i,j}| \leq a; \\ 0, & |y_{i,j}| > a. \end{cases}$$

Then, for any fixed vector $x \in \mathbb{R}^d$, for any failure probability $\delta \in (0, 1/10)$, we have

$$\Pr_A\Big[|\|Ax\|_2^2 - \mathbb{E}[\|Ax\|_2^2]| \geq 1000m\|x\|_2^2(\sigma^2 + a^2)\log^3(m/\delta)\Big] \leq \delta.$$

Further, if $m = \Omega(\epsilon^{-2}\|x\|_2^2(1 + a^2/\sigma^2)\log^3(m/\delta))$,

$$\Pr_A\Big[\frac{1}{m}|\|Ax\|_2^2 - \mathbb{E}[\|Ax\|_2^2]| \geq \epsilon^2\|x\|_2^2\sigma^2\Big] \leq \delta.$$

Proof. We define random variable $b_i = (Ax)_i^2$. We can upper bound $\mathbb{E}[b_i]$

$$\mathbb{E}[b_i] = \mathbb{E}[(Ax)_i^2] \leq \|x\|_2^2 \cdot a^3/\sigma.$$

Similarly,

$$\mathbb{E}[b_i] = \mathbb{E}[(Ax)_i^2] \geq 0.1\|x\|_2^2 \cdot a^3/\sigma.$$

Next, we want to upper bound $\mathbb{E}[b_i^2]$, for simplicity, let u denote the i-th row of matrix A,

$$\begin{aligned} \mathbb{E}[b_i^2] - (\mathbb{E}[b_i])^2 =& \ \mathbb{E}[\langle u, x\rangle^4] - (\mathbb{E}[\langle u, x\rangle^2])^2 \\ =& \ \mathbb{E}\Big[(\sum_{i=1}^d u_i x_i)^4\Big] - \Big(\mathbb{E}\Big[(\sum_{i=1}^d u_i x_i)^2\Big]\Big)^2. \end{aligned}$$

For the first term, we have

$$\begin{aligned} \mathbb{E}\Big[(\sum_{i=1}^d u_i x_i)^4\Big] =& \ \mathbb{E}\Big[\sum_{i=1}^d u_i^4 x_i^4\Big] + 3\,\mathbb{E}\Big[\sum_{i=1}^d \sum_{j\in[d]\backslash\{i\}} u_i^2 x_i^2 u_j^2 x_j^2\Big] \\ \leq& \ \mathbb{E}[u_i^4] \cdot \|x\|_4^4 + 3(\mathbb{E}[u_i^2])^2 \cdot \|x\|_2^4 \\ \leq& \ 4\,\mathbb{E}[u_i^4] \cdot \|x\|_2^4. \end{aligned}$$

For the second term, we have

$$\Big(\mathbb{E}\Big[(\sum_{i=1}^d u_i x_i)^2\Big]\Big)^2 = \Big(\sum_{i=1}^d \mathbb{E}[u_i^2] x_i^2\Big)^2 = (\mathbb{E}[u_i^2])^2 \cdot \|x\|_2^4.$$

Thus, we have

$$\mathbb{E}[b_i^2] - (\mathbb{E}[b_i])^2 \leq 4\,\mathbb{E}[u_i^4] \cdot \|x\|_2^4 \leq 64\sigma^4 \|x\|_2^4.$$

We also need to upper bound $|b_i|$. Apply Lemma B.1, we have, for a fixed $i \in [m]$,

$$|b_i| \leq (10\|x\|_2 a(\sqrt{a/\sigma} + 1)\log(m/\delta))^2 := b_{\max}$$

holds with probability at least $1 - \delta/m$.

Taking a union bound over m coordinates, with probability $1 - \delta$, we have : for all $i \in [m]$, $|b_i| \leq b_{\max}$.

Applying Bernstein inequality (Lemma A.3) on $\sum_{i=1}^m b_i$ again

$$\begin{aligned} \Pr\Big[\Big|\sum_{i=1}^m (b_i - \mathbb{E}[b_i])\Big| \geq t\Big] \leq& \ \exp\Big(-\frac{t^2/2}{\sum_{i=1}^m \mathrm{Var}[b_i] + b_{\max} t/3}\Big) \\ \leq& \ \exp\Big(-\frac{t^2/2}{64m\sigma^4\|x\|_2^4 + b_{\max}t/3}\Big). \end{aligned}$$

Choosing $t = 50m\sigma^2\|x\|_2^2\log(1/\delta) + 50mb_{\max}\log(1/\delta)$, we complete the proof. □

B.2 Application of concentration inequalities, classical random Gaussian

Lemma B.3 (Inner product between a random guassian vector and a fixed vector)**.** *Let $a > 0$. Let $u_1, \cdots, u_d$ denote i.i.d. random guassian variables where $u_i \sim \mathcal{N}(0, \sigma_1^2)$.*

Then, for any fixed vector $e \in \mathbb{R}^d$, for any failure probability $\delta \in (0, 1/10)$, we have

$$\Pr_u\Big[|\langle u, e\rangle| \geq 2\sigma_1\|e\|_2\sqrt{\log(d/\delta)} + \sigma_1\|e\|_\infty \log^{1.5}(d/\delta)\Big] \leq \delta.$$

Proof. First, we can compute $\mathbb{E}[u_i]$

$$\mathbb{E}[u_i] = \mathbb{E}[u_i] = 0.$$

Second, we can compute $\mathbb{E}[(u_i)^2]$

$$\mathbb{E}[(u_i)^2] = \mathbb{E}[u_i^2] = \sigma_1^2.$$

Third, we can upper bound $|u_i|$ and $|u_i e_i|$.

$$\Pr_u[|u_i - \mathbb{E}[u_i]| \geq t_1] \leq \exp\Big(-\frac{t_1^2}{2\sigma_1^2}\Big).$$

Take $t_1 = \sqrt{2\log(d/\delta)}\sigma_1$, then for each fixed $i \in [d]$, we have, $|u_i| \leq \sqrt{2\log(d/\delta)}\sigma_1$ holds with probability $1 - \delta/d$.

Taking a union bound over d coordinates, with probability $1-\delta$, we have : for all $i \in [d]$, $|u_i| \leq \sqrt{2\log(d/\delta)}\sigma_1$.

Let E_1 denote the event that, $\max_{i\in[d]} |u_i e_i|$ is upper bounded by $\sqrt{2\log(d/\delta)}\sigma_1\|e\|_\infty$. $\Pr[E_1] \geq 1-\delta$.

Using Bernstein inequality, we have

$$\begin{aligned}\Pr_u[|\langle u, e\rangle| \geq t] \leq & \exp\Big(-\frac{t^2/2}{\|e\|_2^2\,\mathbb{E}[u_i^2] + \max_{i\in[d]}|u_i e_i| \cdot t/3}\Big)\\ \leq & \exp\Big(-\frac{t^2/2}{\|e\|_2^2\sigma_1^2 + \sqrt{2\log(d/\delta)}\sigma_1\|e\|_\infty \cdot t/3}\Big)\\ \leq & \delta,\end{aligned}$$

where the second step follows from $\Pr[E_1] \geq 1-\delta$ and $\mathbb{E}[u_i^2] = \sigma_1^2$, and the last step follows from choice of t:

$$t = 2\sigma_1\|e\|_2\sqrt{\log(d/\delta)} + \sigma_1\|e\|_\infty \log^{1.5}(d/\delta).$$

Taking a union with event E_1, we have

$$\Pr[|\langle u, e\rangle| \geq t] \leq 2\delta.$$

Rescaling δ completes the proof.

□

Lemma B.4 (Inner product between two random guassian vectors)**.** *Let $a > 0$. Let $u_1, \cdots, u_d$ denote i.i.d. random Gaussian variables where $u_i \sim \mathcal{N}(0, \sigma_1^2)$ and $e_1, \cdots, e_d$ denote i.i.d. random Gaussian variables where $e_i \sim \mathcal{N}(0, \sigma_2^2)$*

Then, for any failure probability $\delta \in (0, 1/10)$, we have

$$\Pr_{u,e}\Big[|\langle u, e\rangle| \geq 10^4\sigma_1\sigma_2\sqrt{d}\log^2(d/\delta)\Big] \leq \delta.$$

Proof. First, using Lemma A.6, we compute the upper bound for $\|e\|_2^2$

$$\Pr_e[\|e\|_2^2 - d\sigma_2^2 \geq (2\sqrt{dt} + 2t)\sigma_2^2] \leq \exp(-t).$$

Take $t = \log(1/\delta)$, then with probability $1-\delta$,

$$\|e\|_2^2 \leq (d + 3\sqrt{d\log(1/\delta)} + 2\log(1/\delta))\sigma_2^2 \leq 4d\log(1/\delta)\sigma_2^2.$$

Thus

$$\Pr_e[\|e\|_2 \leq 4\sqrt{d\log(1/\delta)}\sigma_2] \geq 1-\delta.$$

Second, we compute the upper bound for $\|e\|_\infty$ (the proof is similar to Lemma B.3)

$$\Pr_e[|\|e\|_\infty \leq \sqrt{\log(d/\delta)}\sigma_2] \geq 1-\delta.$$

We define t and t' as follows

$$t = 4 \cdot (\sigma_1 \|e\|_2 \sqrt{\log(d/\delta)} + \sigma_1 \|e\|_\infty \log^{1.5}(d/\delta))$$
$$t' = 8 \cdot (\sigma_1 \sigma_2 \sqrt{d} \log(d/\delta) + \sigma_1 \sigma_2 \log^2(d/\delta)).$$

From the above calculations, we can show

$$\Pr_e[t' \geq t] \geq 1 - 2\delta.$$

By Lemma B.3, for fixed e, we have

$$\Pr_u[|\langle u, e \rangle| \geq t] \leq \delta.$$

Overall, we have

$$\Pr_{e,u}[|\langle u, e \rangle| \geq t'] \leq 3\delta.$$

Rescaling δ completes the proof. □

Lemma B.5 (Concentration of folded Gaussian)**.** *Let matrix $A \in \mathbb{R}^{m \times d}$ be defined as each entry is i.i.d. random variables satisfying $\forall i \in [m]$, $j \in [d]$. $A_{i,j} = y_{i,j}$ where $y_{i,j} \sim \mathcal{N}(0, \sigma_A^2)$. Let $\overline{A} \in \mathbb{R}^{m \times d}$ be defined as, $\forall i \in [m], j \in [d]$, $\overline{A}_{i,j} = y_{i,j} \cdot z_{i,j}$ where*

$$z_{i,j} = \begin{cases} 1, & \text{if } 0 \leq y_{i,j} \leq a; \\ 0, & \text{otherwise }. \end{cases}$$

Let $x \in \mathbb{R}_+^d$ denote a non-negative vector where $\sum_{i=1}^d x_i = 1$.

1) For any failure possibility $\delta \in (0, 1/10)$, we have

$$\Pr\left[\forall i \in [m], (\bar{A}x)_i \geq \sigma_A \cdot C\right] > 1 - \delta,$$

where

$$C := \frac{a^2}{6\sigma_A^2} - (\frac{2a^3}{9\sigma_A^3})^{1/2} \cdot \sqrt{\log(m/\delta)} - \frac{2a}{3\sigma_A} \cdot \log(m/\delta).$$

2) For any failure possibility $\delta \in (0, 1/10)$, if $a/\sigma_A \geq 20 \log(m/\delta)$, then

$$\Pr\left[\forall i \in [m], (\bar{A}x)_i \geq \sigma_A \cdot 0.02 \cdot (a^2/\sigma_A^2)\right] > 1 - \delta.$$

Proof. For a fixed $i \in [m]$, for each $j \in [d]$, we define

$$b_j = \bar{A}_{i,j} x_j.$$

We first calculate $\mathbb{E}[b_j]$, $\mathbb{E}[b_j^2]$ and $\text{Var}[b_j]$.

We provide a lower bound for $\mathbb{E}[b_j]$,

$$\begin{aligned} \mathbb{E}[b_j] =&~ \mathbb{E}[A_{i,j}] x_j \\ =&~ x_j \int_0^a \frac{1}{\sigma_A \sqrt{2\pi}} \exp(-x/\sigma_A^2) x \mathrm{d}x \\ \geq&~ \frac{a^2 x_j}{2\sigma_A \sqrt{2\pi}} \\ \geq&~ \frac{a^2 x_j}{6\sigma_A}. \end{aligned}$$

We give an upper bound for $\mathbb{E}[b_j^2]$,

$$\begin{aligned}\mathbb{E}[b_j^2] =&\ \mathbb{E}[A_{i,j}^2]x_j^2 \\ =&\ x_j^2 \int_0^a \frac{1}{\sigma_A\sqrt{2\pi}} \exp(-x/\sigma_A^2)x^2 \mathrm{d}x \\ \leq&\ \frac{a^3 x_j^2}{3\sigma_A\sqrt{2\pi}} \\ \leq&\ \frac{a^3 x_j^2}{9\sigma_A}.\end{aligned}$$

We can upper bound $\mathrm{Var}[b_j]$,

$$\mathrm{Var}[b_j] = \mathbb{E}[b_j^2] - \mathbb{E}[b_j]^2 \leq \mathbb{E}[b_j^2] \leq \frac{a^3 x_j^2}{9\sigma_A}.$$

Then, we can lower bound $\sum_{j=1}^d \mathbb{E}[b_j]$

$$\sum_{j=1}^d \mathbb{E}[b_j] \geq \frac{a^2}{6\sigma_A} \sum_{j=1}^d x_j = \frac{a^2}{6\sigma_A},$$

where the last step follows from $\sum_{j=1}^d x_j = 1$.

Next, we can upper bound b_j and $\sum_{j=1}^d \mathrm{Var}[b_j]$

$$M := \max_{j\in[d]} b_j \leq \max_{j\in[d]} x_j a \leq a.$$

$$\sum_{j=1}^d \mathrm{Var}[b_j] \leq \sum_{j=1}^d \frac{a^3 x_j^2}{9\sigma_A} \leq \sum_{j=1}^d \frac{a^3 x_j}{9\sigma_A} = \frac{a^3}{9\sigma_A}.$$

Applying Bernstein inequality (Lemma A.3) on $\sum_{j=1}^d (b_j - \mathbb{E}[b_j])$

$$\begin{aligned}\Pr\Big[\sum_{j=1}^d (b_j - \mathbb{E}[b_j]) \leq -t\Big] \leq&\ \exp\Big(-\frac{t^2/2}{\sum_{j=1}^d \mathrm{Var}[b_j] + Mt/3}\Big) \\ \leq&\ \exp\Big(-\frac{t^2/2}{a^3/9\sigma_A + at/3}\Big).\end{aligned}$$

Taking

$$t = \sigma_A \cdot (\sqrt{2a^3/9\sigma_A^3 \cdot \log(m/\delta)} + 2a/3\sigma_A \cdot \log(m/\delta)),$$

then for any $i \in [m]$,

$$\begin{aligned}\Pr\Big[\sum_{j=1}^d b_j \geq a^2/(6\sigma_A) - t\Big] \geq&\ \Pr\Big[\sum_{j=1}^d b_j \geq \sum_{j=1}^d \mathbb{E}[b_j] - t\Big] \\ \geq&\ 1 - \delta,\end{aligned}$$

where the first step holds because $\sum_{j=1}^d \mathbb{E}[b_j] > a^2/(6\sigma_A)$.

Since $(\bar{A}x)_i = \sum_{j=1}^d b_j$, we have for any $i \in [m]$,

$$\Pr\Big[(\bar{A}x)_i \geq \sigma_A \cdot \Big(a^2/6\sigma_A^2 - \sqrt{(2a^3/9\sigma_A^3) \cdot \log(m/\delta)} - (2a/3\sigma_A) \cdot \log(m/\delta)\Big)\Big] > 1 - \delta/m.$$

Taking a union bound over all $i \in [m]$ completes the proof. □

C Anti-concentration

Given a number of independent random variables, the well-known Central Limit Theorem (CLT) states that their sum has good concentration under certain conditions. Such concentration results like the Chernoff bound Chernoff (1952) and Hoeffding's inequality Hoeffding (1963) are among the central tools in Theoretical Computer Science (TCS). From the opposite perspective, we can also ask for *anti-concentration* results. For example, let x be a Rademacher variable (choosing ± 1 with probability $1/2$) and let a denote a vector in $\mathbb{R}^d$. The celebrated Littlewood-Offord Lemma states that any d-variate degree-1 polynomial $p(x) = \sum_{i=1}^d a_i x_i$ does not concentrate on any particular value.

Theorem C.1 (Littlewood and Offord Littlewood & Offord (1943))**.** *Let $C > 0$ denote a universal constant. For any linear form p satisfying $|a_i| \geq 1$, $\forall i \in [d]$, and any open interval I of length 1, we have*

$$\Pr_{x \sim \{-1,+1\}^d}[p(x) \in I] \leq C \cdot \frac{\log d}{\sqrt{d}}.$$

Two years later, Erdös Erdös (1945) removed the $\log d$ factor in Theorem C.1. Recently, Theorem C.1 has been generalized to higher degree polynomials by Costello et al. (2006); Razborov & Viola (2013); Meka et al. (2017).

Instead of considering x_i as $\{-1, +1\}$ random variables, Carbery and Wright Carbery & Wright (2001) showed the anti-concentration result for x_i chosen as i.i.d. Gaussians.

Theorem C.2 (Carbery and Wright Carbery & Wright (2001))**.** *Let $p : \mathbb{R}^d \to \mathbb{R}$ denote a degree-k polynomial with d variables. There is a universal constant $C > 0$ such that*

$$\Pr_{x \sim \mathcal{N}(0, I_d)} \left[|p(x)| \leq \delta \sqrt{\mathrm{Var}[p(x)]} \right] \leq C \cdot \delta^{1/k}.$$

These are worst-case results in the sense that they hold for arbitrary polynomials. For example, Theorem C.2 is tight for any polynomial that is a perfect k-th power.

We can generalize Theorem C.2 into the following[3]:

Lemma C.3 (An variation of Carbery & Wright (2001), Anti-concentration of sum of truncated Gaussians)**.** *Let $x_1, \cdots, x_n$ be n i.i.d. zero-mean Gaussian random variables $\mathcal{N}(0, 1)$. Let $p : \mathbb{R}^n \to \mathbb{R}$ denote a degree-1 polynomial defined as*

$$p(x_1, \cdots, x_n) = \sum_{i=1}^n \alpha_i x_i.$$

Let f denote a truncation function where $f(x) = x$ if $|x| \leq a$, and $f(x) = 0$ if $|x| > a$. Then we have

$$\Pr_{x \sim \mathcal{N}(0, I_d)} \left[|p(f(x))| \leq \min\{a, 0.1\} \cdot \delta \cdot \|\alpha\|_2 \right] \geq C \cdot \delta.$$

Proof. Let $\mu : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ be the truncated Gaussian distribution. We first argue that μ is log-concave. Indeed, for any $x, y \in \mathbb{R}^n$ and $\lambda \in [0, 1]$, if $\mu(x) = 0$ or $\mu(y) = 0$, then we must have

$$\mu(\lambda x + (1 - \lambda) y) \geq 0 = (\mu(x))^\lambda \cdot (\mu(y))^{1-\lambda}.$$

On the other hand, if $\mu(x) > 0$ and $\mu(y) > 0$, then we must have $\mu(\lambda x + (1 - \lambda) y) > 0$, because

$$\|\lambda x + (1 - \lambda) y\|_2 \leq \lambda \|x\|_2 + (1 - \lambda) \|y\|_2,$$

[3] The generalization also has been observed in Song et al. (2020), for the completeness, we provide the proof here.

hence μ would not truncate at $\lambda x+(1-\lambda)y$. Notice that Gaussian distribution is log-concave. Let $\mu' : \mathbb{R}^n \to \mathbb{R}$ be the density function of Gaussian distribution, then $\mu(x) = C_0 \cdot \mu'(x)$ for some universal constant $C_0 > 0$ for all x that is not truncated. so in this case we still have

$$\begin{aligned}\mu(\lambda x + (1-\lambda)y) &= C_0 \cdot \mu'(\lambda x + (1-\lambda)y)\\ &\geq C_0 \cdot (\mu'(x))^{\lambda} \cdot (\mu'(y))^{1-\lambda}\\ &= (C_0\mu'(x))^{\lambda} \cdot (C_0\mu'(y))^{1-\lambda}\\ &= (\mu(x))^{\lambda} \cdot (\mu(y))^{1-\lambda}.\end{aligned}$$

So we conclude that μ is log-concave.

Now we apply Theorem C.5 on μ and p. By setting $q = 2$ and $d = 1$, we have

$$\Big(\int_{x\in\mathbb{R}^n} |p(x)|^2 \mathrm{d}\mu\Big)^{1/2} \cdot \mu(|p(x)| \leq \alpha) \leq C \cdot \alpha. \tag{1}$$

Notice that

$$\begin{aligned}\int_{x\in\mathbb{R}^n} |p(x)|^2 \mathrm{d}\mu = &\mathop{\mathbb{E}}_{x\sim\mu}\Big[\Big(\sum_{i=1}^n \alpha_i x_i\Big)^2\Big]\\ = &\sum_{i=1}^n \alpha_i^2 \mathop{\mathbb{E}}_{x\sim\mu}[x_i^2]\\ = &\sum_{i=1}^n \alpha_i^2 \mathop{\mathrm{Var}}_{x_i\sim\mu_i}[x_i],\end{aligned}$$

where $\mu_i : \mathbb{R} \to \mathbb{R}$ is the distribution on the i-th coordinate, $\forall i \in [n]$. Hence we can rewrite Eq. (1) as

$$\Pr_{x\sim\mathcal{N}(0,I_d)}\left[\Big|\sum_{i=1}^n \alpha_i f(x_i)\Big| \leq \delta\left(\sum_{i=1}^n \alpha_i^2 \mathop{\mathrm{Var}}_{x_i\sim\mu_i}[x_i]\right)^{1/2}\right] \geq C \cdot \delta.$$

By Claim C.4, we have

$$\Pr_{x\sim\mathcal{N}(0,I_d)}\left[|p(f(x))| \leq \delta\left(\sum_{i=1}^n \alpha_i^2 \cdot \left(1 - \sqrt{\frac{2}{\pi}} \cdot \frac{a\cdot e^{-a^2/2}}{\mathrm{erf}(a/\sqrt{2})}\right)\right)^{1/2}\right] \geq C \cdot \delta.$$

For $0 \leq a \ll 1$,

$$1 - \sqrt{\frac{2}{\pi}} \cdot \frac{a\cdot e^{-a^2/2}}{\mathrm{erf}(a/\sqrt{2})} = \frac{5}{6}a^2 + o(a^3).$$

Hence,

$$\Pr_{x\sim\mathcal{N}(0,I_d)}\Big[|p(f(x))| \leq \delta a \|\alpha\|_2\Big] \geq C \cdot \delta.$$

For $a \geq 1$,

$$1 - \sqrt{\frac{2}{\pi}} \cdot \frac{a\cdot e^{-a^2/2}}{\mathrm{erf}(a/\sqrt{2})} = \Theta\left(1 - ae^{-a^2} - e^{-a^2/2}/a\right).$$

Hence,

$$\Pr_{x\sim\mathcal{N}(0,I_d)}\Big[|p(f(x))| \leq \delta(1 - ae^{-a^2} - e^{-a^2/2}/a)^{1/2}\|\alpha\|_2\Big] \geq C \cdot \delta.$$

When $a \geq 1$, we have $0.025 \leq (1 - ae^{-a^2} - e^{-a^2/2}/a) \leq 1$. So we can combine the above two cases to get

$$\Pr_{x\sim\mathcal{N}(0,I_d)}\Big[|p(f(x))| \leq \min\{a, 0.1\} \cdot \delta\|\alpha\|_2\Big] \geq C \cdot \delta.$$

□

Claim C.4. *Let $x \in \mathbb{R}$ be a standard Gaussian random variable $\mathcal{N}(0,1)$. Let f denote a truncation function where $f(x) = x$ if $|x| \leq a$, and $f(x) = 0$ if $|x| > a$. Then, we have*

$$\mathrm{Var}[f(x)] = \ 1 - \sqrt{\frac{2}{\pi}} \cdot \frac{a \cdot e^{-a^2/2}}{\mathrm{erf}(a/\sqrt{2})},$$

where $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \mathrm{d}t$.

Theorem C.5 (Arutyunyan & Kosov (2018))**.** *Let $\mu : \mathbb{R}^n \to \mathbb{R}$ be a log-concave measure over $\mathbb{R}^n$. Let $L^1(\mu) = \int_{x \in \mathbb{R}^n} |\mu(x)| \mathrm{d}x$. For any $q > 0$ and polynomial $p : \mathbb{R}^n \to \mathbb{R}$, define the ℓ_q norm of p with respect to the measure μ as*

$$\|p\|_q = \left(\int p^q \mathrm{d}\mu \right)^{1/q}.$$

Assume p has degree d. Then there exists constant $C(d) > 0$ that only depends on d so that for all $\alpha > 0$ and all $q > 0$,

$$\left(\int |p(x)|^{q/d} \mathrm{d}\mu \right)^{1/q} \cdot \mu(|p(x)| \leq \alpha) \leq C(d) \cdot \alpha^{1/d}.$$

D Sensitivity

D.1 Concentration of folded Gaussian

Lemma D.1 (concentration of folded gaussian)**.** *Let matrix $A \in \mathbb{R}^{m \times d}$ be defined as each entry is i.i.d. random variables satisfying $\forall i \in [m]$, $j \in [d]$. $A_{i,j} = y_{i,j}$ where $y_{i,j} \sim \mathcal{N}(0, \sigma_A^2)$. Let $z_{i,j} = |y_{i,j}|$, then $\forall j \in [d]$,*

$$\Pr\Big[\sum_{i=1}^{m} z_{i,j} - \mathbb{E}[z_{i,j}] > \sigma_A m + 4\sigma_A \sqrt{m} \log^{1.5}(md/\delta)\Big] \leq \delta.$$

Proof. For a fixed j, let $b_i = z_{i,j}$. First we calculate $\mathbb{E}[b_i]$

$$\begin{aligned}\mathbb{E}[b_i] &= \int_0^\infty \frac{2}{\sqrt{2\pi\sigma_A^2}} \exp(-x^2/2\sigma_A^2) x \mathrm{d}x \\ &= \sigma_A \sqrt{2/\pi}.\end{aligned}$$

Second, we calculate $\mathbb{E}[b_i^2]$

$$\mathbb{E}[b_i^2] = \mathbb{E}[z_{i,j}^2] = \mathbb{E}[y_{i,j}^2] = \sigma_A^2.$$

According to Lemma A.4, we can upper bound $z_{i,j}$

$$\Pr[z_{i,j} > t] = \Pr[|y_{i,j}| > t] \leq \exp(-t/\delta_2^2).$$

Taking $t = \sigma_A \sqrt{\log(md/\delta_1)} := M$, we have $\forall i \in [m]$, $j \in [d]$

$$\Pr[\max_{i,j} z_{i,j} > t] \leq \delta_2.$$

Applying Bernstein inequality on $\sum_{i=1}^m b_i$

$$\begin{aligned}\Pr\Big[\Big|\sum_{i=1}^{m}(b_i - \mathbb{E}[b_i])\Big| \geq t\Big] &\leq \exp\Big(-\frac{t^2/2}{\sum_{i=1}^m \mathrm{Var}[b_i] + b_{\max} t/3}\Big) \\ &\leq \exp\Big(-\frac{t^2/2}{m\sigma_A^2 + \sigma_A t \sqrt{\log(md/\delta_2)}/3}\Big).\end{aligned}$$

Choosing $t = \sigma_A m + 4\sigma_A \sqrt{m} \log^{1.5}(md/\delta)$, we have

$$\Pr\Big[\sum_{i=1}^{m} z_{i,j} - \mathbb{E}[z_{i,j}] > t\Big] \leq \delta.$$

□

D.2 ℓ_1-sensitivity functions of single layer neural network

Lemma D.2 (ℓ_1-norm sensitivity of single layer neural network)**.** *Let $x \in [0,1]^d$, fully connected matrix $A \in \mathcal{N}(0, \sigma_A)^{m \times d}$, bias matrix $b \in \mathbb{R}^m$, and ϕ is the ReLU activation function. Let $f(x) = \phi(Ax + b)$ denote a single layer network, then for all neighboring inputs $x_1, x_2 \in \mathbb{R}^d$ that differ at most in one entry, we have*

$$\Pr\Big[\mathrm{GS}_1(f) \leq \sigma_A m + 4\sigma_A \sqrt{m} \log^{1.5}(md/\delta)\Big] \geq 1 - \delta.$$

Proof. Let k denote the index that x_1 and x_2 are different.

$$\begin{aligned}
\mathrm{GS}_1(f) = & \sup_{x_1,x_2\in\mathbb{R}^d} \|f(x_1)-f(x_2)\|_1 \\
= & \sup_{x_1,x_2\in\mathbb{R}^d} \|\phi(Ax_1+b)-\phi(Ax_2+b)\|_1 \\
\leq & \sup_{x_1,x_2\in\mathbb{R}^d} \|(Ax_1+b)-(Ax_2+b)\|_1 \\
= & \sup_{x_1,x_2\in\mathbb{R}^d} \|(A(x_1-x_2)\|_1 \\
= & \|A_{*,k}\|_1 \\
\leq & \sigma_A m + 4\sigma_A\sqrt{m}\log^{1.5}(md/\delta),
\end{aligned}$$

where the fourth step follows that x_1 and x_2 differ in the k-th entry, and the fifth step follows Lemma D.1. □

D.3 ℓ_2-SENSITIVITY FUNCTIONS OF SINGLE LAYER NEURAL NETWORK

Lemma D.3 (ℓ_2-norm sensitivity of single layer neural network)**.** *Let $x \in [0,1]^d$, fully connected matrix $A \in \mathcal{N}(0,\sigma_A)^{m\times d}$, bias matrix $b \in \mathbb{R}^m$, and ϕ is the ReLU activation function. Let $f(x) = \phi(Ax+b)$ denote a single layer network, then for all neighboring inputs $x_1, x_2 \in \mathbb{R}^d$ that differ at most in one entry, we have*

$$\Pr\left[\mathrm{GS}_2(f) \leq 2(\sqrt{md}+\sqrt{\log(1/\delta)})\right] \geq 1-\delta.$$

Proof. Let k denote the index that x_1 and x_2 are different.

$$\begin{aligned}
\mathrm{GS}_2(f) = & \sup_{x_1,x_2\in\mathbb{R}^d} \|f(x_1)-f(x_2)\|_2 \\
= & \sup_{x_1,x_2\in\mathbb{R}^d} \|\phi(Ax_1+b)-\phi(Ax_2+b)\|_2 \\
\leq & \sup_{x_1,x_2\in\mathbb{R}^d} \|(Ax_1+b)-(Ax_2+b)\|_2 \\
= & \sup_{x_1,x_2\in\mathbb{R}^d} \|(A(x_1-x_2)\|_2 \\
= & \|A_{*,k}\|_2 \\
\leq & \sigma_A\left(2\sqrt{md\log(1/\delta)}+2\log(1/\delta)+md\right)^{1/2} \\
\leq & \sigma_A\left(2\sqrt{2md\log(1/\delta)}+2\log(1/\delta)+md\right)^{1/2} \\
= & \sigma_A(\sqrt{md}+\sqrt{2\log 1/\delta}),
\end{aligned}$$

where the fourth step follows that x_1 and x_2 differ in the k-th entry, and the fifth step follows Lemma A.6. □

Table 2: Summary of two results

Statement	$\epsilon_{\rm dp}$	Comment	Pruning
Theorem E.1	$\mathrm{GS}_1(f)/(\sigma\sigma_A)\cdot(m/\delta_{\rm dp})$	General x	Magnitude
Theorem E.2	$\mathrm{GS}_1(f)/(\sigma\sigma_A)\cdot\log(m/\delta_{\rm dp})$	Nonnegative x	Folded Magnitude

E Equivalence between pruning and differential privacy

E.1 Main results

Theorem E.1 (Main result I). *For a single layer neural network $f(x)=\phi(Ax+b)$ where fully connected matrix $A\in\mathcal{N}(0,\sigma_A^2)^{m\times d}$, vector $b\in\mathbb{R}^m$, and ϕ is the ReLU activation function. We assume all the inputs $x\in\mathbb{R}^d$ satisfying that $\|x\|_2=1$. If*

$$m=\Omega(\mathrm{poly}(\epsilon_{\rm ap}^{-1},\log(1/\delta_{\rm ap}),\log(1/\delta_{\rm dp}),a/\sigma_A,\sigma\sigma_A)),$$

then applying magnitude pruning with with truncation threshold $a>0$ on $A\in\mathbb{R}^{m\times d}$ is an $(\epsilon_{\rm ap},\delta_{\rm ap})$-approximation to applying $(\epsilon_{\rm dp},\delta_{\rm dp})$-differential privacy on x, where
$\epsilon_{\rm dp}=2\mathrm{GS}_1(f)(m/\delta_{\rm dp})/(\sigma\sigma_A)$.

Theorem E.2 (Main result II). *For a single layer network $f(x)=\phi(Ax+b)$ where fully connected matrix $A\in\mathcal{N}(0,\sigma_A^2)^{m\times d}$, vector $b\in\mathbb{R}^m$, and ϕ is the ReLU activation function. We assume all the inputs $x\in\mathbb{R}^d$ satisfying that $\|x\|_2=1$ and $x\in\mathbb{R}_+^d$. If*

$$m=\Omega(\mathrm{poly}(\epsilon_{\rm ap}^{-1},\log(1/\delta_{\rm ap}),\log(1/\delta_{\rm dp}),a/\sigma_A,\sigma\sigma_A)),$$

then applying folded magnitude pruning with truncation threshold $a>0$ on $A\in\mathbb{R}^{m\times d}$ is an $(\epsilon_{\rm ap},\delta_{\rm ap})$-approximation to applying $(\epsilon_{\rm dp},\delta_{\rm dp})$-differential privacy on $x\in\mathbb{R}^d$, where
$\epsilon_{\rm dp}=2\mathrm{GS}_1(f)\log(m/\delta_{\rm dp})/(\sigma\sigma_A)$.

Remark E.3. *Note that $\mathrm{GS}_1(f)=\Theta(m\sigma_A)$.*
1) if using folded Gaussian and assume $x\in\mathbb{R}_{\geq 0}$, $\epsilon_{\rm dp}=2\mathrm{GS}_1(f)\cdot\log(m/\delta_{\rm dp})/(\sigma\sigma_A)$,
then we need to pick $\sigma=m$, $\sigma_A=\Theta(1/\sigma)$ and $a=\Theta(\sigma_A)$.
2) if using Gaussian, $\epsilon_{\rm dp}=2\mathrm{GS}_1(f)\cdot(m/\delta_{\rm dp})/(\sigma\sigma_A)$,
then we need to pick $\sigma=m^2$, $\sigma_A=\Theta(1/\sigma)$ and $a=\Theta(\sigma_A)$.

E.2 Differential privacy

Definition E.4 (Differential Privacy, Definition.1 in Dwork et al. (2006b)). *Let $\mathcal{A}:\mathcal{D}^n\to\mathcal{Y}$ be a randomized algorithm. Let $D_1,D_2\in\mathcal{D}^n$ be two databases that differ in at most one entry (we call these databases neighbors). Let $\epsilon>0$. Define $\mathcal{A}$ to be ϵ -differentially private if for all neighboring databases D_1,D_2, and for all (measurable) subsets $Y\subset\mathcal{Y}$, we have*

$$\frac{\Pr\left[\mathcal{A}(D_1)\in Y\right]}{\Pr\left[\mathcal{A}\left(D_2\right)\in Y\right]}\leq\exp(\epsilon).$$

Definition E.5 (Global Sensitivity, Definition 2 in Dwork et al. (2006b)). *Let $f:\mathcal{D}^n\to\mathbb{R}^d$, define $\mathrm{GS}_p(f)$, the ℓ_p global sensitivity of f, for all neighboring databases D_1,D_2 as*

$$\mathrm{GS}_p(f)=\sup_{D_1,D_2\in\mathcal{D}^n}\|f(D_1)-f(D_2)\|_p.$$

Theorem E.6 (Laplace Mechanism Dwork et al. (2006b)). *Let f be defined as before and $\epsilon>0$. Define randomized algorithm $\mathcal{A}$ as*

$$\mathcal{A}(D)=f(D)+(\mathrm{Lap}(\mathrm{GS}_1(f)/\epsilon))^d,$$

where the one-dimensional (zero mean) Laplace distribution $\mathrm{Lap}(b)$ has density $p(x;b)=\frac{1}{2b}\exp(-\frac{|x|}{b})$, and $\mathrm{Lap}(b)^d=(l_1,\ldots,l_d)\in\mathbb{R}^d$ where each l_i i.i.d. is sampled from $\mathrm{Lap}(b)$. Then $\mathcal{A}$ is ϵ-differentially private.

Theorem E.7 (Gaussian Mechanism Dwork & Roth (2014)). *For $c>2\sqrt{\log(1/\delta)}$, the Gaussian Mechanism with parameter $\sigma\geq c\cdot\mathrm{GS}_2(f)/\epsilon$ is (ϵ,δ)-differentially private.*

E.3 Function approximation

Definition E.8 ((ϵ, δ)-approximation)**.** *For a pair of functions $f(x)$ and $g(x)$, we say f is an (ϵ, δ)-approximation of g if for any x*

$$\Pr[\|f(x) - g(x)\|_2 > \epsilon] \leq \delta.$$

E.4 Proof of Theorem E.2

Proof. **Sketch.**

The proof can be splitted into two parts. We use $\widetilde{A} \in \mathbb{R}^{m \times d}$ to denote the weight matrix after magnitude pruning, and $\bar{A} = \widetilde{A} - A \in \mathbb{R}^{m \times d}$. We define vector $e \in \mathbb{R}^m$ as follows

$$e = \mathrm{Lap}(1, \sigma)^m \circ (\bar{A}x).$$

1. Let $B(x) = f(x) + e \in \mathbb{R}^m$, then $B(x)$ is $(\epsilon_{\mathrm{dp}}, \delta_{\mathrm{dp}})$-differential privacy.
2. $\Pr[\frac{1}{\sqrt{m}}\|e - \bar{A}x\|_2 \geq \epsilon_{\mathrm{ap}}] \leq \delta_{\mathrm{ap}}$, as long as $m = \Omega(\mathrm{poly}(\epsilon_{\mathrm{ap}}^{-1}, \log(1/\delta_{\mathrm{ap}}), a/\sigma_A, \sigma\sigma_A))$.

Part 1.

Let $y \in \mathbb{R}^m$ and x_1, x_2 be neighbouring inputs. It is sufficient to bound the ratio $\frac{p(y - f(x_1))}{p(y - f(x_2))}$ where $p(\cdot)$ denotes probability density, because once the densities are bounded, integrating $p(\cdot)$ yields the requirement for differential privacy as defined in E.2.

Since $e_i \sim \mathrm{Lap}(1, \sigma) \cdot (\bar{A}x)_i$, then $p(t : \sigma) = \frac{1}{2\sigma}\exp(-|t/(\bar{A}x)_i - 1|/\sigma)$

$$\begin{aligned}
\frac{p(y - f(x_1))}{p(y - f(x_2))} &= \frac{\prod_{i=1}^m \frac{1}{2\sigma}\exp(-|(y_i - f(x_1)_i)/(\bar{A}x_1)_i - 1|/\sigma)}{\prod_{i=1}^m \frac{1}{2\sigma}\exp(-|(y_i - f(x_2)_i)/(\bar{A}x_2)_i - 1|/\sigma)} \\
&= \frac{\exp(-\sum_{i=1}^m |(y_i - f(x_1)_i)/(\bar{A}x_1)_i - 1|/\sigma)}{\exp(-\sum_{i=1}^m |(y_i - f(x_2)_i)/(\bar{A}x_2)_i - 1|/\sigma)} \\
&= \exp\Big(\frac{1}{\sigma}\sum_{i=1}^m \Big|\frac{y_i - f(x_1)_i}{(\bar{A}x_1)_i} - 1\Big| - \Big|\frac{y_i - f(x_2)_i}{(\bar{A}x_2)_i} - 1\Big|\Big) \\
&\leq \exp\Big(\frac{1}{\sigma}\sum_{i=1}^m \Big|\frac{y_i - f(x_1)_i}{(\bar{A}x_1)_i} - \frac{y_i - f(x_2)_i}{(\bar{A}x_2)_i}\Big|\Big) \\
&\leq \exp\Big(2\sum_{i=1}^m \frac{1}{\sigma\min_{i\in[m]}\{|(\bar{A}x_1)_i|, |(\bar{A}x_2)_i|\}}|f(x_2)_i - f(x_1)_i|\Big) \\
&\leq \exp\Big(2\mathrm{GS}_1(f)/\sigma\min_{i\in[m]}\{|(\bar{A}x_1)_i|, |(\bar{A}x_2)_i|\}\Big) \\
&\leq \exp\Big(2\mathrm{GS}_1(f)/\sigma_A\sigma(1/6 \cdot (a/\sigma_A)^2 - 1/5 \cdot (a/\sigma_A)^{1.5}\log(m/\delta_{dp}))\Big) \\
&\leq \exp\Big(2(\sigma_A m + 4\sigma_A\sqrt{m}\log^{1.5}(md/\delta))/\sigma_A\sigma(1/6 \cdot (a/\sigma_A)^2 - 1/5 \cdot (a/\sigma_A)^{1.5}\log(m/\delta_{dp}))\Big) \\
&\leq \exp\Big(2(m + 4\sqrt{m}\log^{1.5}(md/\delta))/\sigma(1/6 \cdot (a/\sigma_A)^2 - 1/5 \cdot (a/\sigma_A)^{1.5}\log(m/\delta_{dp}))\Big)
\end{aligned}$$

where the first equality is because the noise is independent for each coordinate, and the first inequality is triangle inequality. The third inequality holds because of the definition of $\mathrm{GS}_1(f)$, and the fourth holds because of Lemma B.5. holds with probability $1 - \delta_{\mathrm{dp}}$

According to Lemma D.2,

$$\Pr\Big[\mathrm{GS}_1(f) \leq \sigma_A m + 4\sigma_A\sqrt{m}\log^{1.5}(md/\delta)\Big] \geq 1 - \delta.$$

Part 2.

Let $z_i = (e_i - \bar{A}x_i)^2$, thus $z_i \sim \mathrm{Lap}^2(0, b_i)$, where $b_i = \sigma(\bar{A}x)_i$ We first calculate $\mathbb{E}[z_i^2]$

$$\begin{aligned}
\mathbb{E}[z_i^2] = & \int_{-\infty}^{\infty} \frac{1}{2b_i} \exp(-|x|/b_i)x^4 \mathrm{d}x \\
= & \int_{0}^{\infty} \frac{1}{b_i} \exp(-x/b_i)x^4 \mathrm{d}x \\
= & \, b_i^4 \cdot (-\exp(-x)x^4 - \int_0^\infty -4\exp(-x)x^3 \mathrm{d}x)\Big|_0^\infty \\
= & \, b_i^4 \cdot (-\exp(-x)x^4 + 4(-\exp(-x)x^3 - 3(\exp(-x)x^2 - 2(-\exp(-x)x - \exp(-x))))))|_0^\infty \\
= & \, 24b_i^4 \\
\leq & \, 24\sigma^4 \cdot (10a(\sqrt{a/\sigma_A} + 1)\log(m/\delta))^4,
\end{aligned}$$

where both the third step and the fourth step follow integration by parts. The fifth step follows by plugging in the limits of integration, and the last step follows by Lemma B.1.

Next, we want to bound $\max(z_i)$, since $z_i = e_i^2 \sim \mathrm{Lap}^2(0, b_i)$

$$\begin{aligned}
\Pr[z_i \geq t^2] = & \Pr[|e_i| \geq t] & t > 0 \\
= & 2 \cdot \frac{1}{2}\exp(-t/b_i) \\
= & \exp(-t/b_i),
\end{aligned}$$

where the second step follows by plugging the cumulative distribution function of Laplace distribution.

Take $t = b_i \log(m/\delta)$, then for each fixed $i \in [m]$, we have $\Pr[z_i \leq \sqrt{b_i \log(m/\delta)}] = \delta/m$. Thus, with probability $1-\delta$, we have for all $i \in [m]$,

$$z_i \leq \max_{i \in [m]} \sqrt{b_i \log(m/\delta)} \leq \sqrt{a^3\sigma/\sigma_A \cdot \log(m/\delta)} := M,$$

where the second inequality follows by $(\bar{A}x)_i$'s upper bound in Lemma B.2.

Using Bernstein inequality, we have

$$\begin{aligned}
\Pr\left[\Big|\sum_{i=1}^m (z_i - \mathbb{E}[z_i])\Big| \geq t\right] \leq & \exp\Big(-\frac{t^2/2}{\sum_{i=1}^m \mathbb{E}[z_i^2] + Mt/3}\Big) \\
\leq & \exp\Big(-\frac{t^2/2}{24m\sigma^4 \cdot (10a(\sqrt{a/\sigma_A}+1)\log(m/\delta))^4 + a^3\sigma/\sigma_A \cdot \log(m/\delta)t/3}\Big).
\end{aligned}$$

Since $Mt/3$ is dominated by $\sum_{i=1}^m \mathbb{E}[z_i^2]$, we choose

$$t = m\epsilon^2,$$

then as long as

$$m \geq \epsilon^{-4}\log(1/\delta)\Big(48 \cdot (10a/\sigma_A(\sqrt{a/\sigma_A}+1)\log(m/\delta))^4\Big)(\sigma^4\sigma_A^4 + \sigma\sigma_A^2),$$

we have

$$\Pr\left[\frac{1}{m}\Big|\sum_{i=1}^m (z_i - \mathbb{E}[z_i])\Big| \geq \epsilon^2\right] \leq \delta.$$

which is

$$\Pr\left[\frac{1}{\sqrt{m}}\|e - \bar{A}x\|_2 \geq \epsilon\right] \leq \delta.$$

Note that we need to pick $\sigma = m$, then we need to pick $\sigma_A = 1/m$. □

F Experiment details

F.1 Pruning algorithm

Algorithm 1 describes the full process of magnitude-based pruning.

Algorithm 1 Stochastic Gradient Descent with Magnitude-based Pruning

```
procedure SGDMagPrune($\{x_i, y_i\}_{i \in [n]}, a, \eta$)
                                                        ▷ Loss function $\mathcal{L} : \mathbb{R}^{d_o} \times \mathbb{R}^{d_o} \to [0, 1]$
    Let $W^{(1)}$ denote a random initialization of neural network's weights, and $f(W, x)$ denotes the neural network.
    Let $\mathcal{D} = \{(x_1, y_1), \cdots, (x_n, y_n)\} \subset \mathbb{R}^d \times \mathbb{R}^{d_o}$
    for $t = 1 \to T_{\text{train}}$ do                                          ▷ Training stage
        Sample $(x, y) \sim \mathcal{D}$ uniformly at random
        $W^{(t+1)} \leftarrow W^{(t)} - \eta \cdot \frac{\partial \mathcal{L}(f(W,x),y)}{\partial W}|_{W=W^{(t)}}$
    end for
    for $t = T_{\text{train}} \to T_{\text{train}} + T_{\text{prune}}$ do                              ▷ Pruning stage
        Sample a data $(x, y)$ from $\mathcal{D}$ uniformly at random
        $\widetilde{W}^{(t)} \leftarrow$ ThresholdPrune($W^{(t)}, a^{(t)}$)
        $W^{(t+1)} = \widetilde{W}^{(t)} - \eta \cdot \frac{\partial \mathcal{L}(f(W,x),y)}{\partial W}|_{W=\widetilde{W}^{(t)}}$
    end for
    $T_{\text{end}} \leftarrow T_{\text{train}} + T_{\text{prune}}$
    $\widetilde{W}^{(T_{\text{end}})} \leftarrow$ ThPrune($W^{(T_{\text{end}})}, a^{(T_{\text{end}})}$)
end procedure
procedure ThPrune($W, a$)
    for $l \in [L]$ do
        for $i, j$ do
            $(\widetilde{W_l})_{i,j} \leftarrow \begin{cases} (W_l)_{i,j}, & \text{if } |(W_l)_{i,j}| > a; \\ 0, & \text{otherwise .} \end{cases}$
        end for
    end for
    return $\widetilde{W}$
end procedure
```