APPENDIX

A PRELIMINARY

Theorem A.1 (Theorem 1.1 of (Tropp, 2012)). *Consider a finite sequence $\{X_k\}$ of independent, random, self-adjoint matrices of dimension $\overline{d}$. Assume that each random matrix satisfies*

$$X_k \succeq 0 \quad \text{and} \quad \lambda(X_k) \leq R.$$

Define $\mu_{\min} = \lambda_{\min}(\sum_k \mathbb{E}[X_k])$ and $\mu_{\max} = \lambda_{\max}(\sum_k \mathbb{E}[X_k])$. Then

$$\Pr\left\{\lambda_{\min}(\sum_k X_k) \leq (1-\delta)\mu_{\min}\right\} \leq \overline{d}\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu_{\min}/R} \quad \text{for } \delta \in [0,1], \text{and} \tag{4}$$

$$\Pr\left\{\lambda_{\max}(\sum_k X_k) \geq (1+\delta)\mu_{\max}\right\} \leq \overline{d}\left(\frac{e^{-\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu_{\max}/R} \quad \text{for } \delta \geq 0 \tag{5}$$

Lemma A.2 (Sherman-Morrison Formula). *Let $A \in \mathbb{R}^{n\times n}$ be an invertible matrix, and $u, v \in \mathbb{R}^n$. Suppose that $1 + v^\top A^{-1} u \neq 0$. Then it holds that*

$$(A + uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u}.$$

B TWO LAYER NEURAL NETWORK

Definition B.1 (Two layer neural network). *We define two layer neural network as follows*

$$f_{\mathsf{nn}}(W, a, x) := \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \phi(w_r^\top x) \in \mathbb{R}$$

where $x \in \mathbb{R}^d$ is the input, $w_r \in \mathbb{R}^d, r \in [m]$ is the weight vector of the first layer, $W = [w_1, \cdots, w_m] \in \mathbb{R}^{d\times m}, a_r \in \mathbb{R}, r \in [m]$ is the output weight, $a = [a_1, \cdots, a_m]^\top$ and $\phi(\cdot)$ is the non-linear activation function. Here we consider only training the first layer W with fixed a, so we also write $f_{\mathsf{nn}}(W, x) = f_{\mathsf{nn}}(W, a, x)$. Given training data matrix $X = [x_1, \cdots, x_n] \in \mathbb{R}^{n\times d}$ and labels $Y = [y_1, \cdots, y_n] \in \mathbb{R}^n$, we denote $f_{\mathsf{nn}}(W, X) = [f_{\mathsf{nn}}(W, x_1), \cdots, f_{\mathsf{nn}}(W, x_n)]^\top \in \mathbb{R}^n$.

Definition B.2. *Given $\mathcal{F}_{\mathsf{nn}}$ and the distribution D, let $\{v_1, \ldots, v_{\overline{d}}\}$ be a fixed orthonormal basis of $\mathcal{F}_{\mathsf{nn}}$, where inner products are taken under the distribution D, i.e.,*

$$\mathop{\mathbb{E}}_{x\sim D}[v_i(x) \cdot v_j(x)] = 1, \forall i = j \in [\overline{d}]$$

$$\left|\mathop{\mathbb{E}}_{x\sim D}[v_i(x) \cdot v_j(x)]\right| \leq \rho, \forall i \neq j \in [\overline{d}]$$

Furthermore, for any function $h \in \mathcal{F}_{\mathsf{nn}}$, there exists $\alpha(h) := (\alpha(h)_1, \ldots, \alpha(h)_{\overline{d}})$ under the basis $(v_1, \ldots, v_{\overline{d}})$ such that

$$h(x) = \sum_{i=1}^{\overline{d}} \alpha(h)_i \cdot v_i(x)$$

Remark. *Note that in for linear function family $\mathcal{F}_{\mathsf{nn}}$, we know that $\overline{d} = d$. However, for the neural network function family $\mathcal{F}_{\mathsf{nn}}$, we will have $\overline{d} \gg d$.*

Definition B.3. *Given distribution D and $h \in \mathcal{F}_{\mathsf{nn}}$, we define $\|h(x)\|_D$ as*

$$\|h(x)\|_D^2 := \mathop{\mathbb{E}}_{x\sim D}[|h(x)|^2].$$

Claim B.4. *If the activation ϕ in a neural network f_{nn} satisfies the following conditions,*

- $\phi^{(10d+\log(1/\varepsilon_0)/\log(d))}(x)$ *exists and is continuous.*
- $\phi^{(10d+\log(1/\varepsilon_0)/\log(d))}(x) \leq 1,\ x \in \mathbb{R}$.

then there exists $\{v_1, \ldots, v_{\overline{d}}\}$ *forms a fixed* ρ*-nearly orthonormal basis of* $\mathcal{F}_{\mathsf{nn}}$ *where*

$$\overline{d} \leq \binom{10d + \log(1/\varepsilon_0)/\log(d)}{d}.$$

Furthermore, for any $W \in \mathbb{R}^{d \times m}$*, there exists* $h \in \mathcal{F}_{\mathsf{nn}}$ *such that*

$$\|h(x) - f_{\mathsf{nn}}(W, x)\|_D^2 \leq \varepsilon.$$

Besides, for any $h \in \mathcal{F}_{\mathsf{nn}}$*, there exists* $W \in \mathbb{R}^{d \times m}$ *such that*

$$\|h(x) - f_{\mathsf{nn}}(W, x)\|_D^2 \leq \varepsilon.$$

Proof. For any activation $\phi : \mathbb{R} \to \mathbb{R}$ and input $z_r = w_r^\top x \in \mathbb{R}$, f can be expand by Taylor's theorem

$$\phi(z_r) = \phi(0) + \phi'(0) z_r + \frac{\phi''(0)}{2!} z_r^2 + \cdots + \frac{\phi^{(k)}(0)}{k!} z_r^k + \frac{\phi^{(k+1)}(\xi)}{(k+1)!} z_r^{k+1}$$

where $\xi \in [0, z_r]$.

It's natural to consider the W bound, eg NTK regime. In NTK regime, we have with a high probability that

$$z_r = w_r^\top x \leq \|x\|_1 \leq \sqrt{d}\|x\|_2 \leq \sqrt{d}.$$

We can claim that

$$\frac{\phi^{(k+1)}(\xi)}{(k+1)!} z_r^{k+1} \leq (\frac{e}{k+1} z_r)^{k+1} \leq \varepsilon_0.$$

where the first step follows from $\phi^{(k+1)}(\xi) \leq 1$, the second step follows from $k \geq (e\sqrt{d})^2 + \log(1/\varepsilon_0)/\log(e\sqrt{d})$.

Besides, taking the first $k+1$ terms in the Taylor's theorem. Our neural network f_{nn} can be seen as a polynomial with d variable and at most k degree. So, our neural network f_{nn} can be seen as a polynomial with at most $\binom{k+d}{d}$ terms. As any polynomial can be orthonormal decompose, we have that

$$\overline{d} \leq \binom{k+d}{d} \leq \binom{10d + \log(1/\varepsilon_0)/\log(d)}{d}.$$

where the second step follows from $d \geq e$. □

Claim B.5. *We will claim that polynomial, ReLU, Sigmoid, and Swish hold the condition mentioned in Claim B.4.*

Lemma B.6. *Let* $A \in \mathbb{R}^{\overline{d} \times \overline{d}}$ *be defined as*

$$A(i,j) = 1, \forall i = j \in [\overline{d}]$$
$$|A(i,j)| \leq \rho, \forall i \neq j \in [\overline{d}].$$

We will claim that

$$\lambda_{\max}(A) \leq 1 + \overline{d}\rho, \lambda_{\min}(A) \geq 1 - \overline{d}\rho.$$

Proof. First, we can lower bound $\lambda_{\min}(A)$ as follows

$$\begin{aligned}
\lambda_{\min}(A) &\geq \lambda_{\min}(I) - \|I - A\|_2 \\
&\geq 1 - \|I - A\|_F \\
&\geq 1 - \overline{d}\rho
\end{aligned}$$

Second, we can upper bound $\lambda_{\max}(A)$ as follows

$$\begin{aligned}
\lambda_{\max}(A) &= \|A\|_2 \\
&\leq \|I\|_2 + \|A - I\|_2 \\
&\leq 1 + \|A - I\|_F \\
&\leq 1 + \overline{d}\rho
\end{aligned}$$

□

C NOTATIONS

Claim C.1. *For any function $h \in \mathcal{F}_{\mathsf{nn}}$, we have*

$$C_l \|\alpha(h)\|_2^2 \leq \|h\|_D^2 \leq C_r \|\alpha(h)\|_2^2$$

where $C_l := 1 - \rho$ and $C_r := 1 + \rho(\overline{d} - 1)$.

Proof. We can rewrite $\|h\|_D^2$ as follows:

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim D}[|\sum_{i=1}^{\overline{d}} \alpha(h)_i \cdot v_i(x)|^2] &= \sum_{i=1}^{\overline{d}} \mathop{\mathbb{E}}_{x\sim D}[|\alpha(h)_i \cdot v_i(x)|^2] + 2 \sum_{1\leq i<j\leq \overline{d}} \mathop{\mathbb{E}}_{x\sim D}[\alpha(h)_i \alpha(h)_j \cdot v_i(x) v_j(x)] \\
&= \sum_{i=1}^{\overline{d}} |\alpha(h)_i|^2 + 2\rho \cdot \sum_{1\leq i<j\leq \overline{d}} \alpha(h)_i \alpha(h)_j \\
&= (1-\rho)\|\alpha(h)\|_2^2 + \rho(\sum_{i=1}^{\overline{d}} \alpha(h)_i)^2
\end{aligned}$$

We can provide an upper bound,

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim D}[|\sum_{i=1}^{\overline{d}} \alpha(h)_i \cdot v_i(x)|^2] &= (1-\rho)\|\alpha(h)\|_2^2 + \rho(\sum_{i=1}^{\overline{d}} \alpha(h)_i)^2 \\
&\leq (1-\rho)\|\alpha(h)\|_2^2 + \rho\overline{d}\|\alpha(h)\|_2^2 \\
&= (1-\rho+\rho\overline{d})\|\alpha(h)\|_2^2 \\
&= C_r\|\alpha(h)\|_2^2
\end{aligned}$$

We can provide a lower bound

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim D}[|\sum_{i=1}^{\overline{d}} \alpha(h)_i \cdot v_i(x)|^2] &= (1-\rho)\|\alpha(h)\|_2^2 + \rho(\sum_{i=1}^{\overline{d}} \alpha(h)_i)^2 \\
&\geq (1-\rho)\|\alpha(h)\|_2^2 \\
&= C_l\|\alpha(h)\|_2^2
\end{aligned}$$

Thus, we complete the proof. □

Now, we define condition number. Previous work only consider linear cases and we generalize it to NN.

Definition C.2. *For any distribution D' over the domain G and any function $h : G \to \mathbb{R}$, let $h^{(D')}(x) = \sqrt{\frac{D(x)}{D'(x)}} \cdot h(x)$ such that $\mathop{\mathbb{E}}_{x\sim D'}\left[|h^{(D')}(x)|^2\right] = \mathop{\mathbb{E}}_{x\sim D'}\left[\frac{D(x)}{D'(x)}|h(x)|^2\right] = \mathop{\mathbb{E}}_{x\sim D}\left[|h(x)|^2\right]$. When the neural network function $\mathcal{F}_{\mathsf{nn}}$ and D is clear, we use $K_{D'}$ to denote the condition number of sampling from D', i.e.,*

$$K_{D'} = \sup_x \left\{ \sup_{h\in\mathcal{F}_{\mathsf{nn}}} \left\{ \frac{|h^{(D')}(x)|^2}{\|h^{D'}\|_{D'}^2} \right\} \right\} = \sup_x \left\{ \frac{D(x)}{D'(x)} \cdot \sup_{h\in\mathcal{F}_{\mathsf{nn}}} \{ \frac{|h(x)|^2}{\|h\|_D^2} \} \right\}.$$

Definition C.3. *For any distribution D' over the domain G and any function $h : G \to \mathbb{R}$. When the neural network function $\mathcal{F}_{\mathsf{nn}}$ and D is clear, we use $K_{\alpha,D'}$ to denote the α-condition number of sampling from D', i.e.,*

$$K_{\alpha,D'} = \sup_x \left\{ \frac{D(x)}{D'(x)} \cdot \sup_{h\in\mathcal{F}_{\mathsf{nn}}} \{ \frac{|h(x)|^2}{\|\alpha(h)\|_2^2} \} \right\}.$$

Definition C.4. *Given $\mathcal{F}_{\mathsf{nn}}$ and underlying distribution D, let P be a random sampling procedure that terminates in k iterations (k is not necessarily fixed) and provides a coefficient α_i and a distribution D_i to sample $x_i \sim D_i$ in every iteration $i \in [k]$.*

*We say P is an ε-*importance sampling procedure *if it satisfies the following two properties:*

1. *Let $v_1, \ldots, v_{\overline{d}}$ of $\mathcal{F}_{\mathsf{nn}}$ under D be defined as in Definition B.2. With probability 0.9, the matrix $A(i,j) = \sqrt{u_i} \cdot v_j(x_i) \in \mathbb{R}^{k \times \overline{d}}$ has $\lambda(A^*A) \in [\frac{3}{4}, \frac{5}{4}]$.*

2. *The coefficients always have $\sum_{i=1}^{k} \beta_i \leq \frac{5}{4}$ and $\beta_i \cdot K_{\alpha,D_i} \leq \epsilon/2$.*

Claim C.5. *We will claim that*

$$C_l K_{D'} \leq K_{\alpha,D'} \leq C_r K_{D'}$$

Proof. Since, we have that

$$C_l \frac{1}{\|h\|_D^2} \leq \frac{1}{\|\alpha(h)\|_2^2} \leq C_r \{\frac{1}{\|h\|_D^2}\},$$

we can claim that

$$C_l \cdot \sup_x \left\{ \frac{D(x)}{D'(x)} \cdot \sup_{h \in \mathcal{F}_{\mathsf{nn}}} \{\frac{|h(x)|^2}{\|h\|_D^2}\} \right\} \leq \sup_x \left\{ \frac{D(x)}{D'(x)} \cdot \sup_{h \in \mathcal{F}_{\mathsf{nn}}} \{\frac{|h(x)|^2}{\|\alpha(h)\|_2^2}\} \right\} \leq C_r \cdot \sup_x \left\{ \frac{D(x)}{D'(x)} \cdot \sup_{h \in \mathcal{F}_{\mathsf{nn}}} \{\frac{|h(x)|^2}{\|h\|_D^2}\} \right\}$$

□

Definition C.6. *Given a* importance sampling procedure *P, we say the output of P is* good *only if the samples x_i with weights $u_i = \beta_i \cdot D(x_i)/D_i(x_i)$ satisfy the first property in Definition C.4. Given a joint distribution (D, Y) and an execution of a* importance sampling procedure *P with $x_i \sim D_i$ and u_i of each $i \in [k]$, let the $\widetilde{f}$ be defined as*

$$\widetilde{f} = \arg\min_{h \in \mathcal{F}_{\mathsf{nn}}} \left\{ \sum_{i=1}^{k} u_i \cdot |h(x_i) - y_i|^2 \right\}$$

by querying $y_i \sim (Y|x_i)$ for each point x_i .

D Recovery Guarantee for Importance Samples

D.1 Preliminary

We state a tool from prior work,

Lemma D.1 (Lemma 4.3 of (Chen & Price, 2019))**.** *Let P be a random sampling procedure terminating in k iterations (k is not necessarily fixed) that in every iteration i, it provides a coefficient β_i and a distribution D_i to sample $x_i \sim D_i$. Let the weight $u_i = \beta_i \cdot \frac{D(x_i)}{D_i(x_i)}$ and $A \in \mathbb{R}^{k \times \overline{d}}$ denote the matrix $A(i,j) = \sqrt{u_i} \cdot v_j(x_i)$. Then for $f = \arg\min_{h \in \mathcal{F}_{\mathsf{nn}}} \mathbb{E}_{(x,y) \sim (D,Y)} [|y - h(x)|^2]$,*

$$\mathbb{E}_P \left[\|A^*(\vec{y}_u - \vec{f}_{S,u})\|_2^2 \right] \leq \sup_P \{ \sum_{i=1}^{k} \beta_i \} \cdot \max_j \{ \beta_j \cdot K_{\alpha,D_j} \} \mathbb{E}_{(x,y) \sim (D,Y)} [|y - f(x)|^2],$$

where K_{α,D_i} is the condition number for samples from D_i: $K_{\alpha,D_i} = \sup_x \left\{ \frac{D(x)}{D_i(x)} \cdot \sup_{v \in \mathcal{F}} \{ \frac{|v(x)|^2}{\|\alpha(v)\|_2^2} \} \right\}$.

D.2 Analysis for Importance Sampling Procedure

Lemma D.2. *For any $\varepsilon \in (0,1)$, given $S = (x_1, \ldots, x_k) \subset \mathbb{R}^d$ and their weights $(u_1, \ldots, u_k) \subset \mathbb{R}_{\geq 0}$, let A be the $k \times \overline{d}$ matrix defined as*

$$A(i,j) := \sqrt{u_i} \cdot v_j(x_i).$$

Then,

$$\lambda(A^*A) \in [1 - \varepsilon, 1 + \varepsilon].$$

can imply that

$$\|h\|_{S,u}^2 := \sum_{j=1}^{k} u_j \cdot |h(x_j)|^2 \in [\frac{1-\varepsilon}{C_r}, \frac{1+\varepsilon}{C_l}] \cdot \|h\|_D^2 \quad \text{for every } h \in \mathcal{F}_{\mathsf{nn}}$$

Proof. Notice that

$$A \cdot \alpha(h) = \big(\sqrt{u_1} \cdot h(x_1), \ldots, \sqrt{u_k} \cdot h(x_k)\big). \tag{6}$$

We can show

$$\begin{aligned}
\|h\|_{S,u}^2 &= \sum_{i=1}^{k} u_i \cdot |h(x_i)|^2 \\
&= \|A \cdot \alpha(h)\|_2^2 \\
&= \alpha(h)^* \cdot (A^* \cdot A) \cdot \alpha(h) \\
&\in [\lambda_{\min}(A^* \cdot A), \lambda_{\max}(A^* \cdot A)] \cdot \|\alpha(h)\|_2^2 \\
&\subseteq [1-\varepsilon, 1+\varepsilon] \cdot \|\alpha(h)\|_2^2 \\
&\subseteq [\frac{1-\varepsilon}{C_r}, \frac{1+\varepsilon}{C_l}] \cdot \|h\|_D^2
\end{aligned}$$

where the first step follows from the definition of the norm, the second step follows from the Eq. (6), the last step follows from $C_l\|\alpha(h)\|_2^2 \leq \|h\|_D^2 \leq C_r\|\alpha(h)\|_2^2$.

□

For any $\varepsilon > 0$, given $S = (x_1, \ldots, x_k) \subset \mathbb{R}^d$ and their weights $(u_1, \ldots, u_k) \subset \mathbb{R}_{\geq 0}$. Let A be the $k \times \overline{d}$ matrix defined as

$$A(i,j) := \sqrt{u_i} \cdot v_j(x_i).$$

For any $h \in \mathcal{F}_{\mathsf{nn}}$, let $\|h\|_{S,u}^2$ be defined as

$$\|h\|_{S,u}^2 := \sum_{j=1}^{k} u_j \cdot |h(x_j)|^2$$

Then, for any $h \in \mathcal{F}_{\mathsf{nn}}$

$$A \cdot \alpha(h) = \big(\sqrt{u_1} \cdot h(x_1), \ldots, \sqrt{u_k} \cdot h(x_k)\big). \tag{7}$$

We consider the calculation of the $\widetilde{f}$. Given the weights $(u_1, \cdots, u_k)$ on $(x_1, \ldots, x_k)$ and labels $(y_1, \ldots, y_k)$, let $\vec{y}_u$ denote the vector of weighted labels $(\sqrt{u_1} \cdot y_1, \ldots, \sqrt{u_k} \cdot y_k)$. From Eq. (6), the empirical distance satisfied that for any $h \in \mathcal{F}_{\mathsf{nn}}$

$$\begin{aligned}
\|h(x_i) - y_i\|_{S,u}^2 &= \sum_{i=1}^{k} u_i |h(x_i) - y_i|^2 \\
&= \|A \cdot \alpha(h) - \vec{y}_u\|_2^2
\end{aligned}$$

where the second step follows from Eq. (6) and the definition of $\vec{y}_u$.

Let

$$\begin{aligned}
\widetilde{f} &= \arg\min_{h \in \mathcal{F}_{\mathsf{nn}}} \big\{ \|h(x_i) - y_i\|_{S,u} \big\} \\
&= \arg\min_{h \in \mathcal{F}_{\mathsf{nn}}} \big\{ \|A \cdot \alpha(h) - \vec{y}_u\|_2 \big\}
\end{aligned}$$

Then,

$$\alpha(\widetilde{f}) = (A^* \cdot A)^{-1} \cdot A^* \cdot \vec{y}_u \text{ and } \widetilde{f} = \sum_{i=1}^{\overline{d}} \alpha(\widetilde{f})_i \cdot v_i.$$

Let

$$f = \arg\min_{h \in \mathcal{F}_{\mathsf{nn}}} \big\{ \mathop{\mathbb{E}}_{(x,y) \sim (D,Y)} [|h(x) - y|^2] \big\}$$

Finally, we consider the distance between f and $\widetilde{f}$. For convenience, let $\vec{f}_u = \big(\sqrt{u_1} \cdot f(x_1), \ldots, \sqrt{u_k} \cdot f(x_k)\big)$. Because $f \in \mathcal{F}_{\mathsf{nn}}$ and Eq. (6), we can claim that,

$$\alpha(f) = (A^* \cdot A)^{-1} \cdot A^* \cdot \vec{f}_u.$$

This implies

$$\|\widetilde{f} - f\|_D^2 \in [C_l, C_r] \cdot \|\alpha(\widetilde{f}) - \alpha(f)\|_2^2 = [C_l, C_r] \cdot \|(A^* \cdot A)^{-1} \cdot A^* \cdot (\vec{y}_u - \vec{f}_u)\|_2^2.$$

where the first step follows from $\|h\|_D^2 \in [C_l, C_r]\|\alpha(h)\|_2^2$ and $\alpha(f-g) = \alpha(f) - \alpha(g)$.

Theorem D.3. *Given a neural network function family $\mathcal{F}_{\mathsf{nn}}$, joint distribution (D, Y), and $\varepsilon > 0$, let P be an ε-importance sampling procedure for $\mathcal{F}_{\mathsf{nn}}$ and D, and let $f = \arg\min_{h \in \mathcal{F}} \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y - h(x)|^2]$. Then the $\widetilde{f}$ of a good output of P satisfies*

$$\|f - \widetilde{f}\|_D^2 \leq \epsilon \cdot \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y - f(x)|^2] \text{ in expectation.}$$

Proof. We assume the first property $\lambda(A^* \cdot A) \in [1 - 1/4, 1 + 1/4]$ from Definition C.6. On the other hand,

$$\mathop{\mathbb{E}}_{P}[\|A^* \cdot (\vec{y}_u - \vec{f}_u)\|_2^2] \leq \epsilon \cdot \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y - f(x)|^2]$$

from Lemma D.1 where P is an random sampling procedure. Conditioned on the first property, we know

$$\begin{aligned}\mathop{\mathbb{E}}_{P'}[\|A^* \cdot (\vec{y}_u - \vec{f}_u)\|_2^2] &\leq \frac{1}{0.9} \mathop{\mathbb{E}}_{P}[\|A^* \cdot (\vec{y}_u - \vec{f}_u)\|_2^2] \\ &\leq \frac{\varepsilon}{0.9} \cdot \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y - f(x)|^2]\end{aligned}$$

where P' is the event where importance sampling procedure P good executed.

This implies

$$\begin{aligned}\mathop{\mathbb{E}}_{P'}\left[\|(A^* \cdot A)^{-1} \cdot A^* \cdot (\vec{y}_u - \vec{f}_u)\|_2^2\right] &\leq \mathop{\mathbb{E}}_{P'}\left[\lambda_{\min}(A^* \cdot A)^{-1} \cdot \|A^* \cdot (\vec{y}_u - \vec{f}_u)\|_2^2\right] \\ &\leq 2\epsilon \cdot \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y - f(x)|^2]\end{aligned}$$

where the second step follows from the first property in the definition of a importance sampling procedure P and P' is a good output of P. □

E A Linear-Sample Algorithm for Known Distribution

E.1 Preliminary

We state several tools from prior work.

Lemma E.1 (Lemma 3.3 in (Batson et al., 2012))**.** *For any $j \in [k]$, $\lambda(B_j) \in (l_j, r_j)$.*

Lemma E.2 (A combination of Claim 5.5 and Lemma 5.6 in (Chen & Price, 2019))**.** *After exiting the while loop in Procedure* RandomizedBSS*, we always have*

1. $r_k - l_k \leq 9\overline{d}/\gamma$.
2. $(1 - \frac{0.5\gamma^2}{\overline{d}}) \cdot \sum_{j=1}^k \frac{\gamma}{\phi_j} \leq \mathsf{mid} \leq \sum_{j=1}^k \frac{\gamma}{\phi_j}$.
3. *If* $\frac{r_k}{l_k} \leq 1 + 8\gamma$*, then* $\lambda(A^* \cdot A) \in (1 - 5\gamma, 1 + 5\gamma)$.

Lemma E.3 (Lemma 5.1 in (Chen & Price, 2019))**.** *Given any dimension $\overline{d}$ linear space $\mathcal{F}_{\mathsf{nn}}$, any distribution D over the domain of $\mathcal{F}_{\mathsf{nn}}$, and any $\varepsilon > 0$, there exists an* ε-importance sampling procedure *that terminates in $O(\overline{d}/\varepsilon)$ rounds with probability* 0.99.

Algorithm 2 A importance sampling procedure based on Randomized Sampling

procedure RANDOMIZEDSAMPLING($\mathcal{F}_{\mathsf{nn}}, D, \epsilon$)
  Find an ρ-nearly orthonormal basis $v_1, \ldots, v_{\overline{d}}$ of $\mathcal{F}_{\mathsf{nn}}$ under D;
  $\gamma \leftarrow \sqrt{\epsilon}/C_0$;
  $\mathsf{mid} \leftarrow (4\overline{d}/\gamma)/(1/(1-\gamma) - 1/(1+\gamma))$;
  $B_0 \leftarrow 0$;
  $l_0 \leftarrow -2\overline{d}/\gamma$;
  $r_0 \leftarrow 2\overline{d}/\gamma$;
  $j \leftarrow 0$;
  while $r_{j+1} - l_{j+1} < 8\overline{d}/\gamma$ **do**;
    $\Phi_j \leftarrow \mathrm{tr}[(r_j I - B_j)^{-1}] + \mathrm{tr}[(B_j - l_j I)^{-1}]$;
    $D_j(x) \leftarrow D(x) \cdot \Big(v(x)^\top (r_j I - B_j)^{-1} v(x) + v(x)^\top (B_j - l_j I)^{-1} v(x)\Big)/\Phi_j$;
    Sample $x_j \sim D_j$;
    $s_j \leftarrow \gamma \cdot D(x)/(\Phi_j \cdot D_j(x))$;
    $B_{j+1} \leftarrow B_j + s_j \cdot v(x_j) v(x_j)^\top$;
    $r_{j+1} \leftarrow r_j + \gamma/(\Phi_j(1-\gamma))$;
    $l_{j+1} \leftarrow l_j + \gamma/(\Phi_j(1+\gamma))$;
    $j \leftarrow j + 1$;
  end while
  $k \leftarrow j$;
  for $j \in [k]$ **do**
    $\beta_j \leftarrow \gamma/(\Phi_j \cdot \mathsf{mid})$,
    $u_j \leftarrow s_j/\mathsf{mid}$,
  end for
  Output x, D, u, β
end procedure

E.2 ANALYSIS FOR OUR RANDOMIZED SAMPLING

Our results in Lemma E.4, Lemma E.5, Lemma E.6, and E.7 are different than results in (Allen-Zhu et al., 2015; Lee & Sun, 2018). First, they consider a potential function $\mathrm{tr}[(r_j I - B_j)^{-q}] + \mathrm{tr}[(B_j - l_j I)^{-q}]$ with $q \geq 10$. But, in our cases, $q = 1$. Second, their results is for orthonormal basis but not for ρ-nearly orthonormal basis.

Lemma E.4. *Let $\varepsilon \in (0, 1/2)$. Suppose that $w^\top (uI - A)^{-1} w \leq \varepsilon$ and $w^\top (A - lI)^{-1} w \leq \varepsilon$. It holds that*

$$\mathrm{tr}[(A - lI + ww^\top)^{-1}] \leq \mathrm{tr}[(A - lI)^{-1}] - (1-\varepsilon) w^\top (A - lI)^{-2} w$$
$$\mathrm{tr}[(uI - A - ww^\top)^{-1}] \leq \mathrm{tr}[(uI - A)^{-1}] + (1+2\varepsilon) w^\top (uI - A)^{-2} w.$$

Proof. Let $Y = A - lI$. By the Sherman-Morrison Formula (Lemma A.2), it holds that

$$\mathrm{tr}[(Y + ww^\top)^{-1}] = \mathrm{tr}[Y^{-1} - \frac{Y^{-1} w w^\top Y^{-1}}{1 + w^\top Y^{-1} w}].$$

By the assumption of $w^\top Y^{-1} w \leq \varepsilon$, we have that

$$\begin{aligned}
\mathrm{tr}[(Y + ww^\top)^{-1}] &\leq \mathrm{tr}[Y^{-1} - \frac{Y^{-1} w w^\top Y^{-1}}{1+\varepsilon}] \\
&= \mathrm{tr}[Y^{-1}(I - \frac{Y^{-1/2} w w^\top Y^{-1/2}}{1+\varepsilon})] \\
&\leq \mathrm{tr}[Y^{-1}(I - (1-\varepsilon) Y^{-1/2} w w^\top Y^{-1/2})] \\
&\leq \mathrm{tr}[Y^{-1}] - (1-\varepsilon) w^\top Y^{-2} w.
\end{aligned}$$

Let $Z = uI - A$. By the Sherman-Morrison Formula (Lemma A.2), it holds that

$$\mathrm{tr}[(Z - ww^\top)^{-1}] = \mathrm{tr}[Z^{-1} + \frac{Z^{-1} w w^\top Z^{-1}}{1 - w^\top Z^{-1} w}].$$

By the assumption of $w^\top Z^{-1} w \le \varepsilon$, we have that

$$\begin{aligned}\mathrm{tr}[(Z - ww^\top)^{-1}] &\le \mathrm{tr}[Z^{-1} + \frac{Z^{-1}ww^\top Z^{-1}}{1-\varepsilon}] \\ &= \mathrm{tr}[Z^{-1}(I + \frac{Z^{-1/2}ww^\top Z^{-1/2}}{1-\varepsilon})] \\ &\le \mathrm{tr}[Z^{-1}(I + (1+2\varepsilon)Z^{-1/2}ww^\top Z^{-1/2})] \\ &\le \mathrm{tr}[Z^{-1}] + (1+2\varepsilon)w^\top Z^{-2} w.\end{aligned}$$

where the third step follows from $\varepsilon \le 0.5$. □

Lemma E.5. *Let w_i be defined as*

$$w_j := \sqrt{\frac{\gamma}{v(x_j)^\top (r_j I - B_j)^{-1} v(x_j) + v(x_j)^\top (B_j - l_j I)^{-1} v(x_j)}} \cdot v(x_j).$$

Let $\gamma \le \min\{1/(c\log(\overline{d}/\gamma'), 1/(c(1+\overline{d}\rho)e^2)\}$. Then, it holds that

$$\begin{aligned}\Pr[0 \preceq w_j w_j^\top \preceq \frac{1}{c}\cdot(r_j I - B_j)] \ge 1 - \gamma' \\ \Pr[0 \preceq w_j w_j^\top \preceq \frac{1}{c}\cdot(B_j - l_j I)] \ge 1 - \gamma'\end{aligned}$$

Proof. Let $R_j = v(x_j)^\top (r_j I - B_j)^{-1} v(x_j) + v(x_j)^\top (B_j - l_j I)^{-1} v(x_j)$.
We can claim that

$$\mathop{\mathbb{E}}_{x\sim D_j}[w_j w_j^\top] = \frac{\gamma}{\Phi_j} \mathop{\mathbb{E}}_{x\sim D}[v(x_j)v(x_j)^\top] \preceq \frac{\gamma(1+\overline{d}\rho)}{\Phi_j}\cdot I$$

Let

$$z_j = (r_j I - B_j)^{-1/2} w_j.$$

It holds that

$$\begin{aligned}\mathrm{tr}[z_j z_j^\top] &= \mathrm{tr}[(r_j I - B_j)^{-1/2} w_j w_j^\top (r_j I - B_j)^{-1/2}] \\ &= \frac{\gamma}{R_j}\cdot \mathrm{tr}[(r_j I - B_j)^{-1/2} v_j v_j^\top (r_j I - B_j)^{-1/2}] \\ &= \frac{\gamma}{R_j}\cdot \mathrm{tr}[v_j^\top (r_j I - B_j)^{-1} v_j] \\ &\le \gamma,\end{aligned}$$

and $\lambda_{\max}(z_j z_j^\top) \le \gamma$. Moreover, it holds that

$$\begin{aligned}\mathbb{E}[z_j z_j^\top] &= \frac{\gamma(1+\overline{d}\rho)}{\Phi_j}\cdot (r_j I - B_j)^{-1} \\ &\preceq \frac{\gamma}{\Phi_j}\cdot \lambda_{\max}(\frac{1}{r_j I - B_j})\cdot I.\end{aligned}$$

This implies that

$$\lambda_{\max}(\mathbb{E}[z_j z_j^\top]) \le \frac{\gamma(1+\overline{d}\rho)}{\Phi_j}\cdot \lambda_{\max}(\frac{1}{r_j I - B_j}) =: \mu$$

It holds by the Matrix Chernoff Bound (Lemma A.1) that

$$\Pr[\lambda_{\max}(\mathbb{E}[z_j z_j^\top]) \ge (1+\delta)\mu] \le \overline{d}\cdot (\frac{\exp(\delta)}{(1+\delta)^{(1+\delta)}})^{\mu/\gamma}.$$

Set $1+\delta$ to be

$$\begin{aligned}1+\delta &= 1/(c\mu)\\ &= \frac{\Phi_j}{c\gamma(1+\overline{d}\rho)}\lambda_{\min}(r_jI-B_j)\\ &\geq \frac{1}{c\gamma(1+\overline{d}\rho)}\end{aligned}$$

With probability at least

$$\begin{aligned}1-\overline{d}\cdot(\frac{\exp(\delta)}{(1+\delta)^{(1+\delta)}})^{\mu/\gamma} &\geq 1-\overline{d}\cdot(\frac{e}{1+\delta})^{\mu(1+\delta)/\gamma}\\ &= 1-\overline{d}\cdot(\frac{e}{1+\delta})^{1/c\gamma}\\ &\geq 1-\overline{d}\cdot(ce\gamma(1+\overline{d}\rho))^{1/(c\gamma)}\\ &\geq 1-\overline{d}\cdot\exp(-1/(c\gamma))\\ &\geq 1-\gamma'\end{aligned}$$

where the fourth step follows from $\gamma\leq 1/(c(1+\overline{d}\rho)e^2)$, the fifth step follows from $\gamma\leq 1/(c\log(\overline{d}/\gamma'))$. As a result, we can prove that

$$\Pr[0\preceq w_jw_j^\top\preceq\frac{1}{c}\cdot(r_jI-B_j)]\geq 1-\gamma'.$$

Similarly, we can prove that

$$\Pr[0\preceq w_jw_j^\top\preceq\frac{1}{c}\cdot(B_j-l_jI)]\geq 1-\gamma'.$$

□

Lemma E.6. *It holds that*

$$\mathop{\mathbb{E}}_{x_j\in D_j}[\Phi_{j+1}]\leq\Phi_j.$$

Proof. We claim that

$$w_jw_j^T\preceq\frac{1}{c}\cdot(r_jI-B_j)\preceq\frac{1}{c}\cdot(r_{j+1}I-B_j)$$

We apply Lemma E.4 with $\varepsilon=1/c$ and get that

$$\begin{aligned}&\mathop{\mathbb{E}}_{x_j\in D_j}[\mathrm{tr}[(r_{j+1}I-B_{j+1})^{-1}]]\\ \leq\ &\mathrm{tr}[(r_{j+1}I-B_j)^{-1}]+(1+2/c)\,\mathrm{tr}[(r_{j+1}I-B_j)^{-2}\,\mathbb{E}[w_jw_j^\top]]\\ =\ &\mathrm{tr}[(r_{j+1}I-B_j)^{-1}]+\frac{(1+2/c)\gamma}{\Phi_j}\,\mathrm{tr}[(r_{j+1}I-B_j)^{-2}]\end{aligned}$$

Note that $r_{j+1}-r_j=\gamma/(\Phi_j(1-\gamma))$, we define a function f by

$$f(t)=\mathrm{tr}[((r_j+t\cdot(r_{j+1}-r_j))I-B_j)^{-1}].$$

Notice that

$$\frac{\mathrm{d}f(t)}{\mathrm{d}t}=-\frac{\gamma}{\Phi_j(1-\gamma)}\,\mathrm{tr}[((r_j+t\cdot(r_{j+1}-r_j))I-B_j)^{-2}]$$

Since f is convex, we have that

$$\frac{\mathrm{d}f(t)}{\mathrm{d}t}\big|_{t=1} \geq f(1) - f(0) = \mathrm{tr}[(r_{j+1}I - B_j)^{-1}] - \mathrm{tr}[(r_jI - B_j)^{-1}].$$

We can conclude that

$$\begin{aligned}
\mathop{\mathbb{E}}_{x_j \in D_j}[\mathrm{tr}[(r_{j+1}I - B_{j+1})^{-1}]] &\leq \mathrm{tr}[(r_{j+1}I - B_j)^{-1}] + \frac{(1+2/c)\gamma}{\Phi_j}\mathrm{tr}[(r_{j+1}I - B_j)^{-2}] \\
&\leq \mathrm{tr}[(r_{j+1}I - B_j)^{-1}] \\
&\quad + (1+2/c)(1-\gamma)(\mathrm{tr}[(r_jI - B_j)^{-1}] - \mathrm{tr}[(r_{j+1}I - B_j)^{-1}]) \\
&\leq \mathrm{tr}[(r_jI - B_j)^{-1}]
\end{aligned}$$

where the last step follows from $(1+2/c)(1-\gamma) \leq 1$.

On the other hand, we can proof that

$$w_j w_j^T \preceq \frac{1}{c} \cdot (B_j - l_jI) \preceq \frac{1}{c} \cdot (B_j - l_{j+1}I)$$

We apply Lemma E.4 with $\varepsilon = 1/c$ and get that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_j \in D_j}[\mathrm{tr}[(B_{j+1} - l_{j+1})^{-1}]] \\
&\leq \mathrm{tr}[(B_j - l_{j+1}I)^{-1}] - (1-1/c)\,\mathrm{tr}[(B_j - l_{j+1}I)^{-2}\,\mathbb{E}[w_j w_j^\top]] \\
&= \mathrm{tr}[(B_j - l_{j+1}I)^{-1}] - \frac{(1-1/c)\gamma}{\Phi_j}\mathrm{tr}[(B_j - l_{j+1}I)^{-2}]
\end{aligned}$$

Note that $l_{j+1} - l_j = \gamma/(\Phi_j(1+\gamma))$, we define a function g by

$$g(t) = \mathrm{tr}[(B_j - (l_j + t \cdot (l_{j+1} - l_j))I)^{-1}].$$

Notice that

$$\frac{\mathrm{d}g(t)}{\mathrm{d}t} = \frac{\gamma}{\Phi_j(1+\gamma)}\mathrm{tr}[(B_j - (l_j + t \cdot (l_{j+1} - l_j))I)^{-2}]$$

Since g is convex, we have that

$$\frac{\mathrm{d}g(t)}{\mathrm{d}t}\big|_{t=1} \geq g(1) - g(0) = \mathrm{tr}[(B_j - l_{j+1}I)^{-1}] - \mathrm{tr}[(B_j - l_jI)^{-1}].$$

We can conclude that

$$\begin{aligned}
\mathop{\mathbb{E}}_{x_j \in D_j}[\mathrm{tr}[(B_{j+1} - l_{j+1})^{-1}]] &\leq \mathrm{tr}[(B_j - l_{j+1}I)^{-1}] - \frac{(1-1/c)\gamma}{\Phi_j}\mathrm{tr}[(B_j - l_{j+1}I)^{-2}] \\
&\leq \mathrm{tr}[(B_j - l_{j+1}I)^{-1}] \\
&\quad + (1-1/c)(1+\gamma)(\mathrm{tr}[(B_j - l_jI)^{-1}] - \mathrm{tr}[(B_j - l_{j+1}I)^{-1}]) \\
&\leq \mathrm{tr}[(B_j - l_jI)^{-1}]
\end{aligned}$$

where the last step follows from $(1-1/c)(1+\gamma) \geq 1$ and $\mathrm{tr}[(B_j - l_jI)^{-1}] - \mathrm{tr}[(B_j - l_{j+1}I)^{-1}] \leq 0$. □

Lemma E.7. *There exists a constant C such that with probability at least 0.99, Procedure* RANDOMIZEDSAMPLING *takes at most $k = C \cdot \bar{d}/\gamma^2$ random points $x_1, \ldots, x_k$ and guarantees that $\frac{r_k}{l_k} \leq 1 + 8\gamma$.*

Proof.

$$\begin{aligned}
\Pr[\text{algorithm finishes within } k \text{ iterations}] &\geq \Pr[4\bar{d}/\gamma + \sum_{j=0}^{k-1}(\frac{\gamma}{\Phi_j(1-\gamma)} - \frac{\gamma}{\Phi_j(1+\gamma)}) \geq 8\bar{d}/\gamma] \\
&\geq \Pr[\sum_{j=0}^{k-1}\frac{1}{\Phi_j} \geq \frac{4\bar{d}(1-\gamma^2)}{\gamma^3}] \\
&\geq \Pr[\frac{\gamma^3 k^2}{4\bar{d}(1-\gamma^2)} \geq \sum_{j=0}^{k-1}\Phi_j]
\end{aligned}$$

By Lemma E.5, every picked matrix $w_j w_j^\top$ in iteration j satisfies

$$0 \preceq w_j w_j^\top \preceq \frac{1}{2} \cdot (r_j I - A) \tag{8}$$

with probability at least $(1-\gamma')^k \geq 1 - k\gamma'$. Under the condition of Eq. (8), by Lemma E.6, we have that

$$\mathbb{E}[\sum_{j=0}^{k-1} \Phi_j] \leq k\gamma.$$

Therefore, it holds that

$$\begin{aligned}
&\Pr[\text{algorithm does not finish within } k \text{ iterations}] \\
\leq\ &\Pr[\frac{\gamma^3 k^2}{4\overline{d}(1-\gamma^2)} \leq \sum_{j=0}^{k-1} \Phi_j] \\
\leq\ &\Pr[\sum_{j=0}^{k-1} \Phi_j \geq \frac{\gamma^3 k^2}{4\overline{d}(1-\gamma^2)} \text{ and } \forall j : w_j w_j^\top \preceq \frac{1}{2}(r_j I - A_j)] + \Pr[\exists j : w_j w_j^\top \not\preceq \frac{1}{2}(r_j I - A_j)] \\
\leq\ &\frac{4\overline{d}(1-\gamma^2)}{\gamma^2 k} + k\gamma' \\
\leq\ &\frac{1}{200}
\end{aligned}$$

where the last step follows from $\gamma' \leq 1/(400k)$ and $k \geq 1600\overline{d}(1-\gamma^2)/\gamma^2$.

Let $\Delta_{r,j} = r_j - r_{j-1}, \Delta_{l,j} = l_j - l_{j-1}$. Since

$$\frac{\Delta_{r,j+1} - \Delta_{l,j+1}}{\Delta_{r,j+1}} = \frac{\gamma/(\Phi_j(1-\gamma)) - \gamma/(\Phi_j(1+\gamma))}{\gamma/(\Phi_j(1-\gamma))} = \frac{2\gamma}{1+\gamma}.$$

We can claim that

$$\begin{aligned}
\frac{r_k - l_k}{r_k} &= \frac{4\overline{d}/\gamma + \sum_{j=0}^{k-1}(\Delta_{r,j+1} - \Delta_{l,j+1})}{2\overline{d}/\gamma + \sum_{j=0}^{k-1} \Delta_{r,j+1}} \\
&\leq \frac{4\overline{d}/\gamma + \sum_{j=0}^{k-1}(\Delta_{r,j+1} - \Delta_{l,j+1})}{2\overline{d}/\gamma + (1+\gamma)/(2\gamma) \cdot \sum_{j=0}^{k-1}(\Delta_{r,j+1} - \Delta_{l,j+1})}
\end{aligned}$$

By the ending condition of the algorithm, it holds that $r_k - l_k \geq 8\overline{d}/\gamma$. As a result

$$\frac{r_k - l_k}{r_k} \leq 4\gamma,$$

and

$$\frac{r_k}{l_k} \leq \frac{1}{1-4\gamma} \leq 1 + 8\gamma.$$

where the last step follows from $\gamma \leq 1/8$. □

F Performance of i.i.d. Distributions

Lemma F.1. *Let D' be an arbitrary distribution over G. There exists an absolute constant C such that for any $n \in \mathbb{N}^+$, $\mathcal{F}_{\mathsf{nn}}$ of dimension $\overline{d}$, $\varepsilon \in (0,1)$ and $\delta \in (0,1)$, when $S = (x_1, \ldots, x_k)$ are independently from the distribution D' with $k \geq \frac{C}{\varepsilon^2} \cdot K_{D'} \log \frac{\overline{d}}{\delta}$ and $u_j = \frac{D(x_j)}{k \cdot D'(x_j)}$ for each $j \in [k]$, the $k \times \overline{d}$ matrix $A(i,j) = \sqrt{u_i} \cdot v_j(x_i)$ satisfies*

$$\|A^* A - I\| \leq \varepsilon \text{ with probability at least } 1 - \delta.$$

Proof. At the same time, for any fixed x,

$$\sum_{i\in[\overline{d}]} |v_i^{(D')}(x)|^2 = \sup_{\alpha(h)} \frac{|\sum_{i=1}^{\overline{d}} \alpha(h)_i \cdot v_i^{(D')}(x)|^2}{\|\alpha(h)\|_2^2} = \sup_{h\in\mathcal{F}_{\mathsf{nn}}} \frac{|h^{(D')}(x)|^2}{\|\alpha(h)\|_2^2}$$

by the tightness of the Cauchy Schwartz inequality. Thus

$$K_{\alpha,D'} \stackrel{def}{=} \sup_{x\in G}\{\sup_{h\in\mathcal{F}_{\mathsf{nn}}:h\neq 0} \frac{|h^{(D')}(x)|^2}{\|\alpha(h)\|_2^2}\} \quad \text{indicates} \quad \sup_{x\in G} \sum_{i\in[\overline{d}]} |v_i^{(D')}(x)|^2 = K_{\alpha,D'}. \tag{9}$$

For each point x_j in S with weight $u_j = \frac{D(x_j)}{k\cdot D'(x_j)}$, let A_j denote the jth row of the matrix A. It is a vector in $\mathbb{R}^{\overline{d}}$ defined by $A_j(i) = A(j,i) = \sqrt{u_j}\cdot v_i(x_j) = \frac{v_i^{(D')}(x_j)}{\sqrt{k}}$. So $A^*A = \sum_{j=1}^k A_j^* \cdot A_j$.

For $A_j^* \cdot A_j$, it is always $\succeq 0$. Notice that the only non-zero eigenvalue of $A_j^* \cdot A_j$ is

$$\lambda(A_j^* \cdot A_j) = A_j \cdot A_j^* = \frac{1}{k}\left(\sum_{i\in[\overline{d}]} |v_i^{(D')}(x_j)|^2\right) \leq \frac{K_{\alpha,D'}}{k}$$

from (9).

At the same time, because the expectation of the entry (i,i') in $A_j^* \cdot A_j$ is

$$\begin{aligned}
\mathop{\mathbb{E}}_{x_j\sim D'}[A(j,i)\cdot A(j,i')] &= \mathop{\mathbb{E}}_{x_j\sim D'}[\frac{v_i^{(D')}(x_j)\cdot v_{i'}^{(D')}(x_j)}{k}] \\
&= \mathop{\mathbb{E}}_{x_j\sim D'}[\frac{D(x)\cdot v_i(x_j)\cdot v_{i'}(x_j)}{k\cdot D'(x_j)}] \\
&= \mathop{\mathbb{E}}_{x_j\sim D}[\frac{v_i(x_j)\cdot v_{i'}(x_j)}{k}].
\end{aligned}$$

We can claim that

$$\begin{aligned}
\mathop{\mathbb{E}}_{x_j\sim D'}[A(j,i)\cdot A(j,i')] &= 1/k, \forall i = i', \\
\mathop{\mathbb{E}}_{x_j\sim D'}[A(j,i)\cdot A(j,i')] &\leq \rho/k, \forall i \neq i'.
\end{aligned}$$

As a result, we have that

$$\begin{aligned}
\lambda_{\min}(\sum_{j=1}^k \mathbb{E}[A_j^* \cdot A_j]) &\geq 1-\rho\overline{d}, \\
\lambda_{\max}(\sum_{j=1}^k \mathbb{E}[A_j^* \cdot A_j]) &\leq 1+\rho\overline{d}.
\end{aligned}$$

Now we apply Theorem A.1 on $A^*A = \sum_{j=1}^k (A_j^* \cdot A_j)$:

$$\begin{aligned}
&\Pr\left[\lambda(A^*A) \notin [(1-\varepsilon)(1-\rho\overline{d}), (1+\varepsilon)(1+\rho\overline{d})]\right] \\
&\leq \overline{d}\left(\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right)^{(1-\rho\overline{d})/\frac{K_{\alpha,D'}}{k}} + \overline{d}\left(\frac{e^{-\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right)^{(1-\rho\overline{d})/\frac{K_{\alpha,D'}}{k}} \\
&\leq 2\overline{d}\cdot\exp(-\frac{\varepsilon^2\cdot\frac{k(1-\rho\overline{d})}{K_{\alpha,D'}}}{3}) \\
&\leq \delta.
\end{aligned}$$

where the last step follows from $k \geq \frac{6K_{\alpha,D'}\log\frac{\overline{d}}{\delta}}{\varepsilon^2(1-\rho\overline{d})}$.

Thus, we complete the proof. □

Lemma F.2. *Given any distribution D' with the same support of D and any $\epsilon > 0$, the random sampling procedure with $k = \Theta(K_{\alpha,D'} \log \overline{d} + \frac{K_{D'}}{\varepsilon})$ i.i.d. random samples from D' and coefficients $\beta_i = 1/k$, $\forall i \in [k]$ is an* ε-importance sampling procedure.

Proof. Because the coefficient $\beta_i = 1/k = O(\varepsilon/K_{\alpha,D'})$ and $\sum_i \beta_i = 1$, this indicates the second property of *importance sampling procedure*.

Since $k = \Theta(K_{\alpha,D'} \log \overline{d})$, by Lemma F.1, we know all eigenvalues of $A^* \cdot A$ are in $[1-1/4, 1+1/4]$ with probability $1 - 10^{-3}$. This indicates the first property of *importance sampling procedure*. □

G Results for Active Learning

Previous work (Chen & Price, 2019) only consider linear case, we generalize it into NN.

Lemma G.1. *Consider any dimension $\overline{d}$ linear space $\mathcal{F}_{\mathsf{nn}}$ of functions from a domain G to $\mathbb{R}$. Let (D, Y) be a joint distribution over $G \times \mathbb{R}$ and $f = \arg\min_{h \in \mathcal{F}_{\mathsf{nn}}} \mathbb{E}_{(x,y)\sim(D,Y)}[|y - h(x)|^2]$. Let $K_\alpha = \sup_{h \in \mathcal{F}_{\mathsf{nn}}: h \neq 0} \frac{\sup_{x \in G} |h(x)|^2}{\|\alpha(h)\|_2^2}$ and P be a* importance sampling procedure *terminating in $m_p(\varepsilon)$ rounds with probability $1 - 10^{-3}$ for $\mathcal{F}_{\mathsf{nn}}$, distribution D, and ε. For any $\varepsilon \in (0, 1/10)$, Algorithm 1 takes $O(K \log(\overline{d}) + K/\varepsilon)$ unlabeled samples from D and requests at most $m_p(\varepsilon/8)$ labels to output $\widetilde{f}$ satisfying*

$$\mathbb{E}_{x\sim D}[|\tilde{f}(x) - f(x)|^2] \leq \varepsilon \cdot \mathbb{E}_{(x,y)\sim(D,Y)}[|y - f(x)|^2] \text{ in expectation.}$$

Proof. We still use $\|f\|_{D'}$ to denote $\sqrt{\mathbb{E}_{x\sim D'}[|f(x)|^2]}$.By Lemma F.2 with D and the property of P, with probability at least $1 - 2 \cdot 10^{-3}$,

$$\|h\|_{D_0}^2 \in [\frac{3}{4C_r}, \frac{5}{4C_l}] \cdot \|h\|_D^2 \text{ for every } h \in \mathcal{F}_{\mathsf{nn}}. \tag{10}$$

We condition on Eq. (10) holds from now on.

Let y_i denote a random label of x_i from $Y(x_i)$ for each $i \in [k_0]$ including the unlabeled samples in the algorithm and the labeled samples in Step 6 of Algorithm 1. Let f' be defined as

$$f' = \arg\min_{h \in \mathcal{F}_{\mathsf{nn}}} \mathbb{E}_{x_i \sim D_0, y_i \sim Y(x_i)} \left[|y_i - h(x_i)|^2\right]. \tag{11}$$

Using Eq. (10) and Lemma F.2, we have

$$\mathbb{E}_{(x_1,y_1),\ldots,(x_{k_0},y_{k_0})}[\|f' - f\|_D^2] \leq \varepsilon \cdot \mathbb{E}_{(x,y)\sim(D,Y)}[|y - f(x)|^2] \text{ from the proof of Theorem D.3 .}$$

In the next a few paragraph, we will show that $\widetilde{f}$ of a good output of P with distribution D_0 guarantees $\|\widetilde{f} - f'\|_{D_0}^2 \lesssim \mathbb{E}_{(x,y)\sim(D,Y)} \left[|y - f(x)|^2\right]$ with high probability.

Using Eq. (10) and the guarantee of Procedure P, we have

$$\mathbb{E}_P[\|\widetilde{f} - f'\|_{D_0}^2] \leq \varepsilon \cdot \mathbb{E}_{x\sim D_0} \left[|y_i - f'(x_i)|^2\right]$$

from the proof of Theorem D.3.

Next we bound the right hand side $\mathbb{E}_{x_i\sim D_0} \left[|y_i - f'(x_i)|^2\right]$ by $\mathbb{E}_{(x,y)\sim(D,Y)} \left[|y - f(x)|^2\right]$ over the randomness of $(x_1, y_1), \ldots, (x_{k_0}, y_{k_0})$:

$$\begin{aligned}
&\mathbb{E}_{(x_1,y_1),\ldots,(x_{k_0},y_{k_0})} \left[\mathbb{E}_{x_i\sim D_0} \left[|y_i - f'(x_i)|^2\right]\right] \\
\leq &\mathbb{E}_{(x_1,y_1),\ldots,(x_{k_0},y_{k_0})} \left[2 \mathbb{E}_{x_i\sim D_0} \left[|y_i - f(x_i)|^2\right] + 2\|f - f'\|_{D_0}^2\right] \\
\leq &2 \mathbb{E}_{(x,y)\sim(D,Y)} \left[|y - f(x)|^2\right] + \frac{3}{C_l} \mathbb{E}_{(x_1,y_1),\ldots,(x_{k_0},y_{k_0})} \left[\|f - f'\|_D^2\right].
\end{aligned}$$

where the last step follows from Eq. (10).

Hence

$$\begin{aligned}\mathop{\mathbb{E}}_{(x_1,y_1),\ldots,(x_{k_0},y_{k_0})}\left[\mathop{\mathbb{E}}_{P}[\|\widetilde{f}-f'\|_{D_0}^2]\right] &\lesssim \varepsilon(2+\frac{3\varepsilon}{C_l})\cdot \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}\left[|y-f(x)|^2\right]\\ &\lesssim \varepsilon\cdot \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}\left[|y-f(x)|^2\right].\end{aligned}$$

where the last step follows from $\rho \le 1/2$.

From all discussion above, by rescaling ε, we have

$$\begin{aligned}&\|\widetilde{f}-f\|_D^2\\ &\le 2\|\widetilde{f}-f'\|_D^2+2\|f'-f\|_D^2\\ &\le \frac{8C_r}{3}\|\widetilde{f}-f'\|_{D_0}^2+2\|f'-f\|_D^2\\ &\le \frac{8C_r\varepsilon}{3}\mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y-f(x)|^2]+\frac{\varepsilon}{4}\cdot\mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y-f(x)|^2]\\ &\lesssim \varepsilon(1+C_r)\cdot\mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y-f(x)|^2]\\ &\lesssim \varepsilon(1+\rho\overline{d})\cdot\mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y-f(x)|^2]\end{aligned}$$

□

Theorem G.2. *Let $\mathcal{F}_{\mathsf{nn}}$ be a neural network function family of functions from a domain G to $\mathbb{R}$ with dimension $\overline{d}$, and consider any (unknown) distribution on (x,y) over $G\times\mathbb{R}$. Let D be the marginal distribution over x, and suppose it has bounded "condition number"*

$$K := \sup_{h\in\mathcal{F}_{\mathsf{nn}}:h\neq 0}\frac{\sup_{x\in G}|h(x)|^2}{\|h\|_D^2}. \tag{12}$$

Let $f^*\in\mathcal{F}_{\mathsf{nn}}$ minimize $\mathbb{E}[|f(x)-y|^2]$. For any $\varepsilon<1$, there exists an randomized algorithm that takes $O((1+\rho\overline{d})(K\log(\overline{d})+K/\varepsilon))$ unlabeled samples from D and requires $O(\overline{d}/\varepsilon)$ labels to output $\widetilde{f}$ such that

$$\mathop{\mathbb{E}}_{\widetilde{f}}\mathop{\mathbb{E}}_{x\sim D}[|\widetilde{f}(x)-f^*(x)|^2]\lesssim\varepsilon(1+\rho\overline{d})\cdot\mathop{\mathbb{E}}_{x,y}[|y-f^*(x)|^2].$$

Proof. We can claim that

$$\sup_{h\in\mathcal{F}_{\mathsf{nn}}:h\neq 0}\frac{\sup_{x\in G}|h(x)|^2}{\|h\|_D^2}\le C_r\sup_{h\in\mathcal{F}_{\mathsf{nn}}:h\neq 0}\frac{\sup_{x\in G}|h(x)|^2}{\|\alpha(h)\|_2^2}$$

where $C_r\le 1+\rho\overline{d}$.

By applying Lemma G.1, Lemma E.3, we can finish our proof immediately. □

Theorem G.3. *Let $f_{\mathsf{nn}}(W,x)$ be a neural network as defined in Definition B.1, and consider any (unknown) distribution on (x,y) over $\mathbb{R}^d\times\mathbb{R}$. Let D be the marginal distribution over x, and suppose it has bounded "condition number"*

$$K := \sup_{W\in\mathbb{R}^{d\times m}:W\neq 0}\frac{\sup_{x\in G}|f(W,x)|^2}{\|f(W,x)\|_D^2}. \tag{13}$$

Let $W^*\in\mathbb{R}^{d\times m}$ minimize $\mathbb{E}[|f_{\mathsf{nn}}(W,x)-y|^2]$. There exists $\rho\in(0,1/10),\overline{d}\ge 3,\varepsilon_0\in(0,1/10)$. For any $0<\varepsilon\lesssim 1/\log^3(\overline{d})$, there exists an randomized algorithm P that takes $O((1+\rho\overline{d})(K\log(\overline{d})+K/\varepsilon))$ unlabeled samples from D and requires $O(\overline{d}/\varepsilon)$ labels to output $\widetilde{W}\in\mathbb{R}^{d\times m}$ such that

$$\mathop{\mathbb{E}}_{P}\mathop{\mathbb{E}}_{x\sim D}[|f_{\mathsf{nn}}(\widetilde{W},x)-f_{\mathsf{nn}}(W^*,x)|^2]\lesssim\varepsilon+\varepsilon\cdot\mathop{\mathbb{E}}_{x,y}[|y-f_{\mathsf{nn}}(W^*,x)|^2].$$

Proof. By Claim B.4, we have that there exists $\{v_1, \ldots, v_{\overline{d}}\}$ forms a fixed ρ-nearly orthonormal basis of $\mathcal{F}_{\mathsf{nn}}$, such that there exists h^* satisfied that

$$\|h^*(x) - f_{\mathsf{nn}}(W^*, x)\|_D^2 \leq \varepsilon_0.$$

By Theorem G.2, there exists an randomized algorithm that takes $O((1+\rho\overline{d})(K\log(\overline{d}) + K/\varepsilon))$ unlabeled samples from D and requires $O(\overline{d}/\varepsilon)$ labels to output $\widetilde{h}$ such that

$$\mathop{\mathbb{E}}_{\widetilde{h}} \mathop{\mathbb{E}}_{x\sim D}[|\widetilde{h}(x) - h^*(x)|^2] \lesssim \varepsilon \cdot \mathop{\mathbb{E}}_{x,y}[|y - h^*(x)|^2].$$

By Claim B.4, there exists $\widetilde{W}$ such that

$$\|\widetilde{h}(x) - f_{\mathsf{nn}}(\widetilde{W}, x)\|_D^2 \leq \varepsilon_0.$$

As a result,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{P} \mathop{\mathbb{E}}_{x\sim D}[|f_{\mathsf{nn}}(\widetilde{W}, x) - f_{\mathsf{nn}}(W^*, x)|^2] \\
\leq &\mathop{\mathbb{E}}_{P} \mathop{\mathbb{E}}_{x\sim D}[2|h^*(x) - f_{\mathsf{nn}}(W^*, x)|^2 + 2|f_{\mathsf{nn}}(\widetilde{W}, x) - h^*|^2] \\
\leq &\mathop{\mathbb{E}}_{P} \mathop{\mathbb{E}}_{x\sim D}[2|h^*(x) - f_{\mathsf{nn}}(W^*, x)|^2 + 4|\widetilde{h}(x) - h^*(x)|^2 + 4|\widetilde{h}(x) - f_{\mathsf{nn}}(\widetilde{W}, x)|^2] \\
\lesssim &\varepsilon_0 + \varepsilon \cdot \mathop{\mathbb{E}}_{x,y}[|y - h^*(x)|^2] \\
\leq &\varepsilon_0 + \varepsilon \cdot \mathop{\mathbb{E}}_{x,y}[2|y - f_{\mathsf{nn}}(W^*, x)|^2 + 2|f_{\mathsf{nn}}(W^*, x) - h^*(x)|^2] \\
\lesssim &\varepsilon_0 + \varepsilon \cdot \mathop{\mathbb{E}}_{x,y}[|y - f_{\mathsf{nn}}(W^*, x)|^2].
\end{aligned}$$

□