A Some concentration results for uniform random variables

In this section, we state some concentration results that are useful for the theoretical analysis in Section 3. Let $\widetilde{\boldsymbol{x}}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \overset{\text{i.i.d.}}{\sim} \text{Unif}(\mathcal{B}_{\mathbf{0},\sqrt{d+2}})$ be i.i.d. samples from the uniform distribution over the Euclidean norm ball of radius $\sqrt{d+2}$ in $\mathbb{R}^d$. Let

$$Z_n = \min_{\boldsymbol{x} \in \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}} \|\widetilde{\boldsymbol{x}} - \boldsymbol{x}\|^2. \tag{11}$$

If $n = 1$, $\mathbb{E}Z_1$ is the sum of the variance of each coordinate of $\text{Unif}(\mathcal{B}_{\mathbf{0},\sqrt{d+2}})$. Therefore, $\mathbb{E}Z_n$ provides a generalized measure of concentration. Intuitively, $\mathbb{E}Z_n \to 0$ as $n \to \infty$. The proposition below provides a upper bound on the rate of convergence.

Lemma A.1 (Nearest Neighbor concentration)**.** *Given the assumptions above*

$$\mathbb{E}Z_n \lesssim d^2 \left[\frac{\log\left(n^{1/d}\right)}{n}\right]^{1/d}, \tag{12}$$

where $\lesssim$ means inequality up to an universal constant independent of d and n.

Proof. Define

$$\begin{aligned} \mathcal{E}_1 &= \{Z_n \le \delta^2\}, \\ \mathcal{E}_2 &= \{\delta \le \sqrt{d+2} - \|\widetilde{\boldsymbol{x}}\|\}. \end{aligned} \tag{13}$$

We will compute two probabilities $\mathbb{P}(\mathcal{E}_1|\mathcal{E}_2)$ and $\mathbb{P}(\mathcal{E}_2)$ that will be useful latter.

$$\begin{aligned} \mathbb{P}(\mathcal{E}_1^c|\mathcal{E}_2) &= \mathbb{P}(Z_n \ge \delta^2|\mathcal{E}_2) = \mathbb{P}(\|\widetilde{\boldsymbol{x}} - \boldsymbol{x}_i\| \ge \delta,\ \forall i|\mathcal{E}_2), \\ &= \mathbb{E}_{\widetilde{\boldsymbol{x}}}\mathbb{P}(\|\widetilde{\boldsymbol{x}} - \boldsymbol{x}_i\| \ge \delta|\mathcal{E}_2, \widetilde{\boldsymbol{x}})^n = \mathbb{E}_{\widetilde{\boldsymbol{x}}}(1 - \mathbb{P}(\|\widetilde{\boldsymbol{x}} - \boldsymbol{x}_i\| \le \delta|\mathcal{E}_2, \widetilde{\boldsymbol{x}}))^n, \\ &= \mathbb{E}_{\widetilde{\boldsymbol{x}}}\left[1 - \frac{\text{Vol}(\mathcal{B}_{\widetilde{\boldsymbol{x}},\delta})}{\text{Vol}(\mathcal{B}_{\mathbf{0},\sqrt{d+2}})}\right]^n = \left[1 - \left(\frac{\delta}{\sqrt{d+2}}\right)^d\right]^n, \\ &\le \exp\left[-n\left(\frac{\delta}{\sqrt{d+2}}\right)^d\right]. \end{aligned} \tag{14}$$

Next, we compute $\mathbb{P}(\mathcal{E}_2)$

$$\mathbb{P}(\mathcal{E}_2) = \mathbb{P}(\|\widetilde{\boldsymbol{x}}\| \le \sqrt{d+2} - \delta) = \left(\frac{\sqrt{d+2} - \delta}{\sqrt{d+2}}\right)^d = \left(1 - \frac{\delta}{\sqrt{d+2}}\right)^d. \tag{15}$$

We use $\mathcal{E}_1$ and $\mathcal{E}_2$ to compute the following upper bound

$$\begin{aligned} \mathbb{E}Z_n &= \mathbb{E}(Z_n|\mathcal{E}_1 \cap \mathcal{E}_2)\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) + \mathbb{E}(Z_n|(\mathcal{E}_1 \cap \mathcal{E}_2)^c)P((\mathcal{E}_1 \cap \mathcal{E}_2)^c), \\ &\le \delta^2 + (2\sqrt{d+2})^2\left(1 - \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2)\right), \\ &= \delta^2 + 4(d+2)\left[1 - \mathbb{P}(\mathcal{E}_1|\mathcal{E}_2)\mathbb{P}(\mathcal{E}_2)\right]. \end{aligned} \tag{16}$$

To find an upper bound for $\mathbb{E}Z_n$, we need to find an upper bound for $1 - \mathbb{P}(\mathcal{E}_1|\mathcal{E}_2)\mathbb{P}(\mathcal{E}_2)$.

$$\begin{aligned} 1 - \mathbb{P}(\mathcal{E}_1|\mathcal{E}_2)\mathbb{P}(\mathcal{E}_2) &= 1 - \left[1 - \mathbb{P}(\mathcal{E}_1^c|\mathcal{E}_2)\right]\mathbb{P}(\mathcal{E}_2), \\ &= 1 - \mathbb{P}(\mathcal{E}_2) + \mathbb{P}(\mathcal{E}_1^c|\mathcal{E}_2)\mathbb{P}(\mathcal{E}_2), \\ &\le 1 - \mathbb{P}(\mathcal{E}_2) + \mathbb{P}(\mathcal{E}_1^c|\mathcal{E}_2). \end{aligned} \tag{17}$$

Now choose $\delta = \sqrt{d+2}n^{-1/d}\left[\log\left(n^{1/d}\right)\right]^{1/d}$.

$$\mathbb{P}(\mathcal{E}_1^c|\mathcal{E}_2) \le \exp\left[-n\left(\frac{\delta}{\sqrt{d+2}}\right)^d\right] = \exp\left[-nn^{-1}\log\left(n^{1/d}\right)\right] = n^{-1/d}, \tag{18}$$

and

$$\mathbb{P}(\mathcal{E}_2) = \left(1 - \frac{\delta}{\sqrt{d+2}}\right)^d \geq 1 - d\frac{\delta}{\sqrt{d+2}} = 1 - dn^{-1/d}\left[\log\left(n^{1/d}\right)\right]^{1/d}. \tag{19}$$

Thus

$$1 - \mathbb{P}(\mathcal{E}_1|\mathcal{E}_2)\mathbb{P}(\mathcal{E}_2) \leq 1 - 1 + dn^{-1/d}\left[\log\left(n^{1/d}\right)\right]^{1/d} + n^{-1/d} \lesssim dn^{-1/d}\left[\log\left(n^{1/d}\right)\right]^{1/d}. \tag{20}$$

Combining everything together, we get

$$\begin{aligned}
\mathbb{E}Z_n &\leq (d+2)n^{-2/d}\left[\log\left(n^{1/d}\right)\right]^{2/d} + 4(d+2) \times dn^{-1/d}\left[\log\left(n^{1/d}\right)\right]^{1/d}, \\
&\lesssim d^2 n^{-1/d}\left[\log\left(n^{1/d}\right)\right]^{1/d}, \\
&= d^2\left[\frac{\log\left(n^{1/d}\right)}{n}\right]^{1/d}.
\end{aligned} \tag{21}$$

This completes the proof. □

Proposition A.2 ([47] Corollary 6.20)**.** *Let* $\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim}$ *Unif* $(\mathcal{B}_{\mathbf{0},\sqrt{d+2}})$ *for* $i = 1, \ldots, n$ *be uniformly distributed over a ball of radius* B *in* $\mathbb{R}^d$ *centered at* $\mathbf{0}$*. Let*

$$\boldsymbol{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathsf{T}}$$

be the sample covariance matrix. Then

$$\mathbb{P}(\|\boldsymbol{\Sigma}_n - \boldsymbol{I}\|_{\text{op}} > \varepsilon) \leq 2d\exp\left[-\frac{n\varepsilon^2}{2(d+2)(1+\varepsilon)}\right].$$

B Proof of Theorem 3.3

In this section, we present the proof of Theorem 3.3. In Section B.1, we provide the detail of the decomposition of the risk into T_1 and T_2. Then in Section B.2 we compute an upper bound for T_1, and compute an upper bound for T_2 in Section B.3. Finally, we combine everything together in Section B.4 and completes the proof.

B.1 Decomposition of the test risk

$$
\begin{aligned}
&\mathbb{E}\left[f^{\mathsf{ResMem}}(\widetilde{\boldsymbol{x}}) - f_\star(\widetilde{\boldsymbol{x}})\right]^2 = \mathbb{E}\left[f_n(\widetilde{\boldsymbol{x}}) + r_n(\widetilde{\boldsymbol{x}}) - f_\star(\widetilde{\boldsymbol{x}})\right]^2, \\
&= \mathbb{E}\left[f_n(\widetilde{\boldsymbol{x}}) - f_\star(\widetilde{\boldsymbol{x}}) - f_n(\widetilde{\boldsymbol{x}}_{(1)}) + f_\star(\widetilde{\boldsymbol{x}}_{(1)})\right]^2, \\
&= \mathbb{E}\left[f_n(\widetilde{\boldsymbol{x}}) - f_\infty(\widetilde{\boldsymbol{x}}) + f_\infty(\widetilde{\boldsymbol{x}}) - f_\star(\widetilde{\boldsymbol{x}}) - f_n(\widetilde{\boldsymbol{x}}_{(1)}) + f_\infty(\widetilde{\boldsymbol{x}}_{(1)}) - f_\infty(\widetilde{\boldsymbol{x}}_{(1)}) + f_\star(\widetilde{\boldsymbol{x}}_{(1)})\right]^2, \\
&\le 3 \times [\underbrace{\mathbb{E}(f_n(\widetilde{\boldsymbol{x}}) - f_\infty(\widetilde{\boldsymbol{x}}))^2 + \mathbb{E}(f_n(\widetilde{\boldsymbol{x}}_{(1)}) - f_\infty(\widetilde{\boldsymbol{x}}_{(1)}))^2}_{T_1} + \underbrace{\mathbb{E}(f_\infty(\widetilde{\boldsymbol{x}}) - f_\star(\widetilde{\boldsymbol{x}}) - f_\infty(\widetilde{\boldsymbol{x}}_{(1)}) + f_\star(\widetilde{\boldsymbol{x}}_{(1)}))^2}_{T_2}],
\end{aligned}
\tag{22}
$$

where in the last inequality, we used the fact that $(a+b+c)^2 < 3(a^2+b^2+c^2)$ for any $a, b, c \in \mathbb{R}$.

B.2 Upper bound on T_1.

Since $\mathbb{P}_{\boldsymbol{x}} = \text{Unif}(\mathcal{B}_{\boldsymbol{0},B}\ )$, we apply the bound $\|\widetilde{\boldsymbol{x}}\|, \|\widetilde{\boldsymbol{x}}_{(1)}\| \le B$ to obtain

$$
\begin{aligned}
T_1 &= \mathbb{E}[f_n(\widetilde{\boldsymbol{x}}) - f_\infty(\widetilde{\boldsymbol{x}})]^2 + \mathbb{E}[f_n(\widetilde{\boldsymbol{x}}_{(1)}) - f_\infty(\widetilde{\boldsymbol{x}}_{(1)})]^2, \\
&= \mathbb{E}\langle \boldsymbol{\theta}_n - \boldsymbol{\theta}_\infty, \widetilde{\boldsymbol{x}}\rangle^2 + \mathbb{E}\langle \boldsymbol{\theta}_n - \boldsymbol{\theta}_\infty, \widetilde{\boldsymbol{x}}_{(1)}\rangle^2, \\
&\le \mathbb{E}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_\infty\|^2\|\widetilde{\boldsymbol{x}}\|^2 + \mathbb{E}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_\infty\|^2\|\widetilde{\boldsymbol{x}}_{(1)}\|^2, \\
&\le 2B^2\mathbb{E}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_\infty\|^2.
\end{aligned}
\tag{23}
$$

As n gets large, the empirical covariance matrix $\boldsymbol{\Sigma}_n = \boldsymbol{X}^\mathsf{T}\boldsymbol{X}/n$ is concentrated around its mean $\boldsymbol{I}$. Let $\boldsymbol{\Delta}_n = \boldsymbol{I} - \boldsymbol{\Sigma}_n$ denote this deviation. For some $\varepsilon \in (0,1)$, define the following "good event" over the randomness in $\boldsymbol{\Sigma}_n$

$$
\mathcal{A} = \{\|\boldsymbol{\Delta}_n\|_{\text{op}} < \varepsilon\},
\tag{24}
$$

where $\|\boldsymbol{\Delta}_n\|_{\text{op}}$ denotes the operator norm of the deviation matrix. The high level idea of the proof is to condition on the event $\mathcal{A}$ and deduce and upper bound of $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_\infty\|$ in terms of ε. Then, we use the fact that $\mathcal{A}$ happens with high probability.

Recall that $\boldsymbol{\theta}_\infty = L\boldsymbol{\theta}_\star$, and

$$
\boldsymbol{\theta}_n = \operatorname*{argmin}_{\|\boldsymbol{\theta}\| \le L} \frac{1}{n}\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\|^2.
\tag{25}
$$

Since $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_\star$ by definition, the Lagrangian of the convex program above is

$$
\mathcal{L}(\boldsymbol{\theta}, \lambda) = \frac{1}{n}\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}_\star\|^2 + \lambda(\|\boldsymbol{\theta}\|^2 - L).
\tag{26}
$$

The KKT condition suggests that the primal-dual optimal pair $(\boldsymbol{\theta}_n, \lambda_n)$ is given by

$$
\begin{aligned}
\|\boldsymbol{\theta}_n\| &\le L, \\
\lambda_n &\ge 0, \\
\lambda_n(\|\boldsymbol{\theta}_n\| - L) &= 0,
\end{aligned}
\tag{27}
$$

and at optimality

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}_n, \lambda_n) = 0 &\iff \frac{2}{n}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}(\boldsymbol{\theta} - \boldsymbol{\theta}_\star) + 2\lambda_n\boldsymbol{\theta} = 0, \\
&\iff \boldsymbol{\theta}_n = (\boldsymbol{\Sigma}_n + \lambda_n\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_n\boldsymbol{\theta}_\star.
\end{aligned}
\tag{28}
$$

The complementary slackness condition $\lambda_n(\|\boldsymbol{\theta}_n\| - L) = 0$ suggests that either $\lambda_n = 0$ or $\|\boldsymbol{\theta}_n\| = L$. But if $\lambda_n = 0$, the stationary condition $\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \lambda) = 0$ would suggest that $\boldsymbol{\theta}_n = \boldsymbol{\Sigma}_n^{-1}\boldsymbol{\Sigma}_n\boldsymbol{\theta}_\star = \boldsymbol{\theta}_\star \Rightarrow \|\boldsymbol{\theta}_n\| = 1 > L$, a contradiction. (Note that here $\boldsymbol{\Sigma}_n$ is invertible condition on the event $\mathcal{A}$.) Therefore, we must have $\|\boldsymbol{\theta}_n\| = L$. As a result, the primal and dual pair $(\boldsymbol{\theta}_n, \lambda_n)$ is determined by the system of equations

$$\begin{cases} \boldsymbol{\theta}_n & = (\boldsymbol{\Sigma}_n + \lambda_n \boldsymbol{I})^{-1}\boldsymbol{\Sigma}_n\boldsymbol{\theta}_\star, \\ \|\boldsymbol{\theta}_n\| & = L, \\ \lambda_n & > 0. \end{cases} \tag{29}$$

Next, we proceed to compute the deviation $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_\infty\|$.

$$\begin{aligned} \boldsymbol{\theta}_n &= [(\lambda_n+1)\boldsymbol{I} - \boldsymbol{\Delta}_n]^{-1}\,\boldsymbol{\Sigma}_n\boldsymbol{\theta}_\star, \\ &= (\lambda_n+1)^{-1}\left[\boldsymbol{I} - \frac{\boldsymbol{\Delta}_n}{\lambda_n+1}\right]^{-1}\boldsymbol{\Sigma}_n\boldsymbol{\theta}_\star, \\ &= (\lambda_n+1)^{-1}\left[\boldsymbol{I} + \sum_{k=1}^{\infty}\frac{\boldsymbol{\Delta}_n^k}{(\lambda_n+1)^k}\right](\boldsymbol{I} - \boldsymbol{\Delta}_n)\boldsymbol{\theta}_\star, \\ &= (\lambda_n+1)^{-1}\left[\boldsymbol{I} + \sum_{k=1}^{\infty}\frac{\boldsymbol{\Delta}_n^k}{(\lambda_n+1)^k} - \boldsymbol{\Delta}_n - \sum_{k=1}^{\infty}\frac{\boldsymbol{\Delta}_n^{k+1}}{(\lambda_n+1)^k}\right]\boldsymbol{\theta}_\star, \\ &= (\lambda_n+1)^{-1}\boldsymbol{\theta}_\star + (\lambda_n+1)^{-1}\boldsymbol{\Delta}_n\left[\sum_{k=1}^{\infty}\frac{\boldsymbol{\Delta}_n^{k-1}}{(\lambda_n+1)^k} - \boldsymbol{I} - \sum_{k=1}^{\infty}\frac{\boldsymbol{\Delta}_n^{k}}{(\lambda_n+1)^k}\right]\boldsymbol{\theta}_\star, \\ &= (\lambda_n+1)^{-1}\boldsymbol{\theta}_\star + (\lambda_n+1)^{-1}\boldsymbol{\Delta}_n\left[\sum_{k=1}^{\infty}\frac{\boldsymbol{\Delta}_n^{k-1} - \boldsymbol{\Delta}_n^{k}}{(\lambda_n+1)^k} - \boldsymbol{I}\right]\boldsymbol{\theta}_\star. \end{aligned} \tag{30}$$

Define

$$\boldsymbol{D}_n = \boldsymbol{\Delta}_n\left[\sum_{k=1}^{\infty}\frac{\boldsymbol{\Delta}_n^{k-1} - \boldsymbol{\Delta}_n^{k}}{(\lambda_n+1)^k} - \boldsymbol{I}\right]. \tag{31}$$

Then $\boldsymbol{\theta}_n = (\lambda_n+1)^{-1}\boldsymbol{\theta}_\star + (\lambda_n+1)^{-1}\boldsymbol{D}_n\boldsymbol{\theta}_\star$, and

$$\begin{aligned} \|\boldsymbol{D}_n\| &\le \|\boldsymbol{\Delta}_n\|\left[1 + \sum_{k=1}^{\infty}\frac{\|\boldsymbol{\Delta}_n\|^{k-1} + \|\boldsymbol{\Delta}_n\|^{k}}{(\lambda_n+1)^k}\right], \\ &\le \varepsilon\left[1 + 2(1+\lambda_n)^{-1}\sum_{k=1}^{\infty}\left(\frac{\varepsilon}{1+\lambda_n}\right)^k\right], \\ &= \varepsilon\left(1 + \frac{2}{1+\lambda_n}\frac{1}{1 - \frac{\varepsilon}{1+\lambda_n}}\right) \le 3\varepsilon. \end{aligned} \tag{32}$$

Therefore

$$\begin{aligned} &L = \|\boldsymbol{\theta}_n\|^2 = (\lambda_n+1)^{-2} + (\lambda_n+1)^{-2}\boldsymbol{\theta}_\star^\top\boldsymbol{D}_n^\top\boldsymbol{D}_n\boldsymbol{\theta}_\star + 2(\lambda_n+1)^{-2}\boldsymbol{\theta}_\star\boldsymbol{D}_n\boldsymbol{\theta}_\star, \\ \Rightarrow &(\lambda_n+1)^2L^2 = 1 + \delta_n,\ \delta_n = \boldsymbol{\theta}_\star^\top\boldsymbol{D}_n^\top\boldsymbol{D}_n\boldsymbol{\theta}_\star + 2\boldsymbol{\theta}_\star^\top\boldsymbol{D}_n\boldsymbol{\theta}_\star. \end{aligned} \tag{33}$$

We can obtain the following bound for δ_n:

$$|\delta_n| \le \|\boldsymbol{\theta}_\star\|^2\|\boldsymbol{D}_n\|^2 + 2\|\boldsymbol{\theta}_\star\|^2\|\boldsymbol{D}_n\| \le 9\varepsilon^2 + 6\varepsilon \le 15\varepsilon. \tag{34}$$

Since $1 - \delta_n/2 \le \sqrt{1+\delta_n} \le 1 + \delta_n/2$, and $|\delta_n| \le 15\varepsilon$, we obtain

$$|(\lambda_n+1)L - 1| \le \frac{15\varepsilon}{2} \Rightarrow \left|L - (\lambda_n+1)^{-1}\right| \le \frac{15\varepsilon}{2}(\lambda_n+1)^{-1} \le \frac{15\varepsilon}{2}, \tag{35}$$

where the last inequality follows as we have $\lambda_n > 0$. Finally,

$$\begin{aligned}
\boldsymbol{\theta}_n - \boldsymbol{\theta}_\infty &= (\lambda_n + 1)^{-1}\boldsymbol{\theta}_\star - L\boldsymbol{\theta}_\star + (\lambda_n+1)^{-1}\boldsymbol{D}_n\boldsymbol{\theta}_\star, \\
\Rightarrow \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_\infty\|^2 &= [(1+\lambda_n)^{-1} - L]^2 + (1+\lambda_n)^{-2}\boldsymbol{\theta}_\star\boldsymbol{D}_n^\top\boldsymbol{D}_n\boldsymbol{\theta}_\star + 2(\lambda_n+1)^{-1}[(1+\lambda_n)^{-1} - L]\boldsymbol{\theta}_\star\boldsymbol{D}_n\boldsymbol{\theta}_\star, \\
&\le 64\varepsilon^2 + 9\varepsilon^2 + 45\varepsilon^2 = 118\varepsilon^2, \\
\Rightarrow \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_\infty\|^2 &\lesssim \varepsilon^2.
\end{aligned} \tag{36}$$

Combine the above result with Proposition A.2, we get that

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_\infty\|^2 &= \mathbb{E}(\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_\infty\|^2|\mathcal{A})\mathbb{P}(\mathcal{A}) + \mathbb{E}(\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_\infty\|^2|\mathcal{A}^c)\mathbb{P}(\mathcal{A}^c), \\
&\le \varepsilon^2 + 4L^2 \times 4d\exp\left[-\frac{n\varepsilon^2}{2(d+2)(1+\varepsilon)}\right],
\end{aligned} \tag{37}$$

If we choose $\varepsilon = n^{-1/3}$, we get

$$\mathbb{E}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_\infty\|^2 \lesssim dL^2 n^{-2/3}, \tag{38}$$

which implies that

$$T_1 \lesssim d^2 L^2 n^{-2/3}. \tag{39}$$

B.3 Upper bound on T_2.

Plugging in the formula for $f_\perp(\widetilde{\boldsymbol{x}}) = f_\star(\widetilde{\boldsymbol{x}}) - f_\infty(\widetilde{\boldsymbol{x}}) = \langle \widetilde{\boldsymbol{x}}, \boldsymbol{\theta}_\perp\rangle$, we get

$$\begin{aligned}
T_2 &= \mathbb{E}[f_\perp(\widetilde{\boldsymbol{x}}_{(1)}) - f_\perp(\widetilde{\boldsymbol{x}})]^2, \\
&= \mathbb{E}\langle \boldsymbol{\theta}_\perp, \widetilde{\boldsymbol{x}}_{(1)} - \widetilde{\boldsymbol{x}}\rangle^2, \\
&\le (1-L)^2\|\boldsymbol{\theta}_\star\|^2\mathbb{E}\|\widetilde{\boldsymbol{x}} - \widetilde{\boldsymbol{x}}_{(1)}\|^2, \\
&= (1-L)^2\mathbb{E}\|\widetilde{\boldsymbol{x}} - \widetilde{\boldsymbol{x}}_{(1)}\|^2,
\end{aligned} \tag{40}$$

where in the last inequality, we used the relation that $\boldsymbol{\theta}_\perp = (1-L)\boldsymbol{\theta}_\star$. Proposition A.1 suggests that

$$\mathbb{E}\|\widetilde{\boldsymbol{x}} - \widetilde{\boldsymbol{x}}_{(1)}\|^2 \lesssim d^2\left[\frac{\log\left(n^{1/d}\right)}{n}\right]^{1/d}, \tag{41}$$

which implies

$$T_2 \lesssim d^2(1-L)^2\left[\frac{\log\left(n^{1/d}\right)}{n}\right]^{1/d}. \tag{42}$$

Remark B.1 (Comparison with pure nearest neighbor and ERM). If we rely solely on nearest neighbor method, the prediction error is

$$\mathbb{E}[f_\star(\widetilde{\boldsymbol{x}}) - f_\star(\widetilde{\boldsymbol{x}}_{(1)})]^2 = \mathbb{E}\langle \widetilde{\boldsymbol{x}} - \widetilde{\boldsymbol{x}}_{(1)}, \boldsymbol{\theta}_\star\rangle^2 \le \mathbb{E}\|\widetilde{\boldsymbol{x}} - \widetilde{\boldsymbol{x}}_{(1)}\|^2. \tag{43}$$

On the other hand, if we solely rely on ERM, even with infinite sample, we get

$$\mathbb{E}[f_\star(\widetilde{\boldsymbol{x}}) - f_\infty(\widetilde{\boldsymbol{x}})]^2 = \mathbb{E}\langle \widetilde{\boldsymbol{x}}, \boldsymbol{\theta}_\star - \boldsymbol{\theta}_\infty\rangle^2 \le (1-L)^2\mathbb{E}\|\widetilde{\boldsymbol{x}}\|^2. \tag{44}$$

We can see from the upper bound that ResMem takes advantage of both

- Projecting $f_\star$ onto f_∞, so that the dependence on the prediction function is reduced from 1 to $(1-L)^2$.
- Memorizing the residuals using nearest neighbor, so that the variance is reduced from $\mathbb{E}\|\widetilde{\boldsymbol{x}}\|^2$ to $\mathbb{E}\|\widetilde{\boldsymbol{x}}_{(1)} - \widetilde{\boldsymbol{x}}\|^2$.

B.4 Test loss for ResMem.

If we combine the previous two parts together, we get

$$\mathbb{E}\left[\hat{f}(\widetilde{\boldsymbol{x}}) - f_\star(\widetilde{\boldsymbol{x}})\right]^2 \lesssim d^2 L^2 n^{-2/3} + d^2(1-L)^2 \left[\frac{\log\left(n^{1/d}\right)}{n}\right]^{1/d}. \tag{45}$$

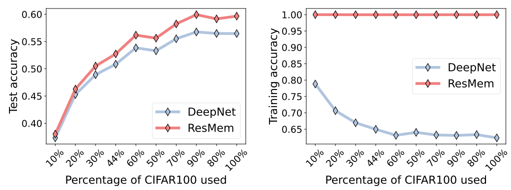
This completes the proof of Theorem 3.3.

C Additional CIFAR100 Results

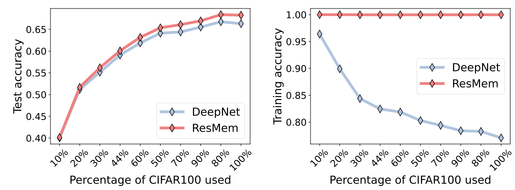
This section includes additional experiment results on applying ResMem to CIFAR100 dataset.

C.1 Additional robustness results

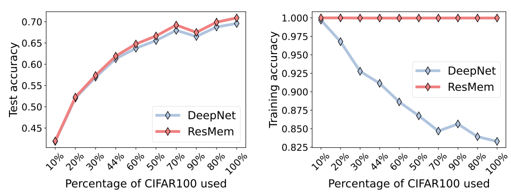
In addition to the results already presented in Section 4.2, we also evaluate ResMem performance for each architecture in CIFAR-ResNet{8, 14, 20, 32, 44, 56} and each subset (10%, 20%, ..., 100%) of CIFAR100 training data. We use the same training hyperparameter and the ResMem hyperparameter as described in Section 4.2. Generally, we see that ResMem yields larger improvement over the baseline DeepNet when the network is small and dataset is large.



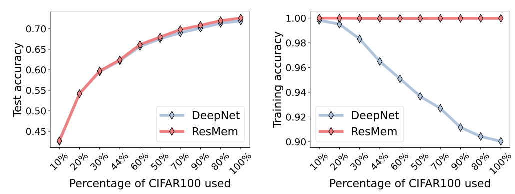
(a) CIFAR-ResNet-8



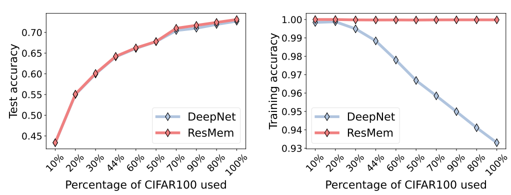
(b) CIFAR-ResNet-14



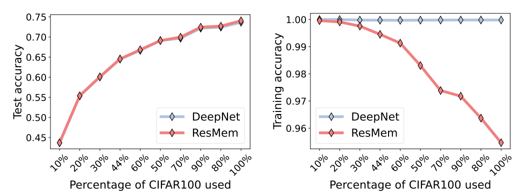
(c) CIFAR-ResNet-20



(d) CIFAR-ResNet-32



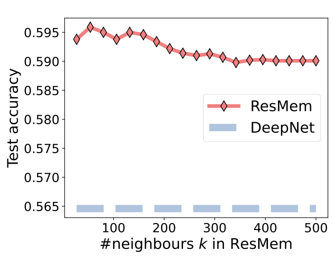
(e) CIFAR-ResNet-44



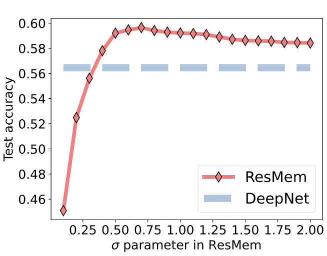
(f) CIFAR-ResNet-56

Figure 4: Test(left)/Training (right) accuracy for different sample sizes.

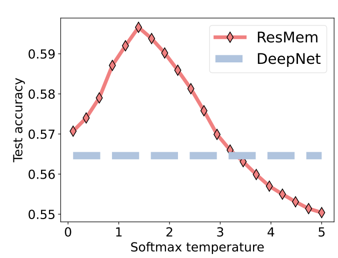
C.2 Sensitivity analysis for CIFAR100



(a) # of neighbours k.



(b) Radius parameter σ.



(c) Temperature T.

Figure 5: Sensitivity analysis of ResMem hyperparameters. The y-axis represents the CIFAR100 test accuracies, and the x-axis represents the sweeping of respective hyperparameters.

Varying locality parameter k and σ. We vary the number of neighbours from $k = 27$ to $k = 500$. We find that ResMem test accuracy is relatively stable across the choice of the number of neighbours (cf. Figure 5(a)). The trend of the curve suggests that as $k \to \infty$, the ResMem test accuracy seems to be converging to a constant level. For σ, we explored different values of $\sigma \in (0.1, 2.0)$. We observe that the test accuracy has a unimodal shape as a function of σ, suggesting that there is an optimal choice of σ (cf. Figure 5(b)).

Varying temperature T and connection to distillation. We tried $T = 0.1$ to $T = 5$, and also identified an unimodal shape for the test accuracy (Figure 5(c)). The fact that we can use different temperatures for (a) training the network and (b) constructing the k-NN predictor reminds us of the well-established knowledge distillation procedure [28]. In knowledge distillation, we first use one model (the teacher network) to generate targets at a higher temperature, and then train a second model (the student network) using the *combination* of the true labels and the output of the first network.

ResMem operates in a reversed direction: Here we have a second model (kNN) that learns the *difference* between true labels and the output of the first model. In both cases, we can tune the temperature of the first model to control how much information is retained. This connection offers an alternative perspective that regards ResMem as a "dual procedure" to knowledge distillation.

D ResMem on ImageNet

This section includes additional experiment results on applying ResMem to ImageNet dataset.

ImageNet. In addition to CIFAR100, we also evaluate the performance of ResMem on ImageNet [42]. We employ a family of pre-trained MobileNet-V2 models [44] from Keras[2], with varying widths controlled by a multiplier a. For ResMem, we again use the second last layer of DeepNet as a 1280-dimensional embedding of an image and rely on the ℓ_2 distance between the embeddings for nearest neighbor search (Step 3, Section 4.1). We specify the ResMem parameter of (k, σ, T) in the table below. We repeat the experiment over several MobileNet-V2 architectures, with MobileNet-V2-a0.35 being the smallest model and MobileNet-V2-a1.3 being the largest one.

Table 1: Test accuracy for ResMem and baseline deep network for ImageNet data.

Architecture	ResMem param.			Test accuracy	
	k	σ	T	DeepNet	ResMem
MobileNet-V2-a0.35	10	0.6	0.4	60.2%	**61.2%**
MobileNet-V2-a0.5	10	0.6	0.4	65.3%	**66.1%**
MobileNet-V2-a0.75	10	0.8	0.6	69.6%	**70.1%**
MobileNet-V2-a1.0	20	0.4	0.4	71.3%	**71.8%**
MobileNet-V2-a1.3	30	0.4	0.4	74.7%	**75.1%**

We can see that (c.f. Table 1) ResMem boosts the test accuracy by 1% on the smallest model and by 0.4% on the largest model.

E Additional details of NLP experiments

The Decoder-Only model used in our experiments is essentially the normal Encoder-Decoder architecture with Encoder and Cross-Attention removed. We pretrained both the T5-small and T5-base model on C4 [41] dataset with auto-regressive language modeling task for 1,000,000 steps, with dropout rate of 0.1 and batch size of 128. The learning rate for the first 10,000 steps is fixed to 0.01 and the rest steps follow a square root decay schedule.

During the inference for retrieval key, query embeddings and residuals, we ensured every token has at least 64 preceding context by adopting a sliding window strategy, where a window of 256 token slides from the beginning to the end on each of the articles, with a stride of $256 - 64 = 192$.

[2] `https://keras.io/api/applications/mobilenet/`

For residuals, we only stored the top 128 residuals measured by the absolute magnitude, as the residual vector is as large as T5 vocabulary size (i.e., 32128), and storing all 32128 residuals for each token is too demanding for storage. However, when weight-combining the residuals, we zero filled the missing residuals so that all the residual vectors have 32128 elements.