DS-GA 1008 HW 2

Long Chen
Center for Data Science, New York University
lc3424@nyu.edu

Fall 2021

Question 1

We have:

$$\|X\|^2 = \sum_{i=1}^{d} x_i^2.$$

Since $X \sim \mathcal{N}\left(0, \frac{I}{d}\right)$, x_i is $i.i.d$ for $\forall i = \{1, \dots, d\}$. Therefore,

$$\begin{aligned} \mu_{\|X\|^2} &= \mathbb{E}\left[\sum_{i=1}^{d} x_i^2\right] \\ &= \sum_{i=1}^{d} \mathbb{E}\left[x_i^2\right] \\ &= \sum_{i=1}^{d} \frac{1}{d} \\ &= 1 \end{aligned}$$

Consider also $\mathrm{Var}[x_i^2] = \mathbb{E}\ x_i^4 - (\mathbb{E}\ x_i^2)^2 = \frac{2}{d^2}$ [1] for $\forall i$. Thus,

$$\mathrm{Var}[\|X\|^2] = \sum_{i=1}^{d} \frac{2}{d^2} = \frac{2}{d}.$$

By Central Limit Theorem, for any distribution $\mathcal{D}$, $\sum_{i=1}^{n} x \xrightarrow{d} \mathcal{N}(\mu, \sigma^2)$ as $n \to \infty$ for $x \in \mathcal{D}$. Therefore,

$$\sup\|X\|^2 = 1 + \sup\left(\sqrt{\frac{2}{d}}\right).$$

[1] Could be easily proven with statistics.

That is:

$$\|X\|^2 = 1 + \mathcal{O}\left(\frac{1}{\sqrt{d}}\right).$$

Question 2

dim	mean	std
10	0.9868	0.4227
100	1.0005	0.1458
1000	0.9972	0.0459
10000	0.9998	0.0143

Table 1: Simulation results for Q2. n=1000.

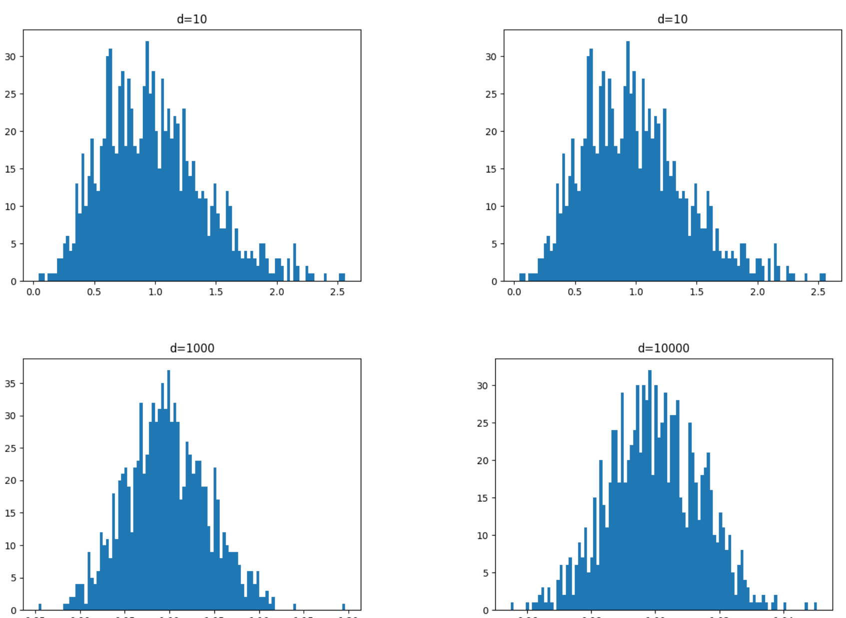

Figure 1: Distribution of 1,000 datapoints generated for each dimension.

The results shown do verify our finding in the previous question.

Question 3

Since $\tilde{X} = RX$, $\mu_{\tilde{X}} = R \times \mu_X = 0 = \mu_X$. Also that R is an unitary matrix, the covariance matrix of $\tilde{X}$ is:

$$\Sigma_{\tilde{X}} = RR^T\Sigma = \Sigma_X.$$

Since all linear combination of normal distributed variables are normally distributed, we could conclude that the pdf of X and $\tilde{X}$ are the same.

Question 4

We have:

$$\begin{aligned}
\mathbb{E}\langle X, X'\rangle &= \mathbb{E}\left[\sum_{i=1}^{d} X_i X_i'\right] \\
&= \sum_{i=1}^{d} \mathbb{E}\left[X_i X_i'\right] \\
&= \sum_{i=1}^{d} \mathbb{E}\left[X_i\right] \mathbb{E}\left[X_i'\right] \\
&= 0.
\end{aligned}$$

Also:

$$\begin{aligned}
\mathrm{Var}[\langle X, X'\rangle] &= \mathrm{Var}\left[\sum_{i=1}^{d} X_i X_i'\right] \\
&= \sum_{i=1}^{d} \mathrm{Var}[X_i X_i'] \\
&= \sum_{i=1}^{d} \mathrm{Var}[X_i]\mathrm{Var}[X_i'] + \mathrm{Cov}(X_i, X_i') \\
&= \sum_{i=1}^{d} \frac{1}{d}\frac{1}{d} \\
&= \sum_{i=1}^{d} \frac{1}{d^2} \\
&= \frac{1}{d}
\end{aligned}$$

Similarly as in Q1, by Central Limit Theorem:

$$|\langle X, X'\rangle| = \mathcal{O}\left(\frac{1}{\sqrt{d}}\right)$$

Consider further:

$$\begin{aligned}\|X - X'\|^2 &= (X - X')^T(X - X') \\ &= X^T X - 2\langle X, X'\rangle + X'^T X' \\ &= \|X\|^2 - 2\langle X, X'\rangle + \|X'\|^2\end{aligned}$$

Combining results from previous, we have:

$$\begin{aligned}\sup\|X - X'\|^2 &= 1 + \mathcal{O}\left(\frac{\sqrt{2}}{\sqrt{d}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{d}}\right) + 1 + \mathcal{O}\left(\frac{\sqrt{2}}{\sqrt{d}}\right) \\ &= 2 + (2\sqrt{2}+1)\mathcal{O}\left(\frac{1}{\sqrt{d}}\right)\end{aligned}$$

Therefore, $\sup\|X - X'\| = \sqrt{2} + \frac{\sqrt{2\sqrt{2}+1}}{\sqrt{d}}$. Similarly, $\inf\|X - X'\| = \sqrt{2} - \frac{\sqrt{2\sqrt{2}+1}}{\sqrt{d}}$. Thus we conclude that:

$$\|X - X'\| \in \left(\sqrt{2} \pm \frac{C}{\sqrt{d}}\right),$$

where $X = \sqrt{2\sqrt{2}+1}$.

Question 5

By definition:

$$\hat{f}_{NN}(x) := f^*(x_j),$$

where $j = \underset{i=1,\dots,n}{\arg\min}\|x - x_i\|$. Since f^* is β-*Lipschitz*:

$$\begin{aligned}|\hat{f}_{NN}(x) - f^*(x)| &= |f^*(x_j) - f^*(x)| \\ &\leq \beta\|x - x_j\| \\ &= \beta \min_i\|x - x_i\|\end{aligned}$$

Therefore (either by Central Limit Theorem or intuition):

$$\mathbb{E}|\hat{f}_{NN}(x) - f^*(x)| \leq \beta\ \mathbb{E}\min_i\|x - x_i\|.$$

Question 6

Consider the two distribution. Since both of them are gaussian distribution, there exists a bijection from the distribution of Y to the distribution of X. Let's assume X_i the bijection of Y_i. By nature of gaussian distribution, we could well define that:

$$X_i = \mu + \sigma Y_i,$$

for $\forall i$. We could verify this:

$$\begin{aligned}\mathbb{E}[X_i] &= \mathbb{E}[\mu + \sigma Y_i] \\ &= \mathbb{E}[\mu] + \sigma \mathbb{E}[Y_i] \\ &= \mathbb{E}[\mu] + \sigma \times 0 \\ &= \mu\end{aligned}$$

$$\begin{aligned}\mathrm{Var}[X_i] &= \mathrm{Var}[\mu + \sigma Y_i] \\ &= \mathrm{Var}[\mu] + \sigma^2 \mathrm{Var}[Y_i] \\ &= 0 + \sigma^2 \times 1 \\ &= \sigma^2\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}\min_i X_i &= \mathbb{E}[\mu + \sigma \min_i Y_i] \\ &= \mu + \sigma E_n.\end{aligned}$$

Question 7

Using the fact that $\|x - x_i\| \perp\!\!\!\perp \|x - x_i\| \mid x$, just like in the previous question, we can construct a bijection $Y_i \mapsto X_i$. Also that we assume asymptotic Gaussianity of $\|x - x_i\|$. Therefore similarly (the $\sim$ symbol is from the asymptotic Gaussianity):

$$\mathbb{E}\min_i \|x - x_i\| \sim \sqrt{2} + \frac{\sqrt{C}}{\sqrt{d}} E_n.$$

Question 8

Recall from previous question that:

$$
\begin{aligned}
\mathbb{E}|\hat{f}_{NN}(x) - f^*(x)| &\leq \beta \ \mathbb{E}\min_i \|x - x_i\| \\
&\sim \beta \left(\sqrt{2} + \frac{\sqrt{C}}{\sqrt{d}} E_n \right) \\
&\approx \beta \left(\sqrt{2} - \frac{\sqrt{C}}{\sqrt{d}} \sqrt{2 \log n} \right) \\
&= \beta \left(\sqrt{2} - \sqrt{2C} \frac{\sqrt{\log n}}{\sqrt{d}} \right) \\
&\approx \sqrt{2}\beta \\
&\approx \beta \mathbb{E}\|x - x_i\|
\end{aligned}
$$

since $\log n << d$. This constraint is natually from the β-*Lipschitz* property of f^*. Therefore, there is no further learning guarantee as of how much deviation of expected predicted value there is.

Question 9

Let $y = \arg\min_{y \in \partial\Omega} \|x - y\|$, $y' = \arg\min_{y' \in \partial\Omega} \|x' - y'\|$. That is, y and y' is the nearest projection of x and x', respectively, to the set Ω. We further have, using triangle inequalities:

$$
\|x - y'\|_\infty \leq \|x - x'\|_\infty + \|x' - y'\|_\infty
$$

$$
\|x' - y\|_\infty \leq \|x - x'\|_\infty + \|x - y\|_\infty
$$

Since y is the nearest projection of x onto Ω, $\|x - y'\|_\infty \geq \|x - y\|_\infty$. Similarly, $\|x' - y\|_\infty \geq \|x' - y'\|_\infty$.

If $\|x - y\|_\infty \geq \|x' - y'\|_\infty$,

$$
\begin{aligned}
|\Psi(x) - \Psi(x')| &= |\|x - y\|_\infty - \|x' - y'\|_\infty| \\
&= \|x - y\|_\infty - \|x' - y'\|_\infty \\
&\leq \|x - x'\|_\infty - (\|x - y'\|_\infty - \|x - y\|_\infty) \\
&\leq \|x - x'\|_\infty.
\end{aligned}
$$

If $\|x - y\|_\infty \leq \|x' - y'\|_\infty$,

$$
\begin{aligned}
|\Psi(x) - \Psi(x')| &= |\|x - y\|_\infty - \|x' - y'\|_\infty| \\
&= \|x' - y'\|_\infty - \|x - y\|_\infty \\
&\leq \|x - x'\|_\infty - (\|x' - y\|_\infty - \|x' - y'\|_\infty) \\
&\leq \|x - x'\|_\infty.
\end{aligned}
$$

Therefore, we conclude that:

$$|\Psi(x) - \Psi(x')| \leq \|x - x'\|_\infty,$$

which is equivalent to say that $\Psi(x)$ is 1-Lipschitz.

Question 10

Since Ω and $\mathcal{B}$ are separable respectively in the standard basis, we first consider for each dimension $i = 1, \ldots, d$. Let Ω_i be the projection of Ω onto ith dimension. With separability in standard basis, Ω_i is essentially $\left[-\frac{1}{2}, \frac{1}{2}\right]$. Easily can we construct $\Omega_i' = \Omega_i + \frac{1}{2} = [0,1]$ and $\Omega_i'' = \Omega_i - \frac{1}{2} = [-1,0]$. The two set is disjoint in ith dimension. Using again the fact that Ω is separable in standard basis, we will have resemble transformation that are orthogonal to transformed cubes in other dimension. Therefore we have 2^d disjoint copies of Ω constructed.

Since for each dimension $i = 1, \ldots, d$, the two set composes the full range of $\mathcal{B}_i$, we verify that:

$$\cup_{i=1}^d \left[\Omega_i' \cup \Omega_i''\right] = \mathcal{B}.$$

Therefore we conclude that we can fit 2^d copies of Ω into $\mathcal{B}$.

Question 11

Assume that for z_i $s.t$ $x - \frac{z_i}{2} \in \left[-\frac{1}{2}, \frac{1}{2}\right]^d$. If we consider such that $\Psi(x) = 0$ for any $x \notin \Omega$, for an arbitrary x, there will only be one $\Psi\left(x - \frac{z_i}{2}\right) > 0$. Assume that $\Psi\left(x - \frac{z_i}{2}\right) > 0$ and $\Psi\left(x' - \frac{z_i}{2}\right) > 0$. Let $y_j,\ j = 1, 2, \ldots, k$ be the intersection of the line segment joining x, x' with boundaries of Ω. Therefore we have:

$$\begin{aligned}
\|x - x'\| &= \|x - y_1 + y_1 - \ldots + y_k - x'\| \\
&\geq \|x - y_1\| + \|y_1 - y_2\| + \ldots + \|y_k - x'\| \\
&\geq \|x - y_1\| + \|y_k - x'\| \\
&\geq \|\Psi\left(x - \frac{z_1}{2}\right)\| + \|\Psi\left(x' - \frac{z_k}{2}\right)\|
\end{aligned}$$

For an arbitrary $k \in \mathbf{Z}$, if $z_1 = z_k$, with the fact that $\Psi(x)$ is 1-Lipschitz:

$$\begin{aligned}
|f^*(x) - f^*(x')| &= |\ (-1)^k \|\Psi\left(x - \frac{z_1}{2}\right)\| - (-1)^k \|\Psi\left(x' - \frac{z_k}{2}\right)\|\ | \\
&= |\ \|\Psi\left(x - \frac{z_1}{2}\right)\| - \|\Psi\left(x' - \frac{z_k}{2}\right)\|\ | \\
&\leq \|x - x'\|.
\end{aligned}$$

If $z_1 \neq z_k$,

$$
\begin{aligned}
|f^*(x) - f^*(x')| &= |\ (-1)^k \|\Psi\left(x - \frac{z_1}{2}\right)\| - (-1)^{k+1}\|\Psi\left(x' - \frac{z_k}{2}\right)\|\ | \\
&= |\ \|\Psi\left(x - \frac{z_1}{2}\right)\| + \|\Psi\left(x' - \frac{z_k}{2}\right)\|\ | \\
&\leq \|x - x'\|.
\end{aligned}
$$

Therefore, we conclude that f^* is 1-Lipschitz.

Question 12

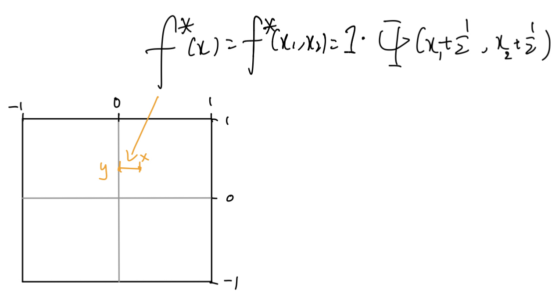

Figure 2: Graph for Q12. Since x is in the first quadrant, $g(z) = (-1, -1)$.

Question 13

For an arbitrary d, we have $n \leq 2^{d-1}$. Consider a point observed $x \in \Omega_i$ s.t. $\{x_i\} \cup \omega_i = \emptyset$. That is, the point x is in a tile without any training point observed. Consider the best that we can do is to return $\hat{f}(x) = 0^d$, that is the center of the tile. Since $x \sim Unif([-1,1]^d)$, this setting will return minimum value for $\mathbb{E}_x|f^*(x) - \hat{f}(x)|$. Intuitively, this also equals to $\mathbb{E}_x|f^*(x)|$, since that x is uniformly distributed. Therefore,

$$
\frac{\mathbb{E}_x|f^*(x) - \hat{f}(x)|}{\mathbb{E}_x|f^*(x)|} = 1,
$$

for x not in a tile with training samples in.

Consider $x \in \Omega_i$ $s.t.$ $\{x_i\} \cup \omega_i \neq \emptyset$, that is, in a tile with training sample(s).Assume that the learning algorithm returns a perfect prediction, i.e. $\mathbb{E}_x|f^*(x) - \hat{f}(x)| = 0$.

Consider that we have $n \leq d^{d-1}$. The probability that x is in a tile with training sample is $\frac{1}{2}$. Assume that x_i the datapoint in tiles with training sample and x_j the datapoint in tiles without. Thus,

$$
\begin{aligned}
\frac{\mathbb{E}_x|f^*(x) - \hat{f}(x)|}{\mathbb{E}_x|f^*(x)|} &= \frac{1}{2}\frac{\mathbb{E}_{x_i}|f^*(x_i) - \hat{f}(x_i)|}{\mathbb{E}_x|f^*(x_i)|} + \left(1 - \frac{1}{2}\right)\frac{\mathbb{E}_{x_j}|f^*(x_j) - \hat{f}(x_j)|}{\mathbb{E}_x|f^*(x_j)|} \\
&\geq \frac{1}{2} \times 0 + \frac{1}{2} \times 1 \\
&= \frac{1}{2},
\end{aligned}
$$

for $x \sim Unif([-1,1]^d)$. Therefore:

$$
\frac{\mathbb{E}_{x \sim Unif([-1,1]^d)}|f^*(x) - \hat{f}(x)|}{\mathbb{E}_{x \sim Unif([-1,1]^d)}|f^*(x)|} \geq \frac{1}{2}.
$$