MAP670L - Validation
Chosen article : "Uniform convergence may be unable to explain generalization in deep learning, Vaishnavh Nagarajan and J. Zico Kolter (2019)"

Group 5
David Admète, Benjamin Cohen

Exercise 1:

Title: Nearly vacuous uniform convergence bound can hide good generalization.

Summary: The goal of this exercise is to present an example of a linear classifier in which uniform convergence bound does not behave like generalization error. More precisely, we will show that the given classifier has a small generalization error, but a tightest algorithm-dependent uniform convergence bound close to 1.

Notations: Let $(\mathcal{X}, \|\cdot\|)$ a $K+D$ dimensional normed vector space, with $\|\cdot\|$ the Euclidean norm. Each element $x \in \mathcal{X}$ is noted (x_1, x_2) with $x_1 \in \mathbb{R}^K$ and $x_2 \in \mathbb{R}^D$. Let $m \in \mathbb{N}^*$ the size of the training sets. All the training sets considered are elements of $(\mathcal{X} \times \{-1,1\})^m$, with $\{-1,1\}$ the two classes of the problem. $u \in \mathbb{R}^K$ determines the center of the classes, and $\|u\|^2 = 1/m$. Let $\mathcal{D}$ a distribution on $\mathcal{X} \times \{-1,1\}$ such that if $(x,y) \sim D$, y is uniformly drawn from $\{-1,1\}$, $x_1 = 2y \cdot u$, and $x_2 \sim \mathcal{N}(0, \frac{16}{D} I_D)$ independently. For a set A, $\mathbb{1}(A)$ is the characteristic function of A.

We consider a linear classifier on $\mathcal{X}$ with weights $w = (w_1, w_2) \in \mathbb{R}^K \times \mathbb{R}^D$ initialized to origin. For $x \in \mathcal{X}$, the hypothesis h learned by the classifier satisfies $h(x) = w_1 x_1 + w_2 x_2$. Given a training set $S = \{(x^{(i)}, y^{(i)})\}_{i \in [\![1,m]\!]}$ of independent samples, the learning algorithm takes one gradient step of learning rate 1 to maximize $y \cdot h(x)$ for each $(x,y) \in S$. h_S denotes the hypothesis function learned on S. We will note $S_i = \{(x^{(j)}, y^{(j)})\}_{j \neq i}$. We denote $\mathcal{L} : \mathbb{R} \times \{-1,1\} \to \{0,1\}$ the $0-1$ loss, taking as argument (y', y) with y' the classifier's output and y the true label, and returning 0 if $yy' > 0$ else 1. The expected loss is $\mathcal{L}_{\mathcal{D}}(h) := \mathbb{E}_{(x,y)\sim D}[\mathcal{L}(h(x), y)]$. The empirical loss on the set S of size m is $\hat{\mathcal{L}}_S(h) := \frac{1}{m}\sum_{i=1}^m \mathcal{L}(h(x^{(i)}), y^{(i)})$. Let's set $\varepsilon > 0$ and $0 < \delta < 1/4$ two small quantities.

Recall that the generalization error is the smallest value $\varepsilon_{\text{gen}}(m,\delta)$ such that $\Pr_{S\sim\mathcal{D}^m}[\mathcal{L}_{\mathcal{D}}(h_S) - \hat{\mathcal{L}}_S(h_S) \leqslant \varepsilon_{\text{gen}}(m,\delta)] \geqslant 1-\delta$. The tightest algorithm-dependent uniform convergence bound, which will be called u-a bound, is the smallest value $\varepsilon_{\text{u-a}}$ such that there exists S_δ satisfying $\varepsilon_{\text{u-a}}(m,\delta) \geqslant \sup_{(S,S')\in S_\delta^2} \left|\mathcal{L}_{\mathcal{D}}(h_{S'}) - \hat{\mathcal{L}}_S(h_{S'})\right|$, with S_δ such that $\Pr_{S\sim\mathcal{D}^m}[S \in S_\delta] \geqslant 1-\delta$.

Accepted lemma A. *For $\alpha > 0$, $a = (a_1, ..., a_n) \in \mathbb{R}^n$ and $\gamma \sim \mathcal{N}(0, I_n)$:*

$$\exists\, C(\alpha) > 0, \quad \Pr_{\gamma\sim\mathcal{N}(0,I_n)}[|a\cdot\gamma| \geqslant C(\alpha)\|a\|] \leqslant \alpha.$$

Accepted lemma B. *For $D \geqslant \frac{8}{\delta}$ and $Z \sim \mathcal{N}(0, \frac{1}{D} I_D)$:*

$$\Pr_{Z\sim\mathcal{N}(0,\frac{1}{D}I_D)}\left[\frac{1}{2} \leqslant \|Z\|^2 \leqslant \frac{3}{2}\right] \geqslant 1-\delta.$$

Questions :

1. The aim of the first question is to prove that for D large enough, $\varepsilon_{\text{gen}}(m,\delta) \leqslant \varepsilon$.

   (a) Suppose that the training set $S = \{(x^{(i)}, y^{(i)})\}_{i\in[\![1,m]\!]}$. Justify that $w_1 = 2m \cdot u$ and $w_2 = \sum_{i=1}^{m} y^{(i)} x_2^{(i)}$. Deduce that for all $i \in [\![1,m]\!]$:

   $$y^{(i)} h_S(x^{(i)}) = 4 + \|x_2^{(i)}\|^2 + y^{(i)} x_2^{(i)} \sum_{j\neq i} y^{(j)} x_2^{(j)}.$$

   (b) Let note $\sigma_i = \sum_{j\neq i} y^{(j)} x_2^{(j)}$. Using the lemma A, show that for D large enough :

   $$\Pr_{S\sim\mathcal{D}^m}\left[\left|y^{(i)} x_2^{(i)} \sigma_i \sqrt{\frac{D}{16(m-1)}}\right| < \|x_2^{(i)}\| C(\delta/3m)\right] \geqslant 1 - \frac{\delta}{3m}.$$

   (c) Using question 1a and then question 1b, deduce that for D large enough,

   $$\Pr_{S\sim\mathcal{D}^m}\left[\hat{\mathcal{L}}_S(h_S) = 0\right] \geqslant 1 - \frac{2\delta}{3}.$$

   Hint : You can condition on the event studied in 1b to bound $\Pr\left[|y^{(i)} x_2^{(i)} \sigma_i| \geqslant 4\right]$.

   (d) Let $(z, \tilde{y}) \sim \mathcal{D}$ independent from S. Prove that there exists $C'(\varepsilon, m)$ such that for $D > C'(\varepsilon, m)$:

   $$\mathcal{L}(h_S(z), \tilde{y}) \leqslant \mathbb{1}\left(\left\{|z_2 \cdot w_2| \geqslant \|w_2\| C(\varepsilon)\sqrt{16/D}\right\}\right) + \mathbb{1}\left(\left\{\|w_2\|^2/16 \notin \left[\frac{m}{2}, \frac{3m}{2}\right]\right\}\right).$$

   Hint : With $\overline{\mathcal{A}}$ *and* $\overline{\mathcal{B}}$ *be the events inside the two indicator functions respectively, study* $\tilde{y} h_S(z)(\omega)$ *for* $\omega \in \mathcal{A} \cap \mathcal{B}$.

   (e) Show from all the previous questions that $\Pr_{S\sim\mathcal{D}^m}[\mathcal{L}_{\mathcal{D}}(h_S) - \hat{\mathcal{L}}_S(h_S) \leqslant \varepsilon] \geqslant 1 - \delta$ for D large enough. Deduce that $\varepsilon_{\text{gen}} \leqslant \varepsilon$.

   Hint : Study $\mathbb{E}_{(z,\tilde{y})\sim\mathcal{D}}[\mathcal{L}(h_S(z), \tilde{y})]$ *by applying the expectation on question 1d.*

2. The aim of the second question is to prove that for D large enough, $\varepsilon_{\text{u-a}}(m,\delta) \geqslant 1 - \varepsilon$. For a training set $S = \{(x^{(1)}, y^{(1)}), ..., (x^{(m)}, y^{(m)})\}$ with $x^{(i)} = (x_1^{(i)}, x_2^{(i)})$, we note $S_{\text{neg}} = \{(x_{\text{neg}}^{(1)}, y^{(1)}), ..., (x_{\text{neg}}^{(m)}, y^{(m)})\}$ with $x_{\text{neg}}^{(i)} = (x_1^{(i)}, -x_2^{(i)})$.

   (a) Using the same reasoning as in question 1c, prove that for D large enough :

   $$\Pr_{S\sim\mathcal{D}^m}\left[\hat{\mathcal{L}}_{S_{\text{neg}}}(h_S) = 1\right] \geqslant 1 - \frac{2\delta}{3}.$$

   (b) Deduce from the previous questions, prove that if S_δ satisfies $\Pr_{S\sim\mathcal{D}^m}[S \in S_\delta] \geqslant 1 - \delta$, we have for D large enough :

   $$\Pr_{S\sim\mathcal{D}^m}\left[\{S \in \mathcal{S}_\delta\} \cap \{S_{\text{neg}} \in \mathcal{S}_\delta\} \cap \{\mathcal{L}_{\mathcal{D}}(h_S) \leqslant \varepsilon\} \cap \{\hat{\mathcal{L}}_{S_{\text{neg}}}(h_S) = 1\}\right] > 1.$$

   Conclude on question 2 and the exercise.

Solution 1:

1. (a) As the weights are initialized to 0 and the learning rate is equal to 1, applying the learning algorithm is equivalent to compute the gradient of $w_j \longmapsto \sum_{i=1}^m y^{(i)}(w_1 x_1^{(i)} + w_2 x_2^{(i)})$ for $j \in \{1,2\}$. As $\sum_{i=1}^m y^{(i)} x_1^{(i)} = 2m \cdot u$, we have $w_1 = 2m \cdot u$. The gradient calculation is immediate for $w_2 = \sum_{i=1}^m y^{(i)} x_2^{(i)}$. Let $i \in [\![1, m]\!]$.

$$\begin{aligned} y^{(i)} h_S(x^{(i)}) &= y^{(i)}(w_1 x_1^{(i)} + w_2 x_2^{(i)}) \\ &= y^{(i)} 2mu \cdot 2y^{(i)} u + y^{(i)} \left(\sum_{j=1}^m y^{(j)} x_2^{(j)} \right) \cdot x_2^{(i)} \\ &= 4m \underbrace{\|u\|^2}_{1/m} + \|x_2\|^2 + y^{(i)} x_2^{(i)} \sum_{j \neq i} y^{(j)} x_2^{(j)}. \end{aligned}$$

(b) Let us note $\tilde{\sigma}_i = \sqrt{D/(16(m-1))}\sigma_i$. Notice that $\tilde{\sigma}_i \sim \mathcal{N}(0, I_n)$, and that σ_i and $x_2^{(i)}$ are independent. First, if we fix $x_2^{(i)}$, thanks to the lemma A, we have :

$$\Pr_{S_i \sim \mathcal{D}^{m-1}} \left[\overbrace{\left| y^{(i)} x_2^{(i)} \tilde{\sigma}_i \right| < \|x_2^{(i)}\| C(\delta/3m)}^{E_i} \right] \geqslant 1 - \frac{\delta}{3m}.$$

We also have $\Pr_S[E_i] = \mathbb{E}_{x_2^{(i)}}[\mathbb{E}_S[\mathbb{1}(E_i) | x_2^{(i)}]] = \mathbb{E}_{x_2^{(i)}}[\Pr_{S_i}[E_i]]$. As $\Pr_{S_i}[E_i] \geqslant 1 - \frac{\delta}{3m}$, we obtain $\Pr_S[E_i] \geqslant 1 - \frac{\delta}{3m}$.

(c) Combining results of questions 1a and questions 1b, we obtain :

$$\begin{aligned} \Pr_{S \sim \mathcal{D}^m} \left[\hat{\mathcal{L}}_S(h_S) = 0 \right] &\geqslant 1 - \sum_{i=1}^m \Pr_{S \sim \mathcal{D}^m} \left[y^{(i)} h(x^{(i)}) \leqslant 0 \right] \\ &= 1 - \sum_{i=1}^m \Pr_{S \sim \mathcal{D}^m} \left[4 + \|x_2^{(i)}\|^2 + y^{(i)} x_2^{(i)} \sigma_i \leqslant 0 \right] \\ &\geqslant 1 - \sum_{i=1}^m \Pr_{S \sim \mathcal{D}^m} \left[\left| y^{(i)} x_2^{(i)} \sigma_i \right| \geqslant 4 \right]. \end{aligned}$$

By taking $D > 16mC(\delta/(3m))^2$, and conditioning on event E_i :

$$\begin{aligned} \Pr_{S \sim \mathcal{D}^m} \left[\left| y^{(i)} x_2^{(i)} \sigma_i \right| \geqslant 4 \right] &= \Pr_{S \sim \mathcal{D}^m} \left[\left| y^{(i)} x_2^{(i)} \sigma_i \right| \geqslant 4 \;\middle\|\; E_i \right] \Pr_{S \sim \mathcal{D}^m} [E_i] \\ &\quad + \Pr_{S \sim \mathcal{D}^m} \left[\left| y^{(i)} x_2^{(i)} \sigma_i \right| \geqslant 4 \;\middle\|\; \overline{E}_i \right] \Pr_{S \sim \mathcal{D}^m} [\overline{E}_i] \\ &\leqslant \Pr_{S \sim \mathcal{D}^m} \left[\|x_2^{(i)}\| \underbrace{\sqrt{\frac{16m}{D}} C(\delta/(3m)) \geqslant 4}_{\leqslant 1} \right] + \Pr_{S \sim \mathcal{D}^m} [\overline{E}_i]. \end{aligned}$$

As $\Pr[\|x_2^{(i)}\| \geqslant 4] \leqslant \frac{\delta}{3m}$, and $\Pr[\overline{E}_i] \leqslant \frac{\delta}{3m}$, we obtain the desired result.

(d) Let $\overline{\mathcal{A}}$ and $\overline{\mathcal{B}}$ be the events inside the two indicator functions respectively, and $D > C'(\varepsilon, m) = 384mC(\varepsilon)^2$. For $\omega \in \overline{\mathcal{A}} \cap \overline{\mathcal{B}}$, omitting the ω dependency for clarity purposes, we have :

$$\begin{aligned} \tilde{y} h_S(z) &= 4 + \tilde{y} z_2 \cdot w_2 \\ &\geqslant 4 - |z_2 \cdot w_2| \\ \text{as } \omega \in \overline{\mathcal{A}} \qquad &\geqslant 4 - \|w_2\| C(\varepsilon) \sqrt{16/D} \\ \text{as } \omega \in \overline{\mathcal{B}} \qquad &\geqslant 4 - C(\varepsilon) \sqrt{384m/D} \geqslant 3 > 0. \end{aligned}$$

Then, $\mathcal{L}(h_S(z), \tilde{y})(\omega) = 0$. So as we can write $\Omega = \{\mathcal{A} \cap \mathcal{B}\} \sqcup \{\overline{\mathcal{A}} \cap \mathcal{B}\} \sqcup \overline{\mathcal{B}}$, we have $\mathcal{L}(h_S(z), \tilde{y}) \leqslant 1 \times \mathbb{1}(\overline{\mathcal{A}} \cap \mathcal{B}) + 0 \times \mathbb{1}(\mathcal{A} \cap \mathcal{B}) + 1 \times \mathbb{1}(\overline{\mathcal{B}})$. Thus, we finally obtain $\mathcal{L}(h_S(z), \tilde{y}) \leqslant \mathbb{1}(\overline{\mathcal{A}}) + \mathbb{1}(\overline{\mathcal{B}})$.

(e) Let take D satisfying all the previous constraints, which is possible because it is a finite number of lower bounds on D.

$$\begin{aligned}\Pr_{S\sim\mathcal{D}^m}[\mathcal{L}_{\mathcal{D}}(h_S)-\hat{\mathcal{L}}_S(h_S)\leqslant\varepsilon] &\geqslant \Pr_{S\sim\mathcal{D}^m}[\{\mathcal{L}_{\mathcal{D}}(h_S)\leqslant\varepsilon\}\cap\{\hat{\mathcal{L}}_S(h_S)=0\}] \\ &= 1-\Pr_{S\sim\mathcal{D}^m}[\{\mathcal{L}_{\mathcal{D}}(h_S)>\varepsilon\}\cup\{\hat{\mathcal{L}}_S(h_S)>0\}] \\ &\geqslant 1-\Pr_{S\sim\mathcal{D}^m}[\mathcal{L}_{\mathcal{D}}(h_S)>\varepsilon]-\Pr_{S\sim\mathcal{D}^m}[\hat{\mathcal{L}}_S(h_S)>0].\end{aligned}$$

By question 1c, $\Pr_{S\sim\mathcal{D}^m}[\hat{\mathcal{L}}_S(h_S)>0]\leqslant 1-\frac{2}{3}\delta$. Using the previous question we have $\mathcal{L}(h_S(z),\tilde{y})\leqslant \mathbb{1}(\overline{\mathcal{A}})+\mathbb{1}(\overline{\mathcal{B}})$. Remark that from the one hand $\mathcal{A}$ is a function of z and S, whereas on the other hand $\mathcal{B}$ is only a function of S. As by the lemma A, $\Pr_{(z,\tilde{y})\sim\mathcal{D}}(\overline{\mathcal{A}})\leqslant\varepsilon$, by applying the expectation on the result question 1d, we have :

$$\mathop{\mathbb{E}}_{(z,\tilde{y})\sim\mathcal{D}}[\mathcal{L}(h_S(z),\tilde{y})]\leqslant \Pr_{(z,\tilde{y})\sim\mathcal{D}}(\overline{\mathcal{A}})+\mathbb{1}(\overline{\mathcal{B}})\leqslant\varepsilon+\mathbb{1}(\overline{\mathcal{B}}).$$

Then $\{\mathbb{E}_{(z,\tilde{y})\sim\mathcal{D}}[\mathcal{L}(h_S(z),\tilde{y})]>\varepsilon\}\cap\mathcal{B}=\varnothing$. So $\Pr_{S\sim\mathcal{D}^m}[\mathbb{E}_{(z,\tilde{y})\sim\mathcal{D}}[\mathcal{L}(h_S(z),\tilde{y})]>\varepsilon]\leqslant \Pr_{S\sim\mathcal{D}^m}[\overline{\mathcal{B}}]\leqslant\frac{\delta}{3}$ by the lemma B. Finally :

$$\Pr_{S\sim\mathcal{D}^m}[\mathcal{L}_{\mathcal{D}}(h_S)-\hat{\mathcal{L}}_S(h_S)\leqslant\varepsilon]\geqslant 1-\frac{2}{3}\delta-\frac{1}{3}\delta=1-\delta.$$

Then by definition, $\varepsilon_{\text{gen}}(m,\delta)\leqslant\varepsilon$.

2. (a) Let $i\in[\![1,m]\!]$. By slightly adapting the beginning of question 1c, we have :

$$\Pr_{S\sim\mathcal{D}^m}\left[\hat{\mathcal{L}}_{S_{\text{neg}}}(h_S)=1\right]\geqslant 1-\sum_{i=1}^{m}\Pr_{S\sim\mathcal{D}^m}\left[4-\|x_2^{(i)}\|^2-y^{(i)}x_2^{(i)}\sigma_i>0\right].$$

Conditioning on the event $\{\|x_2^{(i)}\|^2\geqslant 8\}$ and choosing $D>32m\ln(6m/\delta)$, we have :

$$\begin{aligned}&\Pr_{S\sim\mathcal{D}^m}\left[4-\|x_2^{(i)}\|^2-y^{(i)}x_2^{(i)}\sigma_i>0\right]\\ &=\Pr_{S\sim\mathcal{D}^m}\left[4-\|x_2^{(i)}\|^2-y^{(i)}x_2^{(i)}\sigma_i>0\,\middle\|\,\|x_2^{(i)}\|^2\geqslant 8\right]\Pr_{S\sim\mathcal{D}^m}\left[\|x_2^{(i)}\|^2\geqslant 8\right]\\ &\quad+\Pr_{S\sim\mathcal{D}^m}\left[4-\|x_2^{(i)}\|^2-y^{(i)}x_2^{(i)}\sigma_i>0\,\middle\|\,\|x_2^{(i)}\|^2< 8\right]\Pr_{S\sim\mathcal{D}^m}\left[\|x_2^{(i)}\|^2< 8\right]\\ &\leqslant\Pr_{S\sim\mathcal{D}^m}\left[\left|y^{(i)}x_2^{(i)}\sigma_i\right|>4\right]+\Pr_{S\sim\mathcal{D}^m}\left[\|x_2^{(i)}\|^2<8\right]\leqslant\frac{2\delta}{3m}.\end{aligned}$$

Then we finally find $\Pr_{S\sim\mathcal{D}^m}\left[\hat{\mathcal{L}}_{S_{\text{neg}}}(h_S)=1\right]\geqslant 1-\sum_{i=1}^m\frac{2\delta}{3m}=1-\frac{2\delta}{3}$.

(b) Let S_δ such that $\Pr_{S\sim\mathcal{D}^m}[S\in S_\delta]\geqslant 1-\delta$:

$$\begin{aligned}&\Pr_{S\sim\mathcal{D}^m}\left[\{S\in\mathcal{S}_\delta\}\cap\{S_{\text{neg}}\in\mathcal{S}_\delta\}\cap\{\mathcal{L}_{\mathcal{D}}(h_S)\leqslant\varepsilon\}\cap\{\hat{\mathcal{L}}_{S_{\text{neg}}}(h_S)=1\}\right]\\ &\geqslant 1-\Pr_{S\sim\mathcal{D}^m}[S\notin\mathcal{S}_\delta]-\Pr_{S\sim\mathcal{D}^m}[S_{\text{neg}}\notin\mathcal{S}_\delta]\\ &\qquad-\Pr_{S\sim\mathcal{D}^m}[\mathcal{L}_{\mathcal{D}}(h_S)>\varepsilon]-\Pr_{S\sim\mathcal{D}^m}\left[\hat{\mathcal{L}}_{S_{\text{neg}}}(h_S)\neq 1\right].\end{aligned}$$

The first two probabilities are less than δ by assumption on S_δ, as S_{neg} follows the same distribution than S. The last ones are also less than δ as we respectively showed in questions 1e and 2a. Then the left hand term is greater than $1-4\delta>0$ as $\delta<1/4$. Then, for a given S_δ, there exists S^* such that all conditions in the left hand term are satisfied. Finally, by definition of the u-a bound, $\varepsilon_{\text{u-a}}(m,\delta)\geqslant|\mathcal{L}_{\mathcal{D}}(h_{S^*})-\hat{\mathcal{L}}_{S^*_{\text{neg}}}(h_{S^*})|\geqslant 1-\varepsilon$. In this simple case, we have thus shown that by varying the setup parameters, we can obtain a uniform convergence bound arbitrarily close to 1 (so nearly vacuous), simultaneously with an arbitrarily small generalisation bound. This raises the important point highlighted in the paper: even the tightest uniform convergence does not behave like the generalization error.

Appendix :

Proof of lemma A. First, let us show the Chernoff bound, i.e. that for $z \sim \mathcal{N}(0, \sigma^2)$, we have $\Pr_{z\sim\mathcal{N}(0,\sigma^2)}[|z| > \alpha] \leqslant 2\exp(-\alpha^2/(2\sigma^2))$. By Markov inequality, we have :

$$\Pr_{z\sim\mathcal{N}(0,\sigma^2)}[z > \alpha] = \Pr_{z\sim\mathcal{N}(0,\sigma^2)}[e^{sz} > e^{s\alpha}] \leqslant e^{-s\alpha} \mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,\sigma^2)}[e^{sz}] \leqslant \exp\left[\frac{\sigma^2 s^2}{2} - s\alpha\right].$$

Minimizing the right hand term according to s, and applying the same process on $-z \sim \mathcal{N}(0, \sigma^2)$, we obtain the Chernoff bound.
Then, for $z_1, ..., z_n$ independent with $z_i \sim \mathcal{N}(0, \sigma_i^2)$, we obtain by independence :

$$\Pr_{z_i\sim\mathcal{N}(0,\sigma_i^2)}\left[\left|\sum_{i=1}^n z_i\right| \geqslant \alpha\right] \leqslant 2\exp\left(-\frac{\alpha^2}{2\sum_{i=1}^n \sigma_i^2}\right).$$

Finally, applying this inequality with $z = a \cdot \gamma$, with $\gamma \sim \mathcal{N}(0, I_n)$, we obtain :

$$\Pr_{\gamma\sim\mathcal{N}(0,I_n)}[|\gamma \cdot a| \geqslant \alpha] \leqslant 2\exp\left(-\frac{\alpha^2}{2\|a\|^2}\right).$$

Thus, we find the expected result with $C(\alpha) = \sqrt{2\ln\frac{2}{\alpha}}$.

Proof of lemma B. Let $Z \sim \mathcal{N}(0, \frac{1}{D} I_D)$, we have :

$$\begin{aligned}
\Pr_{Z\sim\mathcal{N}(0,\frac{1}{D}I_D)}\left[\|Z\|^2 \notin \left[\frac{1}{2}, \frac{3}{2}\right]\right] &= \Pr_{Z\sim\mathcal{N}(0,\frac{1}{D}I_D)}\left[\left|\|Z\|^2 - 1\right| > \frac{1}{2}\right] \\
&= \Pr_{Z\sim\mathcal{N}(0,\frac{1}{D}I_D)}\left[\frac{1}{D}\left|\underbrace{\left\|Z\sqrt{D}\right\|^2}_{Y\sim\chi^2(D)} - \underbrace{D}_{\mathbb{E}[Y]}\right| > \frac{1}{2}\right] \\
&\overset{\text{B.T.}}{\leqslant} \frac{\mathbb{V}(Y)/D^2}{(1/2)^2} = \frac{8}{D} \leqslant \delta \quad \text{for} \quad D \geqslant \frac{8}{\delta}.
\end{aligned}$$