Machine Learning Theory **Spring 2021**

Lecture 6: PAC-Bayesian Theory

Lecturer: Liwei Wang *Scribe: Binghui Li, Shiji Xin, Fang Sun, Qizhe Zhang, Mi Yan*

6.1 Review

VC theory focuses on the uniform convergence for all classifiers in hypothesis space $\mathcal{F}$. Specifically, with probability $1-\delta$,

$$P_D \leqslant P_S + O\left(\sqrt{\frac{d\ln n + \ln 1/\delta}{n}}\right) \tag{6.1}$$

Where P_D is the test error w.r.t. $f \in \mathcal{F}$, P_S is the training error w.r.t. $f \in \mathcal{F}$, d is the VC-dim of $\mathcal{F}$.
Notice that the result is independent of the distribution D and the learning algorithm A, and only depends on the hypothesis space $\mathcal{F}$ and sample size n.
However, VC Theory fails to account for the great generalization ability of neural networks. For a neural network, the number of parameters is usually far greater than that of training data. In such a case, the generalization bound yielded by VC Theory is scarcely meaningful. Perhaps randomness plays a vital role behind the huge success of neural networks. After all, stochastic gradient descent (SGD), rather than gradient descent (GD), is the common method for optimizing neural networks. [1–3]

6.2 Comparison of Frequentist and Bayesian

There are two points of view of statistical inference: Frequentist and Bayesian. Here is a brief comparison:

	Frequentist	Bayesian
Views of Probability	Law of Large Numbers	Degree of Belief
Parameters Estimation	Maximum Likelihood Estimate	Maximum a posteriori estimation $P(\theta\|x) = \frac{P(x\|\theta)P(\theta)}{P(x)}$
Output of Learning	a classifier f	a distribution of classifiers $\mathcal{Q}$
Prior	Hypothesis space $\mathcal{F}$	distribution of classifiers $\mathcal{P}$
Performance	$\mathrm{Err}_D(f)$	$\mathbb{E}_{f\sim Q}[\mathrm{Err}_D(f)]$
Generalization	VC Theory uniform convergence for all classifiers in a hypothesis space	PAC-Bayesian Theory for all distributions of classifiers
Gap between $\mathrm{Err}_D(f)$ and $\mathrm{Err}_S(f)$	$O\left(\sqrt{\frac{d\ln n + \ln(1/\delta)}{n}}\right)$	$O\left(\sqrt{\frac{D(\mathcal{Q}\|\mathcal{P}) + \ln(3/\delta)}{n}}\right)$

In the Frequentist view, the gap between test error and training error only depends on hypothesis space $\mathcal{F}$ and sample size, i.e. w/ prob $1-\delta$, $\forall h \in \mathcal{F}$,

$$\mathrm{Err}_D(h) \leqslant \mathrm{Err}_S(h) + O\left(\sqrt{\frac{d \ln n + \ln(1/\delta)}{n}}\right)$$

Where $d = \mathrm{VCD}(\mathcal{F})$.

6.3 The main theorem of PAC-Bayesian Theory

In this section, we are going to present the main theorem in PAC-Bayesian Theory.

Theorem 6.1 (PAC-Bayesian) *For any fixed prior distribution of classifiers $\mathcal{P}$, with probability $1-\delta$ over the random draw of training dataset S of size n, the following inequality holds uniformly for all distributions (i.e. stochastic classifier) $\mathcal{Q}$:*

$$\mathbb{E}_{h\sim\mathcal{Q}}[\mathrm{Err}_D(h)] \leqslant \mathbb{E}_{h\sim\mathcal{Q}}[\mathrm{Err}_S(h)] + \sqrt{\frac{D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P}) + \ln(3/\delta)}{n}} \tag{6.2}$$

Here we denote $\mathbb{P}_{(x,y)\sim D}[y \neq h(x)]$ by $\mathrm{Err}_D(h)$ and $\mathbb{P}_{(x,y)\sim S}[y \neq h(x)]$ by $\mathrm{Err}_S(h)$, $D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P}) \coloneqq \mathbb{E}_{h\sim Q}\left[\ln \frac{q(x)}{p(x)}\right]$ is the KL-divergence between $\mathcal{Q}$ and $\mathcal{P}$.

Intuitively, since $\mathcal{P}$ is independent of the dataset $S \sim D^n$, we can bound the gap between $\mathbb{E}_{h'\sim\mathcal{P}}[\mathrm{Err}_D(h')]$ and $\mathbb{E}_{h'\sim\mathcal{P}}[\mathrm{Err}_S(h')]$ by Chernoff bound, so we only need to bound the gap between $\mathbb{E}_{h\sim Q}[\mathrm{Err}_D(h)]$ and $\mathbb{E}_{h'\sim P}[\mathrm{Err}_D(h')]$. Before proving the theorem, we list some useful lemmas here.

Lemma 6.2 (Change of Measure) *For all distribution $\mathcal{P}, \mathcal{Q}$ over hypothesis space $\mathcal{F}$ and all function $f : \mathcal{F} \to \mathbb{R}$, we have*

$$\mathbb{E}_{h\sim\mathcal{Q}}[f(h)] \leqslant \ln \mathbb{E}_{h'\sim\mathcal{P}}\left[e^{f(h')}\right] + D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P}) \tag{6.3}$$

Proof: In fact,

$$\begin{aligned}
\mathrm{RHS} - \mathrm{LHS} &= \ln \mathbb{E}_{h'\sim\mathcal{P}}\left[e^{f(h')}\right] + D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P}) - \mathbb{E}_{h\sim\mathcal{Q}}[f(h)] \\
&= \mathbb{E}_{h\sim\mathcal{Q}}\left[\ln \frac{q(h)}{p(h)}\right] - \mathbb{E}_{h\sim\mathcal{Q}}[f(h)] + \ln \mathbb{E}_{h'\sim\mathcal{P}}\left[e^{f(h')}\right] \\
&= \mathbb{E}_{h\sim\mathcal{Q}}\left[\ln \frac{q(h)}{\frac{p(h)e^{f(h)}}{\mathbb{E}_{h'\sim\mathcal{P}}e^{f(h')}}}\right] \\
&= D_{\mathrm{KL}}\left(\mathcal{Q}\|\frac{p(h)e^{f(h)}}{\mathbb{E}_{h'\sim\mathcal{P}}e^{f(h')}}\right) \geqslant 0
\end{aligned}$$

■

Lemma 6.3 *For any $\delta > 0$,*

$$\mathbb{P}_{S\sim D^n}\left[\mathbb{E}_{h\sim\mathcal{P}}[e^{n(\mathrm{Err}_D(h)-\mathrm{Err}_S(h))^2}] \geqslant 3/\delta\right] \leqslant \delta \tag{6.4}$$

Proof: We begin our proof by considering a bound for a fixed $h \sim \mathcal{P}$.

$$\mathbb{E}_{S\sim D^n}\left[e^{n(\mathrm{Err}_D(h)-\mathrm{Err}_S(h))^2}\right] \leqslant 3$$

For simplicity, let $\Delta \coloneqq |\mathrm{Err}_D(h) - \mathrm{Err}_S(h)|$. Using Chernoff bound, we have

$$\mathbb{P}_{S\sim D^n}[\Delta \geqslant \varepsilon] \leqslant 2\exp(-2n\varepsilon^2)$$

Then,

$$\begin{aligned}
\mathbb{E}_{S\sim D^n}[e^{n\Delta^2}] &= \int_0^\infty \mathbb{P}_{S\sim D^n}\left[e^{n\Delta^2} \geqslant t\right] \mathrm{d}t \\
&= \int_1^\infty \mathbb{P}_{S\sim D^n}\left[e^{n\Delta^2} \geqslant t\right] \mathrm{d}t + \int_0^1 \mathbb{P}_{S\sim D^n}\left[e^{n\Delta^2} \geqslant t\right] \mathrm{d}t \\
&= \int_1^\infty \mathbb{P}_{S\sim D^n}\left[\Delta \geqslant \sqrt{\frac{\ln t}{n}}\right] \mathrm{d}t + 1 \\
&\leqslant \int_1^\infty 2e^{-2\ln t}\mathrm{d}t + 1 \\
&= 3
\end{aligned}$$

By applying Markov's Inequality, we get

$$\begin{aligned}
\mathbb{P}_{S\sim D^n}\left[\mathbb{E}_{h\sim\mathcal{P}}\left[e^{n(\mathrm{Err}_D(h)-\mathrm{Err}_S(h))^2}\right] \geqslant 3/\delta\right] &= \mathbb{P}_{S\sim D^n}\left[\mathbb{E}_{h\sim\mathcal{P}}\left[e^{n\Delta^2}\right] \geqslant 3/\delta\right] \\
&\leqslant \frac{\mathbb{E}_{S\sim D^n}\left[\mathbb{E}_{h\sim P}\left[e^{n\Delta^2}\right]\right]}{3/\delta} \\
&= \frac{\mathbb{E}_{h\sim P}\left[\mathbb{E}_{S\sim D^n}\left[e^{n\Delta^2}\right]\right]}{3/\delta} \quad \text{(Fubini Theorem)} \\
&\leqslant \frac{\mathbb{E}_{h\sim P}[3]}{3/\delta} = \delta
\end{aligned}$$

■

Now we can prove PAC-Bayesian Theorem (6.1) by applying Jensen's Inequality.
With probability $1-\delta$,

$$\begin{aligned}
(\mathbb{E}_{h\sim\mathcal{Q}}[\mathrm{Err}_D(h) - \mathrm{Err}_S(h)])^2 &\leqslant \frac{1}{n}\mathbb{E}_{h\sim\mathcal{Q}}\, n[\mathrm{Err}_D(h) - \mathrm{Err}_S(h)]^2 \\
&\leqslant \frac{1}{n}\left(\ln \mathbb{E}_{h'\sim\mathcal{P}}\left[e^{n\Delta^2}\right] + D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P})\right) \\
&\leqslant \frac{1}{n}\left(\ln(3/\delta) + D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P})\right)
\end{aligned}$$

Thus,

$$\mathbb{E}_{h\sim\mathcal{Q}}[\mathrm{Err}_D(h)] \leqslant \mathbb{E}_{h\sim\mathcal{Q}}[\mathrm{Err}_S(h)] + \sqrt{\frac{D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P}) + \ln(3/\delta)}{n}}$$

Lemma 6.4 *For any fixed $h \sim \mathcal{P}$, we have*

$$\mathbb{E}_{S\sim D^n}\left[e^{nD_{\mathrm{B}}(\mathrm{Err}_S(h)\|\,\mathrm{Err}_D(h))}\right] \leqslant n+1 \tag{6.5}$$

Proof: For simplicity, we denote $\mathrm{Err}_D(h)$ by p, $\mathrm{Err}_S(h)$ by $\hat{p}_S$. Since $\mathcal{P}$ is the prior, h is independent of S. Therefore,

$$\begin{aligned}\mathbb{E}_{S\sim D^n}\left[e^{nD_{\mathrm{B}}(\mathrm{Err}_S(h)\|\,\mathrm{Err}_D(h))}\right] &= \mathbb{E}_{S\sim D^n}\left[e^{nD_{\mathrm{B}}(\hat{p}_S\|p)}\right]\\ &= \mathbb{E}_{S\sim D^n}\left[\left(\frac{\hat{p}_S}{p}\right)^{n\hat{p}_S}\left(\frac{1-\hat{p}_S}{1-p}\right)^{n(1-\hat{p}_S)}\right]\end{aligned}$$

Note that $S\sim D^n$ implies $\sum_{i=1}^n I[y_i\neq h(x_i)]\sim B(n,p)$, so the equation above can be rewritten as:

$$\begin{aligned}\mathbb{E}_{S\sim D^n}\left[\left(\frac{\hat{p}_S}{p}\right)^{n\hat{p}_S}\left(\frac{1-\hat{p}_S}{1-p}\right)^{n(1-\hat{p}_S)}\right] &= \sum_{k=0}^n\binom{n}{k}p^k(1-p)^{n-k}\left(\frac{k/n}{p}\right)^{n(k/n)}\left(\frac{1-k/n}{1-p}\right)^{n(1-k/n)}\\ &= \sum_{k=0}^n\binom{n}{k}\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k}\\ &\leqslant \sum_{k=0}^n\left(\binom{n}{0}\left(\frac{k}{n}\right)^0\left(1-\frac{k}{n}\right)^{n-0}+\binom{n}{1}\left(\frac{k}{n}\right)^1\left(1-\frac{k}{n}\right)^{n-1}+\cdots\right.\\ &\quad\left.+\binom{n}{k}\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k}+\cdots+\binom{n}{n}\left(\frac{k}{n}\right)^n\left(1-\frac{k}{n}\right)^{n-n}\right)\\ &= \sum_{k=0}^n\sum_{r=0}^n\binom{n}{r}\left(\frac{k}{n}\right)^r\left(1-\frac{k}{n}\right)^{n-r}\\ &= \sum_{k=0}^n\left(\frac{k}{n}+\left(1-\frac{k}{n}\right)\right)^n\\ &= n+1\end{aligned}$$

■

Using Lemma 6.4, we can get a better result:

Theorem 6.5 *With probability* $1-\delta$,

$$D_{\mathrm{B}}\left(\mathbb{E}_{h\sim\mathcal{Q}}[\mathrm{Err}_S(h)]\|\mathbb{E}_{h\sim\mathcal{Q}}[\mathrm{Err}_D(h)]\right)\leqslant\frac{1}{n}\left(D_{\mathrm{B}}(\mathcal{Q}\|\mathcal{P})+\ln\left(\frac{n+1}{\delta}\right)\right)\tag{6.6}$$

Proof: For fixed $h'\sim\mathcal{P}$, by applying Markov Inequality and Lemma 6.4, we have

$$\begin{aligned}\mathbb{P}_{S\sim D^n}\left[\mathbb{E}_{h'\sim\mathcal{P}}\left[e^{nD_{\mathrm{B}}(\mathrm{Err}_S(h')\|\,\mathrm{Err}_D(h'))}\right]\geqslant\frac{n+1}{\delta}\right] &\leqslant \frac{\mathbb{E}_{h'\sim\mathcal{P}}\left[\mathbb{E}_{S\sim D^n}\left[e^{nD_{\mathrm{B}}(\mathrm{Err}_S(h')\|\,\mathrm{Err}_D(h'))}\right]\right]}{\frac{n+1}{\delta}}\\ &\leqslant \frac{n+1}{\frac{n+1}{\delta}}=\delta\end{aligned}$$

Since $D_{\mathrm{B}}(\cdot\|\cdot)$ is convex, by Lemma 6.2 and Jensen's Inequality, with probability $1-\delta$,

$$\begin{aligned} D_{\mathrm{B}}(\mathbb{E}_{h\sim\mathcal{Q}}[\mathrm{Err}_S(h)]\|\mathbb{E}_{h\sim\mathcal{Q}}[\mathrm{Err}_D(h)]) &\leqslant \mathbb{E}_{h\sim\mathcal{Q}}[D_{\mathrm{B}}(\mathrm{Err}_S(h)\|\,\mathrm{Err}_D(h))] \\ &= \frac{1}{n}\mathbb{E}_{h\sim\mathcal{Q}}\left[nD_{\mathrm{B}}(\mathrm{Err}_S(h)\|\,\mathrm{Err}_D(h))\right] \\ &\leqslant \frac{1}{n}\left(\ln\left(\mathbb{E}_{h'\sim\mathcal{P}}\left[e^{nD_{\mathrm{B}}(\mathrm{Err}_S(h')\|\,\mathrm{Err}_D(h'))}\right]\right) + D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P})\right) \\ &\leqslant \frac{1}{n}\left(\ln\left(\frac{n+1}{\delta}\right) + D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P})\right) \end{aligned}$$

Thus proving the claim. ■

6.4 PAC-Bayesian Bound for SVM

Now we will apply PAC-Bayesian Theory to linear classifiers. Suppose $\mathcal{Q}$ is a distribution over classifiers, let us define a deterministic voting classifier $g_{\mathcal{Q}}(\boldsymbol{x})$

$$g_{\mathcal{Q}}(\boldsymbol{x}) = \mathrm{sgn}(\mathbb{E}_{h\sim\mathcal{Q}}h(\boldsymbol{x})) \tag{6.7}$$

We have the following proposition:

Proposition 6.6

$$\mathrm{Err}_D(g_{\mathcal{Q}}) \leqslant 2\mathbb{E}_{h\sim\mathcal{Q}}[\mathrm{Err}_D(h)] \tag{6.8}$$

Proof: For each point $(\boldsymbol{x}, y)\sim D$, $g_{\mathcal{Q}}(\boldsymbol{x}) \neq y$ implies that at least half of hs in $\mathcal{Q}$ satisfy $h(\boldsymbol{x}) \neq y$, which leads to the conclusion that $\mathrm{Err}_D(g_{\mathcal{Q}}) \leqslant 2\mathbb{E}_{h\sim\mathcal{Q}}[\mathrm{Err}_D(h)]$.

■

Now we consider linear classifiers $h(\boldsymbol{x}) = \mathrm{sgn}(\boldsymbol{w}^\top\boldsymbol{x}+b)$, $\boldsymbol{x}\in\mathbb{R}^d, y\in\{\pm1\}, \boldsymbol{w}\in\mathbb{R}^d, b\in\mathbb{R}$.
If we assume $\mathcal{P} = U(S^{d-1})$, the spherical integral of $D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P})$ would be complicated to compute. Alternatively, we assume $\mathcal{P} = \mathcal{N}(\boldsymbol{0}, I_d), \mathcal{Q} = \mathcal{N}(\mu\boldsymbol{w}, I_d)$, here $\|\boldsymbol{w}\|_2 = 1$ and μ is a scale factor. According to PAC-Bayesian Theorem (6.1), we have

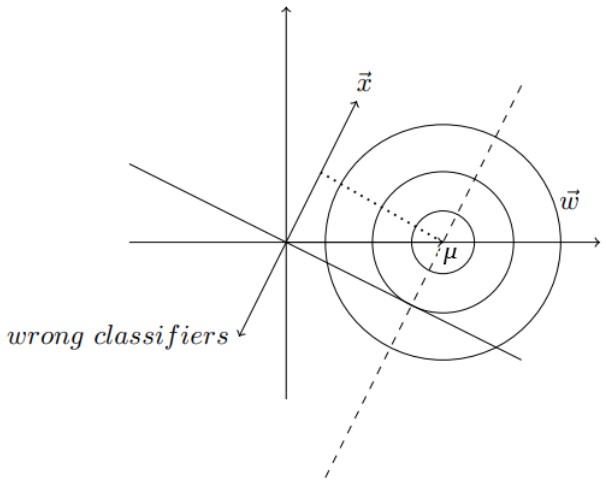
$$\mathrm{Err}_D(g_{\mathcal{Q}}) \leqslant 2\left[\mathrm{Err}_S(\mathcal{N}(\mu\boldsymbol{w}, I_d)) + \sqrt{\frac{D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P}) + \ln(3/\delta)}{n}}\right] \tag{6.9}$$

It suffices to determine $D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P})$ and $\mathrm{Err}_S(\mathcal{N}(\mu\boldsymbol{w}, I_d))$ respectively.

6.4.1 Determine $D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P})$

$$\begin{aligned}
D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P}) &= \int_{\mathbb{R}^d} \frac{1}{\left(\sqrt{2\pi}\right)^d} \exp\left[-\frac{1}{2}\|\boldsymbol{x}-\mu\boldsymbol{w}\|^2\right] \frac{1}{2}\left(\|\boldsymbol{x}\|^2 - \|\boldsymbol{x}-\mu\boldsymbol{w}\|^2\right) \mathrm{d}\boldsymbol{x} \\
&= \int_{\mathbb{R}^d} \frac{1}{\left(\sqrt{2\pi}\right)^d} \exp\left[-\frac{1}{2}\|\boldsymbol{x}-\mu\boldsymbol{w}\|^2\right] \left(\mu\boldsymbol{w}^\top\boldsymbol{x} - \frac{1}{2}\mu^2\right) \mathrm{d}\boldsymbol{x} \\
&= -\frac{1}{2}\mu^2 + \int_{\mathbb{R}^d} \frac{1}{\left(\sqrt{2\pi}\right)^d} \exp\left[-\frac{1}{2}\|\boldsymbol{x}-\mu\boldsymbol{w}\|^2\right] \left(\mu\boldsymbol{w}^\top(\boldsymbol{x}-\mu\boldsymbol{w}) + \mu^2\right) \mathrm{d}\boldsymbol{x} \\
&= \frac{1}{2}\mu^2 + \mu \int_{\mathbb{R}^d} (\boldsymbol{w}^\top\boldsymbol{x}) \frac{1}{\left(\sqrt{2\pi}\right)^d} \exp\left[-\frac{1}{2}\|\boldsymbol{x}\|^2\right] \mathrm{d}\boldsymbol{x} \\
&= \frac{1}{2}\mu^2 + \mu\boldsymbol{w}^\top \mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}}[\boldsymbol{x}] \\
&= \frac{1}{2}\mu^2
\end{aligned}$$

6.4.2 Determine $\mathrm{Err}_S(\mathcal{N}(\mu\boldsymbol{w}, I_d))$



For a fixed point $(\boldsymbol{x}, y)$ and a classifier $\boldsymbol{h} \sim \mathcal{N}(\mu\boldsymbol{w}, I_d)$, with the intuition from the figure above, we have

$$\frac{y\boldsymbol{x}^\top\boldsymbol{h}}{\|\boldsymbol{x}\|} \sim \mathcal{N}\left(\frac{y\boldsymbol{x}^\top(\mu\boldsymbol{w})}{\|\boldsymbol{x}\|}, \frac{y^2\boldsymbol{x}^\top\boldsymbol{x}}{\|\boldsymbol{x}\|^2}\right) = \mathcal{N}\left(\frac{y\mu(\boldsymbol{w}^\top\boldsymbol{x})}{\|\boldsymbol{x}\|}, 1\right)$$

So

$$\begin{aligned}
\mathrm{Err}_S(\mathcal{N}(\mu\boldsymbol{w}, I_d)) &= \mathbb{P}_{\boldsymbol{h}\sim\mathcal{N}(\mu\boldsymbol{w}, I_d)}\left[\frac{y\left(\boldsymbol{x}^\top\boldsymbol{h}\right)}{\|\boldsymbol{x}\|} < 0\right] \\
&= \bar{\Phi}\left(\frac{y\mu(\boldsymbol{w}^\top\boldsymbol{x})}{\|\boldsymbol{x}\|}\right)
\end{aligned}$$

Here, $\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x, \bar{\Phi}(t) = 1 - \Phi(t)$.

6.4.3 Putting them together

By PAC-Bayesian Theorem (6.1), for SVM model

$$\begin{aligned}\mathbb{E}_{h\sim\mathcal{Q}}[\mathrm{Err}_D(h)] &\leqslant \mathbb{E}_{h\sim\mathcal{Q}}[\mathrm{Err}_S(h)] + \sqrt{\frac{D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P}) + \ln\frac{3}{\delta}}{n}} \\ &= \mathbb{E}_{S\sim D^n}\left[\bar{\Phi}\left(\frac{y\mu\left(\boldsymbol{w}^\top\boldsymbol{x}\right)}{\|\boldsymbol{x}\|}\right)\right] + \sqrt{\frac{\frac{\mu^2}{2} + \ln\frac{3}{\delta}}{n}}\end{aligned}$$

Therefore, with probability at least $1-\delta$ over the random draw of n training data, for all μ and $\boldsymbol{w}$, we have

$$\mathrm{Err}_D(g_{\mathcal{Q}}) \leqslant 2\left[\mathbb{E}_{S\sim D^n}\left[\bar{\Phi}\left(\frac{y\mu\left(\boldsymbol{w}^\top\boldsymbol{x}\right)}{\|\boldsymbol{x}\|}\right)\right] + \sqrt{\frac{\frac{\mu^2}{2} + \ln\frac{3}{\delta}}{n}}\right]$$

By optimizing the value of μ, we have

$$\mathrm{Err}_D(g_{\mathcal{Q}}) \leqslant 2\inf_{\mu>0}\left\{\left[\mathbb{E}_{S\sim D^n}\left[\bar{\Phi}\left(\frac{y\mu\left(\boldsymbol{w}^\top\boldsymbol{x}\right)}{\|\boldsymbol{x}\|}\right)\right] + \sqrt{\frac{\frac{\mu^2}{2} + \ln\frac{3}{\delta}}{n}}\right]\right\}$$

Since the Gaussian tail probability is monotonically decreasing w.r.t. μ and $\mu^2/2n$ is monotonically increasing w.r.t. μ, we cannot determine the monotonicity of the bound. Intuitively, the Gaussian tail probability will be small if the margin is large, thus the larger the margin is, the tighter the bound will be, which means the generalization ability of SVM will be good when it has a large margin.

References

[1] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1225–1234, 2016.

[2] Samuel Smith, Erich Elsen, and Soham De. On the generalization benefit of noise in stochastic gradient descent. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 9058–9067, 2020.

[3] Samuel L. Smith and Quoc V. Le. A bayesian perspective on generalization and stochastic gradient descent. In *International Conference on Learning Representations*, 2018.