PAC-Bayes Bound

Ruixin Guo

Department of Computer Science
Kent State University

December 13, 2023

Contents

Recall: Statistics

Point Estimation: Suppose we have a distribution $\mathcal{D}$ with PDF $f(x|\theta)$ where θ is an unknown and fixed parameter. Let $X_1, X_2, ..., X_n$ be the iid random variables from $\mathcal{D}$. Let $\hat{\theta} = W(X_1, X_2, ..., X_n)$ where W is a function of $X_1, X_2..., X_n$. We call $\hat{\theta}$ an point estimator of θ.

Let $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$, we call $W(x_1, x_2, ..., x_n)$ an estimate of θ.

Maximum Likelihood Estimator: Let $L(\theta|\boldsymbol{x})$ be the likelihood function:

$$L(\theta|\boldsymbol{x}) = f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

where $\boldsymbol{x} = [x_1, ..., x_n]$ is the vector of samples from $\mathcal{D}$. Then $\hat{\theta}$ is defined as

$$\hat{\theta} = \underset{\theta}{\text{argmax}}\ L(\theta|\boldsymbol{x})$$

We call such $\hat{\theta}$ as the Maximum Likelihood Estimate of θ. $\hat{\theta}$ is a function of $x_1, ..., x_n$. Replace $x_1, ..., x_n$ by $X_1, ..., X_n$ and we get the estimator.

Recall: Statistics

Bayes Estimator: In classic point estimation approach, θ is considered as an unknown but fixed value. In Bayesian approach, θ is considered as a variable whose variation can be described as a distribution $\pi(\theta)$. The Bayes Estimator is described as

$$\rho(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{m(\boldsymbol{x})}$$

We call $\pi(\theta)$ the prior distribution, $\rho(\theta|\boldsymbol{x})$ the posterior distribution. $f(\boldsymbol{x}|\theta)$ is the likelihood function. $m(\boldsymbol{x})$ is the marginal distribution of $\boldsymbol{x}$:

$$m(\boldsymbol{x}) = \int f(\boldsymbol{x}|\theta)\pi(\theta)\, d\theta$$

A natural estimator of θ is the mean of the posterior distribution, i.e.,

$$\hat{\theta} = \mathbb{E}_\theta[\rho(\theta|\boldsymbol{x})] = \int \theta\rho(\theta|\boldsymbol{x})\, d\theta$$

Recall: Statistics

The Process of Bayes Estimation: Suppose θ is sampled from an unknown distribution u. The steps of Bayes Estimation for θ is as follows [2]:
1. Define the prior distribution $\pi(\theta)$ based on your subject beliefs or knowledge.
2. Define the likelihood function based on the characteristics of the data (e.g., the data looks like normal distribution or exponential distribution).
3. Gather data.
4. Calculate the posterior distribution $\rho(\theta|\boldsymbol{x})$ with the prior distribution and the data. The posterior distribution represents your updated beliefs about the distribution of θ.
5. Construct an estimator of θ based on $\rho(\theta|\boldsymbol{x})$.

The posterior distribution $\rho(\theta|\boldsymbol{x})$ is considered to be closer to the true distribution $u(\theta)$ than the prior distribution $\pi(\theta)$.

Uninformative Priors: If we have no beliefs or prior knowledge about the distribution of θ, we can use some uninformative priors to define $\pi(\theta)$ [3]. One common uninformative priors is uniform prior (i.e., let $\pi(\theta)$ be a uniform distribution).

Recall: Statistics

Note that no matter what distribution $\pi(\theta)$ is chosen, it is not likely to be independent from $u(\theta)$. That is, $\pi(\theta)$ at least contains some information about $u(\theta)$ and gives some prior knowledge.

The Bayes Estimation follows these rules [2]:

- If the prior is uninformative, the posterior is mainly determined by the data.
- If the prior is informative, the posterior is a mixture of the prior and the data.
- The more informative the prior, the more data you need to "change" your beliefs. In this case, the posterior is mainly driven by the prior information.
- If you have a lot of data, the data will dominate the posterior distribution (they will overwhelm the prior).

Bayes Estimation is particularly useful when dealing with limited data or incorporating expert knowledge into the analysis.

Recall: Statistical Learning Theory

Dataset: The dataset is defined as a set $S = \{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$ is the d-dimensional feature vector, $y_i \in \mathbb{R}$ is the label. Suppose each (x_i, y_i) is sampled from an unknown distribution $\mathcal{D}$, and each x_i is mapped to each y_i by an unknown function f^*.

Predictor: Let $f_\theta : \mathbb{R}^d \to \mathbb{R}$ be the predictor parametermized by θ. For the data sample (x_i, y_i), let $\hat{y}_i = f_\theta(x_i)$, then $\hat{y}_i$ is a prediction of y_i. In machine learning, f_θ is a machine learning model and we train it by adjusting θ.

Hypothesis Space and Parameter Space: We wish the predictor f_θ as close to f^* as possible. However, f^* may lie in a set of all measurable functions Ω and difficult to be searched. We usually take $\mathcal{F} \subset \Omega$ and search the f_θ in $\mathcal{F}$ that is closest to f^*. We call $\mathcal{F}$ the hypothesis space. Since f_θ is a function of θ, we define a parameter space Θ such that each $f_\theta \in \mathcal{F}$ corresponds to an $\theta \in \Theta$, and vise versa. Thus, searching f_θ in $\mathcal{F}$ is equivalent to searching θ in Θ.

Loss: We measure the error of $\hat{y}_i$ and y_i by the loss function l. There are many ways to define l. For quadratic loss, $l(\hat{y}_i, y_i) = (\hat{y}_i - y_i)^2$. For absolute loss, $l(\hat{y}_i, y_i) = |\hat{y}_i - y_i|$.

Recall: Statistical Learning Theory

Empirical Risk: The empirical risk is the average loss of all samples in the dataset by the predictor f_θ. Since the dataset S is fixed, f_θ depends only on θ, the empirical risk function can be defined as

$$R^{\text{emp}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(\hat{y}_i - y_i) = \frac{1}{n}\sum_{i=1}^{n} l(f_\theta(x_i) - y_i)$$

Risk: The risk, also called true risk, is the average of empirical risk of all datasets, i.e., taking expectation over the entire distribution $\mathcal{D}$,

$$R^{\text{true}}(\theta) = \mathbb{E}_S[R^{\text{emp}}(\theta)]$$

The Goal of Machine Learning: We want to make the predictor f_θ be able to predict the data from the entire distribution $\mathcal{D}$. That is, we want to find the minimizer of the true risk θ^*,

$$\theta^* = \underset{\theta}{\text{argmin}}\ R^{\text{true}}(\theta)$$

However, we cannot calculate R^{true} since $\mathcal{D}$ is unknown. Instead, we estimate θ^* using an estimator. The most famous estimator of θ^* is the Empirical Risk Minimizer $\hat{\theta}$:

$$\hat{\theta} = \underset{\theta}{\text{argmin}}\ R^{\text{emp}}(\theta)$$

We would like to measure the error between θ^* and $\hat{\theta}$.

Recall: Statistical Learning Theory

Estimation Error: Remember that our goal is to minimize R^{true}. We use the estimation error $R^{\text{true}}(\hat{\theta}) - R^{\text{true}}(\theta^*)$ to measure the error between θ^* and $\hat{\theta}$. The estimation error can be bounded by

$$R^{\text{true}}(\hat{\theta}) - R^{\text{true}}(\theta^*) \leq 2 \sup_{\theta \in \Theta} \left| R^{\text{true}}(\theta) - R^{\text{emp}}(\theta) \right|$$

Since $R^{\text{true}}(\theta) = \mathbb{E}_S[R^{\text{emp}}(\theta)]$, we can bound $\left| R^{\text{true}}(\theta) - R^{\text{emp}}(\theta) \right|$ using concentration inequality. Usually we use Hoeffding Inequality.

Hoeffding Lemma: Suppose a random variable $X \in [a, b]$, then for any $\lambda \in \mathbb{R}$.

$$\mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] \leq e^{\frac{\lambda^2 (b-a)^2}{8}}$$

Suppose each loss function has the range $0 \leq l \leq C$. Let $T = \sum_{i=1}^{n} X_i$. Suppose θ is fixed and we take expectation with respect to S. When $X_i = l(\hat{y}_i - y_i)$, $T = nR^{\text{emp}}(\theta)$. Let $X_1, X_2, ..., X_n$ be independent. For any $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}_S[e^{\lambda n(R^{\text{emp}}(\theta) - R^{\text{true}}(\theta))}] = \mathbb{E}_S[e^{\lambda(T - \mathbb{E}_S[T])}] = \prod_{i=1}^{n} \mathbb{E}_S[e^{\lambda(X_i - \mathbb{E}_S[X_i])}] \leq \prod_{i=1}^{n} e^{\frac{\lambda^2 C^2}{8}} = e^{\frac{n\lambda^2 C^2}{8}}$$

Assign $-\lambda$ to λ, we have

$$\mathbb{E}_S[e^{\lambda n(R^{\text{true}}(\theta) - R^{\text{emp}}(\theta))}] \leq e^{\frac{n\lambda^2 C^2}{8}} \tag{1}$$

Recall: Statistical Learning Theory

Chernoff Inequality: Let X be an random variable and ϵ be a constant. For any $\lambda > 0$,

$$P(X \geq \epsilon) = P(e^{\lambda X} \geq e^{\lambda \epsilon}) \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda \epsilon}}$$

For any $\lambda < 0$,

$$P(X \leq \epsilon) = P(e^{\lambda X} \geq e^{\lambda \epsilon}) \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda \epsilon}}$$

Hoeffding Inequality: By Hoeffding Lemma, for any $\epsilon > 0$, for any $\lambda > 0$, we have

$$\begin{aligned} P\left(R^{\text{emp}}(\theta) - R^{\text{true}}(\theta) > \epsilon\right) &= P\left(e^{n\lambda(R^{\text{emp}}(\theta) - R^{\text{true}}(\theta))} > e^{n\lambda\epsilon}\right) \\ &\leq e^{-n\lambda\epsilon}\mathbb{E}\left[e^{n\lambda(R^{\text{emp}}(\theta) - R^{\text{true}}(\theta))}\right] \leq e^{-n\lambda\epsilon} e^{\frac{n\lambda^2 C^2}{8}} \\ &= e^{\frac{n\lambda^2 C^2}{8} - n\lambda\epsilon} \end{aligned}$$

We can choose λ to get the tightest upper bound. Taking $\lambda = \frac{4\epsilon}{C^2}$, we have

$$P\left(R^{\text{emp}}(\theta) - R^{\text{true}}(\theta) > \epsilon\right) \leq e^{\frac{-2n\epsilon^2}{C^2}}$$

Recall: Statistical Learning Theory

PAC Bound: Suppose $|\Theta| = M$ is finite. We have the union bound:

$$P\left[\sup_{\theta\in\Theta}\left(R^{\text{emp}}(\theta) - R^{\text{true}}(\theta)\right) > \epsilon\right] = P\left[\bigcup_{\theta\in\Theta}\left(R^{\text{emp}}(\theta) - R^{\text{true}}(\theta) > \epsilon\right)\right]$$

$$\leq \sum_{\theta\in\Theta} P\left(R^{\text{emp}}(\theta) - R^{\text{true}}(\theta) > \epsilon\right) \leq Me^{\frac{-2n\epsilon^2}{C^2}} \tag{2}$$

This bound is called Probably Approximately Correct (PAC) bound. The above bound is one-sided.

For $\lambda < 0$, we have

$$P\left(R^{\text{emp}}(\theta) - R^{\text{true}}(\theta) < -\epsilon\right) = P\left(e^{n\lambda(R^{\text{emp}}(\theta)-R^{\text{true}}(\theta))} > e^{-n\lambda\epsilon}\right) = e^{\frac{n\lambda^2C^2}{8}+n\lambda\epsilon}$$

Taking $\lambda = -\frac{4\epsilon}{C^2}$, we have $P\left(R^{\text{emp}}(\theta) - R^{\text{true}}(\theta) < -\epsilon\right) \leq e^{\frac{-2n\epsilon^2}{C^2}}$, thus we get the two-sided bound:

$$P\left[\sup_{\theta\in\Theta}\left(|R^{\text{emp}}(\theta) - R^{\text{true}}(\theta)|\right) > \epsilon\right] \leq 2Me^{\frac{-2n\epsilon^2}{C^2}}$$

Recall: Statistical Learning Theory

In Eq (2), let $\delta = Me^{\frac{-2n\epsilon^2}{C^2}}$, then $\epsilon = C\sqrt{\frac{\log \frac{M}{\delta}}{2n}}$. And Eq (2) can be written as

$$P\left[\sup_{\theta\in\Theta}\left(R^{\mathsf{emp}}(\theta) - R^{\mathsf{true}}(\theta)\right) > C\sqrt{\frac{\log \frac{M}{\delta}}{2n}}\right] \leq \delta$$

That is,

$$P\left[\sup_{\theta\in\Theta}\left(R^{\mathsf{emp}}(\theta) - R^{\mathsf{true}}(\theta)\right) \leq C\sqrt{\frac{\log \frac{M}{\delta}}{2n}}\right] \geq 1 - \delta \tag{3}$$

Recall: Information Theory

KL Divergence: For any two distributions $p(x)$ and $q(x)$ satisfying $p(x) > 0 \Leftrightarrow q(x) > 0$, the KL Divergence of q from p is defined as

$$D(\,p\,||\,q\,) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

In these slides, all the logarithms are of base e.

Donsker-Varadhan Representation of KL Divergence: Let $g(x)$ be any measurable function, $p(x)$ and $q(x)$ be two distributions, and e be the base of the logarithm. The KL Divergence equals to the following variational representation:

$$D(\,p\,||\,q\,) = \sup_{g(x)} \{\mathbb{E}_{p(x)}[g(x)] - \log \mathbb{E}_{q(x)}[e^{g(x)}]\} \tag{4}$$

The above equation fixes p, q and takes g as a variable. In fact, we can fix g, q and take p as a variable:

$$\log \mathbb{E}_{q(x)}[e^{g(x)}] = \sup_{p(x)} \{\mathbb{E}_{p(x)}[g(x)] - D(\,p\,||\,q\,)\} \tag{5}$$

Recall: Information Theory

This is because, when taking $t(x) = \frac{e^{g(x)}q(x)}{\mathbb{E}_{q(x)}[e^{g(x)}]}$. We have

$$D(\,p\,||\,q\,) - \Big(\mathbb{E}_{p(x)}[g(x)] - \log \mathbb{E}_{q(x)}[e^{g(x)}]\Big) = D(\,p\,||\,t\,) \geq 0$$

To make $D(\,p\,||\,q\,) = \big(\mathbb{E}_{p(x)}[g(x)] - \log \mathbb{E}_{q(x)}[e^{g(x)}]\big)$, we need $D(\,p\,||\,t\,) = 0$, which is to make $p = t$. To achieve this, we can either fix p and adjust t (note that adjust t is equivalent to adjust g), or fix t and adjust p. The former case gives Eq (4), and the latter case gives Eq (5).

We call $t(x)$ the Gibbs distribution. For Eq (5), the supremum is obtained when $p = t$, i.e.,

$$p(x) = \frac{e^{g(x)}q(x)}{\mathbb{E}_{q(x)}[e^{g(x)}]} \tag{6}$$

Contents

PAC-Bayes Bound

PAC-Bayes Bound:

The PAC-Bayes Bounds are generalized PAC bounds that allow to deal with the distribution of the parameter space Θ. It bounds the generalization gap using the KL-Divergence $D(\,\rho\,||\,\pi\,)$ where π is the prior distribution of Θ and ρ is the posterior distribution of Θ.

Why PAC-Bayes Bound:

Traditional PAC Bounds only considers the maximum distance between $R^{\text{emp}}(\theta)$ and $R^{\text{true}}(\theta)$, i.e., $\sup_{\theta\in\Theta}\left(R^{\text{emp}}(\theta) - R^{\text{true}}(\theta)\right)$, which tends to be vacuous. PAC-Bayes bound considers the average distance. Suppose θ satisfy a distribution ρ, the average distance can be written as

$$\mathbb{E}_{\theta\sim\rho}[R^{\text{emp}}(\theta) - R^{\text{true}}(\theta)] = \mathbb{E}_{\theta\sim\rho}[R^{\text{emp}}(\theta)] - \mathbb{E}_{\theta\sim\rho}[R^{\text{true}}(\theta)]$$

In machine learning, we consider the parameter vector θ of the machine learning model to be sampled from a parameter space Θ. θ is initialized according to a prior distribution π before training. After the model is trained through some data $\boldsymbol{x}$, θ is supposed to follow a posterior distribution $\rho(\theta|\boldsymbol{x})$.

Another reason for the vacuousness of traditional PAC bound is data-independence. The bound $2Me^{\frac{-2n\epsilon^2}{C^2}}$ only considers the number of samples n of the data, but does not consider the type of the data. For example, suppose we have two groups of image data, each of them has 10000 images. The images in the first group are of size 20×20, and the images in the second group are of size 200×200. The generalization gap of the second group is supposed to be larger than the first group because of Curse of Dimensionality, but traditional PAC considers the generalization gaps on both groups to be the same.

The PAC-Bayes bound uses the KL-Divergence $D(\,\rho\,||\,\pi\,)$ to bound the generalization gap, and $D(\,\rho\,||\,\pi\,)$ is a data-dependent measure. Since ρ is conditioned on $\boldsymbol{x}$, the distance $D(\,\rho\,||\,\pi\,)$ is depends on $\boldsymbol{x}$, and feeding different $\boldsymbol{x}$ can make $D(\,\rho\,||\,\pi\,)$ different.

The History of PAC-Bayes Bound:

The first PAC-Bayes bound is proposed by McAllester [19] in 1999. Since then, different types of PAC-Bayes bounds were proposed. Langford and Seeger [5] proposed the PAC-Bayes Relative Entropy Bound in 2001. This bound is very central in PAC-Bayes theory since many other PAC-Bayes bounds can be derived from this one. Catoni proposed a PAC-Bayes Bound focusing on Gibbs posterior in 2003.

For the application of PAC-Bayes bounds, Langford and Caruana [7] in 2001 showed that the Langford and Seeger's Bound [5] can be non-vacuous in bounding generalization error for two-layer two-hidden-unit neural networks. Later Dziugaite and Roy follow this method and showed that PAC-Bayes Bound can be non-vacuous in bounding generalization error for deep neural networks in 2017, which is a great breakthrough.

Overview of this Slides:

We will first introduce the Langford and Seeger's Bound, then introduce the work of Dziugaite and Roy that shows this bound is tight and non-vacuous in estimating the generalization gap for deep neural networks. We will also inroduce Catoni's Bound in the Appendix.

Langford and Seeger's Bound

Definition 1.1: We define the KL Divergence of two Bernoulli Distribution $\mathsf{Ber}(p)$ and $\mathsf{Ber}(q)$ as

$$D_B(\,p\,||\,q\,) := D(\,\mathsf{Ber}(p)\,||\,\mathsf{Ber}(q)\,) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$$

where a random variable $X \sim \mathsf{Ber}(p)$ means $P(X=1) = p$ and $P(X=0) = 1-p$.

Theorem 1.2 (Langford and Seeger's Bound): Let π be any prior distribution over the parameter space Θ and $\delta \in (0,1)$. Let the loss function $l(f_\theta(x_i) - y_i) \in [0,1]$. Then for any dataset S and any posterior distribution ρ over Θ,

$$P_S\left[D_B(\mathbb{E}_{\theta\sim\rho}[\,R^{\mathsf{emp}}(\theta)]\,||\,\mathbb{E}_{\theta\sim\rho}[\,R^{\mathsf{true}}(\theta)]\,) \le \frac{D(\,\rho\,||\,\pi\,) + \log\frac{2n}{\delta}}{n-1}\right] \ge 1-\delta$$

Proof: To prove Theorem 1.2, we will utilize Lemmas 1.3 and Lemma 1.4:

Lemma 1.3: Let $\beta > 0$, $K > 0$ be constants. Let $p(x)$ and $q(x)$ be two PDFs. If

$$\mathbb{E}_{q(x)}[e^{\beta x}] \leq K$$

then

$$\mathbb{E}_{p(x)}[x] \leq \frac{D(\, p(x) \,||\, q(x)\,) + \log K}{\beta}$$

Proof: This is directly from Donsker-Varadhan Representation of KL Divergence. By Eq (5), for any $p(x)$,

$$\mathbb{E}_{p(x)}[\beta x] - D(\, p(x) \,||\, q(x)\,) \leq \log \mathbb{E}_{q(x)}[e^{\beta x}]$$

Since $\mathbb{E}_{q(x)}[e^{\beta x}] \leq K$, we have

$$\mathbb{E}_{p(x)}[\beta x] - D(\, p(x) \,||\, q(x)\,) \leq \log K$$

Rearrange the above inequality and the theorem is proved. □

Lemma 1.4: For any $\delta > 0$ and any dataset S, suppose θ is sampled from a distribution π, then

$$P_S\left[\mathbb{E}_{\theta\sim\pi}[e^{(n-1)D_B(\,R^{\text{emp}}(\theta)\,||\,R^{\text{true}}(\theta)\,)}] \geq \delta\right] \leq \frac{2n}{\delta}$$

Proof: To prove Lemma 1.4, we will utilize Relative Entropy of Chernoff Bound, as shown in Lemma 1.5.

Lemma 1.5 (Relative Entropy of Chernoff Bound): Assume random variables $X_1, X_2, ..., X_n$ are i.i.d., $X_i \in [0, 1]$ for any i. Let $p = \mathbb{E}[X_i]$ and $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$, then

$$P[\hat{p} \geq p + \epsilon] \leq e^{-D_B(\,p+\epsilon\,||\,p\,)n}$$
$$P[\hat{p} \leq p - \epsilon] \leq e^{-D_B(\,p-\epsilon\,||\,p\,)n}$$

Proof: For the first bound, let $q = p + \epsilon$. For any $\lambda > 0$,

$$P[\hat{p} \geq q] = P[e^{\lambda n\hat{p}} \geq e^{\lambda nq}] \leq e^{-\lambda nq}\mathbb{E}[e^{\lambda n\hat{p}}] = e^{-\lambda nq}\mathbb{E}[e^{\lambda\sum_{i=1}^{n} X_i}]$$
$$= e^{-\lambda nq}\mathbb{E}[\prod_{i=1}^{n} e^{\lambda X_i}] = e^{-\lambda nq}\prod_{i=1}^{n}\mathbb{E}[e^{\lambda X_i}]$$

Now let's use the inequality $e^{\lambda x} \leq 1 - x + xe^{\lambda}$ for any $\lambda > 0$ and $x \in [0, 1]$. This is because $f(x) = e^{\lambda x}$ is a convex function and $(e^{\lambda} - 1)x + 1$ makes the line segment between $(0, f(0))$ and $(1, f(1))$, which will be above $f(x)$.

Since each $X_i \in [0, 1]$,

$$e^{-\lambda nq} \prod_{i=1}^{n} \mathbb{E}[e^{\lambda X_i}] \leq e^{-\lambda nq} \prod_{i=1}^{n} \mathbb{E}[1 - X_i + X_i e^{\lambda}] = e^{-\lambda nq} \prod_{i=1}^{n} (1 - p + pe^{\lambda})$$
$$= e^{-\lambda nq} (1 - p + pe^{\lambda})^n = e^{\phi(\lambda) n}$$

where we define $\phi(\lambda) = \log[e^{-\lambda q}(1 - p + pe^{\lambda})]$. To get the tightest bound, we need to find λ to minimize $\phi(\lambda)$. Solving $\frac{d\phi}{d\lambda} = 0$, we obtain $\lambda = \log \frac{q(1-p)}{p(1-q)}$, thus

$$\min_{\lambda} \phi(\lambda) = -q \log \frac{q}{p} - (1 - q) \log \frac{1 - q}{1 - p} = -D_B(\, q \,||\, p \,)$$

Therefore, we get the bound

$$P[\hat{p} \geq q] \leq e^{-D_B(\, q \,||\, p \,) n}$$

Then we can use the first bound to prove the second bound. Let $X_i \leftarrow 1 - X_i$, such that $\hat{p} \leftarrow 1 - \hat{p}$, $p \leftarrow 1 - p$. Therefore,

$$\begin{aligned} P[\hat{p} \leq p - \epsilon] &= P[1 - \hat{p} \geq 1 - p + \epsilon] \\ &\leq e^{-D_B(\,1-p+\epsilon\,||\,1-p\,)n} = e^{-D_B(\,p-\epsilon\,||\,p\,)n} \end{aligned}$$

□

Let's go back to the proof of Lemma 1.4. By Markov's Inequality,

$$\begin{aligned} P_S\left[\mathbb{E}_{\theta\sim\pi}\left[e^{(n-1)D_B(\,R^{\text{emp}}(\theta)\,||\,R^{\text{true}}(\theta)\,)}\right] \geq \delta\right] &\leq \frac{\mathbb{E}_S\mathbb{E}_{\theta\sim\pi}\left[e^{(n-1)D_B(\,R^{\text{emp}}(\theta)\,||\,R^{\text{true}}(\theta)\,)}\right]}{\delta} \\ &= \frac{\mathbb{E}_{\theta\sim\pi}\mathbb{E}_S\left[e^{(n-1)D_B(\,R^{\text{emp}}(\theta)\,||\,R^{\text{true}}(\theta)\,)}\right]}{\delta} \end{aligned}$$

If we can prove

$$\mathbb{E}_S\left[e^{(n-1)D_B(\,R^{\text{emp}}(\theta)\,||\,R^{\text{true}}(\theta)\,)}\right] \leq 2n$$

Then Lemma 1.4 will be proved. Let's do this!

Note that the only variable in $\mathbb{E}_S\big[e^{(n-1)D_B(\,R^{\text{emp}}(\theta)\,||\,R^{\text{true}}(\theta)\,)}\big]$ is $R^{\text{emp}}(\theta)$, which depends on S. Since $S = \{(x_i, y_i)\}_{i=1}^n$ and $R^{\text{emp}}(\theta) = \frac{1}{n}\sum_{i=1}^n l(f_\theta(x_i) - y_i)$, we can write

$$\begin{aligned}
\mathbb{E}_S\big[e^{(n-1)D_B(\,R^{\text{emp}}(\theta)\,||\,R^{\text{true}}(\theta)\,)}\big] &= \sum_s e^{(n-1)D_B(\,R^{\text{emp}}(\theta)\,||\,R^{\text{true}}(\theta)\,)} P(S = s) \\
&= \sum_r e^{(n-1)D_B(\,R^{\text{emp}}(\theta)\,||\,R^{\text{true}}(\theta)\,)} P(R^{\text{emp}}(\theta) = r) \\
&= \mathbb{E}_{R^{\text{emp}}(\theta)}\big[e^{(n-1)D_B(\,R^{\text{emp}}(\theta)\,||\,R^{\text{true}}(\theta)\,)}\big]
\end{aligned}$$

where we let

$$P(R^{\text{emp}}(\theta) = r) = \sum_{s \in \{\{(x_i, y_i)\}_{i=1}^n \,:\, r = \frac{1}{n}\sum_{i=1}^n l(f_\theta(x_i) - y_i)\}} P(S = s)$$

This means we can convert the expectation of $e^{(n-1)D_B(\,R^{\text{emp}}(\theta)\,||\,R^{\text{true}}(\theta)\,)}$ with respect to S to the one with respect to $R^{\text{emp}}(\theta)$. We obtain the probability of $R^{\text{emp}}(\theta) = r$ by summing up the probabilities of all datasets S that makes $R^{\text{emp}}(\theta) = r$.

Now we can notice that $R^{\text{emp}}(\theta)$ satisfies a new distribution which ranges between $[0, 1]$. We want to prove that $\mathbb{E}_{R^{\text{emp}}(\theta)}[e^{(n-1)D_B(\,R^{\text{emp}}(\theta)\,||\,R^{\text{true}}(\theta)\,)}] \leq 2n$ for any distribution of $R^{\text{emp}}(\theta)$. Let $X = R^{\text{emp}}(\theta)$ and $f(x)$ be the PDF of X, by Lemma 1.5 we know for all $q_1 \geq R^{\text{true}}(\theta)$ and $q_2 \leq R^{\text{true}}(\theta)$,

$$P[x \geq q_1] = \int_{q_1}^{1} f(x)dx \leq e^{-D_B(\,q_1\,||\,R^{\text{true}}(\theta)\,)n}$$

$$P[x \leq q_2] = \int_{0}^{q_2} f(x)dx \leq e^{-D_B(\,q_2\,||\,R^{\text{true}}(\theta)\,)n}$$

The equality is obtained when

$$f(x) = \begin{cases} n\frac{\partial D_B(\,x\,||\,R^{\text{true}}(\theta)\,)}{\partial x} e^{-D_B(\,x\,||\,R^{\text{true}}(\theta)\,)n} & \text{for } x \geq R^{\text{true}}(\theta) \\ -n\frac{\partial D_B(\,x\,||\,R^{\text{true}}(\theta)\,)}{\partial x} e^{-D_B(\,x\,||\,R^{\text{true}}(\theta)\,)n} & \text{for } x \leq R^{\text{true}}(\theta) \end{cases} \tag{7}$$

We prove $x \leq R^{\text{true}}(\theta)$ case for example. Let $F(x)$ be the primitive function of $f(x)$, then by Newton-Leibniz Theorem,

$$\int_{0}^{q_1} f(x)dx = F(q_1) - F(0) = e^{-D_B(\,q_1\,||\,R^{\text{true}}(\theta)\,)n}$$

Taking derivative with respect to q_1 on both sides,

$$f(q_1) = -n \frac{\partial D_B(\, q_1 \,||\, R^{\text{true}}(\theta)\,)}{\partial q_1} e^{-D_B(\, q_1 \,||\, R^{\text{true}}(\theta)\,)n}$$

Replacing q_1 with x, and we proved the bound.

Therefore,

$$\mathbb{E}_{R^{\text{emp}}(\theta)}[e^{(n-1)D_B(\, R^{\text{emp}}(\theta) \,||\, R^{\text{true}}(\theta)\,)}] = \int_0^1 e^{(n-1)D_B(\, x \,||\, R^{\text{true}}(\theta)\,)} f(x)dx$$

$$= \int_0^{R^{\text{true}}(\theta)} e^{(n-1)D_B(\, x \,||\, R^{\text{true}}(\theta)\,)} f(x)dx + \int_{R^{\text{true}}(\theta)}^1 e^{(n-1)D_B(\, x \,||\, R^{\text{true}}(\theta)\,)} f(x)dx$$

Now we want to find $f(x)$ to maximize the last equality. Note that $e^{(n-1)D_B(\, x \,||\, R^{\text{true}}(\theta)\,)}$ increases when x is away from $R^{\text{true}}(\theta)$, either smaller or larger. To maximize the expectation, we use the greedy strategy to assign as greater probability to larger $e^{(n-1)D_B(\, x \,||\, R^{\text{true}}(\theta)\,)}$ as possible. This can be achieved by taking $f(x)$ by Eq (7), the maximum $f(x)$ that reaches Chernoff bound. See Appendix 1.

Therefore,

$$
\begin{aligned}
&\mathbb{E}_{R^{\mathsf{emp}}(\theta)}\left[e^{(n-1)D_B(\,R^{\mathsf{emp}}(\theta)\,||\,R^{\mathsf{true}}(\theta)\,)}\right] \\
&\leq \int_0^{R^{\mathsf{true}}(\theta)} -n\frac{\partial D_B(\,x\,||\,R^{\mathsf{true}}(\theta)\,)}{\partial x} e^{-D_B(\,x\,||\,R^{\mathsf{true}}(\theta)\,)n} e^{(n-1)D_B(\,x\,||\,R^{\mathsf{true}}(\theta)\,)}dx \\
&\quad + \int_{R^{\mathsf{true}}(\theta)}^1 n\frac{\partial D_B(\,x\,||\,R^{\mathsf{true}}(\theta)\,)}{\partial x} e^{-D_B(\,x\,||\,R^{\mathsf{true}}(\theta)\,)n} e^{(n-1)D_B(\,x\,||\,R^{\mathsf{true}}(\theta)\,)}dx \\
&= \int_0^{R^{\mathsf{true}}(\theta)} -n\frac{\partial D_B(\,x\,||\,R^{\mathsf{true}}(\theta)\,)}{\partial x} e^{-D_B(\,x\,||\,R^{\mathsf{true}}(\theta)\,)}dx \\
&\quad + \int_{R^{\mathsf{true}}(\theta)}^1 n\frac{\partial D_B(\,x\,||\,R^{\mathsf{true}}(\theta)\,)}{\partial x} e^{-D_B(\,x\,||\,R^{\mathsf{true}}(\theta)\,)}dx \\
&= n\cdot e^{-D_B(\,x\,||\,R^{\mathsf{true}}(\theta)\,)}\Big|_0^{R^{\mathsf{true}}(\theta)} - n\cdot e^{-D_B(\,x\,||\,R^{\mathsf{true}}(\theta)\,)}\Big|_{R^{\mathsf{true}}(\theta)}^1 \\
&= 2n - n\cdot e^{-D_B(\,0\,||\,R^{\mathsf{true}}(\theta)\,)} - n\cdot e^{-D_B(\,1\,||\,R^{\mathsf{true}}(\theta)\,)} \\
&\leq 2n
\end{aligned}
$$

□

Now let's go back to the proof of Theorem 1.2. Here we show a property of $D_B(\,p\,||\,q\,)$:

Lemma 1.6: Fixing q, $D_B(\,p\,||\,q\,)$ is a convex function of p. Fixing p, $D_B(\,p\,||\,q\,)$ is a convex function of q.

Proof:

$$\frac{\partial D_B(\,p\,||\,q\,)}{\partial p} = \log\frac{p}{q} - \log\frac{1-p}{1-q}$$
$$\frac{\partial^2 D_B(\,p\,||\,q\,)}{\partial p^2} = \frac{1}{p(1-p)}$$
$$\frac{\partial D_B(\,p\,||\,q\,)}{\partial q} = -\frac{p}{q} + \frac{1-p}{1-q}$$
$$\frac{\partial^2 D_B(\,p\,||\,q\,)}{\partial q^2} = \frac{p}{q^2} + \frac{1-p}{(1-q)^2}$$

Thus, $\frac{\partial^2 D_B(\,p\,||\,q\,)}{\partial p^2} \geq 0$ and $\frac{\partial^2 D_B(\,p\,||\,q\,)}{\partial q^2} \geq 0$. □

Therefore, by Jensen's Inequality,

$$D_B(\mathbb{E}_{\theta\sim\rho}[\, R^{\text{emp}}(\theta)] \,||\, \mathbb{E}_{\theta\sim\rho}[\, R^{\text{true}}(\theta)]\,) \leq \mathbb{E}_{\theta\sim\rho} D_B(\, R^{\text{emp}}(\theta) \,||\, \mathbb{E}_{\theta\sim\rho}[\, R^{\text{true}}(\theta)]\,)$$
$$\leq \mathbb{E}_{\theta\sim\rho}\mathbb{E}_{\theta\sim\rho} D_B(\, R^{\text{emp}}(\theta) \,||\, R^{\text{true}}(\theta)\,) = \mathbb{E}_{\theta\sim\rho} D_B(\, R^{\text{emp}}(\theta) \,||\, R^{\text{true}}(\theta)\,) \tag{8}$$

By Lemma 1.3, let $p = \rho$, $\beta = n-1$ and $x = D_B(\, R^{\text{emp}}(\theta) \,||\, R^{\text{true}}(\theta)\,)$,

$$(n-1)\mathbb{E}_{\theta\sim\rho}[D_B(\, R^{\text{emp}}(\theta) \,||\, R^{\text{true}}(\theta)\,)] - D(\,\rho \,||\, \pi\,)$$
$$\leq \log \mathbb{E}_{\theta\sim\pi}[e^{(n-1)D_B(\, R^{\text{emp}}(\theta) \,||\, R^{\text{true}}(\theta)\,)}] \tag{9}$$

By Lemma 1.4,

$$P_S\left[\mathbb{E}_{\theta\sim\pi}[e^{(n-1)D_B(\, R^{\text{emp}}(\theta) \,||\, R^{\text{true}}(\theta)\,)}] \geq \delta\right] \leq \frac{2n}{\delta} \tag{10}$$

Plugging Eq (9) to Eq (10), we get

$$P_S\left[e^{(n-1)\mathbb{E}_{\theta\sim\rho}[D_B(\, R^{\text{emp}}(\theta) \,||\, R^{\text{true}}(\theta)\,)] - D(\,\rho \,||\, \pi\,)} \geq \delta\right] \leq \frac{2n}{\delta}$$

$$P_S\left[\mathbb{E}_{\theta\sim\rho}[D_B(\, R^{\text{emp}}(\theta) \,||\, R^{\text{true}}(\theta)\,)] \geq \frac{D(\,\rho \,||\, \pi\,) + \log\delta}{n-1}\right] \leq \frac{2n}{\delta} \tag{11}$$

And plugging Eq (8) to Eq (11),

$$P_S\left[D_B(\mathbb{E}_{\theta\sim\rho}[\,R^{\mathsf{emp}}(\theta)]\,||\,\mathbb{E}_{\theta\sim\rho}[\,R^{\mathsf{true}}(\theta)]\,) \geq \frac{D(\,\rho\,||\,\pi\,)+\log\delta}{n-1}\right] \leq \frac{2n}{\delta}$$

Now we finished the proof. □

Contents

Non-vacuous Bound for Deep Neural Network

Suppose f_θ is a neural network and θ is its parameters. θ is initialized according to the prior distribution $\pi(\theta)$ and after training with the data $\boldsymbol{x}$, it is assumed to follow the posterior distribution $\rho(\theta|\boldsymbol{x})$.

We can show that f_θ generalizes well by showing $\mathbb{E}_\rho[R^{\text{true}}(\theta)]$ to be small. Here we choose $\mathbb{E}_\rho[R^{\text{true}}(\theta)]$ instead of $R^{\text{true}}(\theta)$ because after training, we are not likely to obtain a θ that make $R^{\text{true}}(\theta)$ large. That is, the θ that makes $R^{\text{true}}(\theta)$ has higher probability to be obtained. $\mathbb{E}_\rho[R^{\text{true}}(\theta)]$ shows the average true risk obtained by a trained neural network whose parameters follows the distribution ρ.

The θs that forms the distribution ρ are obtained by training the neural network with SGD. Dziugaite and Roy assumes that SGD find good solutions only if these solutions are surrounded by a large volume of equally good solutions. That is, the θs that obtained by SGD with high probability are likely to clustered in some flat area on the optimization landscape rather than scattered. Thus we can assume ρ to be a multivariate Gaussian distribution.

$\mathbb{E}_\rho[R^{\text{true}}(\theta)]$ cannot be calculated directly because the data distribution $\mathcal{D}$ is unknown. We in turn analyze the upper bound of $\mathbb{E}_\rho[R^{\text{true}}(\theta)]$. Dziugaite and Roy use the Inverting KL Bound as the upper bound.

Suppose $p, q \in [0, 1]$, we can use the KL Divergence of Bernoulli Distribution $D_B(\, p \,||\, q \,)$ to measure how far q is away from p. Suppose p is known and q is unknown, we can build an upper bound for q. If we know that $D_B(\, p \,||\, q \,)$ is upper bounded by some $c > 0$, then the upper bound of q can be described as follows:

Definition 2.1 (Inverting KL Bound): For some $p \in [0, 1]$ and $c \geq 0$, the Inverting KL bound of q is defined as

$$D_B^{-1}(\, p \,||\, c \,) := \sup\{q \in [0, 1] : D_B(\, p \,||\, q \,) \leq c\}$$

$D_B^{-1}(\, p \,||\, c \,)$ is the upper bound of q for a fixed p under the restriction $D_B(\, p \,||\, q \,) \leq c$. However, it may not be easy to calculate $D_B^{-1}(\, p \,||\, c \,)$, so we keep on upper bounding it.

By Theorem A.2.1, for any $p, q \in [0,1]$, $D_B(\,p\,||\,q\,) \geq 2(p-q)^2$. If $D_B(\,p\,||\,q\,) \leq c$, then $q \leq p + \sqrt{\frac{c}{2}}$. Note that this holds for any q, including the supremum one, thus $D_B^{-1}(\,p\,||\,c\,) \leq p + \sqrt{\frac{c}{2}}$.

By Langford and Seeger's Bound (Theorem 1.2), we know that with at least $1-\delta$ probability,

$$D_B(\mathbb{E}_{\theta\sim\rho}[\,R^{\mathsf{emp}}(\theta)]\,||\,\mathbb{E}_{\theta\sim\rho}[\,R^{\mathsf{true}}(\theta)]\,) \leq \frac{D(\,\rho\,||\,\pi\,) + \log\frac{2n}{\delta}}{n-1}$$

Thus, $\mathbb{E}_{\theta\sim\rho}[\,R^{\mathsf{true}}(\theta)]$ can be bounded by

$$\begin{aligned}\mathbb{E}_{\theta\sim\rho}[\,R^{\mathsf{true}}(\theta)] &\leq D_B^{-1}\left(\,\mathbb{E}_{\theta\sim\rho}[\,R^{\mathsf{emp}}(\theta)]\;\middle\|\;\frac{D(\,\rho\,||\,\pi\,) + \log\frac{2n}{\delta}}{n-1}\,\right)\\ &\leq \mathbb{E}_{\theta\sim\rho}[\,R^{\mathsf{emp}}(\theta)] + \sqrt{\frac{D(\,\rho\,||\,\pi\,) + \log\frac{2n}{\delta}}{2(n-1)}} \qquad (12)\end{aligned}$$

Eq (12) shows that if $D(\,\rho\,||\,\pi\,)$ is smaller, the generalization gap $\mathbb{E}_{\theta\sim\rho}[\,R^{\mathsf{true}}(\theta)] - \mathbb{E}_{\theta\sim\rho}[\,R^{\mathsf{emp}}(\theta)]$ will be smaller. To make $D(\,\rho\,||\,\pi\,)$ computable, we assume ρ and π to be some specific distribution. As we assumed ρ to be multivariate Gaussian, we also assume π to be multivariate Gaussian, for the purpose of simplify the calculation of $D(\,\rho\,||\,\pi\,)$.

Assume ρ to be $\mathsf{N}(w, \mathsf{diag}(s))$ and π to be $\mathsf{N}(w_0, \lambda I_d)$, where $w, w_0, \lambda \in \mathbb{R}$ and $s \in \mathbb{R}_+^d$ are adjustable parameters, we have

$$\sqrt{\frac{D(\,\rho\,||\,\pi\,) + \log\frac{2n}{\delta}}{2(n-1)}} = \sqrt{\frac{D(\,\mathsf{N}(w, \mathsf{diag}(s))\,||\,\mathsf{N}(w_0, \lambda I_d)\,) + \log\frac{2n}{\delta}}{2(n-1)}} \tag{13}$$

where by Corollary A.2.7,

$$D(\,\mathsf{N}(w, \mathsf{diag}(s))\,||\,\mathsf{N}(w_0, \lambda I_d)\,) = \frac{1}{2}[d(\log\lambda - 1) - \sum_{i=1}^{d}(\log s_i - \frac{1}{\lambda}s_i) + \frac{\|w - w_0\|^2}{\lambda}]$$

We initialize w_0 and λ and solve w, s. However, we do not know which λ will make the upper bound tightest, so we need to try different values for λ and see which one gives the tightest bound.

There are different ways to search the optimal λ, one way is to define $\lambda = ce^{-j/b}$ for some $j \in \mathbb{N}^+$ and fixed $b, c > 0$. Here b determines the level of precision and c is the upper bound. We choose different λ by adjusting j.

Let's define the upper bound as

$$B(j; \delta_j) = \sqrt{\frac{D(\,\mathsf{N}(w, \mathsf{diag}(s)) \,||\, \mathsf{N}(w_0, \lambda I_d)\,) + \log \frac{2n}{\delta_j}}{2(n-1)}}, \quad \text{where } \lambda = ce^{-j/b}$$

And define the event

$$E_j : \mathbb{E}_{\theta \sim \rho}[\, R^{\mathsf{true}}(\theta)] > \mathbb{E}_{\theta \sim \rho}[\, R^{\mathsf{emp}}(\theta)] + B(j; \delta_j)$$

If

$$P[E_j] < \frac{6\delta}{\pi^2 j^2} \tag{14}$$

Then

$$P\left[\bigcup_{j=1}^{\infty} E_j\right] < \sum_{j=1}^{\infty} P[E_j] = \delta \frac{6}{\pi^2} \sum_{j=1}^{\infty} \frac{1}{j^2} = \delta$$

That is

$$P\left[\bigcap_{j=1}^{\infty} \neg E_j\right] \geq 1 - \delta$$

This means, if we want to try different js and use Eq (14) as the bound for each j, then with the same probability $1 - \delta$, $\neg E_j$ will be true for any j. Thus we can pick one of the $\neg E_j$s that gives the tightest bound.

Thus, for a specific j, we calculate the upper bound with Eq (14):

$$P[\neg E_j] \geq 1 - \frac{6\delta}{\pi^2 j^2}$$

Let $\delta_j = \frac{6\delta}{\pi^2 j^2}$, the upper bound for $\neg E_j$ is

$$B(j; \frac{6\delta}{\pi^2 j^2}) = \sqrt{\frac{D(\,\mathsf{N}(w, \mathsf{diag}(s)) \,||\, \mathsf{N}(w_0, \lambda I_d)\,) + \log \frac{n\pi^2}{3\delta} + 2\log j}{2(n-1)}}$$

Since $j = b \log \frac{c}{\lambda}$, the upper bound can be represented using only λ:

$$B(\lambda) := \sqrt{\frac{D(\,\mathsf{N}(w, \mathsf{diag}(s)) \,||\, \mathsf{N}(w_0, \lambda I_d)\,) + \log \frac{n\pi^2}{3\delta} + 2\log(b \log \frac{c}{\lambda})}{2(n-1)}}$$

Now we can search λ in $(0, c)$ to minimize the upper bound. Note that j is discrete, so that λ is discrete, which is not easy for optimization. We treat λ to be continuous variable and later round it to the one of integer j [1].

Let's go back to the term $\mathbb{E}_{\theta\sim\rho}[\, R^{\mathsf{emp}}(\theta)]$. Expand it:

$$\mathbb{E}_{\theta\sim\rho}[\, R^{\mathsf{emp}}(\theta)] = \int R^{\mathsf{emp}}(\theta) \frac{1}{\sqrt{(2\pi)^d \prod_{i=1}^n s_i}} e^{-\frac{1}{2}\sum_{i=1}^n \frac{(\theta_i - w_i)^2}{s_i}} d\theta$$

where $R^{\mathsf{emp}}(\theta) = \frac{1}{n}\sum_{i=1}^n l(f_\theta(x_i) - y_i)$. We want to search for the w, s using gradient descent. However, the gradient is difficult to calculate.

[1] What will happen if we treat both λ and j as continuous variables? j will have the continuous distribution $P(j = t) = \frac{1}{t^2}, t \in \mathbb{R}, t \geq 1$ instead of the discrete distribution $P(j = t) = \frac{6}{\pi^2 t^2}, t \in \mathbb{N}^+$. And no round step is required. This can be a way for improvement.

Using the gradient of s for example. Let $\frac{\partial f}{\partial s} = [\frac{\partial f}{\partial s_1}, \frac{\partial f}{\partial s_2}, ..., \frac{\partial f}{\partial s_d}]^T$. For a single θ,

$$\frac{\partial}{\partial s} R^{\text{emp}}(\theta) \frac{1}{\sqrt{(2\pi)^d \prod_{i=1}^n s_i}} e^{-\frac{1}{2}\sum_{i=1}^n \frac{(\theta_i - w_i)^2}{s_i}}$$

$$= R^{\text{emp}}(\theta) \frac{1}{\sqrt{(2\pi)^d \prod_{i=1}^n s_i}} e^{-\frac{1}{2}\sum_{i=1}^n \frac{(\theta_i - w_i)^2}{s_i}} [-\frac{1}{2s} + \frac{(\theta - w)^2}{2s^2}]$$

where the square and division are element-wise. For all the θs,

$$\frac{\partial}{\partial s} \mathbb{E}_{\theta\sim\rho}[\, R^{\text{emp}}(\theta)] = \int \frac{\partial}{\partial s} R^{\text{emp}}(\theta) \frac{1}{\sqrt{(2\pi)^d \prod_{i=1}^n s_i}} e^{-\frac{1}{2}\sum_{i=1}^n \frac{(\theta_i - w_i)^2}{s_i}} d\theta$$

$$= \mathbb{E}_{\theta\sim\rho}\left[R^{\text{emp}}(\theta)[-\frac{1}{2s} + \frac{(\theta - w)^2}{2s^2}]\right]$$

$$= -\frac{1}{2s}\mathbb{E}_{\theta\sim\rho}[\, R^{\text{emp}}(\theta)] + \frac{1}{2s^2}\mathbb{E}_{\theta\sim\rho}[\, R^{\text{emp}}(\theta)(\theta - w)^2]$$

However, calculating $\mathbb{E}_{\theta\sim\rho}[\, R^{\text{emp}}(\theta)]$ and $\mathbb{E}_{\theta\sim\rho}[\, R^{\text{emp}}(\theta)(\theta - w)^2]$ is impractical since we cannot calculate the integral explicitly. Here we can instead compute an unbiased estimate and update with SGD. To achieve this, we can sample θ' according to $\mathsf{N}(w, \mathsf{diag}(s))$, and use the gradient $\frac{\partial}{\partial s} R^{\text{emp}}(\theta')$ in each iteration.

Note that w, s are changing, so we sample θ' from different iteration each time. This process is very like online learning as we calculate the stochastic gradient of each sample and the samples come in a sequence whose distribution may change.

One more thing about $\mathbb{E}_{\theta\sim\rho}[\, R^{\text{emp}}(\theta)]$ is that, in Langford and Caruana's paper, the loss is defined as $0-1$ loss, i.e., $l(f_\theta(x_i), y_i) = \mathbf{1}[f_\theta(x_i) \neq y_i] \in \{0, 1\}$ where $f_\theta(x_i) \in \{-1, 1\}$ and $y_i \in \{-1, 1\}$. Remember that the θ of the posterior distribution is found by training the neural network. The $0-1$ loss is not easy to be trained since its discrete. To show that the upper bound can be applied to a trainable neural network, Dziugaite and Roy replacing the $0-1$ loss with its convex surrogate:

$$\breve{l}(f_\theta(x_i), y_i) = \frac{1}{\log 2} \log(1 + e^{-f_\theta(x_i) y_i}) \tag{15}$$

where $f_\theta(x_i) \in \mathbb{R}$ and $y_i \in \{-1, 1\}$.

In order to apply Theorem 1.2, we need $\breve{l} \in [0, 1]$. How can this be guaranteed? Based on the empirical observation by Zhang et al [13], the overparameterized neural network tends to fit any data (even the data is randomly labeled) and make the training loss close to 0. We assume that $\breve{l} \in [0, 1]$ is true.

Lastly, even we find w and s and determine ρ, we still cannot calculate $\mathbb{E}_{\theta\sim\rho}[\, R^{\text{emp}}(\theta)]$ in Eq (12) effectively since the integral of θ is not explicit. Remember that we can approximate any integral using Monte Carlo method [2]: given n i.i.d. samples $\theta_1, ..., \theta_m$ from ρ, let $\widehat{\mathbb{E}}_{\theta\sim\rho}[\, R^{\text{emp}}(\theta)] = \frac{1}{m}\sum_{i=1}^{m} R^{\text{emp}}(\theta_i)$. Then $\widehat{\mathbb{E}}_{\theta\sim\rho}[\, R^{\text{emp}}(\theta)]$ is the Monte Carlo estimator of $\mathbb{E}_{\theta\sim\rho}[\, R^{\text{emp}}(\theta)]$ and $\mathbb{E}[\widehat{\mathbb{E}}_{\theta\sim\rho}[\, R^{\text{emp}}(\theta)]] = \mathbb{E}_{\theta\sim\rho}[\, R^{\text{emp}}(\theta)]$.

By Theorem A.2.4, we have

$$P\left[D_B(\, \widehat{\mathbb{E}}_{\theta\sim\rho}[\, R^{\text{emp}}(\theta)] \,||\, \mathbb{E}_{\theta\sim\rho}[\, R^{\text{emp}}(\theta)]\,) \leq \frac{\log\frac{1}{\delta'}}{m} \right] \geq 1-\delta' \tag{16}$$

This holds since we assume $\breve{l} \in [0,1]$ such that $R^{\text{emp}}(\theta_i) \in [0,1]$. Theorem A.2.4 holds not only for the Bernouli distribution but also for any distribution in $[0,1]$, because Lemma 1.5 holds for any distribution in $[0,1]$.

[2] https://ib.berkeley.edu/labs/slatkin/eriq/classes/guest_lect/mc_lecture_notes.pdf

Eq (16) implies

$$P\left[\mathbb{E}_{\theta\sim\rho}[\,R^{\mathsf{emp}}(\theta)] \le \widehat{\mathbb{E}}_{\theta\sim\rho}[\,R^{\mathsf{emp}}(\theta)] + \sqrt{\frac{\log\frac{1}{\delta'}}{m}}\right] \ge 1-\delta' \tag{17}$$

Combining Eq (17) with our earlier bound

$$P\left[\mathbb{E}_{\theta\sim\rho}[\,R^{\mathsf{true}}(\theta)] \le \mathbb{E}_{\theta\sim\rho}[\,R^{\mathsf{emp}}(\theta)] + B(\lambda)\right] \ge 1-\delta \tag{18}$$

where

$$B(\lambda) = \sqrt{\frac{D(\,\mathsf{N}(w,\mathsf{diag}(s))\,||\,\mathsf{N}(w_0,\lambda I_d)\,) + \log\frac{n\pi^2}{3\delta} + 2\log(b\log\frac{c}{\lambda})}{2(n-1)}}$$

Define the events

$$E_1: \quad \mathbb{E}_{\theta\sim\rho}[\,R^{\mathsf{emp}}(\theta)] \le \widehat{\mathbb{E}}_{\theta\sim\rho}[\,R^{\mathsf{emp}}(\theta)] + \sqrt{\frac{\log\frac{1}{\delta'}}{m}}$$

$$E_2: \quad \mathbb{E}_{\theta\sim\rho}[\,R^{\mathsf{true}}(\theta)] \le \mathbb{E}_{\theta\sim\rho}[\,R^{\mathsf{emp}}(\theta)] + B(\lambda)$$

Then $P[\neg E_1] \leq \delta'$ and $P[\neg E_2] \leq \delta$. Thus

$$P[\neg E_1 \cup \neg E_2] \leq P[\neg E_1] + P[\neg E_2] = \delta + \delta'$$

which implies $P[E_1 \cap E_2] \geq 1 - \delta - \delta'$.

Define the event

$$E_3 : \quad \mathbb{E}_{\theta \sim \rho}[\, R^{\mathsf{true}}(\theta)] \leq \widehat{\mathbb{E}}_{\theta \sim \rho}[\, R^{\mathsf{emp}}(\theta)] + \sqrt{\frac{\log \frac{1}{\delta'}}{m}} + B(\lambda)$$

Then $E_1 \cap E_2 \Rightarrow E_3$, $P(E_3) \geq P(E_1 \cap E_2)$. Thus we have the final bound

$$P\left[\mathbb{E}_{\theta \sim \rho}[\, R^{\mathsf{true}}(\theta)] \leq \widehat{\mathbb{E}}_{\theta \sim \rho}[\, R^{\mathsf{emp}}(\theta)] + \sqrt{\frac{\log \frac{1}{\delta'}}{m}} + B(\lambda)\right] \geq 1 - \delta - \delta' \qquad (19)$$

To sum up, the algorithm to obtain the tightest PAC-Bayes Bound is as follows:

Algorithm 2.2 (Calculate tightest PAC-Bayes Bound):

1. Using the Bound Eq (19). Give a dataset $S = \{(x_i, y_i)\}_{i=1}^n, x_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$. Initialize $w_0, m, b, c, \delta, \delta'$.
2. Try different λs in $(0, c)$. For each λ, initialize the parameters θ of the neural network according to $\mathsf{N}(w_0, \lambda I_d)$. Then find w, s by optimizing the following problem using SGD

$$\min_{w \in \mathbb{R}^d, s \in \mathbb{R}_+^d} \mathbb{E}_{\theta \sim \rho}[\, R^{\mathsf{emp}}(\theta)] + B(\lambda)$$

where

$$\mathbb{E}_{\theta \sim \rho}[\, R^{\mathsf{emp}}(\theta)] = \frac{1}{n} \breve{l}(f_\theta(x_i), y_i)$$

3. Let the minimum upper bound

$$\min_{\lambda} \ \widehat{\mathbb{E}}_{\theta \sim \rho}[\, R^{\mathsf{emp}}(\theta)] + \sqrt{\frac{\log \frac{1}{\delta'}}{m}} + B(\lambda)$$

be the tightest PAC-Bayes bound.

Experiments

Experiment Settings:

- Dataset: MNIST handwritten digits dataset. 55000 images as training set and 10000 images as test set. Each image is of size 28×28 and each pixel is either 0 (black) or 1 white. The dataset is relabeled: $\{0, 1, ..., 4\}$ is mapped to 1 and $\{5, 6, ..., 9\}$ is mapped to -1.
- Model: Using fully connected neural network with 2 or more layers. The activation function is ReLU. The loss function is defined in Eq (15).
- $\delta = 0.025, \delta' = 0.01, m = 150000$.

There is one problem in Algorithm 2.2. We get the tightest PAC-Bayes Bound together with the posterior ρ, which is defined by w, s. If we simply use SGD to optimize $R^{\text{emp}}(\theta)$ (which is our usual training step), suppose the obtained θ satisfy a posterior distribution ρ', how can we be sure that ρ' is the same as ρ? Remember that ρ is obtained by minimizing the PAC-Bayes bound using SGD.

However, even we cannot guarantee ρ is the same as ρ', we can still make sure that ρ will have the same mean as ρ'. Consider the following experiment steps.

Experiment Steps:

1. Train the neural network by optimizing $R^{\text{emp}}(\theta)$ using SGD. Let w be the obtained θ.
2. Run Algorithm 2.2 by fixing w as the one obtained in step 1, optimizing the PAC-Bayes bound by simply adjusting s.
3. Repeat step 1 and step 2 multiple times and we get different (w, s) pairs and PAC-Bayes Bounds.

In this way, Algorithm 2.2 will generate a posterior distribution which is a multivariate Gaussian with mean w. We will consider this posterior distribution as the possible one that obtained by minimizing $R^{\text{emp}}(\theta)$.

Thus, the PAC-Bayes bound will be restricted to any multivariate Gaussian posterior distribution with mean w. We will use this bound in the experiment.

Moreover, since $s \in \mathbb{R}_+^d$, we could use $\log s'$ to replace s where $s' \in \mathbb{R}^d$. Thus we turn the constrained optimization into an unconstrained one.

Experiment Results:

Below shows the results of the experiment:

Experiment	T-600	T-1200	T-300^2	T-600^2	T-1200^2	T-600^3	R-600
Train error	0.001	0.002	0.000	0.000	0.000	0.000	0.007
Test error	0.018	0.018	0.015	0.016	0.015	0.013	0.508
SNN train error	0.028	0.027	0.027	0.028	0.029	0.027	0.112
SNN test error	0.034	0.035	0.034	0.033	0.035	0.032	0.503
PAC-Bayes bound	0.161	0.179	0.170	0.186	0.223	0.201	1.352
KL divergence	5144	5977	5791	6534	8558	7861	201131
# parameters	471k	943k	326k	832k	2384k	1193k	472k
VC dimension	26m	56m	26m	66m	187m	121m	26m

Optimizing the empirical risk $R^{\text{emp}}(\theta)$ and get the weight $\theta = w$. Train error and test error are obtained on the neural network with weight w.

Optimizing the PAC-Bayes Bound by Algorithm 2.2 with fixed w, we get the tightest PAC-Bayes bound and the s.

We call the exact value of $\mathbb{E}_{\theta\sim\rho}[R^{\text{emp}}(\theta)]$ on training set as SNN train error and $\mathbb{E}_{\theta\sim\rho}[R^{\text{emp}}(\theta)]$ on test set as SNN test error, where SNN means Stochastic Neural Network [7]. Use w, s to construct the distribution $\rho = \mathsf{N}(w, \mathsf{diag}(s))$, and calculate $\mathbb{E}_{\theta\sim\rho}[R^{\text{emp}}(\theta)]$ empirically by sampling θs according to ρ.

Note that test error is not true risk since it is evaluated on another set of samples from $\mathcal{D}$. However, we can consider Test Error $-$ Train Error as an approximation of $R^{\text{true}}(\theta) - R^{\text{emp}}(\theta)$, SNN Test Error $-$ SNN Train Error as an approximation of $\mathbb{E}_{\theta\sim\rho}[R^{\text{true}}(\theta)] - \mathbb{E}_{\theta\sim\rho}[R^{\text{emp}}(\theta)]$, and BAC-Bayes Bound is the upper bound of $\mathbb{E}_{\theta\sim\rho}[R^{\text{true}}(\theta)] - \mathbb{E}_{\theta\sim\rho}[R^{\text{emp}}(\theta)]$.

Each network is trained either on (T) true labels and (R) random labels. The network architecture is expressed as N^L, indicating L hidden layers and N nodes each.

Explaining the Results:

- All trained neural networks achieves nearly perfect training accuracy.
- On true labels, the test error does not change much on different network architectures, which means it is close to the Bayes Risk.
- On true labels, neural network generalizes well because the difference between test error and train error is low. On random labels, test error increases to 0.508 because lacking of generalization. The same as SNN train and test error.
- We call a bound vacuous when it is greater than 1, since $\mathbb{E}_{\theta\sim\rho}[R^{\mathsf{true}}(\theta)] - \mathbb{E}_{\theta\sim\rho}[R^{\mathsf{emp}}(\theta)] \leq 1$. PAC-Bayes bound is non-vacuous on true labels but vacuous on random labels.
- The PAC-Bayes bound does not grow much despite the size of the network grows several times, which is different from the traditional PAC bound.

Conclusion

Based on Langford and Seeger's PAC-Bayes Bound and Langford and Caruana's earlier work on shallow neural networks, Dziugaite and Roy proposed a method to calculate PAC-Bayes Bound for deep neural network, which is the first work showing a theoretical generalization bound can be non-vacuous when applied to deep neural network.

Traditional PAC bounds are vacuous because they do not consider the distribution of parameters. Some parameters, though, can make the generalization gap large, but they are not likely to be obtained in the training process. PAC-Bayes bound is non-vacuous because it considers the distribution of parameters – they are not likely randomly distributed but satisfy a posterior distribution. Those parameters that drawn from the posterior distribution with high probability can usually make the generalization gap small. The posterior distribution depends on the data.

Although the PAC-Bayes bound proposed by Dziugaite and Roy is not very tight and still has room to improve, it has shown to be non-vacuous, and inspired many following papers for tighter PAC-Bayes Bounds, see [1].

Reference

[1] Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. https://arxiv.org/pdf/2110.11216.pdf.

[2] Bayesian Estimation https://stats.stackexchange.com/questions/58564/help-me-understand-bayesian-prior-and-posterior-distributions

[3] Uninformative Priors https://stats.stackexchange.com/questions/20520/what-is-an-uninformative-prior-can-we-ever-have-one-with-truly-no-information

[4] Relative Entropy of Chernoff Bound https://www.cs.princeton.edu/courses/archive/spr08/cos511/scribe_notes/0227.pdf

[5] John Langford and Matthias Seeger. Bounds for averaging classifiers. School of Computer Science, Carnegie Mellon University, 2001. https://www.cs.cmu.edu/~jcl/papers/averaging/averaging_tech.pdf (Important!)

Reference

[6] Gintare Karolina Dziugaite, and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. arXiv preprint arXiv:1703.11008, 2017. https://arxiv.org/pdf/1703.11008.pdf. (Important!)

[7] John Langford and Rich Caruana. (Not) bounding the true error. Advances in Neural Information Processing Systems, 2001. https://proceedings.neurips.cc/paper_files/paper/2001/file/98c7242894844ecd6ec94af67ac8247d-Paper.pdf

[8] Pinsker's Inequality Proof. https://home.ttic.edu/~madhurt/courses/infotheory2014/l5.pdf

[9] Thomas Cover and Thomas Joy. Elements of Information Theory, Second Edition. 2006.

[10] John Langford and Robert Schapire. Tutorial on Practical Prediction Theory for Classification. Journal of machine learning research 6.3, 2005. https://www.jmlr.org/papers/volume6/langford05a/langford05a.pdf

Reference

[11] KL-Divergence of Multivariate Gaussians Proof. https://statproofbook.github.io/P/mvn-kl.html.

[12] Trace of Quadratic Form. https://math.stackexchange.com/questions/2228398/trace-trick-for-expectations-of-quadratic-forms

[13] Chiyuan Zhang, Samy Bengio, Moritz Hardt, et al. Understanding deep learning requires rethinking generalization. International Conference on Learning Representations, 2017. https://arxiv.org/abs/1611.03530

[14] David A. McAllester. PAC-Bayesian model averaging. Proceedings of the twelfth annual conference on Computational learning theory. 1999. https://dl.acm.org/doi/pdf/10.1145/307400.307435

[15] Olivier Catoni. A PAC-Bayesian approach to adaptive classification. preprint LPMA 840, 2003. http://yaroslavvb.com/papers/notes/catoni-pac.pdf

[16] Stochastic Dominance https://en.wikipedia.org/wiki/Stochastic_dominance

Appendix 1: Greedy Strategy for Maximizing Expectation

Theorem A.1.1: Let $x \in [a, b]$ and $h(x)$ be a continuous monotonically increasing function. Let $f(x)$ and $g(x)$ be two PDFs such that $\int_a^b f(x)dx = \int_a^b g(x)dx = 1$. Let $F(x) = \int_x^b f(t)dt$. If for any x, $\int_x^b g(t)dt \leq F(x)$, then $\mathbb{E}_{g(x)}[h(x)] \leq \mathbb{E}_{f(x)}[h(x)]$.

Proof: Suppose there exist $\xi_1, \xi_2, ..., \xi_n$ such that $a = \xi_0 \leq \xi_1 \leq \xi_2 \leq ... \leq \xi_n \leq \xi_{n+1} = b$, $f(\xi_i) = g(\xi_i)$ for $i = 1, 2, ..., n$, and $f(x) - g(x)$ does not change sign in each $[\xi_i, \xi_{i+1}]$ for $i = 0, 1, ..., n$. Then

$$\begin{aligned} \mathbb{E}_{f(x)}[h(x)] - \mathbb{E}_{g(x)}[h(x)] &= \int_a^b [f(x) - g(x)]h(x)dx \\ &= \sum_{i=0}^{n} \int_{\xi_i}^{\xi_{i+1}} [f(x) - g(x)]h(x)dx \end{aligned} \tag{20}$$

For each interval $[\xi_i, \xi_{i+1}]$, by Mean Value Theorem, there exists $x_i \in [\xi_i, \xi_{i+1}]$ such that

$$\int_{\xi_i}^{\xi_{i+1}} [f(x) - g(x)]h(x)dx = h(x_i) \int_{\xi_i}^{\xi_{i+1}} [f(x) - g(x)]dx$$

Since $h(x)$ is monotonically increasing, we have $h(x_i) \leq h(x_{i+1})$. Thus we can write Eq (20) as

$$\mathbb{E}_{f(x)}[h(x)] - \mathbb{E}_{g(x)}[h(x)] = \sum_{i=0}^{n} h(x_i) \int_{\xi_i}^{\xi_{i+1}} [f(x) - g(x)]dx$$

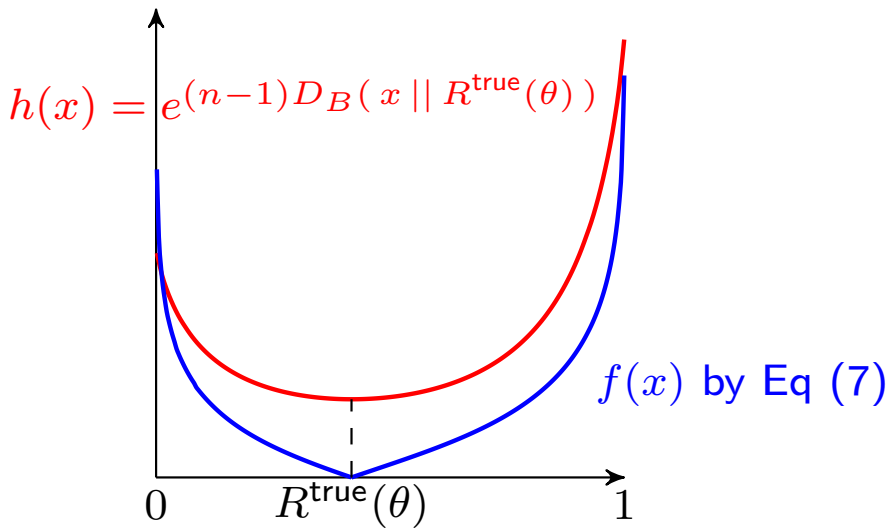
$$= h(x_0) \int_a^b [f(x) - g(x)]dx + \sum_{i=1}^{n} [h(x_i) - h(x_{i-1})] \int_{\xi_i}^{b} [f(x) - g(x)]dx$$

Since $h(x_0) \int_a^b [f(x) - g(x)]dx = 0$, and $\int_{\xi_i}^b [f(x) - g(x)]dx \geq 0$ for any $\xi_i \in [a, b]$, we have

$$\mathbb{E}_{f(x)}[h(x)] - \mathbb{E}_{g(x)}[h(x)] \geq 0$$

□

Theorem A.1.1 indicates the greedy strategy to maximize the expectation: For a sequence of values $x_1, x_2, ..., x_n$ where $x_1 \leq x_2 \leq ... \leq x_n$, start from the largest element x_n, assign as much probability $g(x_n)$ to it as possible until meet the upper bound $f(x_n)$. Then go to x_{n-1}, repeat this step, and go to x_{n-2}... Any probability that reduced from x_i and assigned to x_j where $x_i > x_j$ will always make the expectation $\mathbb{E}_{g(x)}[X]$ smaller. This approach is also known as Stochastic Dominance [16].



In the proof of Lemma 1.4, $h(x)$ is larger when the x is more far away from $R^{\text{true}}(\theta)$. To maximize $\mathbb{E}[h(x)]$, we assign larger probability $f(x)$ to larger $h(x)$.

Appendix 2: Theorems of KL Divergence

Theorem A.2.1: For any $p, q \in [0, 1]$,

$$D_B(\, p \,||\, q \,) \geq 2(p-q)^2$$

Proof: For any p, let $g(q) = D_B(\, p \,||\, q \,) - 2(p-q)^2$,

$$\frac{dg(q)}{dq} = -\frac{p}{q} + \frac{1-p}{1-q} - 4(q-p) = [\frac{1}{q(1-q)} - 4](q-p)$$

Since $q(1-q) \leq \frac{1}{4}$, for any $q \leq p$, we have $\frac{dg(q)}{dq} \leq 0$. $\frac{dg(q)}{q} = 0$. $g(q) = 0$ when $q = p$. Therefore, $g(q) \geq 0$ when $q \leq p$, that is, $D_B(\, p \,||\, q \,) \geq 2(p-q)^2$.

When $q > p$, let $q' = 1 - q$, $p' = 1 - p$, then $q' \leq p'$. We have

$$D_B(\, p \,||\, q \,) = D_B(\, 1-p \,||\, 1-q \,) = D_B(\, p' \,||\, q' \,) \geq 2(p'-q')^2 = 2(p-q)^2$$

□

Theorem A.2.1 has a general case called Pinsker's Inequality, which gives an lower bound of KL-Divergence of any two probability distributions.

Theorem A.2.2 (Pinsker's Inequality): Let $x \in \mathcal{X}$, and $p(X), q(X)$ be two PDFs. Then,

$$D(\, p(X) \,||\, q(X) \,) \geq \frac{1}{2} \|p(X) - q(X)\|_1^2$$

where

$$\|p(X) - q(X)\|_1^2 = \left(\sum_{x \in \mathcal{X}} |p(x) - q(x)| \right)^2$$

Proof: To prove Theorem A.2.2, we will ultilize Lemma A.2.3 and Lemma A.2.4.

Lemma A.2.3 (Chain of KL-Divergence): Let $p(X, Y)$ and $q(X, Y)$ be two distributions for a pair of random variables, then

$$D(\, p(X,Y) \,||\, q(X,Y) \,) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x,y)} = \sum_{x,y} p(y|x)p(x) \log \frac{p(y|x)p(x)}{q(y|x)q(x)}$$

$$= \sum_{x} p(x) \log \frac{p(x)}{q(x)} \sum_{y} p(y|x) + \sum_{x} p(x) \sum_{y} p(y|x) \log \frac{p(y|x)}{q(y|x)}$$

$$= D(\, p(X) \,||\, q(X) \,) + \sum_{x} p(x)\, D(\, p(Y) \,||\, q(Y) \,|X = x)$$

$$= D(\, p(X) \,||\, q(X) \,) + D(\, p(Y) \,||\, q(Y) \,|X) \tag{21}$$

where $D(\, p(Y) \,||\, q(Y) \,|X)$ denote the expectation of $D(\, p(Y|X = x) \,||\, q(Y|X = x) \,)$ with respect to the distribution $p(X)$.

Eq (21) expands $D(\, p(X, Y) \,||\, q(X, Y) \,)$ by conditioning on X. We can also condition on Y:

$$D(\, p(X, Y) \,||\, q(X, Y) \,) = D(\, p(Y) \,||\, q(Y) \,) + D(\, p(X) \,||\, q(X) \,|Y)$$

Thus we have

$$D(\, p(X) \,||\, q(X) \,) + D(\, p(Y) \,||\, q(Y) \,|X) = D(\, p(Y) \,||\, q(Y) \,) + D(\, p(X) \,||\, q(X) \,|Y) \tag{22}$$

Define $A = \{x : p(x) > q(x)\}$ and let $Y = \mathbf{1}[x \in A]$. Then Y can be either 0 or 1, and $p(Y)$ and $q(Y)$ are two different Bernoulli distributions. Let $\tilde{p} = p(Y = 1)$ and $\tilde{q} = q(Y = 1)$, then

$$D(\, p(Y) \,||\, q(Y) \,) = D_B(\, \tilde{p} \,||\, \tilde{q} \,)$$

Consider the conditional distribution $p(Y|X)$, when x is given, we surely know if x is in A or not. If $x \in A$, then $p(Y = 1|X = x) = 1$; otherwise, $p(Y = 1|X = x) = 0$. The distribution is unique, so q will be the same distribution as p.

Therefore,

$$D(\,p(Y)\,||\,q(Y)\,|X) = \sum_{x\colon x\in A} p(x)\,D_B(\,1\,||\,1\,|X=x) + \sum_{x\colon x\notin A} p(x)\,D_B(\,0\,||\,0\,|X=x) = 0$$

Now we can write Eq (22) as

$$D(\,p(X)\,||\,q(X)\,) = D(\,\tilde{p}\,||\,\tilde{q}\,) + D(\,p(X)\,||\,q(X)\,|Y)$$

Since $D(\,p(X)\,||\,q(X)\,|Y) \geq 0$, we have

$$D(\,p(X)\,||\,q(X)\,) \geq D(\,\tilde{p}\,||\,\tilde{q}\,) \tag{23}$$

Lemma A.2.4: Let the $\mathcal{L}_1$ distance of any two distributions $p(x)$ and $q(x)$ be

$$\|p(X) - q(X)\|_1 = \sum_{x\in\mathcal{X}} |p(x) - q(x)|$$

Let $A = \{x : p(x) > q(x)\}$, $Y = \mathbf{1}[x \in A]$, $\tilde{p} = p(Y = 1)$ and $\tilde{q} = q(Y = 1)$, then

$$\|p(X) - q(X)\|_1 = 2(\tilde{p} - \tilde{q}) \tag{24}$$

Proof:

$$\begin{aligned}
\|p(X) - q(X)\|_1 &= \sum_{x \in \mathcal{X}} |p(x) - q(x)| \\
&= \sum_{x \in A} [p(x) - q(x)] + \sum_{x \in A^c} [q(x) - p(x)] \\
&= p(Y = 1) - q(Y = 1) + q(Y = 0) - p(Y = 0) \\
&= 2(p(Y = 1) - q(Y = 1)) = 2(\tilde{p} - \tilde{q})
\end{aligned}$$

Now let's go back to the proof of Theorem A.2.2. We have

$$\begin{aligned}
D(\, p(X) \,||\, q(X) \,) &\geq D(\, \tilde{p} \,||\, \tilde{q} \,) \qquad \text{[by Eq (23)]} \\
&\geq 2(\tilde{p} - \tilde{q})^2 \qquad \text{[by Theorem A.2.1]} \\
&= \frac{1}{2} \|p(X) - q(X)\|_1^2 \qquad \text{[by Eq (24)]}
\end{aligned}$$

□

Theorem A.2.5 (Sample Convergence)[10][6][7]: Assume random variables $X_1, X_2, ..., X_n$ are independent and each X_i is from the Bernoulli distribution Ber(p). Let $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$, then

$$P\left[D_B(\,\hat{p}\,||\,p\,) \leq \frac{\log\frac{1}{\delta}}{n}\right] \geq 1-\delta$$

Proof: First we prove that the CDF of any distribution is a random variable following the uniform distribution.

Let X be a random variable, $F_X(x)$ be the CDF of x. Note that $F_X(x)$ is non-decreasing with x. Let $Y = F_X(x)$, then $0 \leq Y \leq 1$. Define $F_X^{-1}(y) = \max\{x : F_X(x) = y\}$. We have

$$F_Y(y) = P(F_X(x) \leq y) = P(x \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y$$

This holds when both $F_X(x)$ and y are continuous, when $F_X(x)$ is discrete and y is continuous, we have

$$P(F_X(x) \leq y) \leq y$$

Since $\hat{p}$ follows the Binomial distribution, by Lemma 1.5, we have the Chernoff Bound for the Binomial:

$$P[\hat{p} \geq p + \epsilon] \leq e^{-D_B(\,p+\epsilon\,||\,p\,)n}$$
$$P[\hat{p} \leq p - \epsilon] \leq e^{-D_B(\,p-\epsilon\,||\,p\,)n}$$

The Chernoff Bound is an upper bound of the CDF. The bound reaches maximum 1 when $\epsilon = 0$. The bound is a random variable following the uniform distribution on $[a, 1]$ where $a \geq 0$.

Since $\hat{p}$ is discrete, $e^{-D_B(\,\hat{p}\,||\,p\,)n}$ is discrete. Thus we have

$$P\left[D_B(\,\hat{p}\,||\,p\,) \leq \frac{\log\frac{1}{\delta}}{n}\right] = P\left[e^{-D_B(\,\hat{p}\,||\,p\,)n} \geq \delta\right]$$
$$= 1 - P\left[e^{-D_B(\,\hat{p}\,||\,p\,)n} \leq \delta\right] = 1 - \Big(P\left[e^{-D_B(\,\hat{p}\,||\,p\,)n} \leq \delta \,|\, \hat{p} \geq p\right] P[\hat{p} \geq p]$$
$$+ P\left[e^{-D_B(\,\hat{p}\,||\,p\,)n} \leq \delta \,|\, \hat{p} \leq p\right] P[\hat{p} \leq p]\Big)$$
$$\geq 1 - \delta(P[\hat{p} \geq p] + P[\hat{p} \leq p]) = 1 - \delta$$

□

Definition (Multivariate Gaussian): The PDF of the Multivariate Gaussian Distribution $\mathsf{N}(\mu, \Sigma)$ is

$$f(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

where $x \in \mathbb{R}^d$ is the vector of d variables. $\mu = \mathbb{E}[x] \in \mathbb{R}^d$ is the mean.
$\Sigma = (x-\mu)(x-\mu)^T \in \mathbb{R}^{d \times d}$ is the covariance matrix.

Theorem A.2.6 (KL-Divergence of Multivariate Gaussians): Let $p(x)$ be the PDF of $\mathsf{N}(\mu_p, \Sigma_p)$ and $q(x)$ be the PDF of $\mathsf{N}(\mu_q, \Sigma_q)$. $\mu_p, \mu_q \in \mathbb{R}^d$ and $\Sigma_p, \Sigma_q \in \mathbb{R}^{d \times d}$. Then the KL-Divergence of $p(x)$ and $q(x)$ is

$$D(\, p(x) \,||\, q(x) \,) = \frac{1}{2} \left[(\mu_q - \mu_p) \Sigma_q^{-1} (\mu_q - \mu_p)^T + \mathsf{tr}(\Sigma_q^{-1} \Sigma_p) + \log \frac{\det \Sigma_q}{\det \Sigma_p} - d \right]$$

Proof:

$$D(\,p(x)\,||\,q(x)\,) = \mathbb{E}_{p(x)}\left[\log\frac{p(x)}{q(x)}\right]$$

$$= \mathbb{E}_{p(x)}\left[\log\frac{\frac{1}{\sqrt{(2\pi)^d\det\Sigma_p}}e^{-\frac{1}{2}(x-\mu_p)^T\Sigma_p^{-1}(x-\mu_p)}}{\frac{1}{\sqrt{(2\pi)^d\det\Sigma_q}}e^{-\frac{1}{2}(x-\mu_q)^T\Sigma_q^{-1}(x-\mu_q)}}\right]$$

$$= \mathbb{E}_{p(x)}\left[\frac{1}{2}\log\frac{\det\Sigma_q}{\det\Sigma_p} - \frac{1}{2}\left((x-\mu_p)^T\Sigma_p^{-1}(x-\mu_p) - (x-\mu_q)^T\Sigma_q^{-1}(x-\mu_q)\right)\right]$$

Let $a = (x-\mu_p)^T\Sigma_p^{-1}(x-\mu_p)$, then $a \in \mathbb{R}$, $\mathrm{tr}(a) = a$. By the cyclic shift property of trace [3], we have

$$\mathrm{tr}((x-\mu_p)^T\Sigma_p^{-1}(x-\mu_p)) = \mathrm{tr}(\Sigma_p^{-1}(x-\mu_p)(x-\mu_p)^T)$$

And we also have $\mathbb{E}(\mathrm{tr}(A)) = \mathrm{tr}(\mathbb{E}(A))$ for any matrix variable A [4].

---

[3] https://www.quora.com/How-can-I-prove-that-the-trace-of-a-matrix-product-is-invariant-under-cyclic-permutations-text-tr-ABC-text-tr-CAB-text-tr-BCA

[4] https://statproofbook.github.io/P/mean-tr

Therefore,

$$
\begin{aligned}
D(\,p(x)\,||\,q(x)\,) &= \frac{1}{2}[\log\frac{\det\Sigma_q}{\det\Sigma_p} - \mathbb{E}_{p(x)}[\mathsf{tr}(\Sigma_p^{-1}(x-\mu_p)(x-\mu_p)^T) \\
&\quad + \mathsf{tr}(\Sigma_q^{-1}(x-\mu_q)(x-\mu_q)^T)]] \\
&= \frac{1}{2}[\log\frac{\det\Sigma_q}{\det\Sigma_p} - \mathsf{tr}(\Sigma_p^{-1}\mathbb{E}_{p(x)}[(x-\mu_p)(x-\mu_p)^T]) \\
&\quad + \mathsf{tr}(\Sigma_q^{-1}\mathbb{E}_{p(x)}[(x-\mu_q)(x-\mu_q)^T])]
\end{aligned}
$$

Since $\mathbb{E}_{p(x)}[(x-\mu_p)(x-\mu_p)^T]) = \Sigma_p$ and

$$
\begin{aligned}
&\mathbb{E}_{p(x)}[(x-\mu_q)(x-\mu_q)^T] = \mathbb{E}_{p(x)}[xx^T - \mu_q x^T - x\mu_q^T + \mu_q\mu_q^T] \\
&= \mathbb{E}_{p(x)}[(x-\mu_p)(x-\mu_p)^T + x\mu_p^T + \mu_p x^T - \mu_p\mu_p^T - \mu_q x^T - x\mu_q^T + \mu_q\mu_q^T] \\
&= \Sigma_p + \mu_p\mu_p^T - \mu_p\mu_q^T - \mu_q\mu_p^T + \mu_q\mu_q^T \\
&= \Sigma_p + (\mu_p-\mu_q)(\mu_p-\mu_q)^T
\end{aligned}
$$

Thus,

$$
\begin{aligned}
D(\, p(x) \,||\, q(x) \,) &= \frac{1}{2}[\log \frac{\det \Sigma_q}{\det \Sigma_p} - \mathsf{tr}(\Sigma_p^{-1}\Sigma_p) \\
&\quad + \mathsf{tr}(\Sigma_q^{-1}(\Sigma_p + (\mu_p - \mu_q)(\mu_p - \mu_q)^T))] \\
&= \frac{1}{2}[\log \frac{\det \Sigma_q}{\det \Sigma_p} - d + \mathsf{tr}(\Sigma_q^{-1}\Sigma_p) + \mathsf{tr}(\Sigma_q^{-1}(\mu_p - \mu_q)(\mu_p - \mu_q)^T)] \\
&= \frac{1}{2}[\log \frac{\det \Sigma_q}{\det \Sigma_p} - d + \mathsf{tr}(\Sigma_q^{-1}\Sigma_p) + (\mu_p - \mu_q)^T\Sigma_q^{-1}(\mu_p - \mu_q)]
\end{aligned}
$$

□

Corollary A.2.7: Let I_d be the $d \times d$ identity matrix. For any $w, w_0, \lambda \in \mathbb{R}$ and $s = [s_1, s_2, ..., s_d]^T \in \mathbb{R}_+^d$,

$$
D(\, \mathsf{N}(w, \mathsf{diag}(s)) \,||\, \mathsf{N}(w_0, \lambda I_d) \,) = \frac{1}{2}[d(\log \lambda - 1) - \sum_{i=1}^{d}(\log s_i - \frac{1}{\lambda}s_i) + \frac{\|w - w_0\|^2}{\lambda}]
$$

Comparing Hoeffding Bound and Relative Entropy Chernoff Bound:
Let $X_1, X_2, ..., X_n$ be i.i.d. sampled from a distribution $\mathcal{D}$ and each $X_i \in [0, 1]$. Let $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $p = \mathbb{E}[\hat{p}]$, then by Hoeffding Bound,

$$P[\hat{p} - p > \epsilon] \leq e^{-2n\epsilon^2} \tag{25}$$

By Relative Entropy Chernoff Bound (Lemma 1.5),

$$P[\hat{p} - p > \epsilon] \leq e^{-D_B(\, p+\epsilon \,||\, p\,)n} \tag{26}$$

Actually, the Relative Entropy Chernoff Bound is tighter than the Hoeffding Bound because, by Theorem A.2.1,

$$D_B(\, p + \epsilon \,||\, p\,) \geq 2\epsilon^2$$

The only difference in the proofs of the two bounds is, when scaling the term $\mathbb{E}[e^{\lambda X_i}]$, Hoeffding Bound uses the Hoeffding Lemma $\mathbb{E}[e^{\lambda X_i}] \leq e^{\lambda \mathbb{E}[X_i] + \frac{\lambda^2 (b-a)^2}{8}}$ where $b = 1$ and $a = 0$, while the Chernoff Bound uses $e^{\lambda X_i} \leq 1 - X_i + X_i e^{\lambda}$ $\Rightarrow \mathbb{E}[e^{\lambda X_i}] \leq \mathbb{E}[1 - X_i + X_i e^{\lambda}]$.

Hoeffding Bound can be applied to any X_is that bounded in $[a, b]$ where $a \in \mathbb{R}, b \in \mathbb{R}$, while Chernoff Bound can only be applied to X_is bounded in $[0, 1]$.

Appendix 3: Catoni's Bound

Theorem A.3: On a fixed dataset S, for any $\lambda \in \mathbb{R}$ and any $\delta \in (0,1)$, let θ sampled from an unknown distribution ρ, then

$$P_S\left[\sup_{\rho}\{\mathbb{E}_{\theta\sim\rho}[R^{\text{true}}(\theta) - R^{\text{emp}}(\theta)]\} \leq \frac{\lambda C^2}{8n} + \frac{D(\,\rho\,||\,\pi\,) + \log\frac{1}{\delta}}{\lambda}\right] \geq 1-\delta$$

Proof: By Eq (1), for any $\lambda \in \mathbb{R}$,

$$\mathbb{E}_S\big[e^{tn(R^{\text{true}}(\theta)-R^{\text{emp}}(\theta))}\big] \leq e^{\frac{nt^2C^2}{8}}$$

Taking $\lambda = \frac{t}{n}$,

$$\mathbb{E}_S\big[e^{\lambda(R^{\text{true}}(\theta)-R^{\text{emp}}(\theta))}\big] \leq e^{\frac{\lambda^2C^2}{8n}}$$

Since ρ is unknown, we assume a distribution π and taking the expectation of θ with respect to π:

$$\mathbb{E}_{\theta\sim\pi}\mathbb{E}_S\big[e^{\lambda(R^{\text{true}}(\theta)-R^{\text{emp}}(\theta))}\big] \leq e^{\frac{\lambda^2C^2}{8n}} \iff \mathbb{E}_S\mathbb{E}_{\theta\sim\pi}\big[e^{\lambda(R^{\text{true}}(\theta)-R^{\text{emp}}(\theta))}\big] \leq e^{\frac{\lambda^2C^2}{8n}}$$

In Eq (5), let $q = \pi, p = \rho, g = \lambda(R^{\text{true}}(\theta) - R^{\text{emp}}(\theta))$, we have

$$\mathbb{E}_{\theta\sim\pi}\big[e^{\lambda(R^{\text{true}}(\theta)-R^{\text{emp}}(\theta))}\big] = e^{\sup_\rho\{\mathbb{E}_{\theta\sim\rho}[\lambda(R^{\text{true}}(\theta)-R^{\text{emp}}(\theta))]-D(\,\rho\,||\,\pi\,)\}}$$

Thus,

$$\begin{aligned}\mathbb{E}_S\mathbb{E}_{\theta\sim\pi}\big[e^{\lambda(R^{\text{true}}(\theta)-R^{\text{emp}}(\theta))}\big] &= \mathbb{E}_S\big[e^{\sup_\rho\{\mathbb{E}_{\theta\sim\rho}[\lambda(R^{\text{true}}(\theta)-R^{\text{emp}}(\theta))]-D(\,\rho\,||\,\pi\,)\}}\big] \\ &\leq e^{\frac{\lambda^2C^2}{8n}}\end{aligned}$$

That is,

$$\mathbb{E}_S\big[e^{\sup_\rho\{\mathbb{E}_{\theta\sim\rho}[\lambda(R^{\text{true}}(\theta)-R^{\text{emp}}(\theta))]-D(\,\rho\,||\,\pi\,)-\frac{\lambda^2C^2}{8n}\}}\big] \leq 1$$

Therefore, by Chernoff Inequality, for any $\epsilon > 0$,

$$\begin{aligned}&P_S\left[\sup_\rho\{\mathbb{E}_{\theta\sim\rho}[\lambda(R^{\text{true}}(\theta) - R^{\text{emp}}(\theta))] - D(\,\rho\,||\,\pi\,) - \frac{\lambda^2C^2}{8n}\} > \epsilon\right] \\ &= P_S\left[e^{\sup_\rho\{\mathbb{E}_{\theta\sim\rho}[\lambda(R^{\text{true}}(\theta)-R^{\text{emp}}(\theta))]-D(\,\rho\,||\,\pi\,)-\frac{\lambda^2C^2}{8n}\}} > e^\epsilon\right] \\ &\leq e^{-\epsilon}\mathbb{E}_S\big[e^{\sup_\rho\{\mathbb{E}_{\theta\sim\rho}[\lambda(R^{\text{true}}(\theta)-R^{\text{emp}}(\theta))]-D(\,\rho\,||\,\pi\,)-\frac{\lambda^2C^2}{8n}\}}\big] \leq e^{-\epsilon}\end{aligned}$$

Taking $\delta = e^{-\epsilon} \Leftrightarrow \epsilon = \log \frac{1}{\delta}$, then $\delta \in (0, 1)$,

$$P_S \left[\sup_{\rho} \{ \mathbb{E}_{\theta \sim \rho} [\lambda (R^{\mathsf{true}}(\theta) - R^{\mathsf{emp}}(\theta))] - D(\, \rho \,||\, \pi \,) - \frac{\lambda^2 C^2}{8n} \} > \log \frac{1}{\delta} \right] \leq \delta$$

Rearranging terms give

$$P_S \left[\sup_{\rho} \{ \mathbb{E}_{\theta \sim \rho} [R^{\mathsf{true}}(\theta) - R^{\mathsf{emp}}(\theta)] \} > \frac{\lambda C^2}{8n} + \frac{D(\, \rho \,||\, \pi \,) + \log \frac{1}{\delta}}{\lambda} \right] \leq \delta$$

$$P_S \left[\sup_{\rho} \{ \mathbb{E}_{\theta \sim \rho} [R^{\mathsf{true}}(\theta) - R^{\mathsf{emp}}(\theta)] \} \leq \frac{\lambda C^2}{8n} + \frac{D(\, \rho \,||\, \pi \,) + \log \frac{1}{\delta}}{\lambda} \right] \geq 1 - \delta$$

□

Gibbs Posterior:

In Theorem 1.1, the bound can also be written as

$$P_S\left[\forall\rho,\ \mathbb{E}_{\theta\sim\rho}[R^{\mathsf{true}}(\theta)] \le \mathbb{E}_{\theta\sim\rho}[R^{\mathsf{emp}}(\theta)] + \frac{\lambda C^2}{8n} + \frac{D(\,\rho\,||\,\pi\,) + \log\frac{1}{\delta}}{\lambda}\right] \ge 1-\delta \quad (27)$$

We want to find ρ to minimize the right hand side, let

$$\begin{aligned}\hat{\rho}_\lambda &= \operatorname*{argmin}_\rho \left\{\mathbb{E}_{\theta\sim\rho}[R^{\mathsf{emp}}(\theta)] + \frac{D(\,\rho\,||\,\pi\,)}{\lambda}\right\} \\ &= \operatorname*{argmax}_\rho \left\{-\lambda\mathbb{E}_{\theta\sim\rho}[R^{\mathsf{emp}}(\theta)] - D(\,\rho\,||\,\pi\,)\right\}\end{aligned}$$

Note that by Eq (5),

$$\begin{aligned}\sup_\rho \left\{-\lambda\mathbb{E}_{\theta\sim\rho}[R^{\mathsf{emp}}(\theta)] - D(\,\rho\,||\,\pi\,)\right\} &= \sup_\rho \left\{\mathbb{E}_{\theta\sim\rho}[-\lambda R^{\mathsf{emp}}(\theta)] - D(\,\rho\,||\,\pi\,)\right\} \\ &= \log \mathbb{E}_{\theta\sim\pi}[e^{-\lambda R^{\mathsf{emp}}(\theta)}]\end{aligned}$$

where the supremum is obtained by taking

$$\rho(\theta) = \frac{e^{-\lambda R^{\mathsf{emp}(\theta)}}\pi(\theta)}{\mathbb{E}_{\theta\sim\pi}[e^{-\lambda R^{\mathsf{emp}(\theta)}}]}$$

Hence,

$$\hat{\rho}_\lambda(\theta) = \frac{e^{-\lambda R^{\text{emp}(\theta)}}\pi(\theta)}{\mathbb{E}_{\theta\sim\pi}[e^{-\lambda R^{\text{emp}(\theta)}}]} \tag{28}$$

We call such $\hat{\rho}_\lambda$ as Gibbs Posterior.

Why Gibbs Posterior: Multiplying $(\frac{\lambda}{n})^n$ on both denominator and nominator of Eq (28):

$$\hat{\rho}_\lambda(\theta) = \frac{(\frac{\lambda}{n})^n e^{-\lambda R^{\text{emp}(\theta)}}\pi(\theta)}{\mathbb{E}_{\theta\sim\pi}[(\frac{\lambda}{n})^n e^{-\lambda R^{\text{emp}(\theta)}}]}$$

So the likelihood function is

$$f(\boldsymbol{x},\boldsymbol{y}|\theta) = (\frac{\lambda}{n})^n e^{-\lambda R^{\text{emp}(\theta)}} = (\frac{\lambda}{n})^n e^{-\frac{\lambda}{n}\sum_{i=1}^n l(f_\theta(x_i)-y_i)} = \prod_{i=1}^n \frac{\lambda}{n} e^{-\frac{\lambda}{n} l(f_\theta(x_i)-y_i)}$$

We define $f(x_i,y_i|\theta) = \frac{\lambda}{n}e^{-\frac{\lambda}{n}l(f_\theta(x_i)-y_i)}$ such that $f(\boldsymbol{x},\boldsymbol{y}|\theta) = \prod_{i=1}^n f(x_i,y_i|\theta)$. Then $f(x_i,y_i|\theta)$ is an exponential distribution of $l(f_\theta(x_i)-y_i)$.

The exponential distribution is a natural way to describe $l(f_\theta(x_i)-y_i)$ because the loss is always nonnegative. Moreover, if the predictor f_θ is effective, it is not likely to make the loss too large. So we want the probability to go 0 when the loss goes to infinity.

Minimization of Catoni's Bound:

Since $\hat{\rho}$ is the minimizer of the right hand side of Eq (27), we can also write Eq (27) as

$$P_S\left[\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}[R^{\text{true}}(\theta)] \leq \inf_{\rho}\left\{\mathbb{E}_{\theta\sim\rho}[R^{\text{emp}}(\theta)] + \frac{\lambda C^2}{8n} + \frac{D(\,\rho\,||\,\pi\,) + \log\frac{1}{\delta}}{\lambda}\right\}\right] \geq 1-\delta \tag{29}$$

Since ρ can be any distribution of θ on the right hand side of Eq (29), given any $\theta \in \Theta$, let the Dirac Measure[5] δ_θ be

$$\delta_\theta(\eta) = \begin{cases} 1 & \text{if } \eta = \theta \\ 0 & \text{if } \eta \neq \theta \end{cases}$$

Then δ_θ is the PDF of the distribution $P(\eta = \theta) = 1$ and $P(\eta \neq \theta) = 0$. Let $\rho = \delta_\theta$, then we have

$$\mathbb{E}_{\eta\sim\delta_\theta}[R^{\text{emp}}(\eta)] = R^{\text{emp}}(\theta)$$

$$D(\,\delta_\theta\,||\,\pi\,) = \sum_{\eta\in\Theta}\delta_\theta(\eta)\log\frac{\delta_\theta(\eta)}{\pi(\eta)} = \log\frac{1}{\pi(\theta)}$$

[5] https://en.wikipedia.org/wiki/Dirac_measure

Thus, we can write Eq (29) as

$$P_S\left[\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}[R^{\text{true}}(\theta)] \le R^{\text{emp}}(\theta) + \frac{\lambda C^2}{8n} + \frac{\log\frac{1}{\pi(\theta)} + \log\frac{1}{\delta}}{\lambda}\right] \ge 1-\delta \qquad (30)$$

Finite Case: Suppose the cardinality $|\Theta| = M$ is finite, and let $\pi(\theta)$ be the uniform distribution, that is, $\pi(\theta) = \frac{1}{M}$ for all $\theta \in \Theta$. Then we can write Eq (30) as

$$P_S\left[\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}[R^{\text{true}}(\theta)] \le R^{\text{emp}}(\theta) + \frac{\lambda C^2}{8n} + \frac{\log\frac{M}{\delta}}{\lambda}\right] \ge 1-\delta$$

Take $\lambda = \frac{\sqrt{8n\log\frac{M}{\delta}}}{C}$ to minimize the right hand side of the above bound, we get

$$P_S\left[\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}[R^{\text{true}}(\theta)] \le R^{\text{emp}}(\theta) + C\sqrt{\frac{\log\frac{M}{\delta}}{2n}}\right] \ge 1-\delta$$

Hence, we obtain the same upper bound for $\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}[R^{\text{true}}(\theta)]$ as Eq (3). However, this bound can be tighter. If we don't assume $\pi(\theta)$ in Eq (30) as uniform distribution, then for the θ with greater probability $\pi(\theta)$, the right hand side will be smaller.