

PAC-Bayes Bound for Linear Regression

Ruixin Guo

Department of Computer Science
Kent State University

February 29, 2024

- ① Alquier's Bound
- ② PAC-Bayes Bound for Linear Regression

Recall: Statistical Learning Theory

We define the **dataset** as $S = \{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ is the feature vector, and $y_i \in \mathbb{R}$ is the label. Each (x_i, y_i) is i.i.d. sampled from an unknown distribution \mathcal{D} .

The machine learning model is a **predictor** $f_\theta : \mathbb{R}^d \rightarrow \mathbb{R}$ where θ is the vector of parameters. The loss function of the predictor on the sample (x_i, y_i) is defined as $L(f_\theta(x_i), y_i)$. The **empirical risk** is defined as

$$R^{\text{emp}}(\theta) = \frac{1}{n} \sum_{i=1}^n L(f_\theta(x_i), y_i)$$

The true risk is defined as

$$R^{\text{true}}(\theta) = \mathbb{E}_{(x_i, y_i) \sim \mathcal{D}} [L(f_\theta(x_i), y_i)]$$

In this lecture, we will show the PAC-Bayes bound for the linear regression problem.

Recall: Moment Generating Function

Definition 0.1: Let X be a random variable and n be an integer, the n th **moment** of X is $\mathbb{E}[X^n]$.

Definition 0.2: Let X be a random variable, the **Moment Generating Function (MGF)**, denoted by $M_X(t)$, is

$$M_X(t) = \mathbb{E}[e^{tX}]$$

Theorem 0.3: If X has MGF $M_X(t)$, let $M_X^{(n)}(t)$ be the n th derivative of $M_X(t)$, then

$$\mathbb{E}[X^n] = M_X^{(n)}(0)$$

Proof: By Taylor Theorem,

$$\begin{aligned} e^{tX} &= \sum_{k=1}^{\infty} \frac{t^k}{k!} X^k \Rightarrow \mathbb{E}[e^{tX}] = \sum_{k=1}^{\infty} \frac{t^k}{k!} \mathbb{E}[X^k] \Rightarrow \frac{d^n}{dt^n} \mathbb{E}[e^{tX}] = \sum_{k=n}^{\infty} \frac{t^{(k-n)}}{k!/n!} \mathbb{E}[X^k] \\ &\Rightarrow \left. \frac{d^n}{dt^n} \mathbb{E}[e^{tX}] \right|_{t=0} = \mathbb{E}[X^n] \end{aligned}$$



Definition 0.4: Let X_1, X_2, \dots, X_n be random variables iid from $\mathcal{N}(0, 1)$. Then $X = \sum_{i=1}^n X_i^2$ satisfies **Chi-Squared distribution** of n degree of freedom, denoted as $\chi^2(n)$.

Theorem 0.5: If $X \sim \chi^2(n)$, then $M_X(t) = (1 - 2t)^{-\frac{n}{2}}$.

Proof: For $\chi^2(1)$, i.e., when $n = 1$,

$$\begin{aligned} M_X(t) &= \mathbb{E}[e^{tX^2}] = \int e^{tx^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\ &= (1 - 2t)^{-\frac{1}{2}} \int \frac{1}{\sqrt{2\pi} \left(\frac{1}{1-2t}\right)^{\frac{1}{2}}} \exp\left(-\frac{x^2}{\frac{2}{1-2t}}\right) dx \\ &= (1 - 2t)^{-\frac{1}{2}} \end{aligned}$$

For $\chi^2(n)$,

$$M_X(t) = \mathbb{E}[e^{tX}] = \mathbb{E}[e^{t \sum_{i=1}^n X_i}] = \prod_{i=1}^n \mathbb{E}[e^{tX_i}] = \mathbb{E}[e^{tX_i}]^n = (1 - 2t)^{-\frac{n}{2}}$$

□

① Alquier's Bound

② PAC-Bayes Bound for Linear Regression

Alquier's Bound

Theorem 1 (Alquier's Bound) [1]: Let π be a prior distribution of θ and $\lambda > 0$ be a real number. Then for any posterior distribution ρ and $\delta > 0$,

$$P\left(\mathbb{E}_{\theta \sim \rho}[R^{\text{true}}(\theta)] < \mathbb{E}_{\theta \sim \rho}[R^{\text{emp}}(\theta)] + \frac{1}{\lambda} \left[D(\rho \parallel \pi) + \ln \frac{1}{\delta} + \Psi_{L, \pi, \mathcal{D}}(\lambda, n) \right] \right) \geq 1 - \delta$$

where

$$\Psi_{L, \pi, \mathcal{D}}(\lambda, n) = \ln \mathbb{E}_{\theta \sim \pi} \mathbb{E}_{(x_i, y_i) \sim \mathcal{D}} [e^{\lambda(R^{\text{true}}(\theta) - R^{\text{emp}}(\theta))}]$$

Proof: By Donsker-Varadhan representation, for any distribution ρ, π and any function $g(\theta)$,

$$\mathbb{E}_{\theta \sim \rho}[g(\theta)] \leq D(\rho \parallel \pi) + \ln \mathbb{E}_{\theta \sim \pi}[e^{g(\theta)}]$$

Let $g(\theta) = \lambda(R^{\text{true}}(\theta) - R^{\text{emp}}(\theta))$, we have

$$\lambda(\mathbb{E}_{\theta \sim \rho}[R^{\text{true}}(\theta)] - \mathbb{E}_{\theta \sim \rho}[R^{\text{emp}}(\theta)]) \leq D(\rho \parallel \pi) + \ln \mathbb{E}_{\theta \sim \pi}[e^{\lambda(R^{\text{true}}(\theta) - R^{\text{emp}}(\theta))}] \quad (1)$$

Consider Markov's Inequality. For any non-negative random variable X and constant $t > 0$,

$$P[X > t] \leq \frac{\mathbb{E}[X]}{t}$$

Let $\delta = \frac{\mathbb{E}[X]}{t}$, then

$$P[X > \frac{\mathbb{E}[X]}{\delta}] \leq \delta \quad \Longleftrightarrow \quad P[X < \frac{\mathbb{E}[X]}{\delta}] \geq 1 - \delta$$

Let $X = \mathbb{E}_{\theta \sim \pi}[e^{\lambda(R^{\text{true}}(\theta) - R^{\text{emp}}(\theta))}]$, we have

$$P[\mathbb{E}_{\theta \sim \pi}[e^{\lambda(R^{\text{true}}(\theta) - R^{\text{emp}}(\theta))}] < \frac{\mathbb{E}_{(x_i, y_i) \sim \mathcal{D}} \mathbb{E}_{\theta \sim \pi}[e^{\lambda(R^{\text{true}}(\theta) - R^{\text{emp}}(\theta))}]}{\delta}] \geq 1 - \delta \quad \Longleftrightarrow$$

$$P[\ln \mathbb{E}_{\theta \sim \pi}[e^{\lambda(R^{\text{true}}(\theta) - R^{\text{emp}}(\theta))}] < \ln \mathbb{E}_{(x_i, y_i) \sim \mathcal{D}} \mathbb{E}_{\theta \sim \pi}[e^{\lambda(R^{\text{true}}(\theta) - R^{\text{emp}}(\theta))}] + \ln \frac{1}{\delta}] \geq 1 - \delta$$

Plugging Eq (1) in, we have

$$P[\lambda(\mathbb{E}_{\theta \sim \rho}[R^{\text{true}}(\theta)] - \mathbb{E}_{\theta \sim \rho}[R^{\text{emp}}(\theta)]) - D(\rho \parallel \pi) <$$

$$\ln \mathbb{E}_{(x_i, y_i) \sim \mathcal{D}} \mathbb{E}_{\theta \sim \pi}[e^{\lambda(R^{\text{true}}(\theta) - R^{\text{emp}}(\theta))}] + \ln \frac{1}{\delta}] \geq 1 - \delta$$

Rearrange the above inequality and the Theorem is proved. □

Remember that the Moment Generating Function (MGF) of a random variable X is $M_X(\lambda) = \mathbb{E}[e^{\lambda X}]$. In the proof above, λ is introduced by Markov inequality, and

$$\mathbb{E}_{(x_i, y_i) \sim \mathcal{D}} \mathbb{E}_{\theta \sim \pi} [e^{\lambda(R^{\text{true}}(\theta) - R^{\text{emp}}(\theta))}]$$

is indeed the MGF of $R^{\text{true}}(\theta) - R^{\text{emp}}(\theta)$.

Like Catoni's Bound, Alquier's Bound holds for any $\lambda > 0$. Thus we can choose λ to minimize the right hand side of the inequality

$$\mathbb{E}_{\theta \sim \rho} [R^{\text{emp}}(\theta)] + \frac{1}{\lambda} \left[D(\rho \parallel \pi) + \ln \frac{1}{\delta} + \Psi_{L, \pi, \mathcal{D}}(\lambda, n) \right]$$

to get the tightest bound. However, to calculate the minimizer λ^* , we need to know $D(\rho \parallel \pi)$. If we fix π , then λ^* will be a function of ρ .

For convenience, we can let λ be a value independent of ρ , like n or \sqrt{n} . In this case, the bound will not be optimal but applicable to any ρ .

① Alquier's Bound

② PAC-Bayes Bound for Linear Regression

Problem Settings

Consider the Linear Regression problem:

$$L(f_{\theta}(x_i) - y_i) = (y_i - \theta \cdot x_i)^2$$

$$R^{\text{emp}}(\theta) = \frac{1}{n} \sum_{i=1}^n (y_i - \theta \cdot x_i)^2$$

where $\theta \in \mathbb{R}^d$, $\theta \cdot x_i = \theta^T x_i \in \mathbb{R}$.

Assumption 2.1: Suppose $x_i \sim \mathcal{N}_d(0, \sigma_x^2 I)$ where $\sigma_x > 0$ is a constant. Suppose there exists θ^* such that for any i , $y_i = \theta^* \cdot x_i + e_i$ where $e_i \sim \mathcal{N}(0, \sigma_e^2)$ and $\sigma_e > 0$ is a constant. e_i and e_j are independent for any $i \neq j$.

Lemma 2.2: Suppose Assumption 2.1 holds, given $\theta^*, \sigma_x, \sigma_e$, then for any θ , $y_i - \theta \cdot x_i \sim \mathcal{N}(0, v_{\theta})$, where $v_{\theta} = \sigma_x^2 \|\theta - \theta^*\|^2 + \sigma_e^2$.

Proof: By Assumption 2.1, we have

$$y_i - \theta \cdot x_i = (\theta^* - \theta) \cdot x_i + e_i$$

Let $\theta' = (\theta^* - \theta)$, θ'_j be the j th element of θ' , and x_{ij} be the j th element of x_i , then

$$y_i - \theta \cdot x_i = \sum_{j=1}^d \theta'_j x_{ij} + e_i$$

Since x_{ij} s are iid sampled from $\mathcal{N}(0, \sigma_x^2)$, e_i is sampled from $\mathcal{N}(0, \sigma_e^2)$, and θ'_j s are scalars, then $y_i - \theta \cdot x_i$ is a random variable satisfying Gaussian distribution, with

$$\begin{aligned}\mathbb{E}[y_i - \theta \cdot x_i] &= \sum_{j=1}^d \theta'_j \mathbb{E}[x_{ij}] + \mathbb{E}[e_i] = 0 \\ \text{Var}[y_i - \theta \cdot x_i] &= \sum_{j=1}^d \theta_j'^2 \text{Var}[x_{ij}] + \text{Var}[e_i] \\ &= \sigma_x^2 \|\theta^* - \theta\|^2 + \sigma_e^2\end{aligned}$$

□

The predictor θ is an estimator of θ^* . We assume that \mathcal{D} correlates with the predictor by making $p(x_i, y_i | \theta) = p(x_i) p(y_i | \theta, x_i)$ where $p(x_i)$ is $\mathcal{N}(0, \sigma_x^2)$ and $p(y_i | \theta, x_i)$ is $\mathcal{N}(\theta \cdot x_i, \sigma_x^2 \|\theta^* - \theta\|^2 + \sigma_e^2)$.

Bound for Linear Regression

Theorem 2.3 (Shalaeva's Bound) [2]: In Theorem 1, let the loss function be $L(f_\theta(x_i) - y_i) = (y_i - \theta \cdot x_i)^2$. Under Assumption 2.1, we have

$$P\left(\mathbb{E}_{\theta \sim \rho}[R^{\text{true}}(\theta)] < \mathbb{E}_{\theta \sim \rho}[R^{\text{emp}}(\theta)] + \frac{1}{\lambda} \left[D(\rho \parallel \pi) + \ln \frac{1}{\delta} + \Psi_{L, \pi, \mathcal{D}}(\lambda, n) \right] \right) \geq 1 - \delta$$

where

$$\Psi_{L, \pi, \mathcal{D}}(\lambda, n) = \ln \mathbb{E}_{\theta \sim \pi} \frac{e^{\lambda v_\theta}}{\left(1 + \frac{\lambda v_\theta}{\frac{n}{2}}\right)^{\frac{n}{2}}} \leq \ln \mathbb{E}_{\theta \sim \pi} \exp\left(\frac{\lambda^2 v_\theta^2}{\frac{n}{2}}\right)$$

Proof: In Theorem 1,

$$\begin{aligned} \Psi_{L, \pi, \mathcal{D}}(\lambda, n) &= \ln \mathbb{E}_{\theta \sim \pi} \mathbb{E}_{(x_i, y_i) \sim \mathcal{D}} [e^{\lambda(R^{\text{true}}(\theta) - R^{\text{emp}}(\theta))}] \\ &= \ln \mathbb{E}_{\theta \sim \pi} \left(e^{\lambda R^{\text{true}}(\theta)} \mathbb{E}_{(x_i, y_i) \sim \mathcal{D}} [e^{-\lambda R^{\text{emp}}(\theta)}] \right) \end{aligned}$$

We have $R^{\text{true}}(\theta) = \mathbb{E}_{(x_i, y_i) \sim \mathcal{D}}[(y_i - \theta \cdot x_i)^2] = v_\theta$ and

$$\mathbb{E}_{(x_i, y_i) \sim \mathcal{D}}[e^{-\lambda R^{\text{emp}}(\theta)}] = \mathbb{E}_{(x_i, y_i) \sim \mathcal{D}}[e^{-\frac{\lambda v_\theta}{n} \sum_{i=1}^n (\frac{y_i - \theta \cdot x_i}{\sqrt{v_\theta}})^2}] \quad (2)$$

Since $\frac{y_i - \theta \cdot x_i}{\sqrt{v_\theta}} \sim \mathcal{N}(0, 1)$, $\sum_{i=1}^n (\frac{y_i - \theta \cdot x_i}{\sqrt{v_\theta}})^2 \sim \chi^2(n)$. Thus Eq (2) is the MGF of $\chi^2(n)$.

$$\mathbb{E}_{(x_i, y_i) \sim \mathcal{D}}[e^{-\frac{\lambda v_\theta}{n} \sum_{i=1}^n (\frac{y_i - \theta \cdot x_i}{\sqrt{v_\theta}})^2}] = \left(1 + 2\frac{\lambda v_\theta}{n}\right)^{-\frac{n}{2}}$$

Therefore,

$$\Psi_{L, \pi, \mathcal{D}}(\lambda, n) = \ln \mathbb{E}_{\theta \sim \pi} \frac{e^{\lambda v_\theta}}{\left(1 + \frac{\lambda v_\theta}{\frac{n}{2}}\right)^{\frac{n}{2}}}$$

Since for any $x > -1$, $\frac{x}{x+1} \leq \ln(x+1)$, let $k > 0$, we have $e^{\frac{xk}{x+1}} \leq (x+1)^k \Rightarrow e^{\frac{xk}{x+k}} \leq (\frac{x}{k} + 1)^k$. Let $x = \lambda v_\theta$, $k = \frac{n}{2}$, we have

$$\left(1 + \frac{\lambda v_\theta}{\frac{n}{2}}\right)^{\frac{n}{2}} \geq \exp\left(\frac{\lambda v_\theta \frac{n}{2}}{\lambda v_\theta + \frac{n}{2}}\right)$$

Therefore,

$$\begin{aligned}\Psi_{L,\pi,\mathcal{D}}(\lambda, n) &= \ln \mathbb{E}_{\theta \sim \pi} \frac{e^{\lambda v_\theta}}{\left(1 + \frac{\lambda v_\theta}{\frac{n}{2}}\right)^{\frac{n}{2}}} \leq \ln \mathbb{E}_{\theta \sim \pi} \exp \left(\lambda v_\theta - \frac{\lambda v_\theta \frac{n}{2}}{\lambda v_\theta + \frac{n}{2}} \right) \\ &= \ln \mathbb{E}_{\theta \sim \pi} \exp \left(\frac{\lambda^2 v_\theta^2}{\lambda v_\theta + \frac{n}{2}} \right) \leq \ln \mathbb{E}_{\theta \sim \pi} \exp \left(\frac{\lambda^2 v_\theta^2}{\frac{n}{2}} \right)\end{aligned}$$

□

We will show that with proper choice of λ , the bound will converge to 0 as $n \rightarrow \infty$.

(1) When λ does not depend on n , as $n \rightarrow \infty$, we have

$$P \left(\mathbb{E}_{\theta \sim \rho} [R^{\text{true}}(\theta)] < \mathbb{E}_{\theta \sim \rho} [R^{\text{emp}}(\theta)] + \frac{1}{\lambda} \left[D(\rho \parallel \pi) + \ln \frac{1}{\delta} \right] \right) \geq 1 - \delta$$

This is because

$$\lim_{n \rightarrow \infty} \left(1 + \frac{\lambda v_\theta}{\frac{n}{2}} \right)^{\frac{n}{2}} = e^{\lambda v_\theta}$$

such that $\lim_{n \rightarrow \infty} \Psi_{L, \pi, \mathcal{D}}(\lambda, n) = 0$. In this case, even n goes to infinity, there is still a gap $\frac{1}{\lambda} [D(\rho || \pi) + \ln \frac{1}{\delta}]$ that cannot be minimized.

(2) When λ depends on n , we can let $\lambda = n^{\frac{1}{d}}$ such that

$$P \left(\mathbb{E}_{\theta \sim \rho} [R^{\text{true}}(\theta)] < \mathbb{E}_{\theta \sim \rho} [R^{\text{emp}}(\theta)] + \frac{D(\rho || \pi)}{n^{\frac{1}{d}}} + \frac{\ln(\frac{1}{\delta})}{n^{\frac{1}{d}}} \right. \\ \left. + \frac{1}{n^{\frac{1}{d}}} \ln \mathbb{E}_{\theta \sim \pi} \exp \left(\frac{n^{\frac{2}{d}} v_{\theta}^2}{\frac{n}{2}} \right) \right) \geq 1 - \delta$$

Then the gap will converge to 0 as $n \rightarrow \infty$. We show the convergence of the third term:

$$\lim_{n \rightarrow \infty} \frac{1}{n^{\frac{1}{d}}} \ln \mathbb{E}_{\theta \sim \pi} \exp \left(\frac{n^{\frac{2}{d}} v_{\theta}^2}{\frac{n}{2}} \right) = \lim_{n \rightarrow \infty} \ln \left[\mathbb{E}_{\theta \sim \pi} \exp \left(2n^{\frac{2}{d}-1} v_{\theta}^2 \right) \right]^{n^{-\frac{1}{d}}} = \ln [\mathbb{E}_{\theta \sim \pi} 1]^0 = 0$$

Extension to Non-i.i.d. Case

Let's consider a general case that x_i is sampled from a multivariate Gaussian distribution whose dimensions are not i.i.d..

Assumption 2.4: Suppose $x_i \sim \mathcal{N}_d(0, Q_x)$ where $Q_x \in \mathbb{R}^{d \times d}$ is a positive definite matrix. Suppose there exists θ^* such that for any i , $y_i = \theta^* \cdot x_i + e_i$ where $e_i \sim \mathcal{N}(0, \sigma_e^2)$ and $\sigma_e > 0$ is a constant. e_i and e_j are independent for any $i \neq j$.

The reason why we require Q_x to be positive definite is shown in Appendix 3.

Lemma 2.5: Suppose Assumption 2.4 holds, given $\theta^*, \sigma_x, \sigma_e$, then for any θ , $y_i - \theta \cdot x_i \sim \mathcal{N}(0, \check{v}_\theta)$, where $\check{v}_\theta = (\theta^* - \theta)^T Q_x (\theta^* - \theta) + \sigma_e^2$.

Proof: This Lemma is an extension of Lemma 2.2. Since $y_i - \theta \cdot x_i = (\theta^* - \theta)x_i + e_i$, and according to Theorem A.3.5, $(\theta^* - \theta)x_i \sim \mathcal{N}(0, (\theta^* - \theta)^T Q_x (\theta^* - \theta))$, we proved the theorem. □

Theorem 2.6 discusses the case that x_i s are i.i.d. from $\mathcal{N}_d(0, Q_x)$. Theorem 2.7 discusses the case that x_i s are identically but not independently distributed from $\mathcal{N}_d(0, Q_x)$, for example, x_i s may be sampled in a time series where the current sample depends on all the previous samples.

Theorem 2.6: In Theorem 1, let the loss function be $L(f_\theta(x_i), y_i) = (y_i - \theta \cdot x_i)^2$. Under Assumption 2.4, suppose x_i s are i.i.d. sampled from $\mathcal{N}_d(0, Q_x)$, we have

$$P\left(\mathbb{E}_{\theta \sim \rho}[R^{\text{true}}(\theta)] < \mathbb{E}_{\theta \sim \rho}[R^{\text{emp}}(\theta)] + \frac{1}{\lambda} \left[D(\rho \parallel \pi) + \ln \frac{1}{\delta} + \Psi_{L, \pi, \mathcal{D}}(\lambda, n) \right]\right) \geq 1 - \delta$$

where

$$\Psi_{L, \pi, \mathcal{D}}(\lambda, n) \leq \ln \mathbb{E}_{\theta \sim \pi} \frac{e^{\lambda \check{v}_\theta}}{\left(1 + \frac{\lambda \check{v}_\theta}{\frac{n}{2}}\right)^{\frac{n}{2}}}$$

The proof of Theorem 2.6 is exactly the same as Theorem 2.3, just replace v_θ by \check{v}_θ .

If x_i s are identically but not independently distributed from $\mathcal{N}_d(0, Q_x)$, we still have $R^{\text{true}}(\theta) = \mathbb{E}_{(x_i, y_i) \sim \mathcal{D}}[R^{\text{emp}}(\theta)]$, because

$$\begin{aligned}\mathbb{E}_{(x_i, y_i) \sim \mathcal{D}}[R^{\text{emp}}(\theta)] &= \mathbb{E}_{(x_i, y_i) \sim \mathcal{D}}\left[\frac{1}{n} \sum_{i=1}^n L(f_{\theta}(x_i), y_i)\right] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{(x_i, y_i) \sim \mathcal{D}}[L(f_{\theta}(x_i), y_i)] = R^{\text{true}}(\theta)\end{aligned}$$

This does not require $L(f_{\theta}(x_i), y_i)$ to be independent for different i . So the concentration inequality is still applicable. However, $R^{\text{emp}}(\theta)$ may not converge to $R^{\text{true}}(\theta)$ when $n \rightarrow \infty$, as the independency condition of Law of Large Numbers is not satisfied.

In fact, independency of samples is a sufficient but not necessary assumption in statistical learning [7]. The feature samples can be dependent in some cases, for example, in language data, the words in a sentence are dependent. Suppose the feature x and label y comes from an unknown distribution $p(x, y)$, the goal of machine learning is to learn the posterior distribution $p(y|x)$, which is independent from the data distribution $p(x)$.

For language data, the model learns the distribution $p(y|x_k, x_{k-1}, \dots, x_1)$, where k is the size of the window. $p(y|x_k, x_{k-1}, \dots, x_1)$ is independent of the feature distribution $p(x_k, x_{k-1}, \dots, x_1)$. The dependency of x_1, \dots, x_k will only affect $p(x_k, x_{k-1}, \dots, x_1)$ and will not affect $p(y|x_k, x_{k-1}, \dots, x_1)$.

Theorem 2.7: In Theorem 1, let the loss function be $L(f_\theta(x_i), y_i) = (y_i - \theta \cdot x_i)^2$. Under Assumption 2.4, suppose x_i s are identically sampled from $\mathcal{N}_d(0, Q_x)$ but not independent. Let $X = [x_1^T, x_2^T, \dots, x_n^T]^T \in \mathbb{R}^{dn \times 1}$ and let $Q_X = \mathbb{E}[XX^T] \in \mathbb{R}^{dn \times dn}$ be the joint covariance matrix. Let ω be the minimum eigenvalue of Q_X and assume $\omega > 0$. We have

$$P\left(\mathbb{E}_{\theta \sim \rho}[R^{\text{true}}(\theta)] < \mathbb{E}_{\theta \sim \rho}[R^{\text{emp}}(\theta)] + \frac{1}{\lambda} \left[D(\rho || \pi) + \ln \frac{1}{\delta} + \Psi_{L, \pi, \mathcal{D}}(\lambda, n) \right] \right) \geq 1 - \delta$$

where

$$\Psi_{L, \pi, \mathcal{D}}(\lambda, n) \leq \ln \mathbb{E}_{\theta \sim \pi} \frac{e^{\lambda \check{v}_\theta}}{\left(1 + \frac{\lambda \omega_\theta}{2}\right)^{\frac{n}{2}}}$$

and

$$\omega_\theta = \omega(\theta^* - \theta)^T(\theta^* - \theta) + \sigma_e^2$$

Proof: In Theorem 1,

$$\begin{aligned}\Psi_{L,\pi,\mathcal{D}}(\lambda, n) &= \ln \mathbb{E}_{\theta \sim \pi} \mathbb{E}_{(x_i, y_i) \sim \mathcal{D}} [e^{\lambda(R^{\text{true}}(\theta) - R^{\text{emp}}(\theta))}] \\ &= \ln \mathbb{E}_{\theta \sim \pi} \left(e^{\lambda R^{\text{true}}(\theta)} \mathbb{E}_{(x_i, y_i) \sim \mathcal{D}} [e^{-\lambda R^{\text{emp}}(\theta)}] \right)\end{aligned}$$

We have $R^{\text{true}}(\theta) = \mathbb{E}_{(x_i, y_i) \sim \mathcal{D}} [(y_i - \theta \cdot x_i)^2] = \check{v}_\theta$ and

$$\mathbb{E}_{(x_i, y_i) \sim \mathcal{D}} [e^{-\lambda R^{\text{emp}}(\theta)}] = \mathbb{E}_{(x_i, y_i) \sim \mathcal{D}} [e^{-\frac{\lambda}{n} \sum_{i=1}^n (y_i - \theta \cdot x_i)^2}]$$

Let $z_i = y_i - \theta \cdot x_i$ and $Z = [z_1, z_2, \dots, z_n]^T \in \mathbb{R}^{n \times 1}$. Then

$$\sum_{i=1}^n (y_i - \theta \cdot x_i)^2 = \sum_{i=1}^n z_i^2 = Z^T Z$$

Since the z_i s in Z are dependent, we need to convert them to independent random variables. The key idea is to use the covariance matrix. Denote $\mathcal{Q}_Z = \mathbb{E}[ZZ^T] \in \mathbb{R}^{n \times n}$. Since for any z_i and z_j ,

$$\begin{aligned}\mathbb{E}[z_i z_j] &= \mathbb{E}[(\theta^* - \theta)^T x_i (\theta^* - \theta)^T x_j] + \mathbb{E}[e_i e_j] \\ &= (\theta^* - \theta)^T \mathbb{E}[x_i x_j^T] (\theta^* - \theta) + \sigma_e^2 \mathbf{1}_{[i=j]}\end{aligned}$$

we have that $Q_Z = D_\theta^T Q_X D_\theta + \sigma_e^2 I$, where $I \in \mathbb{R}^{n \times n}$ is an identity matrix and

$$D_\theta = \text{diag}(\underbrace{(\theta^* - \theta), (\theta^* - \theta), \dots, (\theta^* - \theta)}_{n \text{ times}}) \in \mathbb{R}^{dn \times n}$$

Thus for any $p \in \mathbb{R}^d / \{0\}$,

$$\begin{aligned} p^T Q_Z p &= (D_\theta p)^T Q_X (D_\theta p) + \sigma_e^2 p^T p \\ &\geq \omega (D_\theta p)^T (D_\theta p) + \sigma_e^2 p^T p \end{aligned} \quad (3)$$

$$= [\omega(\theta^* - \theta)^T (\theta^* - \theta) + \sigma_e^2] p^T p \quad (4)$$

where Eq (3) is because: Suppose $Q_X = Q^T \Lambda Q$ is the eigenvalue decomposition of Q_X where $\Lambda = \text{diag}(\omega_1, \omega_2, \dots, \omega_{dn})$, let $\omega = \min\{\omega_1, \omega_2, \dots, \omega_{dn}\}$, $v = D_\theta p$, $u = Qv$, we have,

$$v^T Q_X v = u^T \Lambda u = \sum_{i=1}^{dn} \omega_i u_i^2 \geq \omega \sum_{i=1}^{dn} u_i^2 = \omega v^T Q^T Q v = \omega v^T v$$

Since $\omega(\theta^* - \theta)^T (\theta^* - \theta) + \sigma_e^2 > 0$, we have $p^T Q_Z p > 0$. Thus Q_Z is positive definite. Hence (1) Z is from $\mathcal{N}_n(0, Q_Z)$; (2) Q_Z must have an inverse Q_Z^{-1} . Let $Q_Z = Q^T \Lambda Q$, then $Q_Z^{-1} = Q^T \Lambda^{-1} Q$.

Let $Q_Z^{-1} = Q_Z^{-1/2} Q_Z^{-1/2}$ where $Q_Z^{-1/2} = Q^T \Lambda^{-1/2} Q$, we can write

$$Z^T Z = Z^T Q_Z^{-1/2} Q_Z Q_Z^{-1/2} Z = (Q_Z^{-1/2} Z)^T Q_Z (Q_Z^{-1/2} Z)$$

Let $S = Q_Z^{-1/2} Z = [s_1, s_2, \dots, s_n] \in \mathbb{R}^{n \times 1}$. By Theorem A.3.5, each s_i is a Gaussian random variable. We have $\mathbb{E}[S] = Q_Z^{-1/2} \mathbb{E}[Z] = 0$ and

$$\mathbb{E}[SS^T] = Q_Z^{-1/2} \mathbb{E}[ZZ^T] Q_Z^{-1/2} = Q_Z^{-1/2} Q_Z Q_Z^{-1/2} = I$$

which means all elements in S are i.i.d. from $\mathcal{N}(0, 1)$. By Eq (4), let $p = S$, then

$$Z^T Z = S^T Q_Z S \geq [\omega(\theta^* - \theta)^T (\theta^* - \theta) + \sigma_e^2] S^T S = \omega_\theta S^T S = \omega_\theta \left(\sum_{i=1}^n s_i^2 \right)$$

where $\sum_{i=1}^n s_i^2 \sim \chi^2(n)$. Therefore,

$$\begin{aligned} \Psi_{L, \pi, \mathcal{D}}(\lambda, n) &= \ln \mathbb{E}_{\theta \sim \pi} \left(e^{\lambda \check{v}_\theta} \mathbb{E}_{(x_i, y_i) \sim \mathcal{D}} [e^{-\frac{\lambda}{n} \sum_{i=1}^n z_i^2}] \right) \\ &\leq \ln \mathbb{E}_{\theta \sim \pi} \left(e^{\lambda \check{v}_\theta} \mathbb{E}_{(x_i, y_i) \sim \mathcal{D}} [e^{-\frac{\lambda \omega_\theta}{n} \sum_{i=1}^n s_i^2}] \right) = \ln \mathbb{E}_{\theta \sim \pi} \frac{e^{\lambda \check{v}_\theta}}{\left(1 + \frac{\lambda \omega_\theta}{n} \right)^{\frac{n}{2}}} \end{aligned}$$

□

Theorem 2.7 implies that when the dimensions of x_i are not i.i.d Gaussian, as $n \rightarrow \infty$, $\Psi_{L,\pi,\mathcal{D}}(\lambda, n)$ will converge but not converge to 0. This is because

$$\lim_{n \rightarrow \infty} \ln \mathbb{E}_{\theta \sim \pi} \frac{e^{\lambda \check{v}_\theta}}{\left(1 + \frac{\lambda \omega_\theta}{2}\right)^{\frac{n}{2}}} = \ln \mathbb{E}_{\theta \sim \pi} e^{\lambda(\check{v}_\theta - \omega_\theta)}$$

And as we have shown in the proof, $\check{v}_\theta \geq \omega_\theta$. The equality is obtained only when all of the eigenvalues of Q_Z are equal, which is not likely to happen.

References

- [1] Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 2016. <https://jmlr.org/papers/volume17/15-290/15-290.pdf>
- [2] Vera Shalaeva et al. Improved PAC-bayesian bounds for linear regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. <https://arxiv.org/pdf/1912.03036.pdf>
- [3] Pascal Germain et al. PAC-Bayesian theory meets Bayesian inference. *Advances in Neural Information Processing Systems* 29, 2016. <https://arxiv.org/pdf/1605.08636.pdf>
- [4] Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. *Concentration inequalities*. Clarendon Press, 2012. <https://www.hse.ru/data/2016/11/24/1113029206/Concentrationinequalities.pdf>
- [5] Lloyd N. Trefethen, David Bau. *Numerical Linear Algebra*. SIAM, 1997. <http://www.stat.uchicago.edu/~lekheng/courses/309/books/Trefethen-Bau.pdf>
- [6] Sum of Non-iid Gaussian Random Variables <https://stats.stackexchange.com/question/s/19948/what-is-the-distribution-of-the-sum-of-non-i-i-d-gaussian-variates>
- [7] Non-iid Assumption of Statistical Learning. <https://stats.stackexchange.com/question/s/213464/on-the-importance-of-the-i-i-d-assumption-in-statistical-learning>

Appendix 1: Sub-Gaussian and Sub-Gamma Distribution

This section introduces some fundamental ideas of concentration inequalities from the book [4]. Concentration inequalities explains under what conditions the random variables will concentrate around their expectations.

Given a random variable X satisfying $\mathbb{E}X = 0$. For any $t > 0$, we say $P(X > t)$ is the **right tail probability** of X and $P(X < -t)$ is the **left tail probability** of X .

Now we show the connection between MGF and tail probabilities. Denote

$$\psi_X(\lambda) = \ln \mathbb{E}[e^{\lambda X}]$$

as the logarithm of the MGF of X .

Definition A.1.1: If $\psi_X(\lambda) \leq \frac{\lambda^2 v}{2}$, then X satisfies **sub-Gaussian distribution** with variance factor v .

Theorem A.1.2: If X satisfies sub-Gaussian distribution with variance v , then for any $t > 0$,

$$P(X > t) \leq e^{-\frac{t^2}{2v}} \quad \text{and} \quad P(X < -t) \leq e^{-\frac{t^2}{2v}}$$

Proof:

Let $\lambda > 0$, applying Chernoff inequality

$$P(X > t) = P(e^{\lambda X} > e^{\lambda t}) \leq e^{-\lambda t} \mathbb{E}[e^{\lambda X}] \leq e^{-\lambda t} e^{\frac{\lambda^2 v}{2}} = e^{\frac{\lambda^2 v}{2} - \lambda t}$$

Now we find λ to minimize the upper bound. When $\lambda = \frac{t}{v}$, we get $\min_{\lambda} \{ \frac{\lambda^2 v}{2} - \lambda t \} = -\frac{t^2}{2v}$. The $P(X < -t)$ case can be proved similarly. Just applying Chernoff bound with $\lambda < 0$. □

Theorem A.1.2 says the tail probability of a sub-Gaussian random variable is upper-bounded by a Gaussian distribution with 0 mean and v variance.

Theorem A.1.3: If $X \sim \mathcal{N}(\mathbb{E}[X], v)$, then $Y = X - \mathbb{E}X$ is sub-Gaussian with variance v .

Proof: Since $Y \sim \mathcal{N}(0, v)$, for any $\lambda > 0$,

$$\begin{aligned} P(Y > t) &\leq -e^{\lambda t} \mathbb{E}[e^{\lambda Y}] = e^{-\lambda t} \int e^{\lambda y} \frac{1}{\sqrt{2\pi v}} e^{-\frac{y^2}{2v}} dy \\ &= e^{-\lambda t + \frac{\lambda^2 v}{2}} \int \frac{1}{\sqrt{2\pi v}} e^{-\frac{(y-v\lambda)^2}{2v}} dy \\ &= e^{-\lambda t + \frac{\lambda^2 v}{2}} \leq e^{-\frac{t^2}{2v}} \end{aligned}$$

$P(Y < -t)$ case can be proved in a similar way. □

Definition A.1.4: If $\psi_X(\lambda) \leq \frac{\lambda^2 v}{2(1-c\lambda)}$ and $0 < \lambda < \frac{1}{c}$, then X satisfies **sub-Gamma distribution** with variance factor v and scale parameter c .

The upper bound of MGF of sub-Gamma distribution is looser than sub-Gaussian distribution. For those random variables that are not quite sub-Gaussian but nearly, we can assume them to be sub-Gamma.

The PDF of the distribution $\text{Gamma}(a, b)$ is

$$f(x) = \frac{x^{a-1} e^{-\frac{x}{b}}}{\Gamma(a) b^a}, \quad x \geq 0$$

If $X \sim \text{Gamma}(a, b)$, then $\mathbb{E}[X] = ab$ and $\text{Var}[X] = ab^2$.

Theorem A.1.5: Let $X \sim \text{Gamma}(a, b)$, $Y = X - \mathbb{E}[X]$, then for any $t > 0$,

$$\psi_Y(\lambda) \leq \frac{\lambda^2 v}{2(1 - c\lambda)}$$

where $v = ab^2$, $c = b$, and $0 < \lambda < \frac{1}{b}$.

Proof:

$$\begin{aligned} \mathbb{E}[e^{\lambda Y}] &= \int_0^\infty e^{\lambda(x-ab)} \frac{x^{a-1} e^{-x/b}}{\Gamma(a) b^a} dx = \frac{e^{-\lambda ab}}{\Gamma(a) b^a} \int_0^\infty x^{a-1} e^{(\lambda - \frac{1}{b})x} dx \\ &= \frac{e^{-\lambda ab}}{\Gamma(a) b^a} \Gamma(a) \left(\frac{1}{\frac{1}{b} - \lambda}\right)^a = \frac{e^{-\lambda ab}}{(1 - b\lambda)^a} = e^{-\lambda ab - a \ln(1 - b\lambda)} \end{aligned} \quad (5)$$

Use the following Lemma:

Lemma A.1.6: For all $u \in (0, 1)$,

$$-\ln(1-u) - u \leq \frac{u^2}{2(1-u)}$$

Proof: By Taylor Theorem,

$$\ln(1-u) = \sum_{k=1}^{\infty} -\frac{u^k}{k} \geq -u - \sum_{k=2}^{\infty} \frac{u^k}{2} = -u - \frac{u^2}{2(1-u)}$$

□

Therefore, let $u = \lambda b$ where $0 < \lambda < \frac{1}{b}$, we have

$$e^{-\lambda ab - a \ln(1-b\lambda)} = \exp\left(\frac{\lambda^2 ab^2}{2(1-\lambda b)}\right)$$

And

$$\psi_Y(\lambda) = \ln \mathbb{E}[e^{\lambda Y}] \leq \frac{\lambda^2 ab^2}{2(1-\lambda b)}$$

□

Theorem shows that if X satisfies the Gamma distribution, then $Y = X - \mathbb{E}X$ satisfies sub-Gamma distribution. This bound holds for both right tail and left tail probability. Note that Y is a shifted Gamma distribution. Its left tail and right tail are not symmetric. In fact, for the left tail, we have a tighter bound.

Corollary A.1.7: Consider the settings of Theorem A.1.5. When $Y < 0$, we have

$$\psi_Y(\lambda) \leq \frac{\lambda^2 v}{2}$$

where $v = ab^2$ and $0 < \lambda < \frac{1}{b}$.

Proof: For any $u < 0$, we have

$$-\ln(1 - u) - u < \frac{u^2}{2} \tag{6}$$

Apply Eq (6) to Eq (5) by letting $u = \lambda b$, and theorem is proved. □

Corollary A.1.7 shows that the left tail probability of Y is sub-Gaussian, which is tighter than sub-Gamma. This means Y is more concentrated on left tail than right tail.

Theorem A.1.8: If a random variable X is of sub-Gamma with variance factor v and scale parameter c , then for any $t > 0$, we have

$$P(X > t) \leq \exp \left(-\frac{v}{c^2} h \left(\frac{ct}{v} \right) \right)$$

where $h(u) = 1 + u - \sqrt{1 + 2u}$ for $u > 0$. Or equivalently, for any $s > 0$,

$$P(X > \sqrt{2vs} + cs) \leq e^{-s}$$

Proof: Given that

$$\psi_X(\lambda) \leq \frac{\lambda^2 v}{2(1 - c\lambda)}$$

By Chernoff inequality, we have

$$P(X > t) \leq e^{-\lambda t} \mathbb{E}[e^{\lambda X}] \leq \exp \left(\frac{\lambda^2 v}{2(1 - c\lambda)} - \lambda t \right)$$

Let $f(\lambda) = \frac{\lambda^2 v}{2(1-c\lambda)} - \lambda t$, we want to find $\lambda \in (0, \frac{1}{c})$ to minimize $f(\lambda)$.

$$f'(\lambda) = -t + \frac{2\lambda v - c\lambda^2 v}{2(1-c\lambda)^2}, \quad f''(\lambda) = \frac{4v(1-c\lambda)^3 + 4\lambda cv(1-c\lambda)(2-c\lambda)}{4(1-c\lambda)^4}$$

Since $f''(\lambda) \geq 0$ on $(0, \frac{1}{c})$, solving $f'(\lambda) = 0$, we get

$$\lambda^* = \frac{1}{c} - \frac{\sqrt{v}}{c} \cdot \frac{1}{\sqrt{2tc+v}}$$

Thus

$$\min f(\lambda) = f(\lambda^*) = -\frac{v}{c^2} - \frac{t}{c} + \frac{\sqrt{v}}{c^2} \sqrt{2tc+v} = -\frac{v}{c^2} h\left(\frac{ct}{v}\right)$$

Since $h(u) = 1 + u - \sqrt{1+2u}$, we know that $h^{-1}(u) = u + \sqrt{2u}$. Thus

$$s = \frac{v}{c^2} h\left(\frac{ct}{v}\right) \iff t = \frac{v}{c} h^{-1}\left(\frac{sc^2}{v}\right) = sc + \sqrt{2sv}$$

□

Appendix 2: Germain's Bound

The Germain's bound is an earlier work of Theorem 2.3 (Shalaeva's Bound) given by Germail et al [3]. This bound is looser than Shalaeva's Bound. Moreover, it does not converge to 0 as $n \rightarrow \infty$ for any $\lambda > 0$.

In Theorem 1 we denote

$$\Psi_{L,\pi,\mathcal{D}}(\lambda, n) = \ln \mathbb{E}_{\theta \sim \pi} \mathbb{E}_{(x_i, y_i) \sim \mathcal{D}} [e^{\lambda(R^{\text{true}}(\theta) - R^{\text{emp}}(\theta))}]$$

Here we can consider $\Psi_{L,\pi,\mathcal{D}}(\lambda, n)$ as the logarithm MGF of the random variable $R^{\text{true}}(\theta) - R^{\text{emp}}(\theta)$, which is a function of both dataset $S = \{(x_i, y_i)\}_{i=1}^n$ and parameter θ . The following theorem shows that **when the loss function L is squared loss, $R^{\text{true}}(\theta) - R^{\text{emp}}(\theta)$ is a sub-Gamma random variable.**

Theorem A.2.1 (Germain's Bound): Under the same settings of Theorem 2.3, assume $\theta \sim \mathcal{N}_d(0, \sigma_\pi^2 I)$ where $\sigma_\pi > 0$ is a constant. Then

$$\Psi_{L,\pi,\mathcal{D}}(\lambda, n) \leq \frac{\lambda^2 v}{2(1 - c\lambda)}$$

where $v = \frac{2}{\lambda} [\sigma_x^2 (\sigma_\pi^2 d + \|w^*\|^2) + \sigma_e^2 (1 - \lambda c)]$ and $c = 2\sigma_x^2 \sigma_\pi^2$.

Proof:

$$\begin{aligned}\Psi_{L,\pi,\mathcal{D}}(\lambda, n) &= \ln \mathbb{E}_{\theta} \mathbb{E}_{(x_i, y_i)} [e^{\lambda(R^{\text{true}}(\theta) - R^{\text{emp}}(\theta))}] \\ &\leq \ln \mathbb{E}_{\theta} \mathbb{E}_{(x_i, y_i)} [e^{\lambda R^{\text{true}}(\theta)}] \end{aligned} \quad (7)$$

$$\begin{aligned} &= \ln \mathbb{E}_{\theta} \mathbb{E}_{(x_i, y_i)} [e^{\lambda \mathbb{E}_{(x_i, y_i)} [(y_i - \theta \cdot x_i)^2]}] \\ &= \ln \mathbb{E}_{\theta} [e^{\lambda \mathbb{E}_{(x_i, y_i)} [(y_i - \theta \cdot x_i)^2]}] \end{aligned} \quad (8)$$

$$= \ln \mathbb{E}_{\theta} [e^{\lambda(\sigma_x^2 \|\theta^* - \theta\|^2 + \sigma_e^2)}] \quad (9)$$

$$= \ln \left[\frac{1}{(1 - 2\lambda\sigma_x^2\sigma_{\pi}^2)^{\frac{d}{2}}} \exp \left(\frac{\lambda\sigma_x^2\|\theta^*\|^2}{1 - 2\lambda\sigma_x^2\sigma_{\pi}^2} + \lambda\sigma_e^2 \right) \right] \quad (10)$$

$$\begin{aligned} &= -\frac{d}{2} \ln(1 - 2\lambda\sigma_x^2\sigma_{\pi}^2) + \frac{\lambda\sigma_x^2\|w^*\|^2}{1 - 2\lambda\sigma_x^2\sigma_{\pi}^2} + \lambda\sigma_e^2 \\ &\leq \frac{\lambda\sigma_x^2\sigma_{\pi}^2 d}{1 - 2\lambda\sigma_x^2\sigma_{\pi}^2} + \frac{\lambda\sigma_x^2\|w^*\|^2}{1 - 2\lambda\sigma_x^2\sigma_{\pi}^2} + \lambda\sigma_e^2 \end{aligned} \quad (11)$$

$$= \frac{\lambda(\sigma_x^2\sigma_{\pi}^2 d + \sigma_x^2\|w^*\|^2 + (1 - 2\lambda\sigma_x^2\sigma_{\pi}^2)\sigma_e^2)}{1 - 2\lambda\sigma_x^2\sigma_{\pi}^2} = \frac{\lambda^2 v}{2(1 - c\lambda)}$$

where we let $v = \frac{2}{\lambda}[\sigma_x^2(\sigma_{\pi}^2 d + \|w^*\|^2) + \sigma_e^2(1 - \lambda c)]$ and $c = 2\sigma_x^2\sigma_{\pi}^2$.

Eq (7) is because $R^{\text{emp}}(\theta) \geq 0$. Eq (8) is because $e^{\lambda \mathbb{E}_{(x_i, y_i)}[(y_i - \theta \cdot x_i)^2]}$ is independent of x_i and y_i . Eq (9) is obtained by Lemma 2.2.

For Eq (10), since the elements of θ are iid,

$$\begin{aligned} \ln \mathbb{E}_{\theta} [e^{\lambda(\sigma_x^2 \|\theta^* - \theta\|^2 + \sigma_e^2)}] &= \ln \left[\mathbb{E}_{\theta} [e^{\lambda \sigma_x^2 \sum_{i=1}^d (\theta_i^* - \theta_i)^2}] e^{\lambda \sigma_e^2} \right] \\ &= \ln \left[\prod_{i=1}^d \mathbb{E}_{\theta} [e^{\lambda \sigma_x^2 (\theta_i^* - \theta_i)^2}] e^{\lambda \sigma_e^2} \right] \end{aligned} \quad (12)$$

where each $\theta_i^* - \theta_i \sim \mathcal{N}(\theta_i^*, \sigma_{\pi}^2)$. Then we will utilize the following Lemma.

Lemma A.2.2: If $Y \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\mathbb{E}_Y [e^{tY^2}] = (1 - 2t\sigma^2)^{-\frac{1}{2}} \exp \left(\frac{t\mu^2}{1 - 2t\sigma^2} \right)$$

Proof: Let $X \sim \mathcal{N}(0, 1)$, by transformation,

$$\begin{aligned}
\mathbb{E}_Y[e^{tY^2}] &= \int e^{ty^2} f_Y(y) dy = \int e^{t(\sigma x + \mu)^2} f_X(x) \frac{d}{dy} \left(\frac{y - \mu}{\sigma} \right) d(\sigma x + \mu) \\
&= \int e^{t(\sigma x + \mu)^2} f_X(x) dx = \mathbb{E}_X[e^{t(\sigma X + \mu)^2}]
\end{aligned}$$

And

$$\begin{aligned}
\mathbb{E}_X[e^{t(\sigma X + \mu)^2}] &= \int e^{t(\sigma x + \mu)^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\
&= \sqrt{\frac{1}{1 - 2t\sigma^2}} \int \frac{1}{\sqrt{2\pi} \sqrt{\frac{1}{1 - 2t\sigma^2}}} \exp \left(-\frac{\left(x - \frac{t\sigma}{\frac{1}{2} - t\sigma^2}\right)^2}{2 \frac{1}{1 - 2t\sigma^2}} \right) dx \exp \left(\frac{t\mu^2}{1 - 2t\sigma^2} \right)
\end{aligned}$$

The blue part of the above equation equals to 1. □

Applying Lemma A.2.2 to Eq (12) by letting $t = \lambda\sigma_x^2$, $\mu = \theta_i^*$, $\sigma = \sigma_\pi$, we get

$$\begin{aligned}
\ln \mathbb{E}_\theta[e^{\lambda(\sigma_x^2 \|\theta^* - \theta\|^2 + \sigma_e^2)}] &= \ln \left[\prod_{i=1}^d \left[(1 - 2\lambda\sigma_x^2\sigma_\pi^2)^{-\frac{1}{2}} \exp \left(\frac{\lambda\sigma_x^2\theta_i^{*2}}{1 - 2\lambda\sigma_x^2\sigma_\pi^2} \right) \right] e^{\lambda\sigma_e^2} \right] \\
&= \ln \left[\frac{1}{(1 - 2\lambda\sigma_x^2\sigma_\pi^2)^{\frac{d}{2}}} \exp \left(\frac{\lambda\sigma_x^2\|\theta^*\|^2}{1 - 2\lambda\sigma_x^2\sigma_\pi^2} + \lambda\sigma_e^2 \right) \right]
\end{aligned}$$

Eq (11) is because $-\ln(1-x) \leq \frac{x}{1-x}$ for $x < 1$ and apply $x = 2\lambda\sigma_x^2\sigma_\pi^2$.



Note that the bound of Theorem A.2.1 does not depend on n . This is because we removed $R^{\text{emp}}(\theta)$ in Eq (7), which is the only term containing n . So the bound will not converge as $n \rightarrow \infty$.

Appendix 3: Multivariate Gaussian Distribution

Definition A.3.1 (Covariance Matrix): Let $x \in \mathbb{R}^d$ be a random vector and $\mu = \mathbb{E}[x]$ be the expectation of x . The **covariance matrix** is defined as $\Sigma = \mathbb{E}[(x - \mu)(x - \mu)^T] \in \mathbb{R}^{d \times d}$, where $\Sigma_{ij} = \text{Cov}(x_i, x_j)$ for $1 \leq i, j \leq d$.

Definition A.3.2: Let $A \in \mathbb{R}^{d \times d}$ be a **symmetric** matrix,

- A is said to be **positive definite** if $x^T A x > 0$ for all $x \in \mathbb{R}^d / \{0\}$.
- A is said to be **positive semi-definite** if $x^T A x \geq 0$ for all $x \in \mathbb{R}^d$.

Theorem A.3.3: Let $A \in \mathbb{R}^{d \times d}$ be a symmetric matrix. A is positive definite if and only if all of its eigenvalues are positive.

Proof: First we show that A must have a eigenvalue decomposition $A = Q \Lambda Q^T$ where $Q \in \mathbb{R}^{d \times d}$ is an orthogonal matrix and $\Lambda \in \mathbb{R}^{d \times d}$ is a diagonal matrix. Since every square matrix A has a Schur factorization $A = Q T Q^T$ where T is an upper-triangular matrix (see Theorem 24.9 of [5]), if A is symmetric, then T is diagonal.

The diagonal matrix Λ must contain all the eigenvalues of A . This is because Q being orthogonal means there are d linearly independent eigenvectors, which implies the sum of geometric multiplicity of the eigenvalues is d . Since the geometric multiplicity of each eigenvalue must be not greater than its algebraic multiplicity and the sum of algebraic multiplicity of all eigenvalues is d , if one eigenvalue is missing in Λ , the sum of geometric multiplicity must be smaller than d , which is contradict.

Now we prove the theorem:

\implies : For any $x \in \mathbb{R}^d / \{0\}$, let $y = Q^T x$, then $y \neq 0$. Hence $x^T A x = y^T \Lambda y = \sum_{i=1}^d \lambda_i y_i^2$. If there exists $\lambda_i \leq 0$ for $i \in \{1, 2, \dots, d\}$, then we can find a non-zero x by letting $x = Qy$, $y_i \neq 0$ and $y_j = 0$ for all $j \neq i$ to make $x^T A x = \lambda_i y_i^2 \leq 0$, which is contradict.

\impliedby : If $\lambda_i > 0$ for any $i = 1, \dots, d$, then for any nonzero x , $x^T A x = \sum_{i=1}^d \lambda_i y_i^2 > 0$, which means A is positive definite.

□

Similar as how we prove Theorem A.3.3, one can prove that A is positive semi-definite if and only if all of its eigenvalues are non-negative.

Definition A.3.4 (Multivariate Gaussian): The PDF of the Multivariate Gaussian Distribution $\mathcal{N}_d(\mu, \Sigma)$ 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. $\Sigma = \mathbb{E}[(x - \mu)(x - \mu)^T] \in \mathbb{R}^{d \times d}$ is the covariance matrix.

By Definition A.3.1, the covariance matrix Σ is positive semi-definite. This is because for any $a \in \mathbb{R}^d$,

$$a^T \Sigma a = a^T \mathbb{E}[(x - \mu)(x - \mu)^T] a = \mathbb{E}[(a^T(x - \mu))^2] \geq 0$$

However, in Definition A.3.4, the Σ for multivariate Gaussian requires to be positive definite. This is because $\det \Sigma = \prod_{i=1}^d \lambda_i$. If there exists $\lambda_i = 0$, then $\det \Sigma = 0$, and the PDF cannot be formulated.

Theorem A.3.5: Let $x = [x_1, x_2, \dots, x_d]^T$ be a random vector of d dimensional multivariate Gaussian distribution $\mathcal{N}_d(\mu, \Sigma)$, and $a = [a_1, a_2, \dots, a_d] \in \mathbb{R}^d$ be a vector. Then $z = a^T x \in \mathbb{R}$ satisfies the Gaussian distribution $\mathcal{N}(a^T \mu, a^T \Sigma a)$.

Proof: The main idea of the proof comes from [6]. Let $X \in \mathbb{R}, t \in \mathbb{R}$, the MGF of $X \sim \mathcal{N}(\nu, \sigma^2)$ is

$$\begin{aligned} M_X(t) &= \mathbb{E}[e^{tX}] = \int e^{tx} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\nu)^2}{2\sigma^2}} dx \\ &= e^{\nu t + \frac{t^2 \sigma^2}{2}} \int \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\nu-t\sigma^2)^2}{2\sigma^2}} dx = e^{\nu t + \frac{t^2 \sigma^2}{2}} \end{aligned}$$

Let $Y \in \mathbb{R}^d, \lambda \in \mathbb{R}^d$, the MGF of $Y \sim \mathcal{N}(\mu, \Sigma)$ is

$$\begin{aligned} M_Y(\lambda) &= \mathbb{E}[e^{\lambda^T Y}] = \int e^{\lambda^T y} \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} e^{-\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)} dy \\ &= \int \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} e^{-\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu) + \lambda^T (y-\mu) + \lambda^T \mu} dy \end{aligned}$$

Let $m \in \mathbb{R}^d$. Since

$$\begin{aligned} & -\frac{1}{2}(y - \mu - m)^T \Sigma^{-1}(y - \mu - m) \\ &= -\frac{1}{2}(y - \mu)^T \Sigma^{-1}(y - \mu) + m^T \Sigma^{-1}(y - \mu) - \frac{1}{2}m^T \Sigma^{-1}m \end{aligned}$$

Let $m^T \Sigma^{-1} = \lambda^T$, then $\frac{1}{2}m^T \Sigma^{-1}m = \frac{1}{2}\lambda^T \Sigma \lambda$. Therefore,

$$\begin{aligned} M_Y(\lambda) &= \mathbb{E}[e^{\lambda^T Y}] = \int \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} e^{-\frac{1}{2}(y - \mu - m)^T \Sigma^{-1}(y - \mu - m) + \lambda^T \mu + \frac{1}{2}\lambda^T \Sigma \lambda} dy \\ &= e^{\lambda^T \mu + \frac{1}{2}\lambda^T \Sigma \lambda} \end{aligned}$$

Define a new random variable $Z = a^T Y$. The MGF of Z is

$$M_Z(t) = \mathbb{E}[e^{tZ}] = \mathbb{E}[e^{ta^T Y}]$$

Let $\lambda^T = ta^T$, then

$$\mathbb{E}[e^{ta^T Y}] = e^{ta^T \mu + \frac{1}{2}a^T \Sigma a t^2}$$

which means Z is of Gaussian distribution with mean $a^T \mu$ and variance $a^T \Sigma a$. \square

Appendix 4: Relationship with Least Squares

This section explains the relationship between the posterior distribution ρ and the least squares solution.

It is well known that for a given dataset $S = \{(x_i, y_i)\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$, the linear regression problem

$$\operatorname{argmin}_{\theta} R^{\text{emp}}(\theta) = \min_{\theta} \frac{1}{n} \sum_{i=1}^n (y_i - \theta \cdot x_i)^2 \quad (13)$$

has a unique solution

$$\hat{\theta} = (X^T X)^{-1} X^T Y \quad (14)$$

where $X = [x_1, x_2, \dots, x_n]^T \in \mathbb{R}^{n \times d}$ and $Y = [y_1, y_2, \dots, y_n]^T \in \mathbb{R}^{n \times 1}$.

We obtain the least squares solution without assuming any distribution for the data (x_i, y_i) . In this case, $\hat{\theta}$ is a constant, and the regressor $\hat{\theta} \cdot x_i$ fits best for the given data S . However, to evaluate the prediction of the regressor on unseen data, we usually assume x_i or y_i or both satisfy a distribution \mathcal{D} . Once x_i or y_i becomes random variable, $\hat{\theta}$ will become a random variable.

The least squares solution that minimizes empirical risk in Eq (13) will only be a point estimator but not a Bayes estimator, where the former estimates a fixed value and the latter estimates a distribution. Since when x_i is given, y_i is fixed, we can consider $\hat{\theta}$ in Eq (14) as a function of x_i s, i.e., $\hat{\theta} = W(x_1, x_2, \dots, x_n)$. When each x_i is considered as a random sample from $p(x_i)$, $\hat{\theta}$ is a point estimator.

The difference between point estimator and Bayes estimator is, point estimator considers θ as an unknown but fixed quantity, while Bayes estimator considers θ as a variable whose variation can be described by a probability distribution. Given dataset S , the Bayes estimation obtains ρ by

$$\rho(\theta|S) = \frac{p(S|\theta)\pi(\theta)}{p(S)}$$

where $p(S) = \int p(S|\theta)\pi(\theta)dS$. The point estimator like MLE treats $\frac{\pi(\theta)}{p(S)}$ as a constant and does not allow us to inject our prior beliefs $\pi(\theta)$ ¹. This means we let $\rho(\theta|S) = \alpha p(S|\theta)$, and the θ^* that maximizes the likelihood $p(S|\theta)$ gains the greatest probability in $\rho(\theta|S)$.

¹<https://stats.stackexchange.com/questions/74082/what-is-the-difference-in-bayesian-estimate-and-maximum-likelihood-estimate>

The posterior ρ exists but unknown, since $p(S|\theta)$ is unknown. Remember that the Alquier's Bound holds for any ρ . In practice, to calculate Alquier's Bound, we usually assume ρ to be the one that gives the tightest bound, i.e.,

$$\operatorname{argmin}_{\rho} \mathbb{E}_{\theta \sim \rho}[R^{\text{emp}}(\theta)] + \frac{1}{\lambda} D(\rho \parallel \pi)$$

The solution of the above problem is

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

which is named as the Gibbs posterior.