A Asymptotic Lower Bound of Shtarkov Complexity for Standard Normal Location Models

We show an asymptotic lower bound of the Shtarkov complexity of standard normal location models.

Lemma 8 *Consider the d-dimensional standard normal location model, given by $f_X(\theta) = \frac{1}{2}\|X-\theta\|_2^2 + \frac{d}{2}\ln 2\pi$, where $X \in \mathcal{X} = \mathbb{R}^d$. Let $\gamma = \lambda\|\theta\|_1$ for $\lambda \geq 0$. Then we have*

$$S(\gamma) \geq d \ln\left(1 + \frac{e^{-\lambda^2/2}}{\sqrt{2\pi}\lambda^3}(1+o(1))\right).$$

Proof By definition of $S(\gamma)$, we have

$$\begin{aligned}
S(\gamma) &= \ln \int e^{-m(f_X+\gamma)}\nu(\mathrm{d}X) \\
&= d\ln \int_{-\infty}^{\infty} \frac{\sup_{t\in\mathbb{R}} \exp\left[-\frac{1}{2}(x-t)^2 - \lambda|t|\right]}{\sqrt{2\pi}}\mathrm{d}x \\
&= d\ln \frac{1}{\sqrt{2\pi}}\left[\int_{-\infty}^{-\lambda} e^{-\lambda(-\lambda-x)-\frac{\lambda^2}{2}}\mathrm{d}x + \right. \\
&\qquad \left. \int_{-\lambda}^{\lambda} e^{-\frac{x^2}{2}}\mathrm{d}x + \int_{\lambda}^{\infty} e^{-\lambda(x-\lambda)-\frac{\lambda^2}{2}}\mathrm{d}x\right] \\
&= d\ln\left[2\Phi(\lambda) - 1 + \frac{2e^{-\lambda^2/2}}{\sqrt{2\pi}}\int_0^{\infty} e^{-\lambda x}\mathrm{d}x\right] \\
&= d\ln\left[2\Phi(\lambda) - 1 + \sqrt{\frac{2}{\pi}}\frac{e^{-\lambda^2/2}}{\lambda}\right],
\end{aligned}$$

where $\Phi(\lambda)$ denotes the standard normal distribution function. Now, by Komatu (1955), $\Phi(\lambda)$ is bounded below with $\Phi(\lambda) > 1 - 2\phi(\lambda)/(\sqrt{2+x^2}+x)$ for $\phi(\lambda)$ being the standard normal density, which yields the lower bound of interest after a few lines of elementary calculation. ■

B Lower Bound on Minimax Regret of Smooth Models

We describe how we adopt the minimax risk lower bound as to show the minimax-regret lower bound.

The story of the proof is based on Donoho and Johnstone (1994). First, the so-called three-point prior is constructed to approximate the least favorable prior. Then, since the approximate prior violates the ℓ_1-constraint, the degree of the violation is shown to be appropriately bounded to derive a valid lower bound.

The goal of our proof is to establish a lower bound on the minimax regret with respect to logarithmic losses, whereas their proof is about the minimax risk with respect to ℓ_q-loss. Therefore, below we present the proof highlighting (i) an approximate least favorable prior for *logarithmic losses* over ℓ_1-balls and (ii) the way to bound *regrets* on the basis of risk bounds.

Let $\mathcal{H} = \{\theta \in \mathbb{R}^d \mid \|\theta\|_1 \leq B\}$ be a ℓ_1-ball. Let $X \sim \mathcal{N}_d[\theta, I_d/L]$ be a d-dimensional normal random variable with mean $\theta \in H$ and precision $L > 0$. We denote the distribution just by $X \sim \theta$ where any confusion is unlikely. Let $h \in \hat{\mathcal{H}}$ be a predictor associated with any sub-probability distribution $P(\cdot|h) \in \mathcal{M}_+(\mathbb{R}^d)$. For notational simplicity, we may write $f_X(\theta) = \frac{L}{2}\|X-\theta\|_2^2 + \frac{d}{2}\ln\frac{2\pi}{L}$ and $f_X(h) = \ln\frac{\mathrm{d}P(X|h)}{\mathrm{d}\nu}$ where ν is the Lebesgue measure over $\mathbb{R}^d$.

Consider the risk function

$$R_d(h,\theta) \overset{\text{def}}{=} \mathbb{E}_{X\sim\theta}\left[f_X(h) - f_X(\theta)\right],$$

and the Bayes risk function

$$R_d(h,\pi) \overset{\text{def}}{=} \mathbb{E}_{\theta\sim\pi}\left[R_d(h,\theta)\right],$$

where $\pi \in \mathcal{P}(\mathcal{H})$ denotes prior distributions on $\mathcal{H}$. Then, the minimax Bayes risk bounds below the minimax regret,

$$\begin{aligned}
\mathrm{REG}^\star(\mathcal{H}) &= \inf_{h\in\hat{\mathcal{H}}} \sup_{\theta\in\mathcal{H}} \sup_{X\in\mathbb{R}^d} f_X(h) - f_X(\theta) \\
&\geq \inf_{h\in\hat{\mathcal{H}}} \sup_{\pi\in\mathcal{P}(\mathcal{H})} \mathbb{E}_{\theta\sim\pi}\mathbb{E}_{X\sim\theta}\left[f_X(h) - f_X(\theta)\right] \\
&= \inf_{h\in\hat{\mathcal{H}}} \sup_{\pi\in\mathcal{P}(\mathcal{H})} R_d(h,\pi).
\end{aligned}$$

The minimax theorem states that there exists a saddle point (h^*, π_*) such that

$$\begin{aligned}
R_d(h^*,\pi_*) &= \inf_{h\in\hat{\mathcal{H}}} \sup_{\pi\in\mathcal{P}(\mathcal{H})} R_d(h,\pi) \\
&= \sup_{\pi\in\mathcal{P}(\mathcal{H})} \inf_{h\in\hat{\mathcal{H}}} R_d(h,\pi) \overset{\text{def}}{=} \sup_{\pi\in\mathcal{P}(\mathcal{H})} R_d(\pi),
\end{aligned}$$

and π_* is referred to as the least favorable prior. We want to approximate π_* to give an analytic approximation of $R_d(\pi_*)$, which is a lower bound of $\mathrm{REG}^\star(\mathcal{H})$.

Let $F_{\epsilon,\mu} \in \mathcal{P}(\mathbb{R})$ be the three-point prior defined by

$$F_{\epsilon,\mu} = (1-\epsilon)\delta_0 + \frac{\epsilon}{2}(\delta_{-\mu} + \delta_\mu)$$

for $\epsilon, \mu > 0$. We show that the corresponding achievable Bayes risk $R_1(F_{\epsilon,\mu})$ tends to be the entropy of the prior $F_{\epsilon,\mu}$ in some limit of small ϵ.

Lemma 9 *Take $\mu = \mu(\epsilon) = \sqrt{2L^{-1}\ln\epsilon^{-1}}$. Let $H_\epsilon = H(F_{\epsilon,\mu}) = (1-\epsilon)\ln(1-\epsilon)^{-1} + \epsilon\ln 2\epsilon^{-1}$ be the entropy of the prior. Then we have*

$$R_1(F_{\epsilon,\mu}) \sim H_\epsilon \sim \epsilon\ln\frac{1}{\epsilon}$$

as $\epsilon \to 0$. Here, $x \sim y$ denotes the asymptotic equality such that $x/y \to 1$.

Proof First, we show the famous inequality on the entropy given by $R_1(F_{\epsilon,\mu}) \le H_\epsilon$. Let $P(\cdot|h) = \mathbb{E}_{\theta\sim F_{\epsilon,\mu}} P(\cdot|\theta) = (1-\epsilon)P(\cdot|0) + \frac{\epsilon}{2}(P(\cdot|-\mu)+P(\cdot|\mu))$ be the Bayes marginal distribution with respect to $F_{\epsilon,\mu}$. Then we have

$$\begin{aligned}
&H_\epsilon - R_1(F_{\epsilon,\mu})\\
&= H_\epsilon - R_1(h, F_{\epsilon,\mu})\\
&= H_\epsilon - \mathbb{E}_{\theta\sim F_{\epsilon,\mu}}\mathbb{E}_{X\sim\theta} \ln \frac{\mathrm{d}P(X|\theta)}{\mathrm{d}P(X|h)}\\
&= H_\epsilon - (1-\epsilon)\mathbb{E}_{P(X|0)} \ln \frac{\mathrm{d}P(X|0)}{\mathrm{d}P(X|h)}\\
&\quad - \epsilon\mathbb{E}_{P(X|\mu)} \ln \frac{\mathrm{d}P(X|\mu)}{\mathrm{d}P(X|h)}\\
&= (1-\epsilon)\mathbb{E}_{P(X|0)} \ln\left(1 + \frac{\epsilon}{1-\epsilon}\frac{\mathrm{d}P(X|\mu)+\mathrm{d}P(X|-\mu)}{2\mathrm{d}P(X|0)}\right)\\
&\quad + \epsilon\mathbb{E}_{P(X|\mu)} \ln\left(1 + \frac{1-\epsilon}{\epsilon}\frac{2\mathrm{d}P(X|0)+\mathrm{d}P(X|-\mu)}{\mathrm{d}P(X|\mu)}\right)\\
&\ge 0.
\end{aligned}$$

Now, we show that, with the specific value of $\mu = \mu(\epsilon)$, the gap is negligible compared to the entropy itself. Applying Jensen's inequality, we have

$$\begin{aligned}
&H_\epsilon - R_1(F_{\epsilon,\mu})\\
&\le \epsilon + \epsilon\mathbb{E}_{P(X|\mu)} \ln\left(1 + (1-\epsilon)\left(2e^{-L\mu X} + \epsilon^3 e^{-2L\mu X}\right)\right)\\
&\le \epsilon(1 + \ln 4 + \mathbb{E}_{P(X|\mu)} \max\{0,\ -2L\mu X\})\\
&= \epsilon\left(1 + \ln 4 + \mathop{\mathbb{E}}_{Z\sim\mathcal{N}[0,1]} \max\left\{0,\ 2\sqrt{L}\mu(Z - \sqrt{L}\mu)\right\}\right)\\
&\quad (\because -\sqrt{L}(X-\mu) = Z)\\
&\le \epsilon\left(1 + \ln 4 + 2\sqrt{L}\mu\epsilon\right)\\
&= \epsilon\left(1 + \ln 4 + 2\epsilon\sqrt{2\ln\frac{1}{\epsilon}}\right) = o(H_\epsilon).
\end{aligned}$$

Thus we get $H_\epsilon \sim R_1(F_{\epsilon,\mu})$. ■

Now we show that the d-th Kronecker product of $F_{\epsilon,\mu}$, $F^d_{\epsilon,\mu}$, can be used to bound the Bayes minimax risk $R_d(\pi_*)$ with an appropriate choice of ϵ and μ. To this end, let $\pi_+ = F^d_{\epsilon,\mu} \mid \mathcal{H}$ be the conditional prior restricted over the ℓ_1-ball $\mathcal{H}$.

Lemma 10 *Take $\epsilon\mu = (1-c)B/d$ and $\mu = \sqrt{2L^{-1}\ln\epsilon^{-1}}$ for $0 < c < 1$. Then, if $\epsilon \to 0$ and $d\epsilon \to \infty$, we have*

$$R_d(\pi_*) \ge R_d(\pi_+) \sim R_d(F^d_{\epsilon,\mu}) \sim d\epsilon \ln\frac{1}{\epsilon}.$$

Proof First of all, the inequality is trivial from the definition of $R_d(\pi)$. Moreover, the second asymptotic equality immediately follows from Lemma 9.

Now we consider the first asymptotic equality. Let h be the Bayesian predictor with respect to the prior $F_{\epsilon,\mu}$ and h^+ be the one with respect to the conditional prior π_+. Then we have

$$\begin{aligned}
&R_d(F^d_{\epsilon,\mu})\\
&= R_d(h, F^d_{\epsilon,\mu})\\
&= \mathbb{E}_{\theta\sim F^d_{\epsilon,\mu}}\left[R_d(h,\theta)\right]\\
&= F^d_{\epsilon,\mu}(\mathcal{H})R_d(h,\pi_+) + \mathbb{E}_{\theta\sim F^d_{\epsilon,\mu}}\left[R_d(h,\theta)\cdot\mathbb{1}\{\theta\notin\mathcal{H}\}\right]\\
&\ge F^d_{\epsilon,\mu}(\mathcal{H})\cdot R_d(\pi_+)
\end{aligned}$$

and

$$\begin{aligned}
R_d(F^d_{\epsilon,\mu}) &\le R_d(h^+, F^d_{\epsilon,\mu})\\
&= \mathbb{E}_{\theta\sim F^d_{\epsilon,\mu}}\left[R_d(h^+,\theta)\right]\\
&= F^d_{\epsilon,\mu}(\mathcal{H})\cdot R_d(\pi_+) +\\
&\quad \mathbb{E}_{\theta\sim F^d_{\epsilon,\mu}}\left[R_d(h^+,\theta)\cdot\mathbb{1}\{\theta\notin\mathcal{H}\}\right].
\end{aligned}$$

Let N be the number of nonzero elements in $\theta \sim F^d_{\epsilon,\mu}$. Then N is subjects to the Binomial distribution $\mathrm{Bin}(d,\epsilon)$. On the other hand, the event $\theta \in \mathcal{H}$ is equal to $\{\|\theta\|_1 \le B\} = \{N \le B/\mu = \mathbb{E}N/(1-c)\}$. Therefore, applying the Chebyshev's inequality, we get

$$\begin{aligned}
P_d \overset{\mathrm{def}}{=} F^d_{\epsilon,\mu}(\mathcal{H}^c) &= \Pr\left\{\frac{N-\mathbb{E}N}{\mathbb{E}N} > \frac{c}{1-c}\right\}\\
&\le \frac{(1-c)^2}{c^2 d\epsilon} \to 0.
\end{aligned}$$

Similarly, we have $\mathbb{E}\left|N-\mathbb{E}N\right|/\mathbb{E}N \to 0$. Now observe that

$$\begin{aligned}
&\mathbb{E}_{\theta\sim F^d_{\epsilon,\mu}}\left[R_d(h^+,\theta)\cdot\mathbb{1}\{\theta\notin\mathcal{H}\}\right]\\
&\le \mathbb{E}_{\theta\sim F^d_{\epsilon,\mu}}\mathbb{E}_{\varphi\sim\pi_+}\left[R_d(\varphi,\theta)\cdot\mathbb{1}\{\theta\notin\mathcal{H}\}\right].\\
&\le 2L\mathbb{E}_{\theta\sim F^d_{\epsilon,\mu}}\mathbb{E}_{\varphi\sim\pi_+}\left[\left(\|\varphi\|_2^2 + \|\theta\|_2^2\right)\cdot\mathbb{1}\{\theta\notin\mathcal{H}\}\right]\\
&\le 2L\mu^2\mathbb{E}\left[P_d N + N\cdot\mathbb{1}\{N > B/\mu\}\right]\\
&\qquad (\because \|\theta\|_2^2 = \mu^2 N)\\
&\le 2L\mu^2\mathbb{E}N\left(2P_d + \frac{\mathbb{E}\left|N-\mathbb{E}N\right|}{\mathbb{E}N}\right)\\
&= 4d\epsilon\ln\frac{1}{\epsilon}\left(2P_d + \frac{\mathbb{E}\left|N-\mathbb{E}N\right|}{\mathbb{E}N}\right).\\
&= o(R_d(F^d_{\epsilon,\mu})).
\end{aligned}$$

Thus, combining all above, we get

$$\begin{aligned}
&(1+o(1))R_d(\pi_+)\\
&= (1-P_d)R_d(\pi_+)\\
&\le R_d(F^d_{\epsilon,\mu})\\
&\le (1-P_d)\cdot R_d(\pi_+) +\\
&\quad \mathbb{E}_{\theta\sim F^d_{\epsilon,\mu}}\left[R_d(h^*,\theta)\cdot\mathbb{1}\{\theta\notin\mathcal{H}\}\right].\\
&= (1-o(1))R_d(\pi_*) + o(R_d(F^d_{\epsilon,\mu})),
\end{aligned}$$

which implies the desired asymptotic equality $R_d(F_{\epsilon,\mu}) \sim R_d(\pi_+)$. ■

Summing these up, we have an asymptotic lower bound on the minimax regret which is the same as the upper bound given by the ST prior within a factor of two (see Theorem 7). This implies that both the regret of the ST prior and the Bayes risk of the prior π_+ are tight with respect to the minimax-regret rate except with a factor of two.

Theorem 11 (Minimax lower bound) *Suppose that* $\omega(1) = \ln(d/\sqrt{L}) = o(L)$. *Then we have*

$$\mathrm{REG}^\star(\mathcal{H}) \gtrsim \frac{B}{2}\sqrt{2L\ln\frac{d}{\sqrt{L}}},$$

where $x \gtrsim y$ *means that there exists* $y' \sim y$ *such that* $x \geq y'$.

Proof The assumptions of Lemma 10 are satisfied for all $0 < c < 1$ since

$$\epsilon \lesssim \epsilon\sqrt{\ln\frac{1}{\epsilon}} = \frac{1-c}{d}\sqrt{\frac{L}{2}} \to 0,$$
$$d\epsilon = (1-c)\sqrt{\frac{L}{2\ln\frac{1}{\epsilon}}} \sim (1-c)\sqrt{\frac{L}{2\ln\frac{d}{\sqrt{L}}}} \to \infty.$$

Thus, we have

$$\mathrm{REG}^\star(\mathcal{H}) \geq R_d(\pi_*) \gtrsim d\epsilon\ln\frac{1}{\epsilon} \sim (1-c)\,\frac{B}{2}\sqrt{2L\ln\frac{d}{\sqrt{L}}}$$

for all $0 < c < 1$. Slowly moving c toward zero completes the theorem. ■

C Existence of Gap between $\mathrm{LREG}^\star$ and $\mathrm{LREG}^{\mathrm{Bayes}}$ under ℓ_1-Penalty

Below we show that, under standard normal location models, the Bayesian luckiness minimax regret is strictly larger than the non-Bayesian luckiness minimax regret if γ is nontrivial and has a non-differentiable point. Here we refer to γ as *trivial* when there exists θ_0 such that $\gamma(\theta) = \infty$ for all $\theta \neq \theta_0$.

Lemma 12 *Let* $f_X(\theta) = \frac{1}{2}(X-\theta)^2 + \frac{1}{2}\ln 2\pi$ *for* $X \in \mathbb{R}$ *and* $\theta \in \mathbb{R}$. *Then, for all nontrivial, convex and non-differentiable penalties* $\gamma : \mathbb{R} \to \overline{\mathbb{R}}$,

$$\mathrm{LREG}^\star(\gamma) < \mathrm{LREG}^{\mathrm{Bayes}}(\gamma).$$

Proof Let $\mathcal{F} = \{f_X \mid X \in \mathbb{R}\}$ and recall that $\mathrm{LREG}^{\mathrm{Bayes}}(\gamma) = \inf_{w\in\mathcal{E}(\mathcal{F}_\gamma)} \ln w\left[e^{-\gamma}\right]$ by Theorem 1. Let $\|\cdot\|_\gamma$ be the metric of pre-priors $w \in \mathcal{M}_+(\mathbb{R})$ given by $\|w\|_\gamma = w\left[e^{-\gamma}\right]$. Owing to the continuity of $w \mapsto \ln w\left[e^{-\gamma}\right]$ and the completeness of $\mathcal{E}(\mathcal{F}_\gamma) \subset \mathcal{M}_+(\mathbb{R})$, it suffices to show that there exists no pre-prior $w \in \mathcal{E}(\mathcal{F}_\gamma)$ such that $\ln w\left[e^{-\gamma}\right] = S(\gamma)$. Let us prove this by contradiction. Now, assume that $\ln w\left[e^{-\gamma}\right] = S(\gamma)$. Observe that

$$\begin{aligned} 0 &= w\left[e^{-\gamma}\right] - \exp S(\gamma) \\ &= w\left[\int e^{-f_X-\gamma}\nu(\mathrm{d}X)\right] - \int e^{-m(f_X+\gamma)}\nu(\mathrm{d}X) \\ &= \int\left\{w\left[e^{-f_X-\gamma}\right] - e^{-m(f_X+\gamma)}\right\}\nu(\mathrm{d}X), \end{aligned}$$

which means $w\left[e^{-f_X-\gamma}\right] = e^{-m(f_X+\gamma)}$ for almost every X since $w \in \mathcal{E}(\mathcal{F}_\gamma)$. Note that $f_X(\theta)$ is continuous with respect to X, and then we have $w\left[e^{-f_X-\gamma}\right] = e^{-m(f_X+\gamma)}$ for all X. After some rearrangement and differentiation, we have

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}X} w\left[e^{-f_X-\gamma+m(f_X+\gamma)}\right] \\ &= w\left[\frac{\mathrm{d}e^{-f_X-\gamma+m(f_X+\gamma)}}{\mathrm{d}X}\right] \\ &= w_\theta\left[(\theta-\theta^*_X)\,e^{-f_X-\gamma+m(f_X+\gamma)}\right], \end{aligned} \tag{13}$$

where $\theta^*_X = \arg m(f_X+\gamma)$. Here we exploited Danskin's theorem at the last equality. One more differentiation gives us

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}X} w_\theta\left[(\theta-\theta^*_X)\,e^{-f_X-\gamma+m(f_X+\gamma)}\right], \\ &= w_\theta\left[\left\{(\theta-\theta^*_X)^2 - \frac{\mathrm{d}\theta^*_X}{\mathrm{d}X}\right\}e^{-f_X-\gamma+m(f_X+\gamma)}\right] \end{aligned}$$

for all $X \in \mathbb{R}$.

Note that we have $\frac{\mathrm{d}\theta^*_X}{\mathrm{d}X}|_{X=t} = 0$ for any non-differentiable points t of γ. Then it implies that $w = c\delta_{\theta^*_t}$ where δ_s denotes the Kronecker delta measure. Then, according to (13), we have

$$\begin{aligned} 0 &= w_\theta\left[(\theta-\theta^*_X)\,e^{-f_X-\gamma+m(f_X+\gamma)}\right]. \\ &= c\left(\theta^*_t-\theta^*_X\right)e^{-f_X(\theta^*_t)-\gamma(\theta^*_t)+m(f_X+\gamma)}, \end{aligned}$$

which means that $\theta^*_X = \theta^*_t$ is a constant independent of X. However, this contradicts to the assumption that γ is nontrivial. ■

As a remark, we note that this lemma is easily extended to multidimensional exponential family of distributions.