# Barron and Cover's Theory in Supervised Learning and its Application to Lasso

**Masanori Kawakita** KAWAKITA@INF.KYUSHU-U.AC.JP
**Jun'ichi Takeuchi** TAK@INF.KYUSHU-U.AC.JP
Kyushu University, 744 Motooka, Nishi-Ku, Fukuoka city, Fukuoka 819-0395, JAPAN

## Abstract

We study Barron and Cover's theory (BC theory) in supervised learning. The original BC theory can be applied to supervised learning only approximately and limitedly. Though Barron & Luo (2008) and Chatterjee & Barron (2014a) succeeded in removing the approximation, their idea cannot be essentially applied to supervised learning in general. By solving this issue, we propose an extension of BC theory to supervised learning. The extended theory has several advantages inherited from the original BC theory. First, it holds for finite sample number $n$. Second, it requires remarkably few assumptions. Third, it gives a justification of the MDL principle in supervised learning. We also derive new risk and regret bounds of lasso with random design as its application. The derived risk bound hold for any finite $n$ without boundedness of features in contrast to past work. Behavior of the regret bound is investigated by numerical simulations. We believe that this is the first extension of BC theory to general supervised learning without approximation.

## 1. Introduction

There have been various techniques to evaluate performance of machine learning methods theoretically. For an example, lasso (Tibshirani, 1996) has been analyzed by nonparametric statistics, empirical process, statistical physics and so on. Most of them require various assumptions like asymptotic assumption (sample number $n$ and/or feature number $p$ go to infinity), boundedness of features or moment conditions. Some of them are much restrictive for practical use. In this paper, we try to develop another way for performance evaluation with as few assumptions as possible. As an important candidate for this purpose, we focus on Barron and Cover's theory (BC theory) (Barron & Cover, 1991), which is one of the most famous results for the minimum description length (MDL) principle. The MDL principle (Rissanen, 1978; Barron et al., 1998; Grünwald, 2007; Takeuchi, 2014) claims that the shortest description of a given set of data leads to the best hypotheses about the data. A famous model selection criterion based on the MDL principle was proposed by Rissanen (1978). This criterion corresponds to a codelength of a two-stage code in which one encodes a statistical model to encode data and then the data are encoded with the model. In this case, an MDL estimator is defined as the minimizer of the total codelength of this two-stage code. BC theory guarantees that a risk based on the Rényi divergence (Rényi, 1961) is tightly bounded above by redundancy of the two-stage code. This result gives a mathematical justification of the MDL principle. Furthermore, BC theory holds for finite $n$ without any complicated technical conditions. However, BC theory has been applied to supervised learning only approximately or limitedly. The original BC theory seems to be widely recognized that it is applicable to both unsupervised and supervised learning. Though it is not false, BC theory actually cannot be applied to supervised learning without a certain condition (2) defined in Section 3. This condition is critical in a sense that a lack of (2) breaks a key technique of BC theory. (Yamanishi, 1992) is the only example of application of BC theory to supervised learning to our knowledge. His work assumed a specific setting, where (2) can be satisfied. However, the risk bound may not be sufficiently tight due to imposing (2) forcedly, which will be explained in Section 3. Another well-recognized disadvantage is the necessity of quantization of parameter space. Barron & Luo (2008) and Chatterjee & Barron (2014b) proposed a way to avoid the quantization and derived a risk bound of lasso. However, their idea cannot be applied to supervised learning in general. This difficulty stems from the above condition (2). It is thus essentially difficult to solve. Actually, their risk bound of lasso was derived with fixed design only (i.e., essentially unsupervised

*Proceedings of the $33^{rd}$ International Conference on Machine Learning*, New York, NY, USA, 2016. JMLR: W&CP volume 48. 

setting). The fixed design, however, is not satisfactory to evaluate generalization error of supervised learning. In this paper, we propose an extension of BC theory to supervised learning without quantization in random design case. The derived risk bound inherits most of advantages of the original BC theory. Furthermore, we can use data-dependent penalties. The main term of the risk bound has again a form of redundancy of two-stage code. Thus, our extension also gives a mathematical justification of the MDL principle in supervised learning. We also derive new risk and regret bounds of lasso with random design as its application under normality of features. This requires much more effort than that for the fixed design case. The derived bounds hold even in case $n \ll p$ without boundedness of features. To our knowledge, no theory has such advantages in the past. This paper is organized as follows. Section 2 introduces an MDL estimator in supervised learning. We review BC theory in Section 3. We extend BC theory to supervised learning and derive new risk and regret bounds of lasso in Section 4. The performance of the regret bound will be investigated by numerical simulations in Section 5.

## 2. MDL Estimator in Supervised Learning

Suppose that training data $(x^n, y^n) := \{(x_i, y_i) \in \mathscr{X} \times \mathscr{Y} | i = 1, 2, \cdots, n\}$ are subject to $\bar{p}_*(x^n, y^n) = q_*(x^n) p_*(y^n|x^n)$, where $\mathscr{X}$ is a domain of feature vector $x$ and $\mathscr{Y}$ could be $\Re$ (regression) or a finite set (classification). Here, the training data are not necessarily independent and identically distributed (i.i.d.) but can be a stochastic process in general. A goal of supervised learning is to estimate $p_*(y^n|x^n)$. We use a parametric model $p_\theta(y^n|x^n)$ with a parameter $\theta \in \Theta$. The parameter space $\Theta$ is a certain continuous space. To define an MDL estimator, we need to encode the model $p_\theta(y^n|x^n)$ (or equivalently the parameter). Since the continuous parameter cannot be encoded, we need to quantize the parameter space $\Theta$ as $\widetilde{\Theta}(x^n)$. Then, let $\tilde{L}(\tilde{\theta}|x^n)$ be a model description length on it. Note that $\tilde{L}$ must satisfy Kraft's inequality $\sum_{\tilde{\theta} \in \widetilde{\Theta}(x^n)} \exp(-\tilde{L}(\tilde{\theta}|x^n)) \le 1$. An MDL estimator is defined by the minimizer of sum of a data description length (minus log-likelihood) and the model description length:

$$\ddot{\theta}(x^n, y^n) := \arg \min_{\tilde{\theta} \in \widetilde{\Theta}(x^n)} \big\{ -\log p_{\tilde{\theta}}(y^n|x^n) + \beta \tilde{L}(\tilde{\theta}|x^n) \big\},$$

where $\beta > 1$. Define the minimum description length attained by the two-stage code as $\tilde{L}_{\beta 2\text{-}p}(y^n|x^n) := -\log p_{\ddot{\theta}}(y^n|x^n) + \beta \tilde{L}(\ddot{\theta}|x^n)$. Because $\tilde{L}_{\beta 2\text{-}p}$ also satisfies Kraft's inequality in terms of $y^n$, it is interpreted as a codelength of the two-stage code. Therefore, $\tilde{p}_{\beta 2\text{-}p}(y^n|x^n) := \exp(-\tilde{L}_{\beta 2\text{-}p}(y^n|x^n))$ is a conditional sub-probability distribution corresponding to the two-stage code.

## 3. Barron and Cover's Theory

We briefly review BC theory and its recent progress in view of supervised learning though they discussed basically unsupervised learning (or supervised learning with fixed design). In BC theory, the Rényi divergence between $p(y^n|x^n)$ and $r(y^n|x^n)$ with order $\lambda \in (0, 1)$

$$d^n_\lambda(p, r) := \frac{-1}{1-\lambda} \log E_{q_*(x^n) p(y^n|x^n)} \Big( \frac{r(y^n|x^n)}{p(y^n|x^n)} \Big)^{1-\lambda} \tag{1}$$

is used as a loss function. Let us introduce a condition that both $\widetilde{\Theta}(x^n)$ and $\tilde{L}(\tilde{\theta}|x^n)$ are independent of $x^n$, i.e.,

$$\widetilde{\Theta}(x^n) = \widetilde{\Theta}, \ \ \tilde{L}(\tilde{\theta}|x^n) = \tilde{L}(\tilde{\theta}). \tag{2}$$

We emphasize that the original BC theory cannot be applied to supervised learning unless the condition (2) is satisfied. Under the condition (2), BC theory gives the following two theorems for supervised learning.

**Theorem 1.** *Let $\beta > 1$. Assume that $\tilde{L}$ satisfies Kraft's inequality and that the condition (2) holds. For any $\lambda \in (0, 1 - \beta^{-1}]$,*

$$E_{\bar{p}_*(x^n, y^n)} d^n_\lambda(p_*, p_{\ddot{\theta}}) \le E_{\bar{p}_*(x^n, y^n)} \log \frac{p_*(y^n|x^n)}{\tilde{p}_{\beta 2\text{-}p}(y^n|x^n)}.$$

**Theorem 2.** *Let $\beta > 1$. Assume that $\tilde{L}$ satisfies Kraft's inequality and that the condition (2) holds. For any $\lambda \in (0, 1 - \beta^{-1}]$,*

$$\Pr\Big( \frac{d^n_\lambda(p_*, p_{\ddot{\theta}})}{n} - \frac{1}{n} \log \frac{p_*(y^n|x^n)}{\tilde{p}_{\beta 2\text{-}p}(y^n|x^n)} \ge \tau \Big) \le e^{-\tau n/\beta}.$$

Recall that the quantized space and the model description length can depend on $x^n$ in their definitions. If we make them independent of $x^n$ for the condition (2), we must make them uniform against $x^n$ (i.e., its worst value), which makes the total codelength longer. This is just a reason why we think the PAC bound by Yamanishi (1992) may not be sufficiently tight. Hence, data-dependent model description lengths is more desirable in view of the MDL principle. In addition, the restriction by (2) excludes a practically important case 'lasso with column normalization' (explained later) from the scope of application. However, it is essentially difficult to remove this restriction as described in Section 1. Another issue is quantization. The quantization for the encoding is natural in view of the MDL principle. Our target, however, is an application to machine learning. A trivial example of such an application is a penalized maximum likelihood estimator (PMLE)

$$\hat{\theta}(x^n, y^n) \ := \ \arg \min_{\theta \in \Theta} \big\{ -\log p_\theta(y^n|x^n) + L(\theta|x^n) \big\},$$

$$p_{\beta 2\text{-}p}(y^n|x^n) \ := \ p_{\hat{\theta}}(y^n|x^n) \cdot \exp(-L(\hat{\theta}|x^n)),$$

where $L : \Theta \times \mathscr{X}^n \to [0, \infty]$ is a certain penalty. PMLE is a wide class of estimators including lasso. If we can accept $\ddot{\theta}$ as an approximation of $\hat{\theta}$, we have a risk bound obtained by BC theory. However, the quantization is unnatural in view of machine learning. Barron et al. (2008) proposed an important notion '*risk validity*' to remove this drawback.

**Definition 3.** *Let $\beta > 1$. For fixed $x^n$, we say that a penalty function $L(\theta|x^n)$ is risk valid if there exist a quantized space $\widetilde{\Theta}(x^n) \subset \Theta$ and a model description length $\tilde{L}(\tilde{\theta}|x^n)$ satisfying Kraft's inequality such that*

$$\forall y^n \in \mathscr{Y}^n, \max_{\theta\in\Theta}\Big\{d_\lambda^n(p_*, p_\theta|x^n) - \log\frac{p_*(y^n|x^n)}{p_\theta(y^n|x^n)} - L(\theta|x^n)\Big\}$$
$$\le \max_{\tilde{\theta}\in\widetilde{\Theta}(x^n)}\Big\{d_\lambda^n(p_*, p_{\tilde{\theta}}|x^n) - \log\frac{p_*(y^n|x^n)}{p_{\tilde{\theta}}(y^n|x^n)} - \beta\tilde{L}(\tilde{\theta}|x^n)\Big\}, \quad (3)$$
$$\text{where } d_\lambda^n(p, r|x^n) := \frac{-1}{1-\lambda}\log E_{p(y^n|x^n)}\Big(\frac{q(y^n|x^n)}{p(y^n|x^n)}\Big)^{1-\lambda}.$$

Here, $d(p, r|x^n)$ is the Rényi divergence with fixed $x^n$ (fixed design). They proved that $\hat{\theta}$ has similar bounds to Theorems 1 and 2 for any risk valid penalty in case of fixed design. Their way is excellent because it requires no additional condition except the risk validity. However, the risk evaluation with fixed design $E_{p_*(y^n|x^n)}[d_\lambda^n(p_*, p_{\hat{\theta}}|x^n)]$ is unsatisfactory for supervised learning to assess the generalization error. We need to evaluate the risk with random design $E_{\bar{p}_*}[d_\lambda^n(p_*, p_{\hat{\theta}})]$. However, it is essentially difficult to apply their idea to random design case. We explain this by using lasso as an example. If we extend the above risk validity to random design straightforwardly, $\widetilde{\Theta}(x^n)$ and $\tilde{L}(\tilde{\theta}|x^n)$ must be independent of $x^n$ due to the condition (2). In addition, (3) is replaced with

$$\forall x^n, y^n \in \mathscr{X}^n \times \mathscr{Y}^n, \max_{\tilde{\theta}\in\widetilde{\Theta}}\Big\{d_\lambda^n(p_*, p_{\tilde{\theta}}) - \log\frac{p_*(y^n|x^n)}{p_{\tilde{\theta}}(y^n|x^n)}$$
$$-\beta\tilde{L}(\tilde{\theta})\Big\} \ge \max_{\theta\in\Theta}\Big\{d_\lambda^n(p_*, p_\theta) - \log\frac{p_*(y^n|x^n)}{p_\theta(y^n|x^n)} - L(\theta|x^n)\Big\}.$$

Note that the above inequality must hold for all $x^n \in \mathscr{X}^n$ in addition to all $y^n \in \mathscr{Y}^n$. Furthermore, $d_\lambda^n(p_*, p_\theta|x^n)$ is replaced with $d_\lambda^n(p_*, p_\theta)$. We can rewrite the above inequality equivalently as

$$\forall x^n \in \mathscr{X}^n, \ \forall y^n \in \mathscr{Y}^n, \ \forall \theta \in \Theta, \quad \min_{\tilde{\theta}\in\widetilde{\Theta}}\Big\{d_\lambda^n(p_*, p_\theta)$$
$$-d_\lambda^n(p_*, p_{\tilde{\theta}}) + \log\frac{p_\theta(y^n|x^n)}{p_{\tilde{\theta}}(y^n|x^n)} + \beta\tilde{L}(\tilde{\theta})\Big\} \le L(\theta|x^n). \quad (4)$$

Let us write the inside part of min of the left side of (4) as $H(\theta, \tilde{\theta}, x^n, y^n)$. To derive a risk valid $L(\theta|x^n)$, we need find an upper bound on $\min_{\tilde{\theta}} H(\theta, \tilde{\theta}, x^n, y^n)$. However, it is difficult to obtain the explicit form of the $\tilde{\theta}$ minimizing $H$. Chatterjee & Barron (2014b) proposed a remedy for fixed design. We can use it in random design case too as follows. Instead of minimization, their idea is to take $\tilde{\theta}$ close to $\theta$. This seems to be meaningful in the following sense. If we quantize $\Theta$ finely, $\ddot{\theta}$ is expected to behave similarly to $\hat{\theta}$. If $\tilde{\theta} \approx \theta$, then $H(\theta, \tilde{\theta}, x^n, y^n) \approx \tilde{L}(\theta)$, which implies that $\tilde{L}(\theta)$ is risk valid and gives a risk bound similar to $\ddot{\theta}$. Note that, however, we cannot make $\tilde{\theta} = \theta$ exactly because $\tilde{\theta} \in \widetilde{\Theta}$. Chatterjee and Barron randomized $\tilde{\theta}$ on $\widetilde{\Theta}(x^n)$ around $\theta$ and took the expectation in terms of $\tilde{\theta}$. This is justified because $\min_{\tilde{\theta}} H \le E_{\tilde{\theta}}[H]$. By tuning the randomization carefully, they succeeded in removing the dependency of $E_{\tilde{\theta}}[H(\theta, \tilde{\theta}, x^n, y^n)]$ on $y^n$. Since this technique can be applied to random design case similarly, we can write $E_{\tilde{\theta}}[H(\theta, \tilde{\theta}, x^n, y^n)]$ as $H'(\theta, x^n)$. By this fact, any risk valid penalties derived in this way should depend on $x^n$. If not, $L(\theta)$ must bound $\max_{x^n} H'(\theta, x^n)$, which makes $L$ much larger. This is unfavorable in view of MDL. In particular, $H'(\theta, x^n)$ includes an unbounded term in terms of $x^n$ in case of lasso, which stems from the likelihood ratio term in (4). Hence, risk valid penalties derived in this way must depend on $x^n$. Though the $\ell_1$ norm used in lasso does not depend on $x^n$, the following weighted $\ell_1$ norm

$$\|\theta\|_{w,1} := \sum_{j=1}^p w_j|\theta_j|, \quad \text{where } w_j := \sqrt{\frac{1}{n}\sum_{i=1}^n x_{ij}^2}$$

plays an important role. The lasso with this penalty is equivalent to the usual lasso with column normalization such that each column of design matrix has the same norm. The column normalization is theoretically and practically important. Hence, we try to find a risk valid penalty of the form $L_1(\theta|x^n) = \mu_1\|\theta\|_{w,1} + \mu_2$. Indeed, there seems to be no other useful penalty dependent on $x^n$ for lasso. However, there are severe difficulties. The main difficulty is caused by (2). Suppose now that $\tilde{\theta}$ is eqeual to $\theta$ almost ideally. This implies that $H(\theta, \tilde{\theta}, x^n, y^n) \approx \tilde{L}(\theta)$. On the other hand, for each fixed $\theta$, $\|\theta\|_{w,1}$ can be arbitrarily small by making $x^n$ small accordingly. Hence, $\mu_1\|\theta\|_{w,1} + \mu_2$ is almost equal to $\mu_2$. This implies that $\mu_2$ must bound $\max_\theta \tilde{L}(\theta)$, which is infinity in general. If $\tilde{L}$ can depend on $x^n$, we could resolve this problem. However, $\tilde{L}$ must be independent of $x^n$. This issue seems not to be limited to lasso. Another major issue is evaluation of the above $H'(\theta, x^n)$ is quite difficult in random design case since $d_\lambda^n(p_*, p_\theta)$ is generally more complicated than $d_\lambda^n(p_*, p_\theta|x^n)$. Hence, their technique seems to be useless in the random design case. We propose a remedy to solve these issues in a lump.

## 4. Main Results

We propose an extension of BC theory to supervised learning and derive new bounds for lasso.

### 4.1. Extension of BC Theory to Supervised Learning

There are several possible approaches to extend BC theory. Despite of our efforts, we can hardly derive a meaningful tight bound for lasso by most of them except the following way. Our key idea is to modify the risk validity by introducing a 'typical set'. Let $\mathscr{S}_x$ be a certain set of stochastic processes $x_1, x_2, \cdots$ and $\mathscr{P}_x^n$ be the set of their marginal distributions of $x_1, x_2, \cdots, x_n$. We assume that we can define a typical set $A_\epsilon^n$ for each $q_* \in \mathscr{P}_x^n$, i.e., $\Pr(x^n \in A_\epsilon^n) \to 1$ as $n \to \infty$. This is possible if $q_*$ is stationary and ergodic for example. For short, $\Pr(x^n \in A_\epsilon^n)$ is written as $P_\epsilon^n$ hereafter. We modify the risk validity as follows.

**Definition 4.** *Let $\beta > 1$ and $\lambda \in (0, 1-\beta^{-1}]$. We say that $L(\theta|x^n)$ is $\epsilon$-risk valid for $(\lambda, \beta, \mathscr{P}_x^n, A_\epsilon^n)$ if, for any $q_* \in \mathscr{P}_x^n$, there exist a quantized subset $\widetilde{\Theta}(q_*) \subset \Theta$ and a model description length $\tilde{L}(\theta|q_*)$ satisfying Kraft's inequality such that $\widetilde{\Theta}(q_*)$ and $\tilde{L}(\theta|q_*)$ satisfy (2) and*

$$\forall x^n, y^n \in A_\epsilon^n \times \mathscr{Y}^n, \max_{\tilde{\theta} \in \widetilde{\Theta}(q_*)} \Big\{ d_\lambda^n(p_*, p_{\tilde{\theta}}) - \log \frac{p_*(y^n|x^n)}{p_{\tilde{\theta}}(y^n|x^n)} - \beta \tilde{L}(\tilde{\theta}|q_*) \Big\} \geq \max_{\theta \in \Theta} \Big\{ d_\lambda^n(p_*, p_\theta) - \log \frac{p_*(y^n|x^n)}{p_\theta(y^n|x^n)} - L(\theta|x^n) \Big\}.$$

A difference from (4) is the restriction of the range of $x^n$ onto the typical set. Therefore, we can possibly avoid the problem described in the previous section. Using the $\epsilon$-risk validity, we can prove the following two main theorems.

**Theorem 5** (risk bound)**.** *Define $E_\epsilon^n$ as a conditional expectation in terms of $\bar{p}_*(x^n, y^n)$ given that $x^n \in A_\epsilon^n$. Let $\beta > 1$, $\epsilon \in (0,1)$ and $\lambda \in (0, 1-\beta^{-1}]$. If $L(\theta|x^n)$ is $\epsilon$-risk valid for $(\lambda, \beta, \mathscr{P}_x^n, A_\epsilon^n)$,*

$$E_\epsilon^n d_\lambda^n(p_*, p_{\hat{\theta}}) \leq E_\epsilon^n \log \frac{p_*(y^n|x^n)}{p_{\beta 2\text{-}p}(y^n|x^n)} + \beta \log \frac{1}{P_\epsilon^n}. \quad (5)$$

**Theorem 6** (regret bound)**.** *Let $\beta > 1$, $\epsilon \in (0,1)$ and $\lambda \in (0, 1-\beta^{-1}]$. If $L(\theta|x^n)$ is $\epsilon$-risk valid for $(\lambda, \beta, \mathscr{P}_x^n, A_\epsilon^n)$,*

$$\Pr\Big( \frac{d_\lambda^n(p_*, p_{\hat{\theta}})}{n} - \frac{1}{n} \log \frac{p_*(y^n|x^n)}{p_{\beta 2\text{-}p}(y^n|x^n)} > \tau \Big) \leq 1 - P_\epsilon^n + \exp(-\tau n/\beta). \quad (6)$$

We describe the proof of Theorem 5 in Appendix A and the proof of Theorem 6 in a supplementary material due to the page restriction. In contrast to the usual BC theory, there is an additional term $\beta \log(1/P_\epsilon^n)$ in the risk bound. Due to the property of the typical set, this term decreases to zero as $n \to \infty$. Hence, the first term is the main term, which has a form of redundancy of a two-stage code like the quantized case. Hence, this theorem gives a justification of the MDL principle in supervised learning. Note that, however, an additional condition on $L$ is required to interpret the first term of (5) as a redundancy exactly. A sufficient condition for it is called 'codelength validity' (Chatterjee & Barron, 2014b). The risk validity does not imply the codelength validity and vice versa in general. Due to the space limitations, we omit more details of the codelength validity.

We note that the conditional expectation in the risk bound (5) is seemingly hard to be replaced with the usual unconditional expectation. The main difficulty arises from the unboundedness of the loss function: the loss function $d_\lambda^n(p_*, p_{\hat{\theta}})$ can be arbitrarily large according to the choice of $x^n$ in general. Our remedy is a typical set. Because $x^n$ lies out of $A_\epsilon^n$ with small probability, the conditional expectation is likely to capture the expectation of almost all cases. In spite of this fact, if one wants to remove the unnatural conditional expectation, Theorem 6 offers a more satisfactory bound. We should remark that the effectiveness of this approach in real situations depends on whether we can show the risk validity of the target penalty and derive a sufficiently small bound for $1 - P_\epsilon^n$. Actually, much effort is required to realize them for lasso.

### 4.2. Risk Bound of Lasso in Random Design

In this section, we derive new risk and regret bounds by Theorems 5 and 6. Assume that training data $\{(x_i, y_i) \in (\Re^p \times \Re) | i = 1, 2, \cdots, n\}$ obey a usual regression model $Y = X\theta^* + \mathscr{E}$, where $Y := (y_1, y_2, \cdots, y_n)^T$, $\mathscr{E}$ is a noise vector subject to $N(\epsilon; 0, \sigma^2 I_n)$, $\theta^*$ is a true parameter and $X = [x_{ij}]$. Here, $x_{ij}$ is the $j$th element of $x_i$ and $N(\cdot; m, \Sigma)$ is a Gaussian distribution with a mean vector $m$ and a covariance matrix $\Sigma$. The dimension $p$ of $\theta$ can be greater than $n$. Under the normality of $x^n$, we can derive a risk valid weighted $\ell_1$ penalty by choosing an appropriate typical set.

**Lemma 1.** *For any $\epsilon \in (0,1)$, define*

$$\mathscr{P}_x^n := \{ q(x^n) = \Pi_{i=1}^n N(x_i; \mathbf{0}, \Sigma) | \textit{ Non-Singular } \Sigma \},$$
$$A_\epsilon^{(n)} := \Big\{ x^n \Big| \forall j, 1 - \epsilon \leq \frac{(1/n)\sum_{i=1}^n x_{ij}^2}{\Sigma_{jj}} \leq 1 + \epsilon \Big\},$$

*where $\Sigma_{jj}$ denotes the $j$th diagonal element of $\Sigma$. Assume a linear regression setting: $p_*(y^n|x^n) = \Pi_{i=1}^n N(y_i|x_i^T\theta^*, \sigma^2)$ and $p_\theta(y^n|x^n) = \Pi_{i=1}^n N(y_i|x_i^T\theta, \sigma^2)$ with $\Theta = \Re^p$. Let $\beta > 1$ and $\lambda \in (0, 1-\beta^{-1}]$. The penalty $L_1(\theta|x^n) = \mu_1\|\theta\|_{w,1} + \mu_2$ is $\epsilon$-risk valid for $(\lambda, \beta, \mathscr{P}_x^n, A_\epsilon^n)$ if*

$$\mu_1 \geq \sqrt{\frac{n\beta \log 4p}{2\sigma^2} \cdot \frac{\lambda + 4\sqrt{1-\epsilon^2}}{(1-\epsilon)}}, \quad \mu_2 \geq \beta \log 2. \quad (7)$$

We describe its proof in Appendix B. The derivation is much more complicated and requires more techniques compared to fixed design case in (Chatterjee & Barron,

2014b). This is because the Rényi divergence is a usual mean square error in the fixed design case, while it is not in the random design case in general. Remarkably, the risk valid penalty in the above theorem also satisfies the code-length validity. This indicates that the main term of the risk bound can always be interpreted as redundancy of a prefix code. Next, we evaluate the convergence rate of $P_\epsilon^n$.

**Lemma 2** (Exponential Bound of Typical Set)**.** *Suppose that $x_i \sim N(0,\Sigma)$ independently. For any $\epsilon \in (0,1)$,*

$$P_\epsilon^n \geq \left(1-2\exp\left(-\frac{n}{2}(\epsilon-\log(1+\epsilon))\right)\right)^p \quad (8)$$
$$\geq 1-2p\exp\left(-\frac{n}{2}(\epsilon-\log(1+\epsilon))\right) \geq 1-2p\exp\left(-\frac{n\epsilon^2}{7}\right).$$

Its proof is described in a supplementary material. For lasso, $n \ll p$ is often assumed. By Lemma 2, $1-P_\epsilon^n$ is bounded above by $O(p\exp(-n\epsilon^2/7))$. Hence, $-\log P_\epsilon^n$ in (5) and $1-P_\epsilon^n$ in (10) can be negligibly small even if $n \ll p$. In this sense, the exponential bound is critical for lasso. From Lemmas 1 and 2, we obtain the following theorem.

**Theorem 7.** *Assume the same setting as Lemma 1. If $L_1(\theta|x^n)=\mu_1\|\theta\|_{w,1}+\mu_2$ satisfies (7), the lasso estimator*

$$\hat{\theta} := \operatorname*{argmin}_{\theta\in\Theta} \frac{1}{2n\sigma^2}\|Y-X\theta\|_2^2 + \mu_1\|\theta\|_{w,1} \quad (9)$$

*has a risk bound*

$$E_\epsilon^n[d_\lambda(p_*,p_{\hat{\theta}})] \leq E_\epsilon^n\left[\inf_{\theta\in\Theta}\left\{\frac{\left(\|y-X\theta\|_2^2-\|y-X\theta^*\|_2^2\right)}{2n\sigma^2} + \mu_1\|\theta\|_{w,1}+\mu_2\right\}\right] - \frac{p\log\left(1-2\exp\left(-\frac{n}{2}(\epsilon-\log(1+\epsilon))\right)\right)}{n\beta}.$$

*and a regret bound*

$$d_\lambda(p_*,p_{\hat{\theta}}) \leq \inf_{\theta\in\Theta}\left\{\frac{\|y-X\theta\|_2^2-\|y-X\theta^*\|_2^2}{2n\sigma^2}+\mu_1\|\theta\|_{w,1}+\mu_2\right\}+\tau \quad (10)$$

*with probability at least*

$$\left(1-2\exp\left(-\frac{n}{2}(\epsilon-\log(1+\epsilon))\right)\right)^p - \exp(-\tau n/\beta). \quad (11)$$

Here, $d_\lambda(p,r)$ denotes $d_\lambda^1(p,r)$. Since $\bar{p}_*(x^n,y^n)$ is i.i.d. in this setting, we presented the risk bound as a single-sample version by dividing the both sides by $n$. Compared to the risk bound in the fixed design case, a coefficient of the weighted $\ell_1$ norm is basically larger. Chatterjee & Barron (2014b) showed that, if $\mu_1 \geq \sqrt{2n\log 4p/\sigma^2}$ and $\mu_2 \geq \beta\log 2$, then the weighted $\ell_1$ norm is risk valid. Ignoring $\epsilon$, the minimum $\mu_1$ in (7) is $(1/2)\sqrt{(\lambda+4)/(1-\lambda)}$ times that for the fixed design case. Hence, the coefficient is always larger than or equal to compared to the fixed design case but its extent is not so large unless $\lambda$ is close to 1.

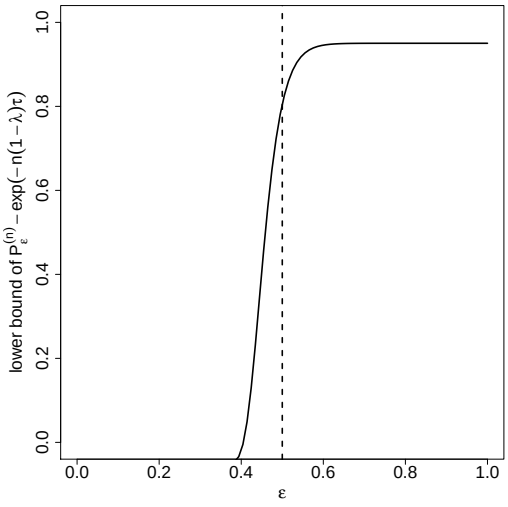


*Figure 1.* Plot of (11) against $\epsilon \in (0,1)$ when $n=200, p=1000$ and $\tau=0.03$. The dotted vertical line indicates $\epsilon=0.5$.

## 5. Numerical Simulations

We investigate the behavior of the regret bound (10). Here, $\mu_1$ and $\mu_2$ are set to their smallest values in (7) and $\lambda = 1-\beta^{-1}$. As described before, the Rényi divergence is no longer a mean square error (MSE) in random design case. The Rényi divergence approaches to KL-divergence when $\lambda \to 1$ which is MSE in this case. If we take $\lambda$ close to 1, however, the risk valid penalty function $L$ (and also the regret bound) tends to diverge unless $n$ is accordingly large enough. That is, we can obtain only the approximate evaluation on the MSE. The precision of that approximation varies according to the sample size $n$. We do not employ the MSE here but another important case $\lambda = 0.5$, that is, Bhattacharyya divergence. Bhattacharyya divergence is an upper bound of two times the squared Hellinger distance

$$d_H^2(p_*,p_\theta) = \int\left(\sqrt{p_*(y|x)}-\sqrt{p_\theta(y|x)}\right)^2 q_*(x)p_*(y|x)dxdy,$$

which is often used to performance evaluation. This can be proved by the fact that $d_\lambda(p_*,p_\theta) \geq \lambda\mathscr{D}^{1-2\lambda}(p_*,p_\theta)$ for any $\theta$ and $\lambda \in (0,1)$, where $\mathscr{D}^\alpha(p,q)$ is $\alpha$-divergence

$$\mathscr{D}_\alpha(p,r) := \frac{4}{1-\alpha^2}\int\left(1-\left(\frac{r(y|x)}{p(y|x)}\right)^{\frac{1+\alpha}{2}}\right)q_*(x)p(y|x)dxdy$$

(Cichocki & Amari, 2010) and $\mathscr{D}^0$ is just four times the squared Hellinger distance. Thus, we can bound $2d_H^2(p_*,p_{\hat{\theta}})$ through Bhattacharyya divergence ($d_{0.5}$). We set $n=200$, $p=1000$ and $\Sigma=I_p$ to mimic a typical situation of sparse learning. The lasso estimator is calculated by a proximal gradient method. To make the regret bound tight, we take $\tau=0.03$ that is close to zero compared to the main term (regret). For this $\tau$, Fig. 1 shows the plot

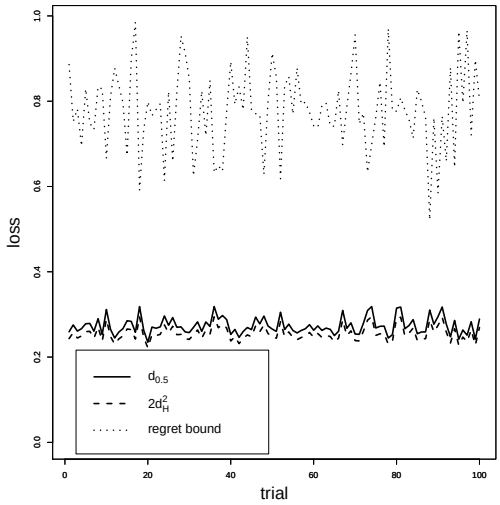


*Figure 2.* Plot of $d_{0.5}$ (Rényi div.), $2d_H^2$ ($\alpha$-div.) and the regret bound with $\tau = 0.03$ in case SN ratio=1.5.

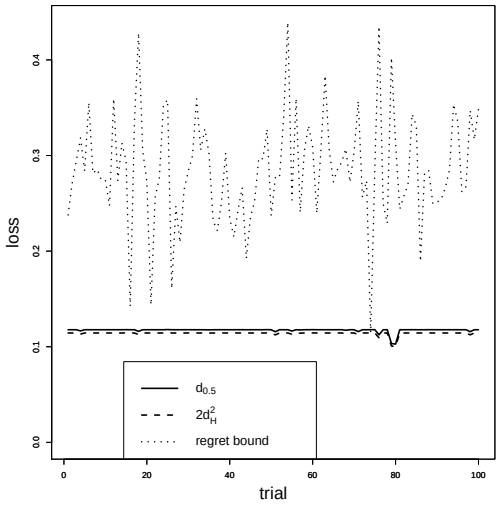


*Figure 3.* Plot of $d_{0.5}$ (Rényi div.), $2d_H^2$ ($\alpha$-div.) and the regret bound with $\tau = 0.03$ in case SN ratio=0.5.

of (11) against $\epsilon$. We should choose the smallest as long as the regret bound holds with large probability. Our choice is $\epsilon = 0.5$ at which the value of (11) is 0.81. We show the results of two cases in Figs. 2 and 3. These plots express the value of $d_{0.5}$, $2d_H^2$ and the regret bound that were obtained in a hundred of repetitions with different SN ratios (SNR) $E_{p_*}[(x^T\theta^*)^2]/\sigma^2$ (that is, different $\sigma^2$). From these figures and other experiments, we observed that $2d_H^2$ almost always equaled $d_{0.5}$ (they are completely overlapped). As the SN ratio got larger, then the regret bound became looser (for example, about six times larger than $2d_H^2$ when SNR is 10). One of the reasons is that the risk validity condition is too strict to bound the loss function when SNR is high. Hence, a possible way to improve the risk bound is to restrict the parameter space $\Theta$ used in $\epsilon$-risk validity to a range of $\hat{\theta}$, which is expected to be considerably narrower than $\Theta$ due to high SNR. In contrast, the regret bound is tight when SNR is 0.5 in Fig. 3. Though the regret bound is probabilistic, the regret bound dominated the Rényi divergence over all trials.

## A. Proof of Theorem 5

*Proof.* Define $F_\lambda^\theta(x^n, y^n) := d_\lambda^n(p_*, p_\theta) - \log \frac{p_*(y^n|x^n)}{p_\theta(y^n|x^n)}$. By the risk validity, we obtain

$$
\begin{aligned}
&E_\epsilon^n\left[\exp\left(\frac{1}{\beta}\max_{\theta\in\Theta}\left\{F_\lambda^\theta(x^n,y^n) - L(\theta|x^n)\right\}\right)\right] \\
\le\ & E_\epsilon^n\left[\exp\left(\frac{1}{\beta}\max_{\tilde{\theta}\in\tilde{\Theta}}\left\{F_\lambda^{\tilde{\theta}}(x^n,y^n) - \beta\tilde{L}(\tilde{\theta}|q_*)\right\}\right)\right] \\
\le & \sum_{\tilde{\theta}\in\tilde{\Theta}(q_*)} E_\epsilon^n\left[\exp\left(\frac{1}{\beta}\left(F_\lambda^{\tilde{\theta}}(x^n,y^n) - \beta\tilde{L}(\tilde{\theta}|q_*)\right)\right)\right]. \quad (12)
\end{aligned}
$$

The following fact is a key technique:

$$
\begin{aligned}
&E_\epsilon^n\left[\exp\left(\frac{1}{\beta}F_\lambda^{\tilde{\theta}}(x^n,y^n)\right)\right] \\
&= \exp\left(\frac{1}{\beta}d_\lambda^n(p_*,p_\theta)\right) E_\epsilon^n\left[\left(\frac{p_{\tilde{\theta}}(y^n|x^n)}{p_*(y^n|x^n)}\right)^{\frac{1}{\beta}}\right] \\
&\le \frac{1}{P_\epsilon^n}\exp\left(\frac{1}{\beta}d_\lambda^n(p_*,p_\theta)\right) E\left[\left(\frac{p_{\tilde{\theta}}(y^n|x^n)}{p_*(y^n|x^n)}\right)^{\frac{1}{\beta}}\right] \\
&= \frac{1}{P_\epsilon^n}\exp\left(\frac{1}{\beta}d_\lambda^n(p_*,p_\theta)\right)\exp\left(-\frac{1}{\beta}d_{1-\beta^{-1}}^n(p_*,p_\theta)\right) \\
&\le \frac{1}{P_\epsilon^n}\exp\left(\frac{1}{\beta}d_\lambda^n(p_*,p_\theta)\right)\exp\left(-\frac{1}{\beta}d_\lambda^n(p_*,p_\theta)\right) = \frac{1}{P_\epsilon^n}.
\end{aligned}
$$

The first inequality holds because $E_{\bar{p}_*(x^n,y^n)}[A] \ge P_\epsilon^n E_\epsilon^n[A]$ for any nonnegative random variable $A$. The second inequality holds because $d_\lambda^n(p_*,p_\theta)$ is monotonically increasing with respect to $\lambda$. Thus, the right side of (12) is bounded by $1/P_\epsilon^n$. By Jensen's inequality,

$$
\begin{aligned}
\frac{1}{P_\epsilon^n} &\ge E_\epsilon^n\left[\exp\left(\frac{1}{\beta}\max_{\theta\in\Theta}\left\{F_\lambda^\theta(x^n,y^n) - L(\theta|x^n)\right\}\right)\right] \quad (13)\\
&\ge \exp\left(E_\epsilon^n\left[\frac{1}{\beta}\max_{\theta\in\Theta}\left\{F_\lambda^\theta(x^n,y^n) - L(\theta|x^n)\right\}\right]\right) \\
&\ge \exp\left(E_\epsilon^n\left[\frac{1}{\beta}\left(F_\lambda^{\hat{\theta}}(x^n,y^n) - L(\hat{\theta}|x^n)\right)\right]\right).
\end{aligned}
$$

Thus, we have

$$
-\beta\log P_\epsilon^n \ge E_\epsilon^n\left[d_\lambda^n(p_*,p_{\hat{\theta}}) - \log\frac{p_*(y^n|x^n)}{p_{\hat{\theta}}(y^n|x^n)} - L(\hat{\theta}|x^n)\right].
$$

Rearranging terms of this inequality, we have the statement. □

## B. Proof of Lemma 1

*Proof.* For convenience, we define $H(\theta, \tilde{\theta}, x^n, y^n)$ as

$$\underbrace{d_\lambda^n(p_*, p_\theta) - d_\lambda^n(p_*, p_{\tilde{\theta}})}_{\text{loss variation part}} + \underbrace{\log \frac{p_\theta(y^n|x^n)}{p_{\tilde{\theta}}(y^n|x^n)} + \beta \tilde{L}(\tilde{\theta}|q_*)}_{\text{codelength validity part}}.$$

We need to find a weighted $\ell_1$ penalty function $L(\theta|x^n)$ that bounds $\min_{\tilde{\theta}\in\widetilde{\Theta}(q_*)} H(\theta, \tilde{\theta}, x^n, y^n)$ from above for any $(\theta, x^n, y^n) \in (\Re^p \times A_\epsilon^n \times \Re^n)$. To bound $\min_{\tilde{\theta}} H(\theta, \tilde{\theta}, x^n, y^n)$, we borrow a nice randomization technique introduced in (Chatterjee & Barron, 2014b) with some modifications. Let us define $w^* := (w_1^*, w_2^*, \cdots, w_p^*)^T$, where $w_j^* = \sqrt{\Sigma_{jj}}$ and $W^* := \mathrm{diag}(w_1^*, \cdots, w_p^*)$. We quantize $\Theta$ as $\widetilde{\Theta}(q_*) := \{\delta(W^*)^{-1}z | z \in \mathscr{Z}^p\}$, where $\delta > 0$ is a quantization width and $\mathscr{Z}$ is a set of all integers. Though $\widetilde{\Theta}$ depends on the data in (Chatterjee & Barron, 2014b), we must remove that dependency to satisfy $\epsilon$-risk validity. A problem is that the minimization of $H(\theta, \tilde{\theta}, x^n, y^n)$ seems to be difficult to evaluate. A key idea here is to bound not $\min_{\tilde{\theta}} H(\theta, \tilde{\theta}, x^n, y^n)$ directly but its expectation $E_{\tilde{\theta}}[H(\theta, \tilde{\theta}, x^n, y^n)]$ with respect to a dexterously randomized $\tilde{\theta}$ because the expectation is larger than the minimum. For each given $\theta$, $\tilde{\theta}$ is randomized as

$$\tilde{\theta}_j = \begin{cases} \frac{\delta}{w_j^*}\lceil m_j \rceil & \text{with prob. } m_j - \lfloor m_j \rfloor \\ \frac{\delta}{w_j^*}\lfloor m_j \rfloor & \text{with prob. } \lceil m_j \rceil - m_j \\ \frac{\delta}{w_j^*} m_j & \text{with prob. } 1 - (\lceil m_j \rceil - \lfloor m_j \rfloor) \end{cases}, \quad (14)$$

where $m_j := w_j^*\theta_j/\delta$ and each component of $\tilde{\theta}$ is statistically independent of each other. Its important properties are $E_{\tilde{\theta}}[\tilde{\theta}] = \theta$ and $E[(\tilde{\theta}_j - \theta_j)(\tilde{\theta}_{j'} - \theta_{j'})] \le I(j = j')\frac{\delta}{w_j^*}|\theta_j|$. Using these, we bound $E_{\tilde{\theta}}[H(\theta, \tilde{\theta}, x^n, y^n)]$ as follows. The loss variation part in $H(\theta, \tilde{\theta}, x^n, y^n)$ is the main concern because it is more complicated than that of fixed design case. Let us consider the following Taylor expansion

$$d_\lambda^n(p_*, p_\theta) - d_\lambda^n(p_*, p_{\tilde{\theta}}) = -\left(\frac{\partial d_\lambda^n(p_*, p_\theta)}{\partial \theta}\right)^T (\tilde{\theta} - \theta) - \frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2 d_\lambda^n(p_*, p_{\theta^\circ})}{\partial\theta\partial\theta^T}(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^T\right), \quad (15)$$

where $\theta^\circ$ is a vector between $\theta$ and $\tilde{\theta}$. The first term in the right side of (15) vanishes after taking expectation w.r.t. $\tilde{\theta}$ because $E_{\tilde{\theta}}[\tilde{\theta} - \theta] = 0$. To bound the second term by the weighted $\ell_1$ norm of $\theta$, we have to bound this term above by a multiple of $\mathrm{Tr}(\Sigma(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^T)$. Nevertheless, it is not an easy task because the dependency of the Hessian of $d_\lambda^n$ on $\tilde{\theta}$ is complicated. Here, Lemma 3 in Appendix C plays a key role. By this lemma and Cauchy-Schwartz inequality, we obtain

$$\begin{aligned}
&\mathrm{Tr}\left(-\frac{\partial^2 d_\lambda^n(p_*, p_{\theta^\circ})}{\partial\theta\partial\theta^T}(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^T\right) \\
&\le \frac{n\lambda}{4\sigma^2}\mathrm{Tr}\left(\left(\frac{\Sigma^{1/2}\bar{\theta}'(\bar{\theta}')^T\Sigma^{1/2}}{\|\bar{\theta}'\|_2^2}\right)\left(\tilde{\theta} - \theta\right)\left(\tilde{\theta} - \theta\right)^T\right) \\
&= \frac{n\lambda}{4\sigma^2}\frac{\left((\bar{\theta}')^T\Sigma^{1/2}(\tilde{\theta} - \theta)\right)^2}{\|\bar{\theta}'\|_2^2} \le \frac{n\lambda}{4\sigma^2}\frac{\|\bar{\theta}'\|_2^2\|\Sigma^{1/2}(\tilde{\theta} - \theta)\|_2^2}{\|\bar{\theta}'\|_2^2} \\
&= \frac{n\lambda}{4\sigma^2}\|\Sigma^{1/2}(\tilde{\theta} - \theta)\|_2^2 = \frac{n\lambda}{4\sigma^2}\mathrm{Tr}\left(\Sigma(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^T\right).
\end{aligned}$$

See Lemma 3 for unknown symbols. Thus, the expectation of the loss variation part with respect to $\tilde{\theta}$ is bounded as

$$E_{\tilde{\theta}}\left[d_\lambda^n(p_*, p_\theta) - d_\lambda^n(p_*, p_{\tilde{\theta}})\right] \le \frac{\delta n\lambda}{8\sigma^2}\|\theta\|_{w^*,1}. \quad (16)$$

The codelength validity part in $H(\theta, \tilde{\theta}, x^n, y^n)$ have the same form as that for the fixed design case in its appearance. However, we need to evaluate it again in our setting because both $\widetilde{\Theta}$ and $\tilde{L}$ are changed. The likelihood ratio term in $H(\theta, \tilde{\theta}, x^n, y^n)$ is calculated as

$$\frac{1}{2\sigma^2}\Big(2(y - X\theta)^T X(\theta - \tilde{\theta}) + \mathrm{Tr}\big(X^T X(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^T\big)\Big).$$

Taking expectation with respect to $\tilde{\theta}$, we have

$$\begin{aligned}
E_{\tilde{\theta}}\left[\log\frac{p_\theta(y^n|x^n)}{p_{\tilde{\theta}}(y^n|x^n)}\right] &= \frac{n}{2\sigma^2}E_{\tilde{\theta}}\Big[\mathrm{Tr}\Big(W^2(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^T\Big)\Big] \\
&\le \frac{\delta n}{2\sigma^2}\sum_{j=1}^p \frac{w_j^2}{w_j^*}|\theta_j|,
\end{aligned}$$

where $W := \mathrm{diag}(w_1, w_2, \cdots, w_p)$. We define a codelength function $C(z) := \|z\|_1 \log 4p + \log 2$ for any $z \in \mathscr{Z}^p$. Note that $C(z)$ satisfies Kraft's inequality. Let us define a penalty function on $\widetilde{\Theta}(q_*)$ as

$$\tilde{L}(\tilde{\theta}|q_*) := C\left(\frac{1}{\delta}W^*\tilde{\theta}\right) = (1/\delta)\|W^*\tilde{\theta}\|_1 \log 4p + \log 2.$$

Note that $\tilde{L}$ satisfies Kraft's inequality and does not depend on $x^n$. By taking expectation w.r.t. $\tilde{\theta}$, we have

$$E_{\tilde{\theta}}\left[\tilde{L}(\tilde{\theta}|q_*)\right] = \frac{\log 4p}{\delta}\|\theta\|_{w^*,1} + \log 2.$$

Thus, the codelength validity part is bounded above by

$$\frac{\delta n}{2\sigma^2}\sum_{j=1}^p \frac{w_j^2}{w_j^*}|\theta_j| + \frac{\beta \log 4p}{\delta}\|\theta\|_{w^*,1} + \beta \log 2$$

Combining with (16), $E_{\tilde{\theta}}[H(\theta, \tilde{\theta}, x^n, y^n)]$ is bounded above by

$$\frac{\delta n\lambda}{8\sigma^2}\|\theta\|_{w^*,1} + \frac{\delta n}{2\sigma^2}\sum_{j=1}^p \frac{w_j^2}{w_j^*}|\theta_j| + \frac{\beta \log 4p}{\delta}\|\theta\|_{w^*,1} + \beta \log 2.$$

Since $x^n \in A_\epsilon^n$, we can bound this by the data-dependent weighted $\ell_1$ norm $\|\theta\|_{w,1}$ as

$$E_{\tilde{\theta}}[H(\theta,\tilde{\theta},x^n,y^n)] \le \frac{\delta n\lambda}{8\sigma^2}\frac{\|\theta\|_{w,1}}{\sqrt{1-\epsilon}}$$
$$+\frac{\delta n\sqrt{1+\epsilon}}{2\sigma^2}\sum_{j=1}^p \frac{w_j^2}{w_j}|\theta_j| + \frac{\beta\log 4p}{\delta}\frac{\|\theta\|_{w,1}}{\sqrt{1-\epsilon}} + \beta\log 2$$
$$=\left(\frac{\delta n}{2\sigma^2}\left(\frac{\lambda}{4\sqrt{1-\epsilon}}+\sqrt{1+\epsilon}\right)+\frac{\beta\log 4p}{\delta\sqrt{1-\epsilon}}\right)\|\theta\|_{w,1}+\beta\log 2.$$

Because this holds for any $\delta > 0$, we can minimize the upper bound with respect to $\delta$, which completes the proof. □

## C. Upper Bound of Negative Hessian

**Lemma 3.** *Let $\bar{\theta} := \theta - \theta^*$ and $\bar{\theta}' := \Sigma^{1/2}\bar{\theta}$. For any $\theta, \theta^*$,*

$$-\frac{\partial^2 d_\lambda(p_*,p_\theta)}{\partial\theta\partial\theta^T} \preceq \frac{\lambda}{4\sigma^2}\left(\frac{\Sigma^{1/2}\bar{\theta}'(\bar{\theta}')^T\Sigma^{1/2}}{\|\bar{\theta}'\|_2^2}\right), \qquad (17)$$

*where $A \preceq B$ implies that $B - A$ is positive semi-definite.*

*Proof.* The Rényi divergence and its derivatives are well interpreted through a distribution

$$\bar{p}_\theta^\lambda(x,y) := q_*(x)p_*(y|x)^\lambda p_\theta(y|x)^{1-\lambda}/Z_\theta^\lambda,$$

where $Z_\theta^\lambda$ is a normalization constant. Here, we show only an explicit form of $q_\theta^\lambda(x) = \int \bar{p}_\theta^\lambda(x,y)dy$ and the Hessian of $d_\lambda(p_*,p_\theta)$ without proof due to the page limit:

$$q_\theta^\lambda(x) = N(x;\mathbf{0},\Sigma_\theta^\lambda),$$
$$\Sigma_\theta^\lambda := \Sigma^{1/2}\left(I_p - \gamma(\|\bar{\theta}'\|_2^2)\left(\frac{\bar{\theta}'}{\|\bar{\theta}'\|_2}\right)\left(\frac{\bar{\theta}'}{\|\bar{\theta}'\|_2}\right)^T\right)\Sigma^{1/2},$$
$$\frac{\partial^2 d_\lambda(p_*,p_\theta)}{\partial\theta\partial\theta^T} = \frac{\lambda}{\sigma^2}\Sigma_\theta^\lambda - \frac{\lambda^2(1-\lambda)}{\sigma^4}\mathrm{Var}_{q_\theta^\lambda(x)}\left(xx^T\bar{\theta}\right),$$
where $c := \dfrac{\sigma^2}{\lambda(1-\lambda)}$, $\gamma(t) := \dfrac{t}{c+t}$.

Here, $I_p$ is an identity matrix of dimension $p$ and $\mathrm{Var}_q(A) := E_q[(A - E_q[A])(A - E_q[A])^T]$. Therefore, we need to evaluate

$$\mathrm{Var}_{q_\theta^\lambda}(xx^T\bar{\theta}) = E_{q_\theta^\lambda}\left[(xx^T\bar{\theta})(xx^T\bar{\theta})^T\right] - (\Sigma_\theta^\lambda\bar{\theta})(\Sigma_\theta^\lambda\bar{\theta})^T.$$

The $(j_1,j_2)$ element of $E_{q_\theta^\lambda}\left[xx^T\bar{\theta}\bar{\theta}^Txx^T\right]$ is calculated as

$$E_{q_\theta^\lambda}\left[\left(xx^T\bar{\theta}\bar{\theta}^Txx^T\right)_{j_1j_2}\right] = \sum_{j_3,j_4=1}^p \bar{\theta}_{j_3}\bar{\theta}_{j_4}E_{q_\theta^\lambda}\left[x_{j_1}x_{j_2}x_{j_3}x_{j_4}\right],$$

where $x_j$ denotes the $j$th element of $x$ only here. We rewrite $\Sigma_\theta^\lambda$ as $S$ to reduce notation complexity hereafter. By the formula of moments of Gaussian distribution,

$$E_{q_\theta^\lambda}\left[x_{j_1}x_{j_2}x_{j_3}x_{j_4}\right] = S_{j_1j_2}S_{j_3j_4}+S_{j_1j_3}S_{j_2j_4}+S_{j_2j_3}S_{j_1j_4}.$$

Therefore, the above quantity is calculated as

$$E_{q_\theta^\lambda}\left[\left(xx^T\bar{\theta}\bar{\theta}^Txx^T\right)_{j_1j_2}\right]$$
$$= \sum_{j_3,j_4=1}^p \bar{\theta}_{j_3}\bar{\theta}_{j_4}(S_{j_1j_2}S_{j_3j_4}+S_{j_1j_3}S_{j_2j_4}+S_{j_2j_3}S_{j_1j_4})$$
$$= \bar{\theta}^TS\bar{\theta}S_{j_1j_2} + 2(S\bar{\theta})_{j_1}(S\bar{\theta})_{j_2}.$$

Summarizing these as a matrix form, we have

$$E_{q_\theta^\lambda}\left[xx^T\bar{\theta}\bar{\theta}^Txx^T\right] = (\bar{\theta}^TS\bar{\theta})S + 2S\bar{\theta}(S\bar{\theta})^T.$$

As a result, $\mathrm{Var}_{q_\theta^\lambda}(xx^T\bar{\theta})$ is obtained as

$$\mathrm{Var}_{q_\theta^\lambda}(xx^T\bar{\theta}) = S\bar{\theta}\bar{\theta}^TS + (\bar{\theta}^TS\bar{\theta})S.$$

We need to survey how this matrix is bounded above in the sense of positive semi-definite. By noticing that $S\bar{\theta} = (1-\gamma(\|\bar{\theta}'\|_2^2))\Sigma^{1/2}\bar{\theta}'$, the first term is calculated as

$$S\bar{\theta}\bar{\theta}^TS = (1-\gamma(\|\bar{\theta}'\|_2^2))^2\|\bar{\theta}'\|_2^2\frac{\Sigma^{1/2}\bar{\theta}'(\bar{\theta}')^T\Sigma^{1/2}}{\|\bar{\theta}'\|_2^2}.$$

Note that $(1-\gamma(t))^2t = c^2t/(c+t)^2 = c^2/(2(c/\sqrt{t}+\sqrt{t})/2)^2 \le c^2/(2\sqrt{c})^2 = c/4$ holds, since $(c/\sqrt{t}+\sqrt{t})/2 \ge \sqrt{c}$. Thus, we have

$$S\bar{\theta}\bar{\theta}^TS \preceq \frac{c}{4}\frac{\Sigma^{1/2}\bar{\theta}'(\bar{\theta}')^T\Sigma^{1/2}}{\|\bar{\theta}'\|_2^2}.$$

As for the second term, we first calculate

$$\bar{\theta}^TS\bar{\theta} = \bar{\theta}^T(1-\gamma(\|\bar{\theta}'\|_2^2))\Sigma^{1/2}\bar{\theta}' = (1-\gamma(\|\bar{\theta}'\|_2^2))\|\bar{\theta}'\|_2^2.$$

Note that $(1-\gamma(t))t = ct/(c+t) = c/(c/t+1) \le c$ holds and that $S$ is positive semi-definite for any $\theta$, the second term is bounded as

$$(\bar{\theta}^TS\bar{\theta})S = f_2(\|\bar{\theta}'\|_2^2)S \preceq cS.$$

Summarizing these, we have

$$\mathrm{Var}_{q_\theta^\lambda}(xx^T\bar{\theta}) \preceq \frac{c}{4}\frac{\Sigma^{1/2}\bar{\theta}'(\bar{\theta}')^T\Sigma^{1/2}}{\|\bar{\theta}'\|_2^2} + cS.$$

Hence, the negative Hessian of $d_\lambda(p_*,p_\theta)$ is bounded as

$$-\frac{\partial^2 d_\lambda(p_*,p_\theta)}{\partial\theta\partial\theta^T} = -\frac{\lambda}{\sigma^2}S + \frac{\lambda}{c\sigma^2}\left(S\bar{\theta}\bar{\theta}^TS + (\bar{\theta}^TS\bar{\theta})S\right)$$
$$\preceq -\frac{\lambda}{\sigma^2}S + \frac{\lambda}{c\sigma^2}\left(\frac{c}{4}\frac{\Sigma^{1/2}\bar{\theta}'(\bar{\theta}')^T\Sigma^{1/2}}{\|\bar{\theta}'\|_2^2} + cS\right)$$
$$= \frac{\lambda}{4\sigma^2}\left(\frac{\Sigma^{1/2}\bar{\theta}'(\bar{\theta}')^T\Sigma^{1/2}}{\|\bar{\theta}'\|_2^2}\right).$$

□

## Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Numbers (25870503) and the Okawa Foundation for Information and Telecommunications. We also thank Mr. Yushin Toyokihara for his support.

## References

Barron, A. R. and Cover, T. M. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37(4):1034–1054, 1991.

Barron, A. R. and Luo, X. MDL procedures with $\ell_1$ penalty and their statistical risk. In *Proceedings of the First Workshop on Information Theoretic Methods in Science and Engineering*, Tampere, Finland, August 18-20 2008.

Barron, A. R., Rissanen, J., and Yu, B. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998.

Barron, A. R., Huang, C., Li, J. Q., and Luo, X. MDL, penalized likelihood and statistical risk. In *Proceedings of IEEE Information Theory Workshop*, Porto, Portugal, May 4-9 2008.

Chatterjee, S. and Barron, A. R. Information theoretic validity of penalized likelihood. *2014 IEEE International Symposium on Information Theory*, pp. 3027–3031, 2014a.

Chatterjee, S. and Barron, A. R. Information theory of penalized likelihoods and its statistical implications. *arXiv'1401.6714v2 [math.ST] 27 Apr.*, 2014b.

Cichocki, A. and Amari, S. Families of alpha- beta- and gamma- divergences: flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

Grünwald, P. D. *The Minimum Description Length Principle*. MIT Press, 2007.

Rényi, A. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:547–561, 1961.

Rissanen, J. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

Takeuchi, J. An introduction to the minimum description length principle. In *A Mathematical Approach to Research Problems of Science and Technology, Springer (book chapter)*, pp. 279–296. Springer, 2014.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58(1):267–288, 1996.

Yamanishi, K. A learning criterion for stochastic rules. *Machine Learning*, 9(2-3):165–203, 1992.