Minimum Description Length Principle in Supervised Learning With Application to Lasso

Masanori Kawakita and Jun'ichi Takeuchi, *Member, IEEE*

***Abstract*— The minimum description length (MDL) principle is extended to supervised learning. The MDL principle is a philosophy that the shortest description of given data leads to the best hypothesis about the data source. One of the key theories for the MDL principle is Barron and Cover's theory (BC theory), which mathematically justifies the MDL principle based on two-stage codes in density estimation (unsupervised learning). Though the codelength of two-stage codes looks similar to the target function of penalized likelihood methods, parameter optimization of penalized likelihood methods is done without quantization of parameter space. Recently, Chatterjee and Barron have provided theoretical tools to extend BC theory to penalized likelihood methods by overcoming this difference. Indeed, applying their tools, they showed that the famous penalized likelihood method 'lasso' can be interpreted as an MDL estimator and enjoys performance guarantee by BC theory. An important fact is that their results assume a fixed design setting, which is essentially the same as unsupervised learning. The fixed design is natural if we use lasso for compressed sensing. If we use lasso for supervised learning, however, the fixed design is considerably unsatisfactory. Only random design is acceptable. However, it is inherently difficult to extend BC theory to the random design regardless of whether the parameter space is quantized or not. In this paper, a novel theoretical tool for extending BC theory to supervised learning (the random design setting and no quantization of parameter space) is provided. Applying this tool, when the covariates are subject to a Gaussian distribution, it is proved that lasso in the random design setting can also be interpreted as an MDL estimator, and that lasso enjoys the risk bound of BC theory. The risk/regret bounds obtained have several advantages inherited from BC theory. First, the bounds require remarkably few assumptions. Second, the bounds hold for any finite sample size n and any finite feature number p even if $n \ll p$. Behavior of the regret bound is investigated by numerical simulations. We believe that this is the first extensions of BC theory to supervised learning (random design).**

***Index Terms*— Lasso, risk bound, regret bound, random design, MDL principle, supervised learning, penalized likelihood.**

Manuscript received May 16, 2016; revised April 19, 2020; accepted May 4, 2020. Date of publication May 29, 2020; date of current version June 18, 2020. This work was supported in part by JSPS KAKENHI Grant Numbers 25870503 and 18H03291 and in part by the Okawa Foundation for Information and Telecommunications. This article was presented in part at the 33rd International Conference on Machine Learning. *(Corresponding author: Masanori Kawakita.)*

Masanori Kawakita was with the Graduate School of Informatics, Nagoya University, Nagoya 464-8601, Japan. He is now with Mie Toyopet Corporation, Tsu city 514-0821, Japan (e-mail: m.kawakita@mietoyopet.co.jp).

Jun'ichi Takeuchi is with the Faculty of Information Science and Electrical Engineering, Kyushu University, Fukuoka 819-0395, Japan (e-mail: tak@inf.kyushu-u.ac.jp).

Communicated by O. Johnson, Associate Editor for Probability and Statistics.

Digital Object Identifier 10.1109/TIT.2020.2998577

I. Introduction

THE minimum description length (MDL) principle is a philosophy that the shortest description of given data leads to the best hypothesis about the data source. Barron and Cover's theory (BC theory) is a seminal work on the MDL principle, which firstly gives a quantitative representation of the MDL principle based on two-stage codes in density estimation (unsupervised learning). More concretely, it states an inequality

$$\text{statistical risk of an estimator induced by a two-stage code} \leq \text{redundancy of the code.} \quad (1)$$

Statistical risk measures the discrepancy between the true distribution generating data and the probability distribution used to encode the data in the two-stage code. This inequality guarantees that finding a code that has small redundancy (description length) leads to a small risk bound. This mathematically justifies the MDL principle. A common goal recognized in the MDL community is to generalize BC theory to wider estimation problems or estimators. Some major progress has been made for penalized likelihood methods. Penalized likelihood methods are now one of the important estimation methods in many fields including statistics, machine learning and information theory. They include many types of estimators, e.g. kernel methods, sparse learning, compressed sensing, graph estimation, learning of neural network and so on. There is a similarity between two-stage codes and penalized likelihood methods as follows. When a parametric model $\{p_\theta(x^n)|\theta \in \Theta\}$ is employed, a two-stage code encodes data x^n as follows. First of all, we prepare a countable subset $\widetilde{\Theta}$ by quantizing the parameter space Θ (i.e., $\widetilde{\Theta} \subset \Theta$). The two-stage code encodes the parameter $\tilde{\theta} \in \widetilde{\Theta}$ in order to specify probability distribution $p_{\tilde{\theta}}(x^n)$ that is used to encode the data x^n. Then the data x^n is encoded using the probability distribution $p_{\tilde{\theta}}(x^n)$. The resulting codelength is the sum of data description length $-\log p_{\tilde{\theta}}(x^n)$ and parameter description (or often called 'model description') length $\tilde{L}(\tilde{\theta})$. In order to minimize the codelength, the parameter $\tilde{\theta}$ and the quantization $\widetilde{\Theta}$ should be chosen such that the codelength is as short as possible. As a result, for a given quantization $\widetilde{\Theta}$, the codelength of the two-stage code can be written as

$$\text{codelength of two-stage code} = \min_{\tilde{\theta} \in \widetilde{\Theta}} \left\{ -\log p_{\tilde{\theta}}(x^n) + \tilde{L}(\tilde{\theta}) \right\}.$$

This codelength looks similar to the minimized target function (penalized likelihood) of penalized likelihood methods

$$\text{negative penalized likelihood} = \min_{\theta\in\Theta}\{-\log p_\theta(x^n) + \text{Pen}(\theta)\},$$

where $\text{Pen}(\theta)$ is a penalty function. A major difference between two-stage codes and penalized likelihood methods is in the quantization of the parameter space for two-stage codes. Several references [3], [4], [17] proposed theoretical tools to extend BC theory to penalized likelihood methods by overcoming the difference. In general, we need to have the following two conditions satisfied in order to extend BC theory to a certain estimator (or estimation problem) other than density estimation.

Definition 1 (BC-Proper MDL Estimator): We say that an estimator is a 'BC-proper MDL estimator' if the following two conditions are satisfied.

Condition 1: The estimator can be exactly interpreted as an MDL estimator. In other words, the target function minimized by the estimator can be interpreted as codelength of a prefix code.

Condition 2: The estimator has a risk bound such that its statistical risk is bounded by redundancy-type quantity like (1) through a kind of BC theory.

Even if a risk bound of the form such as (1) is obtained by applying BC theory, the redundancy-type quantity cannot be exactly interpreted as the redundancy unless Condition 1 is satisfied. However Conditions 1 and 2 are independent in general. Hence we can say that the justification of the MDL principle by BC theory is successfully extended to the estimator if both conditions are satisfied. We refer to such an estimator as 'BC-proper MDL estimator' since BC theory properly holds for the estimator. Chatterjee and Barron provided two convenient sufficient conditions for the above conditions, respectively, in case of penalized likelihood, which are named 'codelength validity' for Condition 1 and 'risk validity' for Condition 2. By these tools, they succeeded in showing that lasso [34] is a BC-proper MDL estimator under certain conditions. Note that codelength validity does not necessarily imply risk validity and vice versa. Lasso is an important estimation method in many areas including variable selection [34], compressed sensing [15], graph estimation [22], learning of neural networks [27] and so on. An important fact is that their results postulate the fixed design setting. In the fixed design setting, the purpose is not estimation of the conditional distribution $p(y^n|x^n)$ but a single density of y^n for a specific x^n. That is, the estimation problem in the fixed design setting is in the framework of unsupervised learning (we use the term 'unsupervised learning' as the problem to estimate an unconditional probability distribution although it is sometimes used for nonprobabilistic learning problems). The fixed design is natural if we use lasso for compressed sensing. If we use lasso for supervised learning, however, the fixed design is considerably unsatisfactory. Only random design is acceptable. However, it is essentially difficult to extend their result to the random design setting. The difficulty arises from a certain property of BC theory itself, which will be explained in Section III-B. In the end, extension of BC theory to supervised setting is difficult regardless of whether the parameter space is quantized or not. To our knowledge, Yamanishi [35] is the only work to apply BC theory to supervised setting in a certain limited situation where the drawback stemming from the above property is not so severe. This history is summarized in Table I. There may be no literature to extend the MDL world (the set of BC-proper MDL estimators) to penalized likelihood in supervised setting. In the remainder part of this paper, we use the word 'supervised learning' as penalized likelihood in the random design setting.

TABLE I
HISTORY OF THE SPREAD OF THE MDL PRINCIPLE. PLM DENOTES PENALIZED LIKELIHOOD METHODS

	unsupervised setting	supervised setting
two-stage code	Barron and Cover (1991)	Yamanishi (1992)
PLM	Chatterjee and Barron (2014)	our target

Our main target is to extend the MDL world to supervised learning. We provide extension of codelength validity and the risk validity to supervised learning. Our extension may give a tight risk bound such as (1) by overcoming the above difficulty ingeniously. Roughly speaking, the obtained risk bound is of the form

$$\text{statistical risk} \le \text{redundancy} + \text{negligibly small term},$$

which also approximately guarantees Condition 2. When the covariates are subject to a Gaussian distribution, it will be proved by this extension that lasso is a BC-proper MDL estimator even in the random design setting. We believe that our work is the first work that extends BC theory to penalized likelihood in the random design setting.

A. Risk Bound for Lasso

In the process of proving Condition 2, a risk bound such as (1) is obtained. We briefly compare this bound with the bounds derived in past studies. There have been many studies about theoretical properties of lasso. Most of such studies have evaluated either the following three aspects of lasso.

1) Prediction accuracy of the estimated regression function $\hat{f}(x)$ for the future sample x, y.
2) Estimation accuracy of the parameter itself.
3) Consistency of feature selection. This is also called 'support estimation'.

Since lasso is known also as a feature selection method, there are many studies about consistency of the feature selection. However, our interest in this paper is in prediction accuracy (statistical risk). Thus, we compare our analysis with several past studies that treated prediction accuracy of lasso, which include [6], [12], [13]. Since accuracy of parameter estimation implies prediction accuracy to some extent, we add a famous reference [38] which treated accuracy of parameter estimation and feature selection. A comparison between our setting and such studies is summarized in Table II. In lasso, the ordinary ℓ_1 norm $\|\theta\|_1 = \sum_{j=1}^m |\theta_j|$ is used as a penalty function $\text{Pen}(\theta)$. In contrast, several references including this paper employ

TABLE II

COMPARISON OF PAST STUDIES ABOUT LASSO WITH OUR ANALYSIS. THE TERM 'DESIGN' INDICATES THE SETTING OF LASS IS EITHER 'FIXED DESIGN' OR 'RANDOM DESIGN'. THE TERM 'BOUNDEDNESS' INDICATES SOME KINDS OF BOUNDEDNESS ON COVARIATES. IN CASE OF THE FIXED DESIGN, THIS TERM INDICATES CONDITIONS THAT ARE IMPOSED ON THE COVARIATES DATA. THE TERM 'MODEL' INDICATES THAT THE ASSUMPTION ABOUT THE REGRESSION FUNCTION $E[Y|X]$

literature	design	boundedness	penalty	model	noise	n, m
Bunea et al. (2007)	random	necessary	weighted ℓ_1 norm	nonparametric	conditioned	finite
Zhang (2009)	fixed	RIP	ℓ_1 norm	nonparametric	sub-Gaussian	finite
Bickel et al. (2009)	fixed	RE	weighted ℓ_1 norm	linear	Gaussian	finite
Bartlett et al. (2012)	random	necessary	ℓ_1 norm	nonparametric	bounded	$n, m \to \infty$
our paper	random	unnecessary	weighted ℓ_1 norm	linear	Gaussian	finite

the weighted ℓ_1 penalty with certain specific weights. We can prove that the lasso with this weighted norm is exactly equivalent to the ordinary lasso after column normalization. The term 'column normalization' means normalizing the data matrix so that each column (covariate) has unit variance. See Section III-B for the details. The column normalization is both practically and theoretically important. It is usual that the given data matrix is normalized as having zero mean and unit variance (column normalization) when we apply lasso to practical data. If not, the computation algorithm may be unstable so that the accuracy of solution may be also unstable. To our knowledge, some kind of column normalization assumption are necessary also for theoretical analysis. Past studies imposed either column normalization or boundedness on covariates in the random design setting. Even in the fixed design setting, the derived bounds is usually somehow sensitive to the sample variance of covariates. More concretely, tightness of risk bound depends severely on that upper bound of covariates. In addition, it is hard to know the true upper bound of covariates in practice. Taking these into account, column normalization (or equivalently the weighted ℓ_1 penalty) is more favorable. Actually, our analysis gives a tighter risk bound to lasso with the weighted ℓ_1 penalty than lasso with the ordinary ℓ_1 penalty. Furthermore, we can't prove Condition 1 for lasso with the ordinary ℓ_1 penalty (though we can't interpret it as an MDL estimator) in contrast to the weighted ℓ_1 penalty. As a result, column normalization is more favorable in view of MDL.

Furthermore, a remarkable property of our risk bound is that the number of required assumptions are quite fewer and simpler compared to other studies. Many of the past studies impose various complicated or technical conditions other than the boundedness of covariates. Some of them are quite hard to check. In contrast, our assumptions are simple. Another remarkable fact is that most past results include several constants that cannot be easily determined. As seen in Table II, Bunea et al. derived an upper bound of loss function for finite n, m. However their result essentially shows just the asymptotic order of the loss function because the bound includes an undetermined scaling factor. The bounds derived by Zhang and Bickel et al. also include constants which cannot be determined easily because of computational difficulty. Although the bound derived by Bartlett et al. has no such constants, their result is not guaranteed to hold for finite n, m. Our bound has also no such constants and holds for finite n, m. This is the reason why we cannot compare our risk bound fairly with other past risk bounds in numerical simulations.

The paper [6] gives a risk bound to lasso in random design with the ordinary ℓ_1 penalty that looks similar to our risk bound. While most of the references before Bartlett *et al.* [6] gives a risk bound that involves the sparsity directly, the risk upper bounds derived in this paper and by Bartlett et al. has the form of the target function of lasso. Bartlett et al. assumes the boundedness of covariates and asymptotic situation ($n, m \to \infty$ such that $\log m = o(n)$) while the regression model is free and the required condition of noise is only sub-Gaussianity. In contrast, we impose the Gaussianity on both covariates and noise while the boundedness of covariates is not necessary and the risk bound holds for any finite n, m. Another major difference is in the loss function. Bartlett et al. used the mean squared error as a loss function, while we use Rényi divergence.

B. Organization of the Paper

This paper is organized as follows. Section II introduces a definition of MDL estimators. We briefly review the original BC theory for density estimation in Section III-A. Section III-B summarizes the extension of BC theory to penalized likelihood estimators. We also explain the reason why the original BC theory and also its extension are difficult to be applied to supervised setting (random design) regardless whether the parameter space is quantized or not. Some devised technical tools will be provided to extend BC theory to the supervised learning in IV-A. Using them, lasso in the random design setting will be proved to be a BC-proper MDL estimator in Section IV-B. Section V contains numerical simulations. All proofs of our results are given in Section VI. A conclusion will appear in Section VII. In this paper, we use many complicated mathematical symbols. For reader's convenience, a glossary will be given in Appendix.

II. MDL ESTIMATORS IN SUPERVISED LEARNING

First we review the definition of MDL estimators for density estimation. Suppose that we have n training data $y^n := \{y_i \in \mathscr{Y} | i = 1, 2, \cdots, n\}$ generated from $p_*(y^n) = \prod_{t=1}^n p_*(y_t)$, where p_* is a probability density function over $\mathscr{Y}$ with a certain reference measure. Here we assume the data source is independently and identically distributed (i.i.d.) for simplicity, but it is not important. We assume that $\mathscr{Y}$ could be continuous or countable. Let us consider the estimation of p_* using a hypothesis set $\{p_\theta | \theta \in \Theta\}$, where p_θ is a density over $\mathscr{Y}$ and the parameter set $\Theta \subset \mathbb{R}^m$ is assumed to be convex.

In general, an MDL estimator for θ based on a two-stage code is defined as

$$\ddot{\theta} = \ddot{\theta}(y^n) = \arg\min_{\tilde{\theta}\in\widetilde{\Theta}} \Big(-\log p_{\tilde{\theta}}(y^n) + \tilde{L}(\tilde{\theta}) \Big). \tag{2}$$

Here $\widetilde{\Theta}$ is a countable set, which is obtained by quantizing Θ appropriately and $\tilde{L}(\tilde{\theta})$ is a codelength function over $\widetilde{\Theta}$ which satisfies the Kraft's inequality:

$$\sum_{\tilde{\theta}\in\widetilde{\Theta}} \exp(-\tilde{L}(\tilde{\theta})) \le 1.$$

The first term of the right side of (2) is called as the data description length and the second the parameter (model) description length. Note that we employ 'nat' for the unit for the codelength, then we have e in the Kraft's inequality. In quantization of typical applications, each coordinate of θ is discretized as its precision is $O(1/\sqrt{n})$, which yields $\tilde{L}(\ddot{\theta}) = (m/2)\log n + O(1)$ and the total codelength of the two-part code is

$$-\log p_{\ddot{\theta}}(y^n) + \frac{m}{2}\log n + O(1).$$

This is known as the MDL criterion for the statistical model selection, in which the second term works as a regularization term among many parametric models.

Let us write the minimum description length attained by the two-stage code as

$$\tilde{L}_2(y^n) := -\log p_{\ddot{\theta}}(y^n) + \tilde{L}(\ddot{\theta}). \tag{3}$$

It is important that $\tilde{L}_2$ satisfies Kraft's inequality with respect to y^n as shown below, where we assume $\mathscr{Y}$ is countable.

$$\begin{aligned}
&\sum_{y^n} \exp\Big(-\tilde{L}_2(y^n)\Big) \\
&= \sum_{y^n} \exp\left(-\min_{\tilde{\theta}\in\widetilde{\Theta}} \Big\{ -\log p_{\tilde{\theta}}(y^n) + \tilde{L}(\tilde{\theta}) \Big\}\right) \\
&= \sum_{y^n} \max_{\tilde{\theta}\in\widetilde{\Theta}} \Big\{ \exp\Big(\log p_{\tilde{\theta}}(y^n) - \tilde{L}(\tilde{\theta})\Big) \Big\} \\
&\le \sum_{y^n} \sum_{\tilde{\theta}\in\widetilde{\Theta}} \exp\Big(\log p_{\tilde{\theta}}(y^n) - \tilde{L}(\tilde{\theta})\Big) \\
&= \sum_{\tilde{\theta}\in\widetilde{\Theta}} \sum_{y^n} p_{\tilde{\theta}}(y^n) \exp\Big(-\tilde{L}(\tilde{\theta})\Big) = \sum_{\tilde{\theta}\in\widetilde{\Theta}} \exp\Big(-\tilde{L}(\tilde{\theta})\Big) \\
&\le 1.
\end{aligned}$$

This implies that $\tilde{p}_2$, which is defined below, is a sub-probability distribution over $\mathscr{Y}^n$.

$$\tilde{p}_2(y^n) = \exp(-\tilde{L}_2(y^n)) = p_{\ddot{\theta}}(y^n)\exp(-\tilde{L}(\ddot{\theta})).$$

Next, we consider MDL estimators for supervised learning. Suppose that we have n training data $(x^n, y^n) := \{(x_i, y_i) \in \mathscr{X} \times \mathscr{Y} | i = 1, 2, \cdots, n\}$ generated from $\bar{p}_*(x^n, y^n) = q_*(x^n)p_*(y^n|x^n)$, where $\mathscr{X}$ is a domain of feature vector x and $\mathscr{Y}$ could be $\Re$ (regression) or a finite set (classification) according to target problems. Here the sequence $(x_1, y_1), (x_2, y_2), \cdots$ is not necessarily i.i.d. but can be a stochastic process in general. We write the jth component of the ith sample as x_{ij}. To define an MDL estimator according to the notion of two-stage code [30], we need to describe data itself and a statistical model used to describe the data too. Letting $\tilde{L}(x^n, y^n)$ be the codelength of the two-stage code to describe (x^n, y^n), $\tilde{L}(x^n, y^n)$ can be decomposed as

$$\tilde{L}(x^n, y^n) = \tilde{L}(x^n) + \tilde{L}(y^n|x^n)$$

by the chain rule. Since a goal of supervised learning is to estimate $p_*(y^n|x^n)$, we need not estimate $q_*(x^n)$. In view of the MDL principle, this implies that $\tilde{L}(x^n)$ (the description length of x^n) can be ignored. Therefore, we only consider the encoding of y^n given x^n hereafter. This corresponds to a description scheme in which an encoder and a decoder share the data x^n. To describe y^n given x^n, we use a parametric model $p_\theta(y^n|x^n)$ with parameter $\theta \in \Theta$. The parameter space Θ is a certain continuous space or a union of continuous spaces. Note that, however, the continuous parameter cannot be encoded. Thus, we need to quantize the parameter space Θ as $\widetilde{\Theta}(x^n)$. According to the notion of the two-stage code, we need to describe not only y^n but also the model used to describe y^n (or equivalently the parameter $\tilde{\theta} \in \widetilde{\Theta}(x^n)$) given x^n. Again by the chain rule, such a codelength can be decomposed as

$$\tilde{L}(y^n, \tilde{\theta}|x^n) = \tilde{L}(y^n|x^n, \tilde{\theta}) + \tilde{L}(\tilde{\theta}|x^n).$$

The term $\tilde{L}(y^n|x^n, \tilde{\theta})$ expresses a codelength to describe y^n using $p_{\tilde{\theta}}(y^n|x^n)$, which is, needless to say, $-\log p_{\tilde{\theta}}(y^n|x^n)$. On the other hand, $\tilde{L}(\tilde{\theta}|x^n)$ expresses a codelength to describe the model $p_{\tilde{\theta}}(y^n|x^n)$ itself. Note that $\tilde{L}(\tilde{\theta}|x^n)$ must satisfy Kraft's inequality

$$\sum_{\tilde{\theta}\in\widetilde{\Theta}(x^n)} \exp(-\tilde{L}(\tilde{\theta}|x^n)) \le 1.$$

The MDL estimator is defined by the minimizer of the above codelength:

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

Let us again write the minimum description length attained by the two-stage code as

$$\tilde{L}_2(y^n|x^n) := -\log p_{\ddot{\theta}}(y^n|x^n) + \tilde{L}(\ddot{\theta}|x^n). \tag{4}$$

Here $\tilde{L}_2$ also satisfies the Kraft's inequality with respect to y^n for each x^n because

$$\begin{aligned}
&\sum_{y^n} \exp\Big(-\tilde{L}_2(y^n|x^n)\Big) \\
&= \sum_{y^n} \exp\left(-\min_{\tilde{\theta}\in\widetilde{\Theta}(x^n)} \Big\{ -\log p_{\tilde{\theta}}(y^n|x^n) + \tilde{L}(\tilde{\theta}|x^n) \Big\}\right) \\
&= \sum_{y^n} \max_{\tilde{\theta}\in\widetilde{\Theta}(x^n)} \Big\{ \exp\Big(\log p_{\tilde{\theta}}(y^n|x^n) - \tilde{L}(\tilde{\theta}|x^n)\Big) \Big\} \\
&\le \sum_{y^n} \sum_{\tilde{\theta}\in\widetilde{\Theta}(x^n)} \exp\Big(\log p_{\tilde{\theta}}(y^n|x^n) - \tilde{L}(\tilde{\theta}|x^n)\Big) \\
&= \sum_{\tilde{\theta}\in\widetilde{\Theta}(x^n)} \sum_{y^n} p_{\tilde{\theta}}(y^n|x^n) \exp\Big(-\tilde{L}(\tilde{\theta}|x^n)\Big) \\
&= \sum_{\tilde{\theta}\in\widetilde{\Theta}(x^n)} \exp\Big(-\tilde{L}(\tilde{\theta}|x^n)\Big) \le 1.
\end{aligned} \tag{5}$$

Hence it is interpreted as a codelength of a prefix two-stage code. Therefore,

$$\tilde{p}_2(y^n|x^n) := \exp(-\tilde{L}_2(y^n|x^n)) \tag{6}$$

is a conditional sub-probability distribution corresponding to the two-stage code. Finally, we remark that any symbols with tilde like $\tilde{L}$ is related to the case where the parameter space is quantized. If the same symbol does not have the tilde, the symbol corresponds to its counterpart in the case where the parameter space is not quantized.

III. Barron and Cover's Theory

We briefly review Barron and Cover's theory (BC theory) in density estimation and its extension to penalized likelihood methods in the fixed design setting. Furthermore, we will explain why such an extension is difficult for the random design setting.

A. Barron and Cover's Theory for Density Estimation

In BC theory, the Rényi divergence [29] between $p(y)$ and $r(y)$ with order $\lambda \in (0,1)$

$$d_\lambda^n(p,r) = -\frac{1}{1-\lambda}\log E_{p(y^n)}\left(\frac{r(y^n)}{p(y^n)}\right)^{1-\lambda} \tag{7}$$

is used as a loss function. See [21] for several properties of the Rényi divergence. The Rényi divergence converges to Kullback-Leibler (KL) divergence

$$\mathscr{D}^n(p,r) := \int p(y^n)\left(\log\frac{p(y^n)}{r(y^n)}\right)dy^n \tag{8}$$

as $\lambda \to 1$, i.e.,

$$\lim_{\lambda\to 1} d_\lambda^n(p,r) = \mathscr{D}^n(p,r) \tag{9}$$

for any p and any r. We also note that the Rényi divergence at $\lambda = 0.5$ is equal to Bhattacharyya divergence [11]

$$d_{0.5}^n(p,r) = -2\log\int\sqrt{p(y^n)r(y^n)}dy^n. \tag{10}$$

We drop n of each divergence like $d_\lambda(p,r)$ if it is defined with a single random variable, i.e.,

$$d_\lambda(p,r) = -\frac{1}{1-\lambda}\log E_{p(y)}\left(\frac{r(y)}{p(y)}\right)^{1-\lambda}.$$

BC theory requires the model description length to satisfy a little bit stronger Kraft's inequality defined as follows.

Definition 2: Let β be a real number in $(0,1)$. We say that a function $h(\tilde{\theta})$ satisfies β-stronger Kraft's inequality if

$$\sum_{\tilde{\theta}}\exp(-\beta h(\tilde{\theta})) \le 1,$$

where the summation is taken over a range of $\tilde{\theta}$ in its context.

BC theory [2] gives the following two theorems. Though these theorems were shown only for the case of Hellinger distance in the original paper [2], we state them with the Rényi divergence.

Theorem 1: Let β be a real number in $(0,1)$. Assume that $\tilde{L}$ satisfies β-stronger Kraft's inequality. Then

$$\begin{aligned}\frac{1}{n}E_{p_*(y^n)}d_\lambda^n(p_*,p_{\ddot{\theta}}) &\le \frac{1}{n}E_{p_*(y^n)}\left[\inf_{\tilde{\theta}\in\tilde{\Theta}}\left\{\log\frac{p_*(y^n)}{p_{\tilde{\theta}}(y^n)}+\tilde{L}(\tilde{\theta})\right\}\right] &(11)\\ &= \frac{1}{n}E_{p_*(y^n)}\log\frac{p_*(y^n)}{\tilde{p}_2(y^n)} &(12)\end{aligned}$$

for any $\lambda \in (0, 1-\beta]$.

Theorem 2: Let β be a real number in $(0,1)$. Assume that $\tilde{L}$ satisfies β-stronger Kraft's inequality. Then

$$\Pr\left(\frac{d_\lambda^n(p_*,p_{\ddot{\theta}})}{n} - \frac{1}{n}\log\frac{p_*(y^n)}{\tilde{p}_2(y^n)} \ge \tau\right) \le e^{-n\tau\beta} \tag{13}$$

for any $\lambda \in (0, 1-\beta]$.

The right side of (12) is the redundancy of the two-stage code. Similarly, the second term of the left side of (13) is the regret of the two-stage code. Therefore, both theorems claim that small redundancy/regret leads to the small bound on statistical risk of the estimator $\ddot{\theta}$. This is the first and the most essential mathematical justification of the MDL principle as ever. Prominently, both theorems hold in a variety of situation because no noisy assumptions are necessary.

Remark: About (11) of Theorem 1, by exchanging expectation and infimum, we have

$$\begin{aligned}&\frac{1}{n}E_{p_*(y^n)}\left[\inf_{\tilde{\theta}\in\tilde{\Theta}}\left\{\log\frac{p_*(y^n)}{p_{\tilde{\theta}}(y^n)}+\tilde{L}(\tilde{\theta})\right\}\right]\\ &\le \inf_{\tilde{\theta}\in\tilde{\Theta}}\frac{1}{n}E_{p_*(y^n)}\left\{\log\frac{p_*(y^n)}{p_{\tilde{\theta}}(y^n)}+\tilde{L}(\tilde{\theta})\right\}\\ &= \inf_{\tilde{\theta}\in\tilde{\Theta}}\left\{\frac{\mathscr{D}^n(p_*,p_{\tilde{\theta}})}{n}+\frac{\tilde{L}(\tilde{\theta})}{n}\right\}\end{aligned}$$

The right side is "the index of resolvability of p_*" introduced in [2], which expresses the trade off between approximation error and the complexity of the used model. This is particularly important when the true distribution p_* does not belong to the model $\{p_\theta\}$, but the situation is out of scope of this paper. Hence, we do not discuss it in detail.

Both theorems can be proved by the following lemma.

Lemma 1: Let β be a real number in $(0,1)$. Assume that $\tilde{L}$ satisfies β-stronger Kraft's inequality. Let $\dot{\theta}$ be an arbitrary function from $\mathscr{Y}^n$ to $\tilde{\Theta}$. Then,

$$E_{p_*(y^n)}\exp\left(\beta d_\lambda^n(p_*,p_{\dot{\theta}}) - \beta\log\frac{p_*(y^n)}{p_{\dot{\theta}}(y^n)} - \beta\tilde{L}(\dot{\theta})\right) \le 1 \tag{14}$$

holds for any $\lambda \in (0, 1-\beta]$.

By letting $\dot{\theta}$ be $\ddot{\theta}$ and by using the Jensen's inequality, we have Theorem 1. Similarly, usage of Markov's inequality yields Theorem 2.

Proof of Lemma 1: We can assume $\lambda = 1-\beta$, since the Rényi divergence is increasing in λ. We have

$$\begin{aligned}
&E_{p_*(y^n)} \exp\Big(\beta d_\lambda^n(p_*, p_{\dot\theta}) - \beta \log \frac{p_*(y^n)}{p_{\dot\theta}(y^n)} - \beta \tilde L(\dot\theta)\Big) \\
\le& E_{p_*(y^n)} \sum_{\tilde\theta} \exp\Big(\beta d_\lambda^n(p_*, p_{\tilde\theta}) - \beta \log \frac{p_*(y^n)}{p_{\tilde\theta}(y^n)} - \beta \tilde L(\tilde\theta)\Big) \\
\le& \sum_{\tilde\theta} \exp\Big(-\beta \tilde L(\tilde\theta) + \beta d_\lambda^n(p_*, p_{\tilde\theta})\Big) E_{p_*(y^n)} \Big(\frac{p_{\tilde\theta}(y^n)}{p_*(y^n)}\Big)^{1-\lambda} \\
\le& \sum_{\tilde\theta} \exp\Big(-\beta \tilde L(\tilde\theta)\Big) \le 1.
\end{aligned}$$

In the last line, we have used

$$E_{p_*(y^n)} \Big(\frac{p_{\tilde\theta}(y^n)}{p_*(y^n)}\Big)^{1-\lambda} = \exp\Big(-\beta d_\lambda^n(p_*, p_{\tilde\theta})\Big), \tag{15}$$

which cancels the factor $\exp(\beta d_\lambda^n(p_*, p_{\tilde\theta}))$ in the second last line. This is a key trick of BC theory. □

We would like to note that Lemma 1 can be stated with exponential stochastic inequality (ESI) notation that was introduced by [25]. For any pair of random variables (U, U') subject to a probability distribution p and any positive real number $b \in \Re$, the notation $U \preceq_b^p U'$ denotes $E_p[\exp(b(U-U'))] \le 1$. Using the definition, (14) can be written as

$$d_\lambda^n(p_*, p_{\dot\theta}) \preceq_\beta^{p_*} \log \frac{p_*(y^n)}{\tilde p_2(y^n)}. \tag{16}$$

Since Theorems 1 and 2 are immediately obtained from Lemma 1, ESI notation (16) is a simple device to summarize both theorems. However, in the main content of this paper, our results are a little more complicated statements in the risk bound. Hence, we cannot use ESI notations as itself. Hence we do not employ the ESI notation hereafter.

The theorems and their proofs in this subsection are more sophisticated ones compared to the original form by Barron and Cover. In particular, the expectation form (Theorem 1) was established in 1999 in Ph.D thesis of Li [28], who is a former student of Barron. Also in 2004, Zhang [36] gave the essentially same result and proof with certain generalized results. In 2009, Zhang [37] gave the more general results including the probability form (Theorem 2). See a historical note by Grünwald (in p. 483, [23]) for more detailed history.

B. Barron and Cover's Theory for Penalized Likelihood

Since 2008, Barron *et al.* [3], Barron and Luo [4], and Chatterjee and Barron [17] have been developing a framework to enhance BC theory to penalized likelihood methods without quantization of parameters, essentially for unsupervised setting. Here we briefly review their contributions. For their purpose, they introduced notions called codelength validity and risk validity of penalty functions. When the employed penalty function satisfies both notions of validity, we can show that the considered penalized likelihood estimator

$$\hat\theta(y^n) := \arg\min_{\theta\in\Theta} \{-\log p_\theta(y^n) + L(\theta)\} \tag{17}$$

is a BC-proper MDL estimator. The penalty function is written as $L(\theta)$ here instead of $\mathrm{Pen}(\theta)$ because we try to interpret this term as a counterpart of codelength $\tilde L$. However Θ is a continuous space so that θ cannot be encoded. Thus, $L(\theta)$ is not codelength (needless to say, it does not satisfy Kraft's inequality) but just a penalty function. This implies that $p_{\hat\theta}(y^n)\exp(-L(\hat\theta))$ is not necessarily a sub-probability density function in general.

Codelength validity is defined as follows.

Definition 3 (Codelength Validity): Suppose that there exist a quantized parameter space $\tilde\Theta$ (a countable set) and a codelength function $\tilde L$ over $\tilde\Theta$ which satisfies Kraft's inequality, such that

$$\begin{aligned}
&\forall n \ge 1,\ \forall y^n \in \mathscr{Y}^n, \\
&\min_{\theta\in\Theta} \{-\log p_\theta(y^n) + L(\theta)\} \ge \min_{\tilde\theta\in\tilde\Theta} \Big\{-\log p_{\tilde\theta}(y^n) + \tilde L(\tilde\theta)\Big\}.
\end{aligned}$$

Then, L is said to be codelength valid.

Suppose that L is codelength valid. Since the right side satisfies Kraft's inequality as a function of y^n, the left side also satisfies Kraft's inequality. Hence this is a sufficient condition for Condition 1. This implies that $p_{\hat\theta}(y^n)\exp(-L(\hat\theta))$ is a sub-probability density over $\mathscr{Y}^n$. See Section IV-A for more details about how the codelength validity works. Though the explanation there is described for supervised learning, its essence is the same.

The following is risk validity.

Definition 4 (Risk Validity for Density Estimation): Suppose that there exist a quantized parameter space $\tilde\Theta$ (a countable set) and a codelength function $\tilde L$ over $\tilde\Theta$ which satisfies β-stronger Kraft's inequality ($0 < \beta < 1$), such that

$$\begin{aligned}
&\forall n \ge 1,\ \forall y^n \in \mathscr{Y}^n, \\
&\quad \max_{\theta\in\Theta} \Big\{ d_\lambda^n(p_*, p_\theta) - \log \frac{p_*(y^n)}{p_\theta(y^n)} + L(\theta) \Big\} \\
&\quad \le \max_{\tilde\theta\in\tilde\Theta} \Big\{ d_\lambda^n(p_*, p_{\tilde\theta}) - \log \frac{p_*(y^n)}{p_{\tilde\theta}(y^n)} + \tilde L(\tilde\theta) \Big\}.
\end{aligned} \tag{18}$$

Then, L is said to be risk valid.

Now suppose that L is risk valid. Let $V(y^n, \theta)$ and $\tilde V(y^n, \tilde\theta)$ denote the inside of $\max$ of the left and right sides of (18), respectively. For any function $\dot\theta : \mathscr{Y}^n \to \Theta$, we have

$$\begin{aligned}
E_{p_*(y^n)} \exp(\beta V(y^n, \dot\theta)) &\le E_{p_*(y^n)} \exp(\beta \max_{\theta\in\Theta} V(y^n, \theta)) \\
&\le E_{p_*(y^n)} \exp(\beta \max_{\tilde\theta\in\tilde\Theta} \tilde V(y^n, \tilde\theta)) \\
&\le 1.
\end{aligned}$$

The second inequality is obtained by the risk validity, while the last inequality is obtained by Lemma 1. This yields the similar risk bound for $\hat\theta$ as Theorems 1 and 2 by the same way. Further if L is codelength valid, the right sides of the bounds are interpreted as redundancy and regret, respectively. Then, we have an extension of BC theory for penalized likelihood methods. Similarly as for codelength functions, the corresponding penalty functions are desirable to be as small as possible. Hence, the remainder task is to obtain small penalty functions which are codelength and risk valid. In [3], [4], and [17], various penalized maximum likelihood estimators are considered. For example in [17], they obtain penalty functions with codelength and risk validity for the graphical lasso, which

is an estimator for multivariate Gaussian distributions with ℓ_1-penalty.

They also considered the lasso too. We guess that many readers can hardly understand the difference between their analysis and our analysis of this paper immediately. Thus we introduce their result first by using notation which makes the difference clear. As is well known, lasso can be interpreted as a penalized likelihood method. The readers unfamiliar to lasso can refer to the head of Section IV-B. Its likelihood is not $p_\theta(y^n)$ but a conditional likelihood $p_\theta(y^n|x^n)$. In the analysis, they fixed x^n to a specific value through learning and validation processes. This setting is called *supervised learning in fixed design setting* in machine learning. It means that x^n does not play nothing more than the role of index which specifies the hypothesis class $\{p_\theta(y^n|x^n)|\theta \in \Theta\}$. Hence the fixed design setting is equivalent to the unsupervised learning (density estimation). Thus, the above discussion for density estimation holds for the the fixed design setting as follows. For evaluation of lasso in [17], the Rényi divergence was defined for given x^n as

$$d^n_\lambda(p, r|x^n) = -\frac{1}{1-\lambda}\log\int p(y^n|x^n)\left(\frac{r(y^n|x^n)}{p(y^n|x^n)}\right)^{1-\lambda} dy^n. \tag{19}$$

Further the risk validity condition (18) is simply modified as follows.

Definition 5 (Risk Validity for Fixed Design): For given x^n, suppose that there exist a quantized parameter space $\tilde{\Theta}$ (a countable set) and a codelength function $\tilde{L}$ over $\tilde{\Theta}$ which satisfies β-stronger Kraft's inequality ($0 < \beta < 1$), such that

$$\begin{aligned}&\forall n \geq 1,\ \forall y^n \in \mathscr{Y}^n,\\ &\max_{\theta\in\Theta}\left\{d^n_\lambda(p_*, p_\theta|x^n) - \log\frac{p_*(y^n|x^n)}{p_\theta(y^n|x^n)} - L(\theta|x^n)\right\}\\ &\leq \max_{\tilde{\theta}\in\tilde{\Theta}(x^n)}\left\{d^n_\lambda(p_*, p_{\tilde{\theta}}|x^n) - \log\frac{p_*(y^n|x^n)}{p_{\tilde{\theta}}(y^n|x^n)} - \tilde{L}(\tilde{\theta}|x^n)\right\}.\end{aligned} \tag{20}$$

Then, L is said to be risk valid.

Note that their original definition in [17] was presented only for the case where $\lambda = 1-\beta$. We define a penalized likelihood estimator by

$$\hat{\theta}(x^n, y^n) = \min_{\theta\in\Theta}\left\{-\log p_\theta(y^n|x^n) + L(\theta|x^n)\right\}. \tag{21}$$

As described in Section II, it is natural that the codelength of parameter depends on x^n. By analogy we consider penalty functions of the form $L(\theta|x^n)$ though widely used penalty functions are usually independent of x^n. As the result of the paper, the data-dependent penalty functions are more suitable in view of the MDL principle. Similarly to (4) and (6) in the quantized case, we can formally define

$$L_2(y^n|x^n) := \min_{\theta\in\theta}\left\{-\log p_\theta(y^n|x^n) + L(\theta|x^n)\right\}, \tag{22}$$

$$p_2(y^n|x^n) := \exp(-p_2(y^n|x^n)) = p_{\hat{\theta}}(y^n|x^n)\cdot\exp(-L(\hat{\theta})). \tag{23}$$

However we again note that $L_2(y^n|x^n)$ is not necessarily codelength and that $p_2(y^n|x^n)$ is not necessarily a sub-probability distribution due to the same reason as the unsupervised case. By this risk validity, it is easy to obtain a similar risk bound

$$\begin{aligned}&E_{p_*(y^n|x^n)} d^n_\lambda(p_*, p_{\hat{\theta}}|x^n)\\ &\leq E_{p_*(y^n|x^n)}\left[\inf_{\theta\in\Theta}\left\{\log\frac{p_*(y^n|x^n)}{p_\theta(y^n|x^n)} + L(\theta|x^n)\right\}\right]\\ &= E_{p_*(y^n|x^n)}\left[\log\frac{p_*(y^n|x^n)}{p_2(y^n|x^n)}\right]\end{aligned}$$

by the same way as before. Codelength validity can be directly extended without any problem. Thus, we can say that penalized likelihood estimators in the fixed design setting are BC-proper MDL estimators under the codelength and risk validity.

However it is clear that $d^n_\lambda(p_*, p_{\hat{\theta}}|x^n)$ cannot measure generalization error in supervised learning. In the random design setting, x^n is also stochastic (subject to a certain distribution $q_*(x^n)$). For supervised learning, we employ the following conditional Rényi divergence as a loss function.

$$\begin{aligned}&d^n_\lambda(p, r) :=\\ &-\frac{1}{1-\lambda}\log\int q_*(x^n)p(y^n|x^n)\left(\frac{r(y^n|x^n)}{p(y^n|x^n)}\right)^{1-\lambda} dx^n dy^n.\end{aligned} \tag{24}$$

Note that $d^n_\lambda(p, r)$ depends on the distribution q_* of the covariates. This is a natural generalization of (7) and can measure generalization error. Similarly to the unsupervised version, it converges to conditional Kullback-Leibler (KL) divergence

$$\mathscr{D}^n(p, r) := \int q_*(x^n)p(y^n|x^n)\left(\log\frac{p(y^n|x^n)}{r(y^n|x^n)}\right) dx^n dy^n, \tag{25}$$

as $\lambda \to 1$, i.e.,

$$\lim_{\lambda\to 1} d^n_\lambda(p, r) = \mathscr{D}^n(p, r) \tag{26}$$

for any p and r. In addition, the Rényi divergence at $\lambda = 0.5$ is equal to Bhattacharyya divergence [11]

$$d^n_{0.5}(p, r) = -2\log\int q_*(x^n)\sqrt{p(y^n|x^n)r(y^n|x^n)}dx^n dy^n. \tag{27}$$

Again, we drop n of each divergence like $d_\lambda(p, r)$ if $n = 1$, i.e., $d_\lambda(p, r) := d^1_\lambda(p, r)$. Hereafter we see why the above idea in the fixed design setting is significantly difficult to be extended to the random design setting. By extending the definition of risk validity to random design straightforwardly, we obtain the following definition.

Definition 6 (Risk Validity in Random Design): Let β be a real number in $(0, 1)$ and λ be a real number in $(0, 1-\beta]$. 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 β-stronger Kraft's inequality for each x^n such that

$$\begin{aligned}&\forall n \geq 1,\ \forall x^n \in \mathscr{X}^n,\ y^n \in \mathscr{Y}^n,\\ &\max_{\theta\in\Theta}\left\{d^n_\lambda(p_*, p_\theta) - \log\frac{p_*(y^n|x^n)}{p_\theta(y^n|x^n)} - L(\theta|x^n)\right\}\\ &\leq \max_{\tilde{\theta}\in\widetilde{\Theta}(x^n)}\left\{d^n_\lambda(p_*, p_{\tilde{\theta}}) - \log\frac{p_*(y^n|x^n)}{p_{\tilde{\theta}}(y^n|x^n)} - \tilde{L}(\tilde{\theta}|x^n)\right\}.\end{aligned} \tag{28}$$

In contrast to the fixed design case, (20) must hold not only for a fixed $x^n \in \mathscr{X}^n$ but also for all $x^n \in \mathscr{X}^n$. By this risk validity, let us try to recover Lemma 1, which is the essence of BC theory. Assume $\beta = 1 - \lambda$ for the same reason as before. Let $V(x^n, y^n, \theta)$ and $\tilde{V}(x^n, y^n, \tilde{\theta})$ denote the inside of $\max$ of the left and right sides of (28), respectively. For any function $\dot{\theta} : \mathscr{X}^n \times \mathscr{Y}^n \to \Theta$,

$$\begin{aligned}
&E_{\bar{p}_*(x^n,y^n)} \exp\left(\beta V(x^n, y^n, \dot{\theta})\right) \\
\le& E_{\bar{p}_*(x^n,y^n)} \exp\left(\max_{\theta\in\Theta}\{\beta V(x^n, y^n, \theta)\}\right) \\
\le& E_{\bar{p}_*(x^n,y^n)} \exp\left(\max_{\tilde{\theta}\in\widetilde{\Theta}(x^n)}\left\{\beta \tilde{V}(x^n, y^n, \tilde{\theta})\right\}\right) \qquad (29)\\
\le& E_{p_*(y^n|x^n)q_*(x^n)} \sum_{\tilde{\theta}\in\widetilde{\Theta}(x^n)} \exp\left(\beta \tilde{V}(x^n, y^n, \tilde{\theta})\right) \\
=& E_{q_*(x^n)} \sum_{\tilde{\theta}\in\widetilde{\Theta}(x^n)} \exp\left(-\beta \tilde{L}(\tilde{\theta}|x^n)\right) \exp(\beta d^n_\lambda(p_*, p_{\tilde{\theta}})) \\
&\cdot E_{p_*(y^n|x^n)} \exp\left(-\beta \log \frac{p_*(y^n|x^n)}{p_{\tilde{\theta}}(y^n|x^n)}\right) \\
=& E_{q_*(x^n)} \sum_{\tilde{\theta}\in\widetilde{\Theta}(x^n)} \exp\left(-\beta \tilde{L}(\tilde{\theta}|x^n)\right) \exp(\beta d^n_\lambda(p_*, p_{\tilde{\theta}})) \\
&\cdot F(\tilde{\theta}, x^n),
\end{aligned}$$

where we let

$$F(\tilde{\theta}, x^n) = E_{p_*(y^n|x^n)} \exp\left(-\beta \log \frac{p_*(y^n|x^n)}{p_{\tilde{\theta}}(y^n|x^n)}\right).$$

Since

$$E_{q_*(x^n)} F(\tilde{\theta}, x^n) = \exp(-\beta d^n_\lambda(p_*, p_{\tilde{\theta}})), \qquad (30)$$

if we can apply the expectation about x^n directly to $F(\tilde{\theta}, x^n)$, then the Rényi divergence factor will vanish (a key trick of BC theory) and we can utilize the Kraft's inequality about $\tilde{\theta}$, which yields the same result as Lemma 1. However, it seems impossible in general. A simple solution is to assume the following condition.

Condition 3 (Independent Condition): Both the quantized space and the model description length are independent of x^n, i.e.,

$$\widetilde{\Theta}(x^n) = \widetilde{\Theta}, \quad \tilde{L}(\tilde{\theta}|x^n) = \tilde{L}(\tilde{\theta}). \qquad (31)$$

We can easily check that this condition overcomes the above issue. Also, it seems difficult to find another substitute. This is not specific to penalized likelihood. The independent condition is necessary also for the two-stage code in the random design setting. If we try to recover Lemma 1 for this case, we have for any $\dot{\theta} : \mathscr{X}^n \times \mathscr{Y}^n \to \widetilde{\Theta}(x^n)$

$$\begin{aligned}
&E_{\bar{p}_*(x^n,y^n)} \exp\left(\beta d^n_\lambda(p_*, p_{\dot{\theta}}) - \beta \log \frac{p_*(y^n|x^n)}{p_{\dot{\theta}}(y^n|x^n)} - \beta \tilde{L}(\dot{\theta}|x^n)\right) \\
&= E_{\bar{p}_*(x^n,y^n)} \exp\left(\beta \tilde{V}(x^n, y^n, \dot{\theta})\right) \\
&\le E_{\bar{p}_*(x^n,y^n)} \exp\left(\max_{\tilde{\theta}\in\widetilde{\Theta}(x^n)}\left\{\beta \tilde{V}(x^n, y^n, \tilde{\theta})\right\}\right).
\end{aligned}$$

Since the right side is the same as (29), the remaining step is exactly the same. Thus, the independent condition is necessary regardless if the parameter space is quantized or not. In fact, Yamanishi [35] employed the quantization and the codelength function which satisfy the independence condition. In general, the restriction caused by the independence condition prevents us from optimizing the quantization and the parameter codelength according to x^n.

In addition, we face another difficulty caused in applying the idea of risk validity for penalized likelihood estimators (no quantization). Let us explain the main difficulties by using lasso as an example. Under the independent condition, (28) can be equivalently rewritten as

$$\begin{aligned}
&\forall n \ge 1,\ \forall x^n \in \mathscr{X}^n,\ \forall y^n \in \mathscr{Y}^n,\ \forall \theta \in \Theta, \\
&\min_{\tilde{\theta}\in\widetilde{\Theta}}\left\{ d^n_\lambda(p_*, p_\theta) - d^n_\lambda(p_*, p_{\tilde{\theta}}) + \log \frac{p_\theta(y^n|x^n)}{p_{\tilde{\theta}}(y^n|x^n)} + \tilde{L}(\tilde{\theta})\right\} \\
&\le L(\theta|x^n). \qquad (32)
\end{aligned}$$

For simplicity, we write the inside part of the minimum of the left side of (32) as $H(\theta, \tilde{\theta}, x^n, y^n)$. We want to evaluate $\min_{\tilde{\theta}}\{H(\theta, \tilde{\theta}, x^n, y^n)\}$ in order to find its upper bound (risk valid penalties). Chatterjee and Barron [17] provided a fairly nice technique to obtain a tight upper bound on the minimum for the fixed design case. Below, we roughly explain their idea assuming x^n is fixed and using the notation $L(\theta)$ instead of $L(\theta|x^n)$.

The idea of Chatterjee and Barron is to evaluate $E_{\tilde{\theta}} H(\theta, \tilde{\theta}, x^n, y^n)$ by randomizing $\tilde{\theta}$ dexterously instead of evaluating the minimum since $\min_{\tilde{\theta}} H(\theta, \tilde{\theta}, x^n, y^n) \le E_{\tilde{\theta}} H(\theta, \tilde{\theta}, x^n, y^n)$. In addition, the randomization of $\tilde{\theta}$ was constructed based on the following idea. Assume that $L(\theta)$ is a smooth function (for example class C^1) and that $L(\theta) = \tilde{L}(\tilde{\theta}) + C$ holds over $\widetilde{\Theta}$ with a certain real number C. If the parameter space is quantized finely enough, the penalized likelihood estimator $\hat{\theta}(x^n, y^n)$ is expected to behave similarly to $\ddot{\theta}(x^n, y^n)$ and is expected to have a similar risk bound. In order to make L similar to $\tilde{L}$ risk valid, taking $\tilde{\theta}$ close to θ seems to be a good choice for given $\theta \in \Theta$. If $\tilde{\theta}$ is close to θ, the first three terms of $H(\theta, \tilde{\theta}, x^n, y^n)$ is expected to be small so that H is close to $\tilde{L}(\theta)$. In the end, it is expected that $L(\theta)$ which is similar to $\tilde{L}(\theta)$ satisfies (32). Indeed by randomizing $\tilde{\theta}$ on $\widetilde{\Theta}$ around θ nicely, Chatterjee and Barron succeeded in bounding the first three terms of H from above by a small value and in showing that the weighted ℓ_1 penalty (see (33) for its definition) is risk valid under an appropriate condition. Note that the distribution of $\tilde{\theta}$ is designed as it naturally reduces to the point mass at θ when $\theta \in \widetilde{\Theta}$ holds. We also employ their technique for the random design setting in this paper, since there seems no other significantly better way.

Let us see what will happen if we apply the technique of Chatterjee and Barron (called CB technique hereafter) to random design setting. For any choice of $\widetilde{\Theta} \subset \Theta$ and $\tilde{L}(\tilde{\theta})$, we consider the following two cases separately: Case (a): $\theta \notin \widetilde{\Theta}$, and Case (b): $\theta \in \widetilde{\Theta}$.

Consider Case (a) first. In contrast to Case (b), $E_{\tilde{\theta}} H(\theta, \tilde{\theta}, x^n, y^n)$ depends on not only θ but also x^n, y^n. Though Chatterjee and Barron succeeded in removing the dependency of $E_{\tilde{\theta}} H(\theta, \tilde{\theta}, x^n, y^n)$ on y^n by carefully tuned

randomization in case of linear regression (including lasso), the dependency on x^n is hard to eliminate. Let us write the resultant expectation as $H'(\theta, x^n) := E_{\tilde{\theta}}[H(\theta, \tilde{\theta}, x^n, y^n)]$. Since the risk valid penalties derived by CB technique are upper bounds on $H'(\theta, x^n)$ for any x^n, the above fact implies that the obtained risk valid penalties depend on x^n in general. If not ($L(\theta|x^n) = L(\theta)$), $L(\theta)$ must satisfy

$$\max_{x^n} H'(\theta, x^n) \leq L(\theta),$$

which makes $L(\theta)$ much larger. Indeed the resultant $H(\theta, x^n)$ is unbounded with respect to x^n in case of linear regression despite the effort to mimic the case $\theta \in \widetilde{\Theta}$. The unboundedness of $H'(\theta, x^n)$ originates from the third term of the left side of (32). This can be seen by checking Section III of [17]. Though their setting is fixed design, this fact is also true for the random design. This is again unfavorable in view of the MDL principle. Hence we should design the risk valid penalties as they depend on x^n in general.

However penalty functions used by penalized likelihood are often independent of x^n. Indeed the ℓ_1 norm used in the usual lasso does not depend on x^n. Hence, at first sight, the risk validity seems to be useless for lasso. However the following weighted ℓ_1 norm

$$\|\theta\|_{w,1} := \sum_{j=1}^{m} w_j |\theta_j|,$$

$$\text{where} \quad w := (w_1, \cdots, w_m)^T, \quad w_j := \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2} \tag{33}$$

plays an important role here. The superscript T denotes the vector/matrix transpose. The lasso with this weighted ℓ_1 norm is equivalent to an ordinary lasso with column normalization such that each column of the design matrix has the same norm.

Lemma 2 (Column Normalization): The ordinary lasso with column normalization is equal to the lasso with weighted ℓ_1 penalty. More formally, provided that W is a diagonal matrix which has no zero diagonal element,

$$\arg\min_{\theta \in \Re^m} \left\{ \frac{1}{2n\sigma^2} \|Y - X\theta\|_2^2 + \mu_1 \|\theta\|_{w,1} \right\}$$
$$= \arg\min_{\theta \in \Re^m} \left\{ \frac{1}{2n\sigma^2} \|Y - X'\theta\|_2^2 + \mu_1 \|\theta\|_1 \right\},$$

where $W = \mathrm{diag}\{w_1, w_2, \cdots, w_m\}$, $X' := XW^{-1}$ and $\|\theta\|_1 := \sum_{j=1}^m |\theta_j|$.

See Section IV-B for the definitions of the above matrix/vector notations.

Proof:

$$\arg\min_{\theta \in \Re^m} \left\{ \frac{1}{2n\sigma^2} \|Y - X\theta\|_2^2 + \mu_1 \|\theta\|_{w,1} \right\}$$
$$= \arg\min_{\theta \in \Re^m} \left\{ \frac{1}{2n\sigma^2} \|Y - XW^{-1}W\theta\|_2^2 + \mu_1 \|W\theta\|_1 \right\}$$
$$= \arg\min_{\theta \in \Re^m} \left\{ \frac{1}{2n\sigma^2} \|Y - X'W\theta\|_2^2 + \mu_1 \|W\theta\|_1 \right\}$$
$$= \arg\min_{\theta' \in \Re^m} \left\{ \frac{1}{2n\sigma^2} \|Y - X'\theta'\|_2^2 + \mu_1 \|\theta'\|_1 \right\}.$$

The column normalization is theoretically and practically important as mentioned in Section I-A. Hence we try to find a risk valid penalty of the form $L_1(\theta|x^n) = \mu_1\|\theta\|_{w,1} + \mu_2$, where μ_1 and μ_2 are real coefficients. Indeed, there seems to be no other useful penalty dependent on x^n for the usual lasso. Chatterjee and Barron succeeded in deriving a condition such that $L_1(\theta|x^n)$ is risk valid in the fixed design setting.

However, as long as we employ CB technique, we cannot find any 'useful' risk valid weighted ℓ_1 penalty in the random design setting. Even if $L(\theta|x^n)$ actually depends on x^n like the weighted ℓ_1 norm, the form of $H'(\theta, x^n)$ differs from that of desirable penalty functions in general. Thus it is not easy to bound $H'(\theta, x^n)$ by penalty function $L(\theta|x^n)$ for the whole range of x^n in general. Indeed it seems desperately difficult in case of lasso.

In fact, we can show, by focusing on Case (b), that the weighted ℓ_1 norm type penalties derived by CB technique cannot be risk valid. Recall that the distribution of $\tilde{\theta}$ used in CB technique reduces to the point mass at θ when $\theta \in \widetilde{\Theta}$. Hence we have $E_{\tilde{\theta}}[H(\theta, \tilde{\theta}, x^n, y^n)] = H(\theta, \theta, x^n, y^n) = \tilde{L}(\theta)$, which means

$$\tilde{L}(\theta) \leq L(\theta|x^n) \tag{34}$$

must hold for any $\theta \in \widetilde{\Theta}$ and any x^n to make $L(\theta|x^n)$ to be risk valid. Recalling the definition of $\|\theta\|_{w,1}$ (33) and $\tilde{L}$, we have

$$\tilde{L}(\theta) \leq \min_{x^n} L(\theta|x^n) = \min_{x^n} \mu_1\|\theta\|_{w,1} + \mu_2 = \mu_2$$

for any $\theta \in \widetilde{\Theta}$. That is, $\tilde{L}(\theta)$ must be bounded and cannot satisfy Kraft's inequality over $\widetilde{\Theta}$, since $\widetilde{\Theta}$ is countably infinite. (This case is equivalent to the condition that $e^{-\tilde{L}(\theta)}$ cannot be normalized over $\widetilde{\Theta}$.) Since the definition of the risk validity requires $\tilde{L}$ to satisfy β-stronger Kraft's inequality, we cannot obtain any risk valid penalty by using CB technique. Here we should note that $\min_{x^n} \|\theta\|_{w,1}$ is attained when $w_j = 0$ for all j. It implies that the design matrix X is zero. Such extreme cases including non full rank X should be avoided. However, even if we assume $w_j > 0$ for all j, we still have $\inf_{x^n} \|\theta\|_{w,1} = 0$ for all $\theta \in \widetilde{\Theta}$, and the same conclusion follows. If $\tilde{L}$ could depend on x^n, the choice $\tilde{L}(\theta|x^n) = L(\theta|x^n) = L_1(\theta|x^n)$ solved the above issue. In this sense, the main difficulty is caused by Condition 3. This issue does not seem to be specific to lasso.

Another major issue is the Rényi divergence $d_\lambda^n(p_*, p_\theta)$. In the fixed design case, the Rényi divergence $d_\lambda^n(p_*, p_\theta|x^n)$ is a simple convex function in terms of θ, which makes the analysis easy. In contrast, the Rényi divergence $d_\lambda^n(p_*, p_\theta)$ in case of random design is not convex and more complicated than that of fixed design cases, which makes it significantly difficult to evaluate $H'(\theta, x^n)$. We will describe why the non-convexity of loss function makes the derivation difficult in Section VI-G. In order to solve this issue, another version of Rényi divergence

seems to be a good alternative.

$$\begin{aligned}&\mathrm{altd}^n_\lambda(p,r|q_*)\\&:=-\frac{1}{1-\lambda}\int q_*(x^n)\log\int p(y^n|x^n)\left(\frac{r(y^n|x^n)}{p(y^n|x^n)}\right)^{1-\lambda}dy^n dx^n\\&=\int q_*(x^n)d^n_\lambda(p,r|x^n)dx^n.\end{aligned}$$

It is immediate to see that altd^n_λ satisfies definition of statistical divergence. If the Rényi divergence of the fixed design case $d^n_\lambda(p,r|x^n)$ has a simple form, then altd^n_λ is likely to be simple because altd^n_λ is obtained by taking its expectation with respect to x^n. Indeed, altd^n_λ is just a sort of mean squared error of $\hat{\theta}$ for lasso setting. Thus, it is easy to analyze and to interpret. However, this try fails due to another reason. Since the divergence $\mathrm{altd}^n_\lambda(p_*,p_{\hat{\theta}})$ does not satisfy (30), the key trick of BC theory is not applicable to derive a risk bound. There is a possibility that a similar risk bound can be derived by other technique than BC theory. However we see that $\mathrm{altd}^n_\lambda(p,r|q_*)$ is an upper bound on $d^n_\lambda(p,r)$ by Jensen's inequality. Hence, the statistical risk with $\mathrm{altd}^n_\lambda(p,r|q_*)$ is harder than that with $d^n_\lambda(p,r)$ (or impossible) to be bounded from above by the same redundancy.

The difficulties that we face when we use the techniques of [17] in the random design case are not limited to them. We do not explain them here because it requires the readers to understand their techniques in detail. However, we only remark that these difficulties seem to make their techniques useless for supervised learning with random design. We propose a remedy to solve all these issues in the next section.

IV. Main Results

First, we extend some theoretical tools given by Chatterjee and Barron to supervised learning. Using them, we show that the lasso estimator is a BC-proper MDL estimator.

A. Extension of BC Theory to Supervised Learning

In order to extend the MDL world (the set of BC-proper MDL estimator) to supervised learning, the idea of codelength validity and risk validity seems to be basically promising. On one hand, the definition of codelength validity (Definition 3) is applicable to the random design setting almost as itself, which enables us to interpret $\hat{\theta}$ as an MDL estimator. We will see its exact definition later in this section. On the other hand, the straightforward extension of the risk validity to supervised learning is useless as explained in Section III-B. Thus we need some essential modifications. As was seen in Section III-B, the difficulties basically stems from the fact x^n can freely move in $\mathscr{X}^n$, which is unbounded in general. Our idea can be roughly summarized as follows. A key concept is a so-called typical set of x^n, which is determined by the true distribution $q_*(x^n)$. We employ a definition of typical sets which is different from the usual one, but the following property remains. Though the typical set is significantly smaller than $\mathscr{X}^n$, its probability is close to one. Further, their statistics (average and empirical covariance matrix) are close to their expectation. We modify the definition of the risk validity such that the inequality (32) is required to hold not for all x^n but only for x^n in the typical set. By restricting the range of x^n to the typical set, we can find a weighted ℓ_1 norm that bounds $H'(\theta,x^n)$ from above.

Let us go into details. We postulate that a probability distribution of stochastic process $x_1,x_2,\cdots$, is a member of a certain class $\mathscr{P}_x$. Furthermore, we define $\mathscr{P}^n_x$ by the set of marginal distributions of the first n elements $x_1,x_2,\cdots,x_n$. For each $q\in\mathscr{P}_x$,

$$\sum_{x_{n+1}} q(x^{n+1})=q(x^n)$$

holds for each n. We assume that we can define a collection of typical set $\{A^n_\epsilon\subset\mathscr{X}^n|\epsilon>0,n=1,2,\cdots\}$ for each $q_*\in\mathscr{P}^n_x$, i.e., $\Pr(x^n\in A^n_\epsilon)\to 1$ as $n\to\infty$ for any $\epsilon>0$ and $A^n_\epsilon\subset A^n_{\epsilon'}$ for any real positive numbers ϵ,ϵ' such that $\epsilon<\epsilon'$. This is possible if q_* is stationary and ergodic for example. See [19] for detail. To be concise, $\Pr(x^n\in A^n_\epsilon)$ is written as P^n_ϵ hereafter. We modify the risk validity by using the typical set.

Definition 7 (Restricted Risk Validity): Let β,λ,ϵ be real numbers such that $\beta,\epsilon\in(0,1),\lambda\in(0,1-\beta]$. We say that $L(\theta|x^n)$ is restricted risk valid for $(\lambda,\beta,\epsilon,\mathscr{P}^n_x,A^n_\epsilon)$ if for any $q_*\in\mathscr{P}^n_x$, there exist a quantized subset $\widetilde{\Theta}(q_*)\subset\Theta$ and a model description length $\tilde{L}(\tilde{\theta}|q_*)$ satisfying β-stronger Kraft's inequality such that $\widetilde{\Theta}(q_*)$ and $\tilde{L}(\tilde{\theta}|q_*)$ satisfy Condition 3 and

$$\begin{aligned}&\forall x^n\in A^n_\epsilon,\ \forall y^n\in\mathscr{Y}^n,\\&\max_{\theta\in\Theta}\Big\{d^n_\lambda(p_*,p_\theta)-\log\frac{p_*(y^n|x^n)}{p_\theta(y^n|x^n)}-L(\theta|x^n)\Big\}\\&\le\max_{\tilde{\theta}\in\widetilde{\Theta}(q_*)}\Big\{d^n_\lambda(p_*,p_{\tilde{\theta}})-\log\frac{p_*(y^n|x^n)}{p_{\tilde{\theta}}(y^n|x^n)}-\tilde{L}(\tilde{\theta}|q_*)\Big\}.\end{aligned}$$

Note that both $\widetilde{\Theta}$ and $\tilde{L}$ can depend on the unknown distribution $q_*(x^n)$. This is not problematic because the final penalty L does not depend on the unknown $q_*(x^n)$. A difference from (32) is the restriction of the range of x^n onto the typical set. From here to the next section, we will see how this small change solves the problems described in the previous section. First, we show what can be proved for restricted risk valid penalties.

Theorem 3 (Risk Bound): Let β,ϵ be arbitrary real numbers in $(0,1)$. Define E^n_ϵ as a conditional expectation with regard to $\bar{p}_*(x^n,y^n)$ given that $x^n\in A^n_\epsilon$. For any $\lambda\in(0,1-\beta]$, if $L(\theta|x^n)$ is restricted risk valid for $(\lambda,\beta,\epsilon,\mathscr{P}^n_x,A^n_\epsilon)$,

$$E^n_\epsilon d^n_\lambda(p_*,p_{\hat{\theta}})\le E^n_\epsilon\log\frac{p_*(y^n|x^n)}{p_2(y^n|x^n)}+\frac{1}{\beta}\log\frac{1}{P^n_\epsilon}.\tag{35}$$

In addition, if $L(\theta|x^n)$ is codelength valid,

$$\begin{aligned}E^n_\epsilon d^n_\lambda(p_*,p_{\hat{\theta}})&\le E^n_\epsilon\log\frac{p_*(y^n|x^n)}{p_2(y^n|x^n)}+\frac{1}{\beta}\log\frac{1}{P^n_\epsilon}\\&\le\frac{1}{P^n_\epsilon}E_{\bar{p}_*}\log\frac{p_*(y^n|x^n)}{p_2(y^n|x^n)}+\frac{1}{\beta}\log\frac{1}{P^n_\epsilon}.\end{aligned}\tag{36}$$

Theorem 4 (Regret Bound): Let β,ϵ be arbitrary real numbers in $(0,1)$. For any $\lambda\in(0,1-\beta]$, if $L(\theta|x^n)$ is restricted

risk valid for $(\lambda, \beta, \epsilon, \mathscr{P}_x^n, A_\epsilon^n)$,

$$\Pr\Big(\frac{d_\lambda^n(p_*, p_{\hat{\theta}})}{n} - \frac{1}{n}\log\frac{p_*(y^n|x^n)}{p_2(y^n|x^n)} \geq \tau\Big) \leq \exp(-n\tau\beta) + 1 - P_\epsilon^n. \tag{37}$$

A proof of Theorem 3 is described in Section VI-A, while a proof of Theorem 4 is described in Section VI-B. Note that both bounds become tightest when $\lambda = 1 - \beta$ because the Rényi divergence $d_\lambda^n(p, r)$ is monotonically increasing in terms of λ (see [23] for example). We call the quantity $-\log(1/p_2(y^n|x^n)) - (-\log(1/p_*(y^n|x^n)))$ in Theorem 4 'regret' of the two-stage code p_2 on the given data (x^n, y^n) in this paper, though the ordinary regret is defined as the codelength difference from $\log(1/p_{\hat{\theta}_{\mathrm{mle}}}(y^n|x^n))$, where $\hat{\theta}_{\mathrm{mle}}$ denotes the maximum likelihood estimator. Compared to the usual BC theory, there are additional terms, i.e., $(1/\beta)\log(1/P_\epsilon^n)$ in (35) and (36) or $1 - P_\epsilon^n$ in (37). Due to the property of the typical set, these terms decrease to zero as $n \to \infty$. Therefore, the first term of the right side of (35)-(37) is the main term, which has a form of redundancy/regret of two-stage codes like the original BC theory. In order to interpret the main term exactly as redundancy/regret, we need to prove that $-\log p_2(y^n|x^n)$ satisfies Kraft's inequality. The condition for it is 'codelength validity' as mentioned before. We can use its definition of fixed design case almost as itself. The exact definition of codelength validity in random design cases is given as follows.

Definition 8 (Codelength Validity): We say that $L(\theta|x^n)$ is codelength valid if there exist a quantized subset $\tilde{\Theta}(x^n) \subset \Theta$ and a model description length $\tilde{L}(\tilde{\theta}|x^n)$ satisfying Kraft's inequality such that

$$\forall y^n \in \mathscr{Y}^n, \quad \min_{\theta\in\Theta}\Big\{-\log p_\theta(y^n|x^n) + L(\theta|x^n)\Big\} \geq \min_{\tilde{\theta}\in\tilde{\Theta}(x^n)}\Big\{-\log p_{\tilde{\theta}}(y^n|x^n) + \tilde{L}(\tilde{\theta}|x^n)\Big\} \tag{38}$$

for each x^n.

We note that both the quantized parameter space and the model description length on it can depend on x^n in contrast to the definition of restricted risk validity. This is because x^n can be fixed in order to justify the redundancy/regret interpretation. Let us see that $L_2(y^n|x^n) = -\log p_2(y^n|x^n)$ can be exactly interpreted as a codelength if $L(\theta|x^n)$ is codelength valid. First, we assume that $\mathscr{Y}$, the range of y, is discrete. By (4) and (22), the codelength validity means that

$$L_2(y^n|x^n) \geq \tilde{L}_2(y^n|x^n).$$

As shown by (5), the codelength $\tilde{L}_2(y^n|x^n)$ satisfies Kraft's inequality. Since $L_2(y^n|x^n)$ is equal to or larger than $\tilde{L}_2(y^n|x^n)$ for any x^n, y^n, $L_2(y^n|x^n)$ must satisfy Kraft's inequality. Recall that $-\log p_2(y^n|x^n)$ can be exactly interpreted as a codelength of a prefix code. Next, we consider the case where $\mathscr{Y}$ is a continuous space. The Kraft's inequality trivially holds by replacing the sum with respect to y^n with an integral. Thus, $p_2(y^n|x^n)$ is guaranteed to be a sub-probability density function. Needless to say, $-\log p_2(y^n|x^n)$ cannot be interpreted as a codelength as itself in continuous cases. As is well known, however, difference $(-\log p_2(y^n|x^n)) - (-\log p_*(y^n|x^n))$ can be exactly interpreted as difference of codelength by way of quantization. See Section III of [2] for details. This indicates that both the redundancy interpretation of the fist term of (35) and the regret interpretation of the (negative) second term in the left side of the inequality in the first line of (37) are justified by the codelength validity. Note that the restricted risk validity does not imply the codelength validity and vice versa in general.

We discuss about the conditional expectation in the risk bounds (35) or (36). This conditional expectation seems to be hard to be replaced with the usual (unconditional) expectation. The main difficulty arises from the unboundedness of the loss function. Indeed, we can immediately show a similar risk bound with unconditional expectation for bounded loss functions. As an example, let us consider a class of divergence, called α-divergence [18]

$$\mathscr{D}_\alpha^n(p, r) := \frac{4}{1-\alpha^2}\int\left(1 - \left(\frac{r(y^n|x^n)}{p(y^n|x^n)}\right)^{\frac{1+\alpha}{2}}\right) q_*(x^n)p(y^n|x^n)dx^n dy^n. \tag{39}$$

The α-divergence approaches KL divergence as $\alpha \to \pm 1$ [1]. More exactly,

$$\lim_{\alpha\to-1}\mathscr{D}_\alpha^n(p, r) = \mathscr{D}^n(p, r), \quad \lim_{\alpha\to 1}\mathscr{D}_\alpha^n(p, r) = \mathscr{D}^n(r, p). \tag{40}$$

Furthermore, α-divergence with $\alpha = 0$ is four times the squared Hellinger distance

$$d_H^{2,n}(p, r) := \int\Big(\sqrt{p(y^n|x^n)} - \sqrt{r(y^n|x^n)}\Big)^2 q_*(x^n)p(y^n|x^n)dx^n dy^n, \tag{41}$$

which has been studied and used in statistics for a long time. We focus here on the following two properties of α-divergence:

(i) The α-divergence is bounded:

$$\mathscr{D}_\alpha^n(p, r) \in [0, 4/(1-\alpha^2)] \tag{42}$$

for any p, r and $\alpha \in (-1, 1)$.

(ii) The α-divergence is bounded by the Rényi divergence as

$$d_{(1-\alpha)/2}^n(p, r) \geq \frac{1-\alpha}{2}\mathscr{D}_\alpha^n(p, r) \tag{43}$$

for any p, r and $\alpha \in (-1, 1)$. See [32] for its proof.

As a corollary of Theorem 3, we obtain the following risk bound.

Corollary 5: Let β, ϵ be arbitrary real numbers in $(0, 1)$. Define a function $\lambda(t) := (1-t)/2$. For any $\alpha \in [2\beta - 1, 1)$, if $L(\theta|x^n)$ is restricted risk valid for $(\lambda(\alpha), \beta, \epsilon, \mathscr{P}_x^n, A_\epsilon^n)$ and $p_2(y^n|x^n)$ is a sub-probability distribution,

$$E_{\bar{p}_*}[\mathscr{D}_\alpha^n(p_*, p_{\hat{\theta}})] \leq \frac{1}{\lambda(\alpha)}E_{\bar{p}_*}\left[\log\frac{p_*(y^n|x^n)}{p_2(y^n|x^n)}\right] + \frac{P_\epsilon^n}{\lambda(\alpha)\beta}\log\frac{1}{P_\epsilon^n} + \frac{(1-P_\epsilon^n)}{\lambda(\alpha)(\lambda(\alpha)+\alpha)},$$

In particular, taking $\beta = (\alpha+1)/2$ yields the tightest bound

$$E_{\bar{p}_*}[\mathscr{D}^n_\alpha(p_*, p_{\hat{\theta}})] \le \frac{1}{\lambda(\alpha)} E_{\bar{p}_*}\left[\log \frac{p_*(y^n|x^n)}{p_2(y^n|x^n)}\right] + \frac{P^n_\epsilon}{\lambda(\alpha)(\lambda(\alpha)+\alpha)} \log \frac{1}{P^n_\epsilon} + \frac{(1-P^n_\epsilon)}{\lambda(\alpha)(\lambda(\alpha)+\alpha)}. \quad (44)$$

Its proof will be described in Section VI-C. Though it is not so obvious when the condition "$p_2(y^n|x^n)$ is a sub-probability distribution" is satisfied, we remark that the code-length validity of $L(\theta|x^n)$ is its simple sufficient condition. The second and the third terms of the right side vanish as $n \to \infty$ due to the property of the typical set. The boundedness of loss function is indispensable for the proof. On the other hand, it seems to be impossible to bound the risk for unbounded loss functions. Our remedy for this issue is the risk evaluation based on the conditional expectation on the typical set. Because x^n lies out of A^n_ϵ 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 4 offers a more satisfactory bound.

We remark the relationship of our result with KL divergence $\mathscr{D}^n(p, r)$. Because of (26) or (40), it seems to be possible to obtain a risk bound with KL divergence. However, it is impossible because taking $\lambda \to 1$ in (35) or $\alpha \to \pm 1$ in (44) makes the bounds diverge to the infinity. That is, we cannot derive a risk bound for the risk with KL divergence by BC theory, though we can do it for the Rényi divergence and the α-divergence. It sounds somewhat strange because KL divergence seems to be related the most to the notion of the MDL principle because it has a clear information theoretical interpretation. This issue originates from the original BC theory and has been cast as an open problem for a long time. We remark one possibility based on recent progress. In [25], some developments were made for unbounded losses. If we apply their idea to our problem for lasso, we may have a risk bound for the KL divergence, which is an interesting future work.

Finally, we remark that the effectiveness of our proposal in real situations depends on whether we can show the risk validity of the target penalty and derive a sufficiently small bound for $\log(1/P^n_\epsilon)$ and $1 - P^n_\epsilon$. Actually, much effort is required to realize them for lasso.

B. Application to Lasso in Random Design Setting

By using the tools in the previous section, we will show that lasso is a BC-proper MDL estimator. In the setting of lasso, training data $\{(x_i, y_i) \in \Re^m \times \Re | i = 1, 2, \cdots, n\}$ obey a usual regression model $y_i = x_i^T \theta^* + \upsilon_i$ for $i = 1, 2, \cdots, n$, where $\theta^* \in \Theta = \Re^m$ is a true parameter and $\{\upsilon_i\}$ is i.i.d. Gaussian noise having zero mean and a known variance σ^2. That is, our parametric model is written as

$$p_\theta(y^n|x^n) = \Pi^n_{i=1} N(y_i|x_i^T\theta, \sigma^2)$$

where $N(x|\mu, \Sigma)$ is the density of Gaussian distribution with a mean vector μ and a covariance matrix Σ. By introducing $Y := (y_1, y_2, \cdots, y^n)^T$, $\Upsilon := (\upsilon_1, \upsilon_2, \cdots, \upsilon_n)^T$ and an $n \times m$ matrix $X := [x_1\ x_2\ \cdots x_n]^T$, we have a vector/matrix expression of the regression model $Y = X\theta^* + \Upsilon$. In the original lasso setting [34], the penalty function is defined by the usual ℓ_1 norm

$$L(\theta|x^n) = \mu_1 \|\theta\|_1 + \mu_2,$$

where μ_1, μ_2 are positive real numbers. However we employ the penalty function with the weighted ℓ_1 norm

$$L(\theta|x^n) = \mu_1 \|\theta\|_{w,1} + \mu_2$$

due to the reason explained in Section I-A and Section III-B. The constant μ_2 plays no role in estimation of θ^* while it influences risk bounds. The resultant lasso estimator is defined as a penalized likelihood estimator

$$\hat{\theta}(x^n, y^n) := \arg\min_{\theta\in\Theta} \{-\log p_\theta(y^n|x^n) + L(\theta|x^n)\} = \arg\min_{\theta\in\Theta} \left\{ \frac{1}{2n\sigma^2} \|Y - X\theta\|_2^2 + \mu_1 \|\theta\|_{w,1} + \mu_2 \right\}. \quad (45)$$

Note that the dimension p of parameter θ can be greater than n. When x^n is Gaussian with zero mean, we can derive a risk valid weighted ℓ_1 penalty by choosing an appropriate typical set.

Lemma 3: Let $\mathscr{S}(m)$ be a set of all positive definite symmetric $m \times m$ matrices. For any $\epsilon \in (0, 1)$, define

$$\mathscr{P}^n_x := \{q(x^n) = \Pi^n_{i=1} N(x_i|\mathbf{0}, \Sigma) | \Sigma \in \mathscr{S}(m)\},$$
$$A^n_\epsilon := \left\{ x^n \,\middle|\, \forall j, 1-\epsilon \le \frac{(1/n)\sum_{i=1}^n x_{ij}^2}{\Sigma_{jj}} \le 1+\epsilon \right\}. \quad (46)$$

Here, Σ_{jj} denotes the jth diagonal element of Σ and x_{ij} denotes the jth element of x_i. Assume a linear regression setting:

$$p_*(y^n|x^n) = \Pi^n_{i=1} N(y_i|x_i^T\theta^*, \sigma^2),$$
$$p_\theta(y^n|x^n) = \Pi^n_{i=1} N(y_i|x_i^T\theta, \sigma^2).$$

Let β be a real number in $(0, 1)$ and λ be a real number in $(0, 1-\beta]$. The weighted ℓ_1 norm $L_1(\theta|x^n) = \mu_1\|\theta\|_{w,1} + \mu_2$ is restricted risk valid for $(\lambda, \beta, \epsilon, \mathscr{P}^n_x, A^n_\epsilon)$ if

$$\mu_1 \ge \sqrt{\frac{n \log 4m}{\beta\sigma^2(1-\epsilon)} \cdot \frac{\lambda + 8\sqrt{1-\epsilon^2}}{4}}, \quad \mu_2 \ge \frac{\log 2}{\beta}. \quad (47)$$

Furthermore, if the variance of each x_j is additionally assumed to be bounded from above such as

$$\forall j,\ \Sigma_{jj} \le M,$$

then the ordinary ℓ_1 norm $L_1(\theta|x^n) = \mu_1\|\theta\|_1 + \mu_2$ is restricted risk valid for $(\lambda, \beta, \epsilon, \mathscr{P}^n_x, A^n_\epsilon)$ if

$$\mu_1 \ge M\sqrt{\frac{n \log 4m}{\beta\sigma^2} \cdot \frac{\lambda + 8(1+\epsilon)}{4}}, \quad \mu_2 \ge \frac{\log 2}{\beta}. \quad (48)$$

We describe its proof in Section VI-F. The derivation is much more complicated and requires more techniques, compared to the fixed design setting in [17]. This is because the Rényi divergence is a usual mean square error (MSE) in

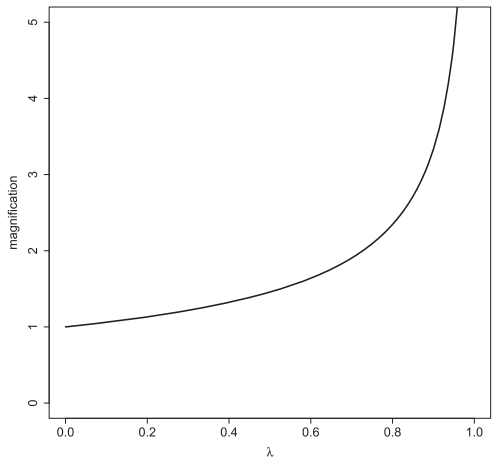


Fig. 1. Plot of $\sqrt{(\lambda+8)/8(1-\lambda)}$ against λ.

the fixed design setting, while it is not in the random design setting in general. In addition, it is important for the risk bound derivation to choose an appropriate typical set in a sense that we can show that P_ϵ^n approaches to one sufficiently fast and we can also show the restricted risk validity of the target penalty with the chosen typical set. In case of lasso with normal design, the typical set A_ϵ^n defined in (46) satisfies such properties.

Let us compare the coefficient of the risk valid weighted ℓ_1 penalty with the fixed design setting in [17]. They showed that the weighted ℓ_1 norm satisfying

$$\mu_1 \geq \sqrt{\frac{2n\log 4m}{\sigma^2}}, \quad \mu_2 \geq \frac{\log 2}{\beta} \tag{49}$$

is risk valid in the fixed design setting. The condition for μ_2 is the same, while the condition for μ_1 in (47) is more strict than that of the fixed design setting. We compare them by taking $\beta = 1-\lambda$ (the tightest choice) and $\epsilon = 0$ in (47) because ϵ can be negligibly small for sufficiently large n. The minimum μ_1 for the risk validity in the random design setting is

$$\sqrt{\frac{\lambda+8}{8(1-\lambda)}}$$

times that for the fixed design setting. Hence, the smallest value of regularization coefficient μ_1 for which the risk bound holds in the random design is always larger than that of the fixed design setting for any $\lambda \in (0,1)$ but its extent is not so large unless λ is extremely close to 1 (See Fig. 1).

Next, we show that P_ϵ^n exponentially approaches to one as n increases.

Lemma 4 (Exponential Bound of Typical Set): Suppose that $x_i \sim N(x_i|0,\Sigma)$ independently. For any $\epsilon \in (0,1)$,

$$\begin{aligned} P_\epsilon^n &\geq \left(1-2\exp\left(-\frac{n}{2}(\epsilon-\log(1+\epsilon))\right)\right)^m \\ &\geq 1-2m\exp\left(-\frac{n}{2}(\epsilon-\log(1+\epsilon))\right) \\ &\geq 1-2m\exp\left(-\frac{n\epsilon^2}{7}\right). \end{aligned} \tag{50}$$

See Section VI-H for its proof. In the lasso case, it is often postulated that p is much greater than n. Due to Lemma 4, $1-P_\epsilon^n$ is $O(m\cdot\exp(-n\epsilon^2/7))$, which also implies that the second term in (35) can be negligibly small even if $n \ll m$. In this sense, the exponential bound is important for lasso.

Another remaining task is to show the codelength validity of the weighted ℓ_1 penalty in order to interpret lasso as an MDL estimator. Interestingly, the weighted ℓ_1 penalties derived in Lemma 3 are not only restricted risk valid but also codelength valid.

Lemma 5: Assume a linear regression setting:

$$\begin{aligned} p_*(y^n|x^n) &= \Pi_{i=1}^n N(y_i|x_i^T\theta^*,\sigma^2), \\ p_\theta(y^n|x^n) &= \Pi_{i=1}^n N(y_i|x_i^T\theta,\sigma^2). \end{aligned}$$

If μ_1 and μ_2 satisfy (47), then the weighted ℓ_1 norm $L(\theta|x^n) = \mu_1\|\theta\|_{w,1} + \mu_2$ is codelength valid.

Its proof will be described in Section VI-I. On the contrary, the ordinary ℓ_1 penalty satisfying (48) cannot be shown to satisfy (38) for every x^n. By this fact, lasso with column normalization can be interpreted as an MDL estimator because it is obtained by minimizing the codelength. That is, Condition 1 is satisfied for lasso with column normalization. It also indicates that we can obtain the unconditional risk bound with respect to α-divergence for those weighted ℓ_1 penalties by Corollary 5 without any additional condition. Combining Lemmas 3, 4 and 5 with Theorems 3 and 4, our main result about lasso with 'column normalization' is summarized in the following theorem.

Theorem 6: Let $\mathscr{S}(m)$ be a set of all positive definite symmetric $m\times m$ matrices. For any $\epsilon \in (0,1)$, define

$$\begin{aligned} \mathscr{P}_x^n &:= \{q(x^n) = \Pi_{i=1}^n N(x_i|\mathbf{0},\Sigma) | \Sigma \in \mathscr{S}(m)\}, \\ A_\epsilon^n &:= \left\{x^n \,\middle|\, \forall j, 1-\epsilon \leq \frac{(1/n)\sum_{i=1}^n x_{ij}^2}{\Sigma_{jj}} \leq 1+\epsilon\right\}. \end{aligned}$$

Assume a linear regression setting:

$$\begin{aligned} p_*(y^n|x^n) &= \Pi_{i=1}^n N(y_i|x_i^T\theta^*,\sigma^2), \\ p_\theta(y^n|x^n) &= \Pi_{i=1}^n N(y_i|x_i^T\theta,\sigma^2). \end{aligned}$$

Let β be a real number in $(0,1)$. For any $\lambda \in (0, 1-\beta]$, if

$$\mu_1 \geq \sqrt{\frac{\log 4m}{n\beta\sigma^2(1-\epsilon)}\cdot\frac{\lambda+8\sqrt{1-\epsilon^2}}{4}}, \quad \mu_2 \geq \frac{\log 2}{n\beta}, \tag{51}$$

the weighted ℓ_1 penalty $L(\theta|x^n) = \mu_1\|\theta\|_{w,1} + \mu_2$ is both restricted risk valid for $(\lambda,\beta,\epsilon,\mathscr{P}_x^n,A_\epsilon^n)$ and codelength valid. As a result, the lasso estimator $\hat{\theta}(x^n,y^n)$ in (45) has a risk bound

$$\begin{aligned} &E_\epsilon^n[d_\lambda(p_*, p_{\hat{\theta}(x^n,y^n)})] \\ &\leq E_\epsilon^n\left[\inf_{\theta\in\Theta}\left\{\frac{(\|Y-X\theta\|_2^2 - \|Y-X\theta^*\|_2^2)}{2n\sigma^2} + \mu_1\|\theta\|_{w,1} + \mu_2\right\}\right] \\ &\quad - \frac{m\log\left(1-2\exp\left(-\frac{n}{2}(\epsilon-\log(1+\epsilon))\right)\right)}{n\beta}, \\ &\leq \frac{1}{1-\rho_n} \end{aligned} \tag{52}$$

$$\cdot E_{\bar{p}_*}\left[\inf_{\theta\in\Theta}\left\{\frac{\left(\|Y-X\theta\|_2^2-\|Y-X\theta^*\|_2^2\right)}{2n\sigma^2}+\mu_1\|\theta\|_{w,1}+\mu_2\right\}\right] - \frac{m\log\left(1-2\exp\left(-\frac{n}{2}(\epsilon-\log(1+\epsilon))\right)\right)}{n\beta}, \quad (53)$$

$$\rho_n := 2m\exp(-n\epsilon^2/7)$$

and a regret bound

$$d_\lambda(p_*, p_{\hat{\theta}(x^n,y^n)}) \le \inf_{\theta\in\Theta}\left\{\frac{\left(\|Y-X\theta\|_2^2-\|Y-X\theta^*\|_2^2\right)}{2n\sigma^2}+\mu_1\|\theta\|_{w,1}+\mu_2\right\}+\tau \quad (54)$$

with probability at least

$$\left(1-2\exp\left(-\frac{n}{2}(\epsilon-\log(1+\epsilon))\right)\right)^m-\exp(-\tau n\beta), \quad (55)$$

which is bounded from below by

$$1-O\left(m\cdot\exp\left(-n\kappa\right)\right)$$

with $\kappa := \min\{\epsilon^2/7, \tau\beta\}$.

Remark: The main terms of these bounds are approximately less than

$$\mu_1\|\theta^*\|_{w,1}+\mu_2,$$

which replaces $(\bar{m}/2)\log n$ term in usual MDL estimators. Here $\bar{m}$ is ℓ_0 norm of θ^*. Hence, our bounds are related to ℓ_1 norm of θ^* instead of the number of true parameters $\bar{m}$. This is related to the fact that lasso is different from ℓ_0-regularization algorithms.

Since $\bar{p}_*(x,y)$ is i.i.d. now, $d_\lambda^n(p,r) = nd_\lambda(p,r)$. Hence, we presented the risk bound as a single-sample version in (52) by dividing the both sides by n. Theorem 6 clearly proved that lasso with column normalization is a BC-proper MDL estimator if (47) holds. Furthermore, the statistical risk of lasso is bounded from above by the redundancy of the two-stage code associated with lasso. That is, the right side of (53) and (54) can be exactly interpreted as the redundancy and the regret. This also implies that 'lasso is an MDL estimator that attains the minimum description length'. On the other hand, (53) provides the risk bound of the form

$$\text{statistical risk of lasso} \le \text{redundancy of lasso} + \text{negligibly small term}.$$

Hence, the justification of the MDL principle by BC theory was successfully extended to lasso.

We finally remark that what we can know about the lasso without column normalization (the ordinary ℓ_1 norm) when we assume the boundedness of covariates like Bartlett *et al.* [6]. Note that we need the boundedness of Σ_{jj} in order to show the restricted risk validity of the ordinary ℓ_1 norm as seen in Lemma 3 while Bartlett et al. assumes the boundedness of covariates x^n directly. However the boundedness of Σ_{jj} implies the boundedness of x^n through the typical set in our analysis. Hence our boundedness assumption is essentially similar to that of [6]. When each Σ_{jj} is bounded, we can have the same result as Theorem 6 except replacing (51) with (48). The restricted risk valid penalties and the resultant redundancy/regret bounds directly depend on the constant of the boundedness M in a similar way with the result of [6]. This makes the risk bound loose in general because the weight is replaced with its worst value M^2. Furthermore, we cannot show the codelength validity. That is, lasso without column normalization cannot be interpreted as an MDL estimator. As a result, lasso with column normalization is favorable in terms of both BC-properness and the tightness of the risk bound.

TABLE III
NOISE VARIANCE AND REGULARIZATION CONSTANTS ACCORDING TO SNR

SNR	σ^2	minimum μ_1	minimum μ_2
6	2.35	$3.62\cdot10^{-1}$	$6.93\cdot10^{-3}$
2	7.05	$2.09\cdot10^{-1}$	$6.93\cdot10^{-3}$
0.5	28.21	$1.05\cdot10^{-1}$	$6.93\cdot10^{-3}$

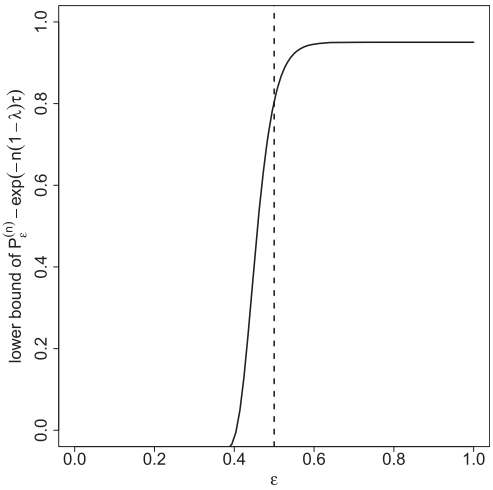
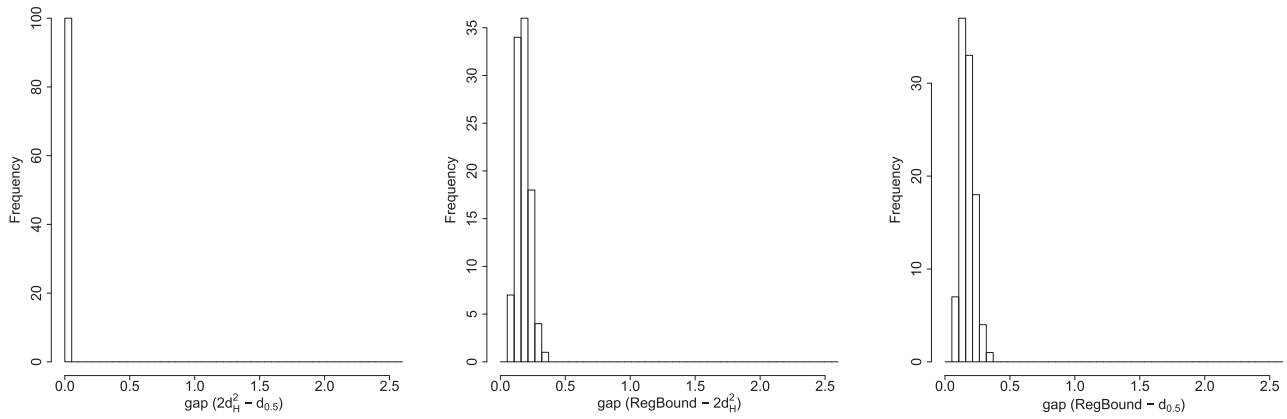


Fig. 2. Plot of (55) against $\epsilon \in (0,1)$ when $n = 200, m = 1000$ and $\tau = 0.03$. The dotted vertical line indicates $\epsilon = 0.5$.

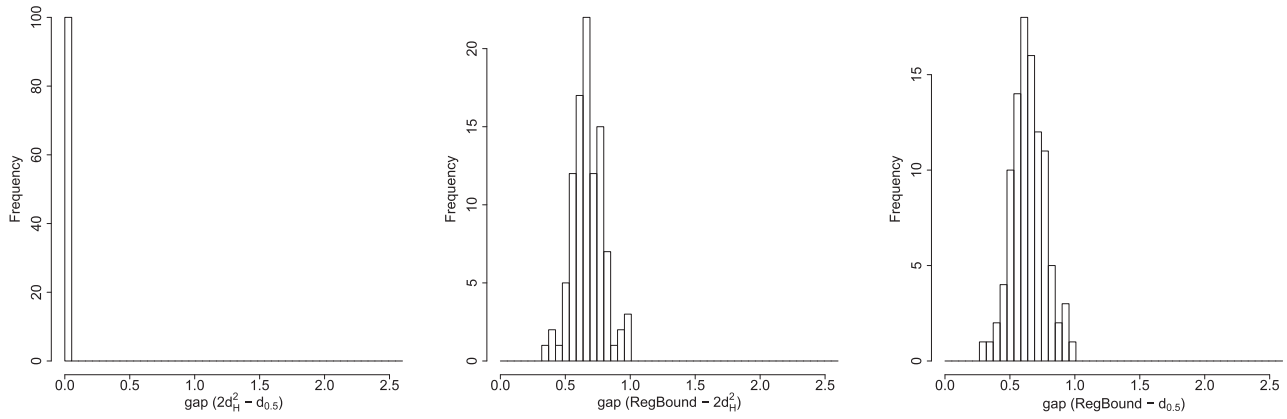
V. NUMERICAL SIMULATIONS

We investigate the behavior of the risk/regret bounds of lasso, i.e., (52) and (54). In both bounds, we let $\beta = 1-\lambda$ in order to make the bounds tightest. Furthermore, we let μ_1 and μ_2 be their smallest values in (47). The Rényi divergence in the left side of (54) does not include the well-known KL divergence (the mean square error in this case) but includes the Bhattacharyya divergence ($d_{0.5}$) in (27). By (43), the Bhattacharyya divergence is an upper bound on two times the squared Hellinger distance $2d_H^2$. The Hellinger distance was defined in (41) as n sample version (i.e., $d_H^2 = d_H^{2,1}$). Since the Hellinger distance is also a common loss function, we investigate the behavior of $d_H^2 = d_H^{2,1}$ and risk/regret bounds (52) and (54) through (43). We set $n = 200$, $m = 1000$ and $\Sigma = I_m$ to mimic a typical situation of sparse learning where I_m is an identity matrix of order m. The true parameter θ^* has only 100 nonzero components that are generated from the Gaussian distribution with mean 0 and variance 1 independently. The lasso estimator is calculated by a proximal gradient method [10]. To make the regret bound tight, we take $\tau = 0.03$ that is close to zero compared to the



(a) gap '$d_{0.5} - 2d_H^2$' (b) gap 'RegBound-$d_{0.5}$' (c) gap 'RegBound-$2d_H^2$'

Fig. 3. Histograms of the gap among $d_{0.5}$ (Bhattacharyya div.), $2d_H^2$ (Hellinger dist.) and RegBound (the RegBound with $\tau = 0.03$) in case that SNR=0.5.



(a) gap '$d_{0.5} - 2d_H^2$' (b) gap 'RegBound-$d_{0.5}$' (c) gap 'RegBound-$2d_H^2$'

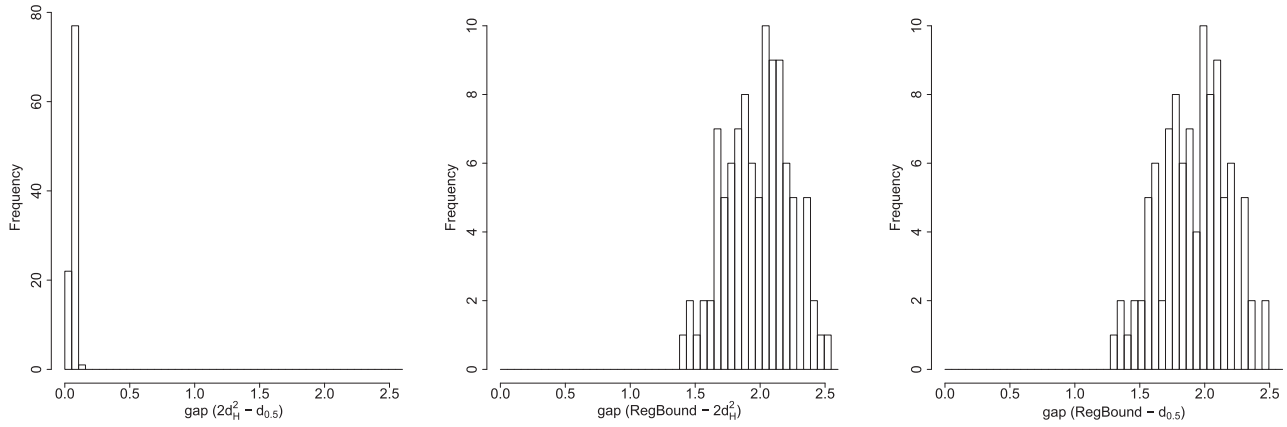
Fig. 4. Histograms of the gap among $d_{0.5}$ (Bhattacharyya div.), $2d_H^2$ (Hellinger dist.) and RegBound (the regret bound with $\tau = 0.03$) in case that SNR=2.

main term (regret). For this τ, Fig. 2 shows the plot of (55) against ϵ. We should choose the smallest ϵ as long as the regret bound holds with large probability. Our choice is $\epsilon = 0.5$ at which the value of (55) is 0.81.

First, we consider the three cases in which the signal-to-noise ratios (SNR) $E_{q_*}[(x^T\theta^*)^2]/\sigma^2$ varies among $6, 2, 0.5$ (i.e., σ^2 varies). For each SNR, the noise variance σ^2 and the minimum value of μ_1 and μ_2 satisfying (47) are summarized in Table III. In the above setting we have the following relationship

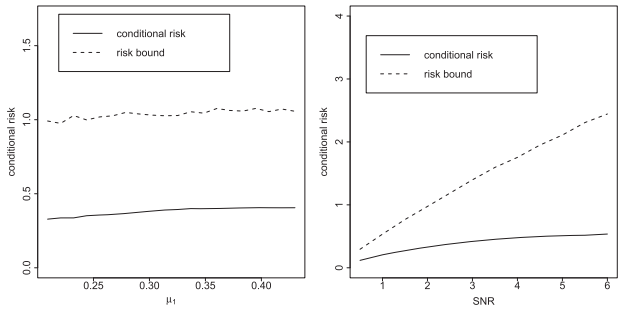
$$2d_H^2(p_*, p_{\hat{\theta}}) \le d_{0.5}(p_*, p_{\hat{\theta}(x^n, y^n)}) \le \inf_{\theta \in \Theta} \left\{ \frac{\left(\|Y - X\theta\|_2^2 - \|Y - X\theta^*\|_2^2\right)}{2n\sigma^2} + \mu_1 \|\theta\|_{w,1} + \mu_2 \right\} + \tau.$$

We sometimes abbreviate it $2d_H^2 \le d_{0.5} \le$RegBound hereafter. We show the results about the regret bound in Figs. 3-5. In each figure, three panels (a)-(c) show histograms of the gaps $d_{0.5} - 2d_H^2$, RegBound-$d_{0.5}$ and RegBound$-2d_H^2$ that were obtained by hundred repetitions. First of all, we remark that the regret bound dominated the Rényi divergence over all trials, though the regret bound is probabilistic. One of the reason is the looseness of the lower bound (55) of the probability for the regret bound to hold. This suggests that ϵ can be reduced more in order to derive a tighter bound. The panel (a) in the three figures show that the gap between $2d_H^2$ and $d_{0.5}$ is negligible relative to the gap from the regret bound. In contrast magnitude of the gap 'RegBound-$d_{0.5}$' depends on SN ratio. As SN ratio got larger, the gap had larger mean and larger variance. To make sure this aptitude, we investigate the behavior of the conditional risk and risk bounds in (52) against SN ratio. The expectation in the conditional risk and the risk bound was calculated by taking conditional mean in hundred repetitions. The panel (b) of Fig. 6 shows the result. At first glance it seems to be strange that the value of loss function got larger as SN ratio got larger. The reason is that our measure of goodness is not the error of parameter estimation but the Rényi information (discrepancy of distribution). Clearly the gap between the conditional risk and its risk bound gets larger as SNR gets larger. When SN ratio gets small, the penalty term gets large relative to the likelihood term since the likelihood term is of order $O(1/\sigma^2)$ while μ_1 is $O(1/\sqrt{\sigma^2})$. As a result, the lasso estimator $\hat{\theta}$ is shrunk to the zero vector to minimize the penalty term. Hence the penalty term $\mu_1\|\hat{\theta}\|_{w,1}$ is also close to zero. The resultant regret bound behaves like a constant and gets close to the loss. However since the value of the loss function also behaves like a constant because $\hat{\theta}$ is close to the constant



(a) gap '$d_{0.5} - 2d_H^2$' (b) gap 'RegBound-$d_{0.5}$' (c) gap 'RegBound-$2d_H^2$'

Fig. 5. Histograms of the gap among $d_{0.5}$ (Bhattacharyya div.), $2d_H^2$ (Hellinger dist.) and RegBound (the regret bound with $\tau = 0.03$) in case that SNR=6.



(a) Plot of the risk/risk bound against μ_1 when SNR is 2. (b) Plot of the risk/risk bound against SNR when μ_1 is the minimum value as shown in (51).

Fig. 6. Plot of the conditional risk and its bound against μ_1 and SNR.

(zero) vector, the regret bound can get close to the value of the loss function safely (no risk of breaking the bound).

Finally we investigate how sensitively the gap between the conditional risk and the risk bound (the mean of regret bounds) depends on the choice of the regularization parameter μ_1. Panel (a) of Fig. 6 shows their gap against μ_1. The value of μ_1 moved within the range that does not violate the condition (51). Though the conditional risk and its risk bound got larger slowly as μ_1 got larger, the gap was not sensitive to the choice of μ_1.

VI. Proofs of Theorems, Lemmas and Corollary

We give all proofs to the theorems, the lemmas and the corollary in the main results.

A. Proof of Theorem 3

Here we prove our main theorem. The proof proceeds along with the same line as [17] though some modifications are necessary.

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 restricted risk validity, we obtain

$$\begin{aligned}
&E_\epsilon^n\left[\exp\left(\beta \max_{\theta\in\Theta}\left\{F_\lambda^\theta(x^n,y^n) - L(\theta|x^n)\right\}\right)\right] \\
&\le E_\epsilon^n\left[\exp\left(\beta \max_{\tilde\theta\in\widetilde\Theta}\left\{F_\lambda^{\tilde\theta}(x^n,y^n) - \tilde L(\tilde\theta|q_*)\right\}\right)\right] \\
&\le \sum_{\tilde\theta\in\widetilde\Theta(q_*)} E_\epsilon^n\left[\exp\left(\beta\Big(F_\lambda^{\tilde\theta}(x^n,y^n) - \tilde L(\tilde\theta|q_*)\Big)\right)\right] \\
&= \sum_{\tilde\theta\in\widetilde\Theta(q_*)} \exp(-\beta\tilde L(\tilde\theta|q_*)) E_\epsilon^n\left[\exp\left(\beta F_\lambda^{\tilde\theta}(x^n,y^n)\right)\right]. \quad (56)
\end{aligned}$$

The following fact is an extension of the key technique of BC theory:

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

The first inequality holds because $E_{\bar p_*(x^n,y^n)}[A] \ge P_\epsilon^n E_\epsilon^n[A]$ for any non-negative random variable A. The second inequality holds because of the monotonically increasing property of $d_\lambda^n(p_*,p_\theta)$ in terms of λ. Thus, the right side of (56) is bounded as

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

Hence we have an important inequality

$$\frac{1}{P_\epsilon^n} \ge E_\epsilon^n\left[\exp\left(\beta\max_{\theta\in\Theta}\left\{F_\lambda^\theta(x^n,y^n) - L(\theta|x^n)\right\}\right)\right]. \quad (57)$$

Applying Jensen's inequality to (57), we have

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

Thus we have

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

Rearranging the terms of this inequality, we have (35). Furthermore, if $L(\theta|x^n)$ is codelength valid, $p_2(y^n|x^n)$ is a sub-probability distribution. In this case, the right side of (35) is bounded from above by the ordinary (unconditional) redundancy as follows. Let further $I_A(x^n)$ be an indicator function of a set $A \subset \mathscr{X}^n$. By the decomposition of expectation, we have

$$\begin{aligned}
&E_\epsilon^n\left[\log\frac{p_*(y^n|x^n)}{p_2(y^n|x^n)}\right] \\
&= \frac{1}{P_\epsilon^n} E_{\bar p_*(x^n,y^n)}\left[I_{A_\epsilon^n}(x^n)\log\frac{p_*(y^n|x^n)}{p_2(y^n|x^n)}\right] \\
&= \frac{1}{P_\epsilon^n} E_{q_*(x^n)}\left[I_{A_\epsilon^n}(x^n) E_{p_*(y^n|x^n)}\left[\log\frac{p_*(y^n|x^n)}{p_2(y^n|x^n)}\right]\right] \\
&\leq \frac{1}{P_\epsilon^n} E_{q_*(x^n)}\left[E_{p_*(y^n|x^n)}\left[\log\frac{p_*(y^n|x^n)}{p_2(y^n|x^n)}\right]\right].
\end{aligned}$$

Since the conditional expectation part is non-negative, removing the indicator function $I_{A_\epsilon^n}(x^n)$ cannot decrease this quantity, which gives the last inequality.

B. Proof of Theorem 4

It is not necessary to start from scratch. We reuse the proof of Theorem 3.

Proof: We can start from (57). For convenience, we define

$$\begin{aligned}
&\xi(x^n,y^n) \\
&= \frac{1}{n}\max_{\theta\in\Theta}\left\{F_\lambda^\theta(x^n,y^n) - L(\theta|x^n)\right\} \\
&= \max_{\theta\in\Theta}\left\{\frac{d_\lambda^n(p_*,p_\theta)}{n} - \frac{1}{n}\log\frac{p_*(y^n|x^n)}{p_\theta(y^n|x^n)} - \frac{L(\theta|x^n)}{n}\right\}.
\end{aligned}$$

By Markov's inequality and (57),

$$\begin{aligned}
&\Pr\left(\xi(x^n,y^n) \geq \tau | x^n \in A_\epsilon^n\right) \\
&= \Pr\left(\exp\left(n\beta\xi(x^n,y^n)\right) \geq \exp(n\beta\tau) | x^n \in A_\epsilon^n\right) \\
&\leq \frac{\exp(-n\tau\beta)}{P_\epsilon^n}.
\end{aligned}$$

Hence, we obtain

$$\begin{aligned}
&\Pr\left(\xi(x^n,y^n) \geq \tau\right) \\
&= P_\epsilon^n \Pr\left(\xi(x^n,y^n) \geq \tau | x^n \in A_\epsilon^n\right) \\
&\quad + (1-P_\epsilon^n)\Pr\left(\xi(x^n,y^n) \geq \tau | x^n \notin A_\epsilon^n\right) \\
&\leq P_\epsilon^n \Pr\left(\xi(x^n,y^n) \geq \tau | x^n \in A_\epsilon^n\right) + (1-P_\epsilon^n) \\
&\leq \exp(-n\tau\beta) + (1-P_\epsilon^n).
\end{aligned}$$

The proof completes by noticing that

$$(1/n)\left(F_\lambda^{\hat\theta}(x^n,y^n) - L(\hat\theta|x^n)\right) \leq \xi(x^n,y^n)$$

for any x^n and y^n.

C. Proof of Corollary 5

The proof is obtained immediately from Theorem 3.

Proof: Let again E_ϵ^n denote a conditional expectation with regard to $\bar p_*(x^n,y^n)$ given that $x^n \in A_\epsilon^n$. The unconditional risk is bounded as

$$\begin{aligned}
&E_{\bar p_*}[\mathscr{D}_\alpha^n(p_*,p_{\hat\theta})] \\
&= E_{\bar p_*}[I_{A_\epsilon^n}(x^n)\mathscr{D}_\alpha^n(p_*,p_{\hat\theta})] \\
&\quad + E_{\bar p_*}[(1-I_{A_\epsilon^n}(x^n))\mathscr{D}_\alpha^n(p_*,p_{\hat\theta})] \\
&\leq P_\epsilon^n E_\epsilon^n[\mathscr{D}_\alpha^n(p_*,p_{\hat\theta})] + (1-P_\epsilon^n)\cdot\frac{4}{1-\alpha^2} \\
&\leq \frac{P_\epsilon^n}{\lambda(\alpha)} E_\epsilon^n[d_{\lambda(\alpha)}^n(p_*,p_{\hat\theta})] + \frac{(1-P_\epsilon^n)}{\lambda(\alpha)(\lambda(\alpha)+\alpha)} \\
&\leq \frac{P_\epsilon^n}{\lambda(\alpha)}\left(\frac{1}{P_\epsilon^n}E_{\bar p_*}\log\frac{p_*(y^n|x^n)}{p_2(y^n|x^n)} + \frac{1}{\beta}\log\frac{1}{P_\epsilon^n}\right) \\
&\quad + \frac{(1-P_\epsilon^n)}{\lambda(\alpha)(\lambda(\alpha)+\alpha)} \\
&= \frac{1}{\lambda(\alpha)}E_{\bar p_*}\left[\log\frac{p_*(y^n|x^n)}{p_2(y^n|x^n)}\right] + \frac{P_\epsilon^n}{\lambda(\alpha)\beta}\log\frac{1}{P_\epsilon^n} \\
&\quad + \frac{(1-P_\epsilon^n)}{\lambda(\alpha)(\lambda(\alpha)+\alpha)}.
\end{aligned}$$

The first and second inequalities follow from the two properties of α-divergence in (42) and (43) respectively. The third inequality follows from Theorem 3 because $\lambda(\alpha) \in (0, 1-\beta)$ by the assumption. The final part of the statement follows from the fact that taking $\lambda = 1-\beta$ makes the bound in (35) tightest because of the monotonically increasing property of Rényi divergence with regard to λ.

As mentioned in the main text, we remark that the sub-probability condition of $p_2(y^n|x^n)$ can be replaced with its sufficient condition "$L(\theta|x^n)$ is codelength valid." In addition, the sub-probability condition can be relaxed to

$$\sup_{x^n\in\mathscr{X}^n}\int p_2(y^n|x^n)dy^n < \infty,$$

under which the bound increases by

$$(1-P_\epsilon^n)\log\sup_{x^n\in\mathscr{X}^n}\int p_2(y^n|x^n)dy^n.$$

D. Rényi Divergence and Its Derivatives

In this section and the next section, we prove a series of lemmas, which will be used to derive restricted risk valid penalties for lasso. First, we show that the Rényi divergence can be understood by defining $\bar p_\theta^\lambda(x,y)$ in Lemma 6. Then, their explicit forms in the lasso setting are calculated in Lemma 7.

Lemma 6: Define a probability distribution $\bar p_\theta^\lambda(x,y)$ by

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

where Z_θ^λ is a normalization constant. Then, the Rényi divergence and its first and second derivatives are written as

$$\begin{aligned} d_\lambda(p_*, p_\theta) &= \frac{-1}{1-\lambda}\log Z_\theta^\lambda, \\ \frac{\partial d_\lambda(p_*, p_\theta)}{\partial \theta} &= -E_{\bar{p}_\theta^\lambda}\left[s_\theta(y|x)\right], \qquad (58) \\ \frac{\partial^2 d_\lambda(p_*, p_\theta)}{\partial\theta\partial\theta^T} &= -E_{\bar{p}_\theta^\lambda}\left[G_\theta(x,y)\right] \\ &\quad -(1-\lambda)\mathrm{Var}_{\bar{p}_\theta^\lambda}\left(s_\theta(y|x)\right), \qquad (59) \end{aligned}$$

where $\mathrm{Var}_p(A)$ denotes a covariance matrix of A with respect to p and

$$s_\theta(y|x) := \frac{\partial \log p_\theta(y|x)}{\partial\theta}, \quad G_\theta(x,y) := \frac{\partial^2 \log p_\theta(y|x)}{\partial\theta\partial\theta^T}.$$

Proof: The normalizing constant is rewritten as

$$\begin{aligned} Z_\theta^\lambda &= \int q_*(x)p_*(y|x)\left(\frac{p_\theta(y|x)}{p_*(y|x)}\right)^{1-\lambda}dxdy \\ &= E_{\bar{p}_*(x,y)}\left[\left(\frac{p_\theta(y|x)}{p_*(y|x)}\right)^{1-\lambda}\right]. \end{aligned}$$

Thus the Rényi divergence is written as

$$d_\lambda(p_*, p_\theta) = -\frac{1}{1-\lambda}\log Z_\theta^\lambda.$$

Next, we calculate the partial derivative of $\log Z_\theta^\lambda$ as

$$\begin{aligned} &\frac{\partial \log Z_\theta^\lambda}{\partial\theta} \\ &= \frac{1}{Z_\theta^\lambda}\frac{\partial Z_\theta^\lambda}{\partial\theta} \\ &= \frac{1}{Z_\theta^\lambda}E_{\bar{p}_*}\left[\left(\frac{p_\theta(y|x)}{p_*(y|x)}\right)^{1-\lambda}\frac{\partial}{\partial\theta}\log\left(\frac{p_\theta(y|x)}{p_*(y|x)}\right)^{1-\lambda}\right] \\ &= \frac{1-\lambda}{Z_\theta^\lambda}E_{\bar{p}_*}\left[\left(\frac{p_\theta(y|x)}{p_*(y|x)}\right)^{1-\lambda}\frac{\partial \log p_\theta(y|x)}{\partial\theta}\right] \\ &= \frac{1-\lambda}{Z_\theta^\lambda}\int q_*(x)p_*(y|x)^\lambda p_\theta(y|x)^{1-\lambda}s_\theta(y|x)dxdy \\ &= (1-\lambda)E_{\bar{p}_\theta^\lambda}[s_\theta(y|x)]. \end{aligned}$$

Therefore, the first derivative is

$$\frac{\partial d_\lambda(p_*, p_\theta)}{\partial\theta} = -\frac{1}{1-\lambda}\frac{\partial \log Z_\theta^\lambda}{\partial\theta} = -E_{\bar{p}_\theta^\lambda}\left[s_\theta(y|x)\right].$$

Furthermore, we have

$$\begin{aligned} \frac{\partial \log \bar{p}_\theta^\lambda(x,y)}{\partial\theta} &= \frac{\partial}{\partial\theta}\log\left(\frac{q_*(x)p_*(y|x)^\lambda p_\theta(y|x)^{1-\lambda}}{Z_\theta^\lambda}\right) \\ &= (1-\lambda)\frac{\partial \log p_\theta(y|x)}{\partial\theta} - \frac{\partial \log Z_\theta^\lambda}{\partial\theta} \\ &= (1-\lambda)s_\theta(y|x) - (1-\lambda)E_{\bar{p}_\theta^\lambda}[s_\theta(y|x)] \\ &= (1-\lambda)\left(s_\theta(y|x) - E_{\bar{p}_\theta^\lambda}[s_\theta(y|x)]\right). \end{aligned}$$

Hence,

$$\begin{aligned} &\frac{\partial^2 d_\lambda(p_*, p_\theta)}{\partial\theta\partial\theta^T} \\ &= -\int s_\theta(y|x)\bar{p}_\theta^\lambda(x,y)\left(\frac{\partial \log \bar{p}_\theta^\lambda(x,y)}{\partial\theta}\right)^T \\ &\quad + \bar{p}_\theta^\lambda(x,y)\frac{\partial s_\theta(y|x)}{\partial\theta^T}dxdy \\ &= -E_{\bar{p}_\theta^\lambda}\Bigg[(1-\lambda)s_\theta(y|x)\left(s_\theta(y|x) - E_{\bar{p}_\theta^\lambda}[s_\theta(y|x)]\right)^T \\ &\quad + \frac{\partial^2 \log p_\theta(y|x)}{\partial\theta\partial\theta^T}\Bigg] \\ &= -E_{\bar{p}_\theta^\lambda}\left[\frac{\partial^2 \log p_\theta(y|x)}{\partial\theta\partial\theta^T}\right] - (1-\lambda) \\ &\cdot E_{\bar{p}_\theta^\lambda}\left[\left(s_\theta(y|x) - E_{\bar{p}_\theta^\lambda}\left[s_\theta(y|x)\right]\right)\left(s_\theta(y|x) - E_{\bar{p}_\theta^\lambda}[s_\theta(y|x)]\right)^T\right] \\ &= -E_{\bar{p}_\theta^\lambda}\left[\frac{\partial^2 \log p_\theta(y|x)}{\partial\theta\partial\theta^T}\right] - (1-\lambda)\mathrm{Var}_{\bar{p}_\theta^\lambda}\left(s_\theta(y|x)\right). \end{aligned}$$

Lemma 7: Let

$$\theta(\lambda) := \lambda\theta^* + (1-\lambda)\theta,\ \bar{\theta} := \theta - \theta^*,\ \bar{\theta}' := \Sigma^{1/2}\bar{\theta},$$
$$c := \frac{\sigma^2}{\lambda(1-\lambda)},\ q_\theta^\lambda(x) := \int \bar{p}_\theta^\lambda(x,y)dy,\ p_\theta^\lambda(y|x) := \frac{\bar{p}_\theta^\lambda(x,y)}{q_\theta^\lambda}.$$

If we assume that $p_*(y|x) = N(y|x^T\theta^*, \sigma^2)$ (i.e., linear regression setting),

$$\begin{aligned} p_\theta^\lambda(y|x) &= N(y|x^T\theta(\lambda), \sigma^2), \\ q_\theta^\lambda(x) &= \frac{q_*(x)\exp\left(-\frac{1}{2c}(x^T\bar{\theta})^2\right)}{Z_\theta^\lambda}, \\ \frac{\partial d_\lambda(p_*, p_\theta)}{\partial\theta} &= \frac{\lambda}{\sigma^2}E_{q_\theta^\lambda}[xx^T]\bar{\theta}, \\ \frac{\partial^2 d_\lambda(p_*, p_\theta)}{\partial\theta\partial\theta^T} &= \frac{\lambda}{\sigma^2}E_{q_\theta^\lambda}[xx^T] - \frac{\lambda}{\sigma^2 c}\mathrm{Var}_{q_\theta^\lambda}\left(xx^T\bar{\theta}\right). \qquad (60) \end{aligned}$$

If we additionally assume that $q_*(x) = N(x|\mathbf{0}, \Sigma)$ with a non-singular covariance matrix Σ,

$$\begin{aligned} q_\theta^\lambda(x) &= N(x|\mathbf{0}, \Sigma_\theta^\lambda), \\ \frac{\partial d_\lambda(p_*, p_\theta)}{\partial\theta} &= \frac{\lambda}{\sigma^2}\left(\frac{c}{c + \|\bar{\theta}'\|_2^2}\right)\Sigma^{1/2}\bar{\theta}', \qquad (61) \\ \frac{\partial^2 d_\lambda(p_*, p_\theta)}{\partial\theta\partial\theta^T} &= \frac{\lambda}{\sigma^2}\left(\frac{c}{c + \|\bar{\theta}'\|_2^2}\right)\Sigma \\ &\quad - \frac{2\lambda}{\sigma^2}\left(\frac{c}{(c + \|\bar{\theta}'\|_2^2)^2}\right)\Sigma^{1/2}\bar{\theta}'\left(\bar{\theta}'\right)^T\Sigma^{1/2}, \qquad (62) \end{aligned}$$

where

$$\Sigma_\theta^\lambda := \Sigma - \frac{\Sigma^{1/2}\bar{\theta}'(\bar{\theta}')^T\Sigma^{1/2}}{c + \|\bar{\theta}'\|_2^2}.$$

Proof: By completing squares, we can rewrite $\bar{p}_\theta^\lambda(x,y)$ as

$$\begin{aligned}
&\bar{p}_\theta^\lambda(x,y)\\
&= \frac{q_*(x)}{(2\pi\sigma^2)^{\frac{1}{2}}\, Z_\theta^\lambda}\\
&\quad \cdot \exp\left(-\frac{\lambda(y-x^T\theta^*)^2+(1-\lambda)(y-x^T\theta)^2}{2\sigma^2}\right)\\
&= \frac{q_*(x)}{(2\pi\sigma^2)^{\frac{n}{2}}\, Z_\theta^\lambda}\\
&\quad \cdot \exp\left(-\frac{(y-x^T\theta(\lambda))^2+\lambda(1-\lambda)(x^T(\theta^*-\theta))^2}{2\sigma^2}\right)\\
&= \frac{q_*(x)}{Z_\theta^\lambda}\cdot \exp\left(-\frac{\lambda(1-\lambda)(x^T\bar{\theta})^2}{2\sigma^2}\right) N(y|x^T\theta(\lambda),\sigma^2).
\end{aligned}$$

Hence, $p_\theta^\lambda(y|x)$ is $N(y|x^T\theta(\lambda),\sigma^2)$. Integrating y out, we also have

$$q_\theta^\lambda(x) = \frac{q_*(x)\exp\left(-\frac{1}{2c}(x^T\bar{\theta})^2\right)}{Z_\theta^\lambda}.$$

When $q_*(x)=N(\mathbf{0},\Sigma)$,

$$\begin{aligned}
q_\theta^\lambda(x) &= \frac{\exp\left(-\frac{1}{2}x^T\Sigma^{-1}x-\frac{1}{2c}x^T\bar{\theta}\bar{\theta}^Tx\right)}{(2\pi)^{m/2}|\Sigma|^{1/2}Z_\theta^\lambda}\\
&= \frac{\exp\left(-\frac{1}{2}x^T\left(\Sigma^{-1}+\frac{1}{c}\bar{\theta}\bar{\theta}^T\right)x\right)}{(2\pi)^{m/2}|\Sigma|^{1/2}Z_\theta^\lambda}. \qquad (63)
\end{aligned}$$

Since Σ is strictly positive definite by the assumption, $\Sigma^{-1}+(1/c)\bar{\theta}\bar{\theta}^T$ is non-singular. Hence, by the inverse formula (Lemma 10 in Appendix),

$$\begin{aligned}
\Sigma_\theta^\lambda &= \left(\Sigma^{-1}+\frac{1}{c}\bar{\theta}\bar{\theta}^T\right)^{-1} = \Sigma-\frac{\Sigma\bar{\theta}\bar{\theta}^T\Sigma}{c+\bar{\theta}^T\Sigma\bar{\theta}}\\
&= \Sigma-\frac{\Sigma^{1/2}\bar{\theta}'(\bar{\theta}')^T\Sigma^{1/2}}{c+\|\bar{\theta}'\|_2^2}. \qquad (64)
\end{aligned}$$

Therefore, $q_\theta^\lambda(x) = N(x|\mathbf{0},\Sigma_\theta^\lambda)$. The score function and Hessian of $\log p_\theta(y|x)$ are

$$\begin{aligned}
s_\theta(y|x) &= \frac{1}{\sigma^2}x(y-x^T\theta),\\
\frac{\partial^2\log p_\theta(y|x)}{\partial\theta\partial\theta^T} &= -\frac{1}{\sigma^2}xx^T. \qquad (65)
\end{aligned}$$

Using (58), the first derivative is obtained as

$$\begin{aligned}
\frac{\partial d_\lambda(p_*,p_\theta)}{\partial\theta} &= -E_{\bar{p}_\theta^\lambda}[s_\theta(y|x)]\\
&= -E_{q_\theta^\lambda}\left[E_{p_\theta^\lambda}[s_\theta(y|x)]\right]\\
&= -E_{q_\theta^\lambda}\left[E_{p_\theta^\lambda}\left[\frac{1}{\sigma^2}x(y-x^T\theta)\right]\right]\\
&= -E_{q_\theta^\lambda}\left[\frac{1}{\sigma^2}xx^T(\theta(\lambda)-\theta)\right]\\
&= \frac{\lambda}{\sigma^2}E_{q_\theta^\lambda}\left[xx^T\right]\bar{\theta}
\end{aligned}$$

because $\theta(\lambda)-\theta=-\lambda\bar{\theta}$. When $q_*(y|x)=N(\mathbf{0},\Sigma)$,

$$\frac{\partial d_\lambda(p_*,p_\theta)}{\partial\theta} = \frac{\lambda}{\sigma^2}\Sigma_\theta^\lambda\bar{\theta}.$$

From (64), we have

$$\begin{aligned}
\Sigma_\theta^\lambda\bar{\theta} &= \Sigma\bar{\theta}-\frac{\Sigma^{1/2}\bar{\theta}'(\bar{\theta}')^T\Sigma^{1/2}\bar{\theta}}{c+\|\bar{\theta}'\|_2^2}\\
&= \Sigma^{\frac{1}{2}}\bar{\theta}'-\left(\frac{\|\bar{\theta}'\|_2^2}{c+\|\bar{\theta}'\|_2^2}\right)\Sigma^{1/2}\bar{\theta}'\\
&= \left(\frac{c}{c+\|\bar{\theta}'\|_2^2}\right)\Sigma^{1/2}\bar{\theta}', \qquad (66)
\end{aligned}$$

which gives (61). Though (62) can be obtained by differentiating (61), we derive it by way of (59) here. To calculate the covariance matrix of s_θ in terms of $\bar{p}_\theta^\lambda$, we decompose s_θ as

$$\begin{aligned}
s_\theta(y|x) &= \frac{1}{\sigma^2}x(y-x^T\theta(\lambda)+x^T\theta(\lambda)-x^T\theta)\\
&= \frac{1}{\sigma^2}x(y-x^T\theta(\lambda))-\frac{\lambda}{\sigma^2}xx^T\bar{\theta}.
\end{aligned}$$

Note that the covariance of $(1/\sigma^2)x(y-x^T\theta(\lambda))$ and $-(\lambda/\sigma^2)xx^T\bar{\theta}$ vanishes since

$$\begin{aligned}
&E_{\bar{p}_\theta^\lambda}[x(y-x^T\theta(\lambda))(xx^T\bar{\theta})^T]\\
&= E_{q_\theta^\lambda}\left[xx^T(x^T\bar{\theta})E_{p_\theta^\lambda}\left[(y-x^T\theta(\lambda))\right]\right] = 0.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\mathrm{Var}_{\bar{p}_\theta^\lambda}(s_\theta)\\
&= \mathrm{Var}_{\bar{p}_\theta^\lambda}\left(\frac{1}{\sigma^2}x(y-x^T\theta(\lambda))\right)+\mathrm{Var}_{\bar{p}_\theta^\lambda}\left(\frac{\lambda}{\sigma^2}xx^T\bar{\theta}\right)\\
&= \frac{1}{\sigma^4}E_{\bar{p}_\theta^\lambda}\left[(y-x^T\theta(\lambda))^2xx^T\right]+\frac{\lambda^2}{\sigma^4}\mathrm{Var}_{q_\theta^\lambda}\left(xx^T\bar{\theta}\right)\\
&= \frac{1}{\sigma^2}E_{q_\theta^\lambda}\left[xx^T\right]+\frac{\lambda^2}{\sigma^4}\mathrm{Var}_{q_\theta^\lambda}\left(xx^T\bar{\theta}\right)
\end{aligned}$$

By (59) combined with (65), the Hessian of Rényi divergence is calculated as

$$\begin{aligned}
\frac{\partial^2 d_\lambda(p_*,p_\theta)}{\partial\theta\partial\theta^T} &= \frac{1}{\sigma^2}E_{\bar{p}_\theta^\lambda}[xx^T]\\
&\quad -(1-\lambda)\left(\frac{1}{\sigma^2}E_{q_\theta^\lambda}[xx^T]+\frac{\lambda^2}{\sigma^4}\mathrm{Var}_{q_\theta^\lambda}\left(xx^T\bar{\theta}\right)\right)\\
&= \frac{\lambda}{\sigma^2}E_{q_\theta^\lambda}[xx^T]-\frac{\lambda^2(1-\lambda)}{\sigma^4}\mathrm{Var}_{q_\theta^\lambda}\left(xx^T\bar{\theta}\right)\\
&= \frac{\lambda}{\sigma^2}E_{q_\theta^\lambda}[xx^T]-\frac{\lambda}{\sigma^2c}\mathrm{Var}_{q_\theta^\lambda}\left(xx^T\bar{\theta}\right).
\end{aligned}$$

When $q_*(x)=N(\mathbf{0},\Sigma)$, $\mathrm{Var}_{q_\theta^\lambda}\left(xx^T\bar{\theta}\right)$ is calculated as follows. Note that

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

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

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

where x_j denotes the jth element of x only here. Thus, we need all the fourth-moments of $q_\theta^\lambda(x)$. We rewrite Σ_θ^λ as S to reduce notation complexity hereafter. By the formula of moments of Gaussian distribution, we have

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

Therefore, the above quantity is calculated as

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

Summarizing these as a matrix form, we have

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

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

$$\begin{aligned}\mathrm{Var}_{q_\theta^\lambda}(xx^T\bar{\theta}) &= (\bar{\theta}^T S\bar{\theta})S+2S\bar{\theta}\bar{\theta}^T S-S\bar{\theta}\bar{\theta}^T S\\&= S\bar{\theta}\bar{\theta}^T S+(\bar{\theta}^T S\bar{\theta})S. \qquad (67)\end{aligned}$$

Using (66), the first and second terms of (67) are calculated as

$$\begin{aligned}S\bar{\theta}\bar{\theta}^T S &= \left(\frac{c^2}{(c+\|\bar{\theta}'\|_2^2)^2}\right)\Sigma^{1/2}\bar{\theta}'\left(\bar{\theta}'\right)^T\Sigma^{1/2},\\\bar{\theta}^T S\bar{\theta} &= \left(\frac{c}{c+\|\bar{\theta}'\|_2^2}\right)(\bar{\theta})^T\Sigma^{1/2}\bar{\theta}'=\frac{c\|\bar{\theta}'\|_2^2}{c+\|\bar{\theta}'\|_2^2}.\end{aligned}$$

Combining these,

$$\begin{aligned}&\frac{\partial^2 d_\lambda(p_*,p_\theta)}{\partial\theta\partial\theta^T}\\&=\frac{\lambda}{\sigma^2}S\\&\quad-\frac{\lambda}{\sigma^2 c}\left(\frac{c^2}{(c+\|\bar{\theta}'\|_2^2)^2}\Sigma^{1/2}\bar{\theta}'\left(\bar{\theta}'\right)^T\Sigma^{1/2}+\frac{c\|\bar{\theta}'\|_2^2}{c+\|\bar{\theta}'\|_2^2}S\right)\\&=\frac{\lambda}{\sigma^2}\cdot\frac{c}{c+\|\bar{\theta}'\|_2^2}S-\frac{\lambda}{\sigma^2}\cdot\frac{c}{(c+\|\bar{\theta}'\|_2^2)^2}\Sigma^{1/2}\bar{\theta}'\left(\bar{\theta}'\right)^T\Sigma^{1/2}\\&=\frac{\lambda}{\sigma^2}\cdot\frac{c}{c+\|\bar{\theta}'\|_2^2}\left(\Sigma-\frac{\Sigma^{1/2}\bar{\theta}'(\bar{\theta}')^T\Sigma^{1/2}}{c+\|\bar{\theta}'\|_2^2}\right)\\&\quad-\frac{\lambda}{\sigma^2}\cdot\frac{c}{(c+\|\bar{\theta}'\|_2^2)^2}\Sigma^{1/2}\bar{\theta}'\left(\bar{\theta}'\right)^T\Sigma^{1/2}\\&=\frac{\lambda}{\sigma^2}\cdot\frac{c}{c+\|\bar{\theta}'\|_2^2}\Sigma-\frac{2\lambda}{\sigma^2}\cdot\frac{c}{(c+\|\bar{\theta}'\|_2^2)^2}\Sigma^{1/2}\bar{\theta}'\left(\bar{\theta}'\right)^T\Sigma^{1/2}.\end{aligned}$$

E. Upper Bound of Negative Hessian

Using Lemma 7 in Section VI-D, we show that the negative Hessian of the Rényi divergence is bounded from above.

Lemma 8: Assume that $q_*(x)=N(x|\mathbf{0},\Sigma)$ and $p_*(y|x)=N(y|x^T\theta^*,\sigma^2)$, where Σ is non-singular. For any θ,θ^*,

$$-\frac{\partial^2 d_\lambda(p_*,p_\theta)}{\partial\theta\partial\theta^T}\preceq\frac{\lambda}{8\sigma^2}\Sigma, \qquad (68)$$

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

Proof: By Lemma 7, we have

$$\begin{aligned}-\frac{\partial^2 d_\lambda(p_*,p_\theta)}{\partial\theta\partial\theta^T}&=\frac{2\lambda}{\sigma^2}\left(\frac{c}{(c+\|\bar{\theta}'\|_2^2)^2}\right)\Sigma^{1/2}\bar{\theta}'\left(\bar{\theta}'\right)^T\Sigma^{1/2}\\&\quad-\frac{\lambda}{\sigma^2}\left(\frac{c}{c+\|\bar{\theta}'\|_2^2}\right)\Sigma.\end{aligned}$$

For any nonzero vector $v\in\Re^m$,

$$\begin{aligned}v^T\Sigma^{1/2}\bar{\theta}'\left(\bar{\theta}'\right)^T\Sigma^{1/2}v&=\left(v^T\Sigma^{1/2}\bar{\theta}'\right)^2\\&\le\|\Sigma^{1/2}v\|_2^2\cdot\|\bar{\theta}'\|_2^2=v^T(\|\bar{\theta}'\|_2^2\,\Sigma)v\end{aligned}$$

by Cauchy-Schwartz inequality. Hence, we have

$$\Sigma^{1/2}\bar{\theta}'\left(\bar{\theta}'\right)^T\Sigma^{1/2}\preceq\|\bar{\theta}'\|_2^2\,\Sigma.$$

Thus,

$$\begin{aligned}&-\frac{\partial^2 d_\lambda(p_*,p_\theta)}{\partial\theta\partial\theta^T}\\&\preceq\frac{2\lambda}{\sigma^2}\left(\frac{c\|\bar{\theta}'\|_2^2}{(c+\|\bar{\theta}'\|_2^2)^2}\right)\Sigma-\frac{\lambda}{\sigma^2}\left(\frac{c}{c+\|\bar{\theta}'\|_2^2}\right)\Sigma\\&=\frac{\lambda}{\sigma^2}\left(\frac{c(\|\bar{\theta}'\|_2^2-c)}{(c+\|\bar{\theta}'\|_2^2)^2}\right)\Sigma.\end{aligned}$$

Define

$$f(t):=\frac{c(t-c)}{(c+t)^2}$$

for $t\ge 0$. Checking the properties of $f(t)$, we have

$$f(0)=-1,\ f(c)=0,\ f(\infty)=0,\ \frac{df(t)}{dt}=\frac{c(3c-t)}{(t+c)^3}.$$

Therefore, $\max_{t\in[0,\infty)}f(t)=f(3c)=1/8$. As a result, we obtain

$$-\frac{\partial^2 d_\lambda(p_*,p_\theta)}{\partial\theta\partial\theta^T}\preceq\frac{\lambda}{8\sigma^2}\Sigma.$$

F. Proof of Lemma 3

We are now ready to derive restricted risk valid weighted ℓ_1 penalties.

Proof: Similarly to the rewriting from (20) to (32), we can rewrite the condition for restricted risk validity as

$$\begin{aligned}&\forall x^n\in A_\epsilon^n,\forall y^n\in\mathscr{Y}^n,\forall\theta\in\Theta,\\&\min_{\tilde{\theta}\in\widetilde{\Theta}(q_*)}\Big\{\underbrace{d_\lambda^n(p_*,p_\theta)-d_\lambda^n(p_*,p_{\tilde{\theta}})}_{\text{loss variation part}}+\underbrace{\log\frac{p_\theta(y^n|x^n)}{p_{\tilde{\theta}}(y^n|x^n)}+\tilde{L}(\tilde{\theta}|q_*)}_{\text{codelength validity part}}\Big\}\\&\le L(\theta|x^n). \qquad (69)\end{aligned}$$

We again write the inside part of the minimum in (69) as $H(\theta,\tilde{\theta},x^n,y^n)$. As described in Section III-B, the direct minimization of $H(\theta,\tilde{\theta},x^n,y^n)$ seems to be difficult. Instead of evaluating the minimum explicitly, we borrow a nice randomization technique introduced in [17] with some modifications. Their key idea is to evaluate 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 $\tilde{\theta}$ around θ because the expectation is larger than the minimum. Let us define $w^*:=(w_1^*,w_2^*,\cdots,w_m^*)^T$, where $w_j^*=\sqrt{\Sigma_{jj}}$ and $W^*:=\mathrm{diag}(w_1^*,\cdots,w_m^*)$. We quantize Θ as

$$\widetilde{\Theta}(q_*):=\{\delta(W^*)^{-1}z|z\in\mathscr{Z}^m\}, \qquad (70)$$

where $\delta>0$ is a quantization width and $\mathscr{Z}$ is the set of all integers. Though $\widetilde{\Theta}$ depends on x^n in fixed design cases [17],

we must remove the dependency to satisfy the restricted risk validity as above. For each θ, $\tilde{\theta}$ is randomized as

$$\tilde{\theta}_j = \begin{cases} \frac{\delta}{w_j^*}\lceil a_j \rceil & \text{with prob. } a_j - \lfloor a_j \rfloor \\ \frac{\delta}{w_j^*}\lfloor a_j \rfloor & \text{with prob. } \lceil a_j \rceil - a_j \\ \frac{\delta}{w_j^*} a_j & \text{with prob. } 1 - (\lceil a_j \rceil - \lfloor a_j \rfloor) \end{cases}, \quad (71)$$

where $a_j := w_j^* \theta_j / \delta$ and each component of $\tilde{\theta}$ is statistically independent of each other. Its important properties are

$$\begin{aligned} &E_{\tilde{\theta}}[\tilde{\theta}] = \theta, \quad \text{(unbiasedness)} \\ &E_{\tilde{\theta}}[|\tilde{\theta}|] = |\theta|, \\ &E_{\tilde{\theta}}[(\tilde{\theta}_j - \theta_j)(\tilde{\theta}_{j'} - \theta_{j'})] \le I(j = j') \frac{\delta}{w_j^*} |\theta_j|, \end{aligned} \quad (72)$$

where $|\tilde{\theta}|$ denotes a vector whose jth component is the absolute value of $\tilde{\theta}_j$ and similarly for $|\theta|$. Using these, we can bound $E_{\tilde{\theta}}[H(\theta, \tilde{\theta}, x^n, y^n)]$ as follows. The loss variation part in (69) is the main concern because it is more complicated than squared error of fixed design cases. Let us consider the following Taylor expansion

$$\begin{aligned} &d_\lambda^n(p_*, p_\theta) - d_\lambda^n(p_*, p_{\tilde{\theta}}) \\ &= -\left(\frac{\partial d_\lambda^n(p_*, p_\theta)}{\partial \theta} \right)^T (\tilde{\theta} - \theta) \\ &\quad - \frac{1}{2} \mathrm{Tr} \left(\frac{\partial^2 d_\lambda^n(p_*, p_{\theta^\circ})}{\partial \theta \partial \theta^T} (\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^T \right), \end{aligned} \quad (73)$$

where θ° is a vector between θ and $\tilde{\theta}$. The first term in the right side of (73) vanishes after taking expectation with respect to $\tilde{\theta}$ because $E_{\tilde{\theta}}[\tilde{\theta} - \theta] = 0$. As for the second term, we obtain

$$\begin{aligned} &\mathrm{Tr} \left(-\frac{\partial^2 d_\lambda^n(p_*, p_{\theta^\circ})}{\partial \theta \partial \theta^T} (\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^T \right) \\ &\le \frac{n\lambda}{8\sigma^2} \mathrm{Tr} \left(\Sigma \left(\tilde{\theta} - \theta \right) \left(\tilde{\theta} - \theta \right)^T \right) \end{aligned}$$

by Lemma 8. Thus, expectation of the loss variation part with respect to $\tilde{\theta}$ is bounded as

$$E_{\tilde{\theta}} \left[d_\lambda^n(p_*, p_\theta) - d_\lambda^n(p_*, p_{\tilde{\theta}}) \right] \le \frac{\delta n \lambda}{16\sigma^2} \|\theta\|_{w^*,1}. \quad (74)$$

The codelength validity part in (69) 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 different from those of [17]. The likelihood term is calculated as

$$\frac{1}{2\sigma^2} \left(2(Y - X\theta)^T X(\theta - \tilde{\theta}) + \mathrm{Tr} \left(X^T X (\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^T \right) \right).$$

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

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

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

$$\tilde{L}(\tilde{\theta}|q_*) := \frac{1}{\beta} C \left(\frac{1}{\delta} W^* \tilde{\theta} \right) = \frac{1}{\beta\delta} \|W^* \tilde{\theta}\|_1 \log 4m + \frac{\log 2}{\beta}. \quad (75)$$

By this definition, $\tilde{L}$ satisfies β-stronger Kraft's inequality and does not depend on x^n but depends on $q_*(x)$ through W^*. By taking expectation with respect to $\tilde{\theta}$, we have

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

because of (72). Thus the codelength validity part is bounded from above by

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

Combining with the loss variation part, we obtain an upper bound of $E_{\tilde{\theta}}[H(\theta, \tilde{\theta}, x^n, y^n)]$ as

$$\frac{\delta n \lambda}{16\sigma^2} \|\theta\|_{w^*,1} + \frac{\delta n}{2\sigma^2} \sum_{j=1}^m \frac{w_j^2}{w_j^*} |\theta_j| + \frac{\log 4m}{\beta\delta} \|\theta\|_{w^*,1} + \frac{\log 2}{\beta}.$$

Since $x^n \in A_\epsilon^n$, we have

$$\sqrt{(1-\epsilon)} w_j^* \le w_j \le \sqrt{(1+\epsilon)} w_j^*. \quad (76)$$

Thus we can bound $E_{\tilde{\theta}}[H(\theta, \tilde{\theta}, x^n, y^n)]$ by the data-dependent weighted ℓ_1 norm $\|\theta\|_{w,1}$ as

$$\begin{aligned} &E_{\tilde{\theta}}[H(\theta, \tilde{\theta}, x^n, y^n)] \\ &\le \frac{\delta n \lambda}{16\sigma^2} \frac{\|\theta\|_{w,1}}{\sqrt{1-\epsilon}} + \frac{\delta n \sqrt{1+\epsilon}}{2\sigma^2} \sum_{j=1}^m \frac{w_j^2}{w_j} |\theta_j| + \frac{\log 4m}{\beta\delta} \frac{\|\theta\|_{w,1}}{\sqrt{1-\epsilon}} \\ &\quad + \frac{\log 2}{\beta} \\ &= \left(\frac{\delta n}{\sigma^2} \left(\frac{\lambda}{16\sqrt{1-\epsilon}} + \frac{\sqrt{1+\epsilon}}{2} \right) + \frac{\log 4m}{\delta\beta\sqrt{1-\epsilon}} \right) \|\theta\|_{w,1} \\ &\quad + \frac{\log 2}{\beta}. \end{aligned}$$

Because this holds for any $\delta > 0$, we can minimize the upper bound with respect to δ. The minimum value of the upper bound is attained by

$$\delta = 4 \sqrt{\frac{\sigma^2 \log 4m}{n\beta} \cdot \frac{\sqrt{1-\epsilon}}{\lambda + 8\sqrt{1-\epsilon^2}}},$$

which completes the proof of (47). The latter part of the lemma can be proved easily by modifying the above proof. By (76), $w_j^2 \le (1+\epsilon)(w_j^*)^2$. Since $w_j^* \le M$ for all j,

$$\|\theta\|_{w^*,1} = \sum_{j=1}^m w_j^* |\theta_j| \le M \sum_{j=1}^m |\theta_j| = M \|\theta\|_1.$$

Thus, we can bound $E_{\tilde{\theta}}[H(\theta,\tilde{\theta},x^n,y^n)]$ by the ordinary ℓ_1 norm $\|\theta\|_{w,1}$ as

$$\begin{aligned}
&E_{\tilde{\theta}}[H(\theta,\tilde{\theta},x^n,y^n)]\\
&\le \frac{\delta n\lambda}{16\sigma^2}\|\theta\|_{w^*,1}+\frac{\delta n(1+\epsilon)}{2\sigma^2}\sum_{j=1}^m w_j^*|\theta_j|+\frac{\log 4m}{\beta\delta}\|\theta\|_{w^*,1}\\
&\quad+\frac{\log 2}{\beta}\\
&=\left(\frac{\delta n}{\sigma^2}\left(\frac{\lambda}{16}+\frac{1+\epsilon}{2}\right)+\frac{\log 4m}{\delta\beta}\right)\|\theta\|_{w^*,1}+\frac{\log 2}{\beta}\\
&\le M\left(\frac{\delta n}{\sigma^2}\left(\frac{\lambda}{16}+\frac{1+\epsilon}{2}\right)+\frac{\log 4m}{\delta\beta}\right)\|\theta\|_1+\frac{\log 2}{\beta}.
\end{aligned}$$

Again, minimizing the upper bound with respect to δ, we have

$$\delta=4\sqrt{\frac{\sigma^2\log 4m}{n\beta(\lambda+8(1+\epsilon))}},$$

which gives (48).

G. Some Remarks on the Proof of Lemma 3

The main difference of the proof from the fixed design case is in the loss variation part. In the fixed design case, the Rényi divergence $d_\lambda(p_*,p_\theta|x^n)$ is convex in terms of θ. When the Rényi divergence is convex, the negative Hessian is negative semi-definite for all θ. Hence, the loss variation part is trivially bounded from above by zero. On the other hand, $d_\lambda(p_*,p_\theta)$ is not convex in terms of θ. This can be intuitively seen by deriving the explicit form of $d_\lambda(p_*,p_\theta)$ instead of checking the positive semi-definiteness of its Hessian. From (63), we have

$$\begin{aligned}
Z_\theta^\lambda&=\int\frac{\exp\left(-\frac{1}{2}\left(x^T(\Sigma_\theta^\lambda)^{-1}x\right)\right)}{(2\pi)^{m/2}|\Sigma|^{1/2}}dx=|\Sigma|^{-1/2}|\Sigma_\theta^\lambda|^{1/2}\\
&=|\Sigma^{-1/2}\Sigma_\theta^\lambda\Sigma^{-1/2}|^{1/2}\\
&=\left|I_m-\left(\frac{1}{c+\|\bar{\theta}'\|_2^2}\right)\bar{\theta}'\left(\bar{\theta}'\right)^T\right|^{1/2}\\
&=\left|I_m-\left(\frac{\|\bar{\theta}'\|_2^2}{c+\|\bar{\theta}'\|_2^2}\right)\left(\frac{\bar{\theta}'}{\|\bar{\theta}'\|_2}\right)\left(\frac{\bar{\theta}'}{\|\bar{\theta}'\|_2}\right)^T\right|^{1/2}. \quad (77)
\end{aligned}$$

Prof. A. R. Barron suggested in a private discussion that Z_θ^λ can be simplified more as follows. Let $Q:=[q_1,q_2,\cdots,q_m]$ be an orthogonal matrix such that $q_1:=\bar{\theta}'/\|\bar{\theta}'\|_2$. Using this, we have

$$\begin{aligned}
&I_m-\left(\frac{\|\bar{\theta}'\|_2^2}{c+\|\bar{\theta}'\|_2^2}\right)\left(\frac{\bar{\theta}'}{\|\bar{\theta}'\|_2}\right)\left(\frac{\bar{\theta}'}{\|\bar{\theta}'\|_2}\right)^T\\
&=QQ^T-\left(\frac{\|\bar{\theta}'\|_2^2}{c+\|\bar{\theta}'\|_2^2}\right)q_1q_1^T\\
&=\left(1-\left(\frac{\|\bar{\theta}'\|_2^2}{c+\|\bar{\theta}'\|_2^2}\right)\right)q_1q_1^T+\sum_{j=2}^m q_jq_j^T\\
&=\left(\frac{c}{c+\|\bar{\theta}'\|_2^2}\right)q_1q_1^T+\sum_{j=2}^m q_jq_j^T\\
&=Q\begin{pmatrix}c/(c+\|\bar{\theta}'\|_2^2) & 0 & 0 & \cdots & 0\\ 0 & 1 & 0 & \cdots & 0\\ 0 & 0 & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & 1\end{pmatrix}Q^T.
\end{aligned}$$

Hence, the resultant Z_θ^λ is obtained as

$$\begin{aligned}
Z_{\tilde{\theta}}^\lambda&=\left|I_m-\gamma(\|\bar{\theta}'\|_2^2)\left(\frac{\bar{\theta}'}{\|\bar{\theta}'\|_2}\right)\left(\frac{\bar{\theta}'}{\|\bar{\theta}'\|_2}\right)^T\right|^{\frac{1}{2}}\\
&=\left(\frac{c}{c+\|\bar{\theta}'\|_2^2}\right)^{\frac{1}{2}}.
\end{aligned}$$

Thus, we have a simple expression of the Rényi divergence as

$$d_\lambda(p_*,p_\theta)=\frac{1}{2(1-\lambda)}\log\left(1+\frac{\|\bar{\theta}'\|_2^2}{c}\right). \quad (78)$$

From this form, we can easily know that the Rényi divergence is not convex. When the Rényi divergence is non-convex, it is unclear in general whether and how the loss variation part is bounded from above. This is one of the main reasons why the derivation becomes more difficult than that of the fixed design case.

We also mention an alternative proof of Lemma 3 based on (78). We provided Lemma 6 to calculate Hessian of the Rényi divergence. However, the above simple expression of the Rényi divergence is somewhat easier to differentiate, while the expression based on (77) is somewhat hard to do it. Therefore, we can twice differentiate the above Rényi divergence directly in order to obtain Hessian instead of Lemma 7 in our Gaussian setting. However, there is no guarantee that such a simplification is always possible in general setting. In our proof, we tried to give a somewhat systematic way which is easily applicable to other settings to some extent. Suppose now, for example, we are aim at deriving restricted risk valid ℓ_1 penalties for lasso when $q_*(x)$ is subject to non-Gaussian distribution. By (60) in Lemma 7, it suffices only to bound $\mathrm{Var}_{q_\theta^\lambda}(xx^T\bar{\theta})$ in the sense of positive semi-definiteness because $-E_{q_\theta^\lambda}[xx^T]$ is negative semi-definite. In general, it seemingly depends on a situation which is better, the direct differential or using (60). In our Gaussian setting, we imagine that the easiest way to calculate Hessian for most readers is to calculate the first derivative by the formula (58) and then to differentiate it directly, though this depends on readers' background knowledge. For other settings, we believe that providing Lemmas 6 and 7 would be useful in some cases.

H. Proof of Lemma 4

Here we show that x^n distributes out of A_ϵ^n with exponentially small probability with respect to n.

Proof: The typical set A_ϵ^n can be decomposed covariate-wise as

$$\begin{aligned}
A_\epsilon^n&=\Pi_{j=1}^m A_\epsilon^n(j),\\
A_\epsilon^n(j)&:=\left\{\boldsymbol{x}_j\in\Re^n\,\middle|\,\left|(w_j^*)^2-(\|\boldsymbol{x}_j\|_2^2/n)\right|\le\epsilon(w_j^*)^2\right\}\\
&=\left\{\boldsymbol{x}_j\in\Re^n\,\middle|\,\left|(w_j^*)^2-w_j^2\right|\le\epsilon(w_j^*)^2\right\},
\end{aligned}$$

where $\boldsymbol{x}_j := (x_{1j}, x_{2j}, \cdots, x_{nj})^T$ and the above Π denotes a direct product of sets. We write w_j^2 as z and $(w_j^*)^2$ as s (the index j is dropped for legibility). From its definition, w_j^2 is subject to a Gamma distribution $\mathrm{Ga}((n/2), (2s)/n)$ because $\boldsymbol{x}_j \sim \Pi_{i=1}^n N(x_j|0, s)$. We rewrite the Gamma distribution $g(z; s)$ in the form of exponential family:

$$\begin{aligned} g(z;s) &:= \mathrm{Ga}\left(\frac{n}{2}, \frac{2s}{n}\right) = \frac{\Gamma(\frac{n}{2})}{z^{\frac{n}{2}-1}} \cdot \exp\left(-\frac{nz}{2s}\right)\left(\frac{2s}{n}\right)^{\frac{n}{2}} \\ &= \exp\left(\frac{n-2}{2}\log z - \frac{nz}{2s} - \log\left(\frac{2s}{n}\right)^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)\right) \\ &= \exp\left(C(z) + \nu z - \psi(\nu)\right), \end{aligned}$$

where

$$\begin{aligned} &C(z) := \left(\frac{n-2}{2}\right)\log z, \quad \nu := -\frac{n}{2s}, \\ &\psi(\nu) := \log(-\nu)^{-n/2}\Gamma(n/2). \end{aligned}$$

That is, ν is a natural parameter and z is a sufficient statistic, so that the expectation parameter $\eta(s)$ is $E_{g(z;s)}[z]$. The relationship between the variance parameter s and natural/expectation parameters are summarized as

$$\nu(s) := -\frac{n}{2s}, \quad \eta(\nu) = -\frac{n}{2\nu}.$$

For exponential families, there is a useful Sanov-type inequality (Lemma 9 in Appendix). Using this Lemma, we can bound $\Pr(\boldsymbol{x}_j \notin A_\epsilon^n(j))$ as follows. For this purpose, it suffices to bound the probability of the event $|w_j^2 - w_j^{*2}| \le w_j^{*2}\epsilon$. When $s = (w_j^*)^2$ and $s' = s(1 \pm \epsilon)$,

$$\begin{aligned} &\mathscr{D}(\nu(s \pm \epsilon s), \nu) \\ &= \left(-\frac{n}{2s(1\pm\epsilon)} - \left(-\frac{n}{2s}\right)\right) s(1\pm\epsilon) - \frac{n}{2}\log(1\pm\epsilon) \\ &= \left(-\frac{n}{2s}\right)\left(\frac{1}{(1\pm\epsilon)} - 1\right) s(1\pm\epsilon) - \frac{n}{2}\log(1\pm\epsilon) \\ &= \left(-\frac{n}{2}\right)(1 - (1\pm\epsilon)) - \frac{n}{2}\log(1\pm\epsilon) \\ &= \frac{n}{2}\left(\pm\epsilon - \log(1\pm\epsilon)\right), \end{aligned}$$

where $\mathscr{D}$ is the single data version of the KL-divergence defined by (8). It is easy to see that $\epsilon - \log(1+\epsilon) \le -\epsilon - \log(1-\epsilon)$ for any $0 < \epsilon < 1$. By Lemma 9, we obtain

$$\begin{aligned} &\Pr(|w_j^2 - w_j^{*2}| \le \epsilon w_j^{*2}) \\ &= 1 - \Pr(w_j^2 - w_J^{*2} \ge \epsilon w_j^{*2} \text{ or } w_J^{*2} - w_j^2 \ge \epsilon w_j^{*2}) \\ &= 1 - \Pr(w_j^2 - w_J^{*2} \ge \epsilon w_j^{*2}) - \Pr(w_J^{*2} - w_j^2 \ge \epsilon w_j^{*2}) \\ &\ge 1 - \exp\left(-\frac{n}{2}(\epsilon - \log(1+\epsilon))\right) \\ &\quad - \exp\left(-\frac{n}{2}(-\epsilon - \log(1-\epsilon))\right) \\ &\ge 1 - 2\exp\left(-\frac{n}{2}(\epsilon - \log(1+\epsilon))\right). \end{aligned}$$

Hence P_ϵ^n can be bounded from below as

$$\begin{aligned} P_\epsilon^n &= \Pr(x^n \in A_\epsilon^n) = \Pi_{j=1}^m (1 - \Pr(\boldsymbol{x}_j \notin A_\epsilon^n(j))) \\ &\ge \left(1 - 2\exp\left(-\frac{n}{2}(\epsilon - \log(1+\epsilon))\right)\right)^m \\ &\ge 1 - 2m\exp\left(-\frac{n}{2}(\epsilon - \log(1+\epsilon))\right). \end{aligned}$$

The last inequality follows from $(1-t)^m \ge 1 - mt$ for any $t \in [0,1]$ and $m \ge 1$. To simplify the bound, we can do more. The maximum positive real number u such that, for any $\epsilon \in [0,1]$, $u\epsilon^2 \le (1/2)(\epsilon - \log(1+\epsilon))$ is $(1 - \log 2)/2$. Then, the maximum integer u_1 such that $(1-\log 2)/2 \ge 1/u_1$ is 7, which gives the last inequality in the statement.

I. Proof of Lemma 5

We can prove this lemma by checking the proof of Lemma 3.

Proof: Let

$$L_1(\theta|x^n) := \mu_1\|\theta\|_{w,1} + \mu_2.$$

Similarly to the rewriting from (28) to (32), we can restate the codelength validity condition for $L_1(\theta|x^n)$ as "there exist a quantize subset $\widetilde{\Theta}(x^n)$ and a model description length $\tilde{L}(\tilde{\theta}|x^n)$ satisfying the usual Kraft's inequality, such that

$$\begin{aligned} &\forall x^n \in \mathscr{X}^n,\ \forall y^n \in \mathscr{Y}^n,\ \forall \theta \in \Theta, \\ &\min_{\tilde{\theta}\in\widetilde{\Theta}(x^n)} \left\{ \log\frac{p_\theta(y^n|x^n)}{p_{\tilde{\theta}}(y^n|x^n)} + \tilde{L}(\tilde{\theta}|x^n) \right\} \le L_1(\theta|x^n)." \end{aligned} \tag{79}$$

Recall that (47) is a sufficient condition for the restricted risk validity of L_1, in fact, it was derived as a sufficient condition for the proposition that $L_1(\theta|x^n)$ bounds from above

$$\begin{aligned} E_{\tilde{\theta}}[H(\theta, \tilde{\theta}, v^n, y^n)] = &\underbrace{E_{\tilde{\theta}}\left[d_\lambda^n(p_*, p_\theta) - d_\lambda^n(p_*, p_{\tilde{\theta}})\right]}_{\text{(i)}} \\ &+ \underbrace{E_{\tilde{\theta}}\left[\log\frac{p_\theta(y^n|v^n)}{p_{\tilde{\theta}}(y^n|v^n)} + \tilde{L}(\tilde{\theta}|q_*)\right]}_{\text{(ii)}} \end{aligned} \tag{80}$$

for any $q_* \in \mathscr{P}_{\mathscr{X}}^n$, $v^n \in A_\epsilon^n$, $y^n \in \mathscr{Y}^n$, $\theta \in \Theta$, where $\tilde{\theta}$ was randomized on $\widetilde{\Theta}(q_*)$ and $(\widetilde{\Theta}(q_*), \tilde{L}(\tilde{\theta}|q_*))$ were defined by (70) and (75), in particular, $\tilde{L}(\tilde{\theta}|q_*)$ satisfies β-stronger Kraft's inequality. Recall that $H(\theta, \tilde{\theta}, x^n, y^n)$ is the inside part of the minimum in (69). Here, we used v^n instead of x^n so as to discriminate from the above fixed x^n. To derive the sufficient condition, we obtained upper bounds on the terms (i) and (ii) of (80) respectively, and shown that $L_1(\theta|v^n)$ with $v^n \in A_\epsilon^n$ is not less than the sum of both upper bounds if (47) is satisfied. A point is that the upper bound on the term (i) we derived is a non-negative function of θ (see (74)). Hence, if $v^n \in A_\epsilon^n$ and (47) hold, $L_1(\theta|v^n)$ is an upper bound on the term (ii), which is not less than

$$\min_{\tilde{\theta}\in\widetilde{\Theta}(q^*)} \left\{ \log\frac{p_\theta(y^n|v^n)}{p_{\tilde{\theta}}(y^n|v^n)} + \tilde{L}(\tilde{\theta}|q^*) \right\}.$$

Now assume (47) and let us take $q_* \in \mathscr{P}_x^n$ given x^n, such that Σ_{jj} is equal to $(1/n)\sum_{i=1}^n x_{ij}^2$ for all j. Then we have $x^n \in A_\epsilon^n$, which implies

$$L_1(\theta|x^n) \ge \min_{\tilde{\theta}\in\widetilde{\Theta}(q^*)} \left\{ \log\frac{p_\theta(y^n|x^n)}{p_{\tilde{\theta}}(y^n|x^n)} + \tilde{L}(\tilde{\theta}|q^*) \right\}.$$

Since q^* is determined by x^n and $\tilde{L}(\tilde{\theta}|q^*)$ satisfies Kraft's inequality, the codelength validity condition holds for L_1.

VII. Conclusion

We proposed a way to extend BC theory to supervised learning. Our extension does not require any additional complicated assumptions. This is important in order to show the correctness of the MDL principle in a wide range of situations. As an interesting application, we proved that lasso is a BC-proper MDL estimator. That is, lasso can be interpreted as an estimator obtained by minimizing the description length of the given data using a two-stage code. Furthermore, lasso is shown to have a new risk and regret bounds in which its statistical risk is bounded from above by its redundancy or regret. That is, compressing the data more leads to the more favorable estimator. The derived bounds hold for any n and p without any complicated assumptions including bounded assumptions. In this sense, our risk bound is unique compared to other past studies. Numerical simulations illustrate how tight our regret bound is. Our next challenges are extending our result to wider situations including non-normal covariates, non-normal noise, other penalties and other machine learning methods. According to our naive trials, none of them is easily obtained only by tracing the way of this paper as itself.

Appendix
Sanov-type Inequality

The following lemma is a special case of the result in [20]. Below, we give a simpler proof. In the lemma, we denote a random variable of one dimension by X and denote its corresponding one dimensional variable by x.

Lemma 9: Let

$$x \sim p_\theta(x) := \exp(\theta x - \psi(\theta)),$$

where x and θ are of one dimension. Then,

$$\begin{aligned} \Pr_\theta(X \geq \eta') &\leq \exp(-\mathscr{D}(\theta', \theta)) \quad \text{if } \eta' \geq \eta, \\ \Pr_\theta(X \leq \eta') &\leq \exp(-\mathscr{D}(\theta', \theta)) \quad \text{if } \eta' \leq \eta, \end{aligned}$$

where η is the expectation parameter corresponding to the natural parameter θ and similarly for η'. The symbol $\mathscr{D}$ denotes the single sample version of the KL-divergence defined by (25).

Proof: In this setting, the KL divergence is calculated as

$$\mathscr{D}(\theta, \theta') = E_{p_\theta}\left[\log\left(\frac{p_\theta(X)}{p_{\theta'}(X)}\right)\right] = (\theta - \theta')\eta - \psi(\theta) + \psi(\theta').$$

Assume $\eta' - \eta \geq 0$. Because of the monotonicity of natural parameter and expectation parameter of exponential family,

$$\begin{aligned} X \geq \eta' &\Leftrightarrow (\theta' - \theta)X \geq (\theta' - \theta)\eta' \\ &\Leftrightarrow \exp\left((\theta' - \theta)X\right) \geq \exp\left((\theta' - \theta)\eta'\right). \end{aligned}$$

By Markov's inequality, we have

$$\begin{aligned} &\Pr_\theta\left(\exp\left((\theta' - \theta)X\right) \geq \exp\left((\theta' - \theta)\eta'\right)\right) \\ &\leq \frac{E_{p_\theta}\left[\exp\left((\theta' - \theta)X\right)\right]}{\exp\left((\theta' - \theta)\eta'\right)} \\ &= \int \exp(\theta x - \psi(\theta)) \exp((\theta' - \theta)x) dx \cdot \exp(-(\theta' - \theta)\eta') \\ &= \int \exp(\theta' x - \psi(\theta)) dx \cdot \exp(-(\theta' - \theta)\eta') \\ &= \exp(\psi(\theta')) \exp(-\psi(\theta)) \cdot \exp(-(\theta' - \theta)\eta') \\ &= \exp(-\left((\theta' - \theta)\eta' - \psi(\theta') + \psi(\theta)\right)). \end{aligned}$$

The other inequality can also be proved in the same way.

TABLE IV

Glossary That Contains Almost All Symbols Appearing in the Sections Except Proof Section and Appendices

$\mathscr{X}$	the domain of feature (or covariate) x
$\mathscr{Y}$	the domain of teacher signal y
$x^n \in \mathscr{X}^n$	$x^n = (x_1, x_2, \cdots, x_n)$ is a covariate sample sequence
$y^n \in \mathscr{Y}^n$	$y^n = (y_1, y_2, \cdots, x_n)$ is a teacher signal sequence
$\bar{p}_*(x, y) = q_*(x)p_*(y\|x)$	the underlying joint, marginal, conditional distribution generating the data
$p_\theta(y\|x)$	a certain parametric model with parameter $\theta \in \Theta$
Θ	$\Theta \subset \Re^m$ is a parameter space
$\widetilde{\Theta}$	$\widetilde{\Theta} \subset \Theta$ is a quantized parameter space
$d^n_\lambda(p_*, p_\theta)$	the Rényi divergence between $p_*(y^n\|x^n)$ and $p_\theta(y^n\|x^n)$
$d_\lambda(p_*, p_\theta)$	the Rényi divergence between $p_*(y\|x)$ and $p_\theta(y\|x)$
$\mathscr{D}^n(p_*, p_\theta)$	the KL divergence between $p_*(y^n\|x^n)$ and $p_\theta(y^n\|x^n)$
$\mathscr{D}^n_\alpha(p_*, p_\theta)$	the α-divergence between $p_*(y^n\|x^n)$ and $p_\theta(y^n\|x^n)$
$d^{2,n}_H(p_*, p_\theta)$	the squared Hellinger distance between $p_*(y^n\|x^n)$ and $p_\theta(y^n\|x^n)$
$\|\theta\|_1$	$\|\theta\|_1 = \sum_{j=1}^m \|\theta_j\|$ the usual ℓ_1 norm
$\|\theta\|_{w,1}$	$\|\theta\|_{w,1} = \sum_{j=1}^m w_j\|\theta_j\|$ the weighted ℓ_1 norm with the weight vector w
$\tilde{\theta}$	$\tilde{\theta}$ is a dummy variable moving over $\widetilde{\Theta}$
θ	θ is a dummy variable moving over Θ
$\ddot{\theta}$	$\ddot{\theta}$ is the parameter attaining the minimum description length of the two-stage code
$\hat{\theta}$	$\hat{\theta}$ is a penalized likelihood estimator
$\tilde{L}_2(y^n\|x^n)$	the codelength of the two-stage code
$\tilde{L}(\theta\|x^n)$	the model description length of the two-stage code
$L(\theta\|x^n)$	the penalty function of penalized likelihood estimator
$\tilde{L}_2(y^n\|x^n)$	total codelength of the two-stage code
$L_2(y^n\|x^n)$	a the minimized penalized likelihood
$\tilde{p}_2(y^n\|x^n)$	the sub-probability distribution of the two-stage code
$p_2(y^n\|x^n)$	a counterpart of $\tilde{p}_2(y^n\|x^n)$ for penalized likelihood estimator
A^n_ϵ	a typical set
P^n_ϵ	the probability that x^n belongs to the typical set A^n_ϵ
μ_1, μ_2	μ_1 is a regularization coefficient while μ_2 is a constant
w_j	$w_j = \sqrt{(1/n)\sum_{i=1}^n x_{ij}^2}$
w^*_j	$w^*_j = \sqrt{\Sigma_{jj}}$
Y	$Y := (y_1, y_2, \cdots, y_n)^T$
X	$X = [x_{ij}]$ is the matrix whose i, jth element is the jth element of x_i
Υ	noise vector subject to $N(\epsilon\|\mathbf{0}, \sigma^2 I_n)$
I_m	the identity matrix of dimension m
$I_A(x^n)$	the indicator function of a set $A \subset \mathscr{X}^n$
$N(x\|\mu, \Sigma)$	the density function of Gaussian random vector with mean vector μ and covariance Σ

Inverse Matrix Formula

Lemma 10: Let A be a non-singular $m \times m$ matrix. If c and d are both $m \times 1$ vectors and $A + cd$ is non-singular, then

$$(A + cd^T)^{-1} = A^{-1} - \frac{A^{-1}cd^T A^{-1}}{1 + d^T A^{-1} c}.$$

See, for example, Corollary 1.7.2 in [31] for its proof.

Glossary

In this paper, many complicated symbols appeared. We provide a glossary for the reader's convenience.

ACKNOWLEDGMENT

The authors would like to thank Prof. Andrew Barron for fruitful discussion. The form of Rényi divergence (78) is the result of simplification suggested by him. Furthermore, they learned the simple proof of Lemma 9 from him. They thank Mr. Yushin Toyokihara for his support.

REFERENCES

[1] S. Amari and H. Nagaoka, *Methods of Information Geometry*. Providence, RI, USA: AMS & Oxford Univ. Press, 2000.
[2] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inf. Theory*, vol. 37, no. 4, pp. 1034–1054, Jul. 1991.
[3] A. R. Barron, C. Huang, J. Q. Li, and X. Luo, "MDL, penalized likelihood, and statistical risk," in *Proc. IEEE Inf. Theory Workshop*, Porto, Portugal, May 2008, pp. 247–257.
[4] A. R. Barron and X. Luo, "MDL procedures with ℓ_1 penalty and their statistical risk," in *Proc. 1st Workshop Inf. Theoretic Methods Sci. Eng.*, Tampere, Finland, Aug. 2008.
[5] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2743–2760, 1998.
[6] P. L. Bartlett, S. Mendelson, and J. Neeman, "ℓ_1-regularized linear regression: Persistence and oracle inequalities," *Probab. Theory Rel. Fields*, vol. 154, nos. 1–2, pp. 193–224, 2012.
[7] M. Bayati, J. Bento, and A. Montanari, "The LASSO risk: Asymptotic results and real world examples," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 145–153.
[8] M. Bayati, M. Erdogdu, and A. Montanari, "Estimating LASSO risk and noise level," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1–9.
[9] M. Bayati and A. Montanari, "The LASSO risk for Gaussian matrices," *IEEE Trans. Inf. Theory*, vol. 58, no. 4, pp. 1997–2017, Apr. 2012.
[10] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, Jan. 2009.
[11] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Math. Soc.*, vol. 35, no. 1, pp. 99–109, 1943.
[12] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *Ann. Statist.*, vol. 37, no. 4, pp. 1705–1732, Aug. 2009.
[13] F. Bunea, A. Tsybakov, and M. Wegkamp, "Sparsity oracle inequalities for the Lasso," *Electron. J. Statist.*, vol. 1, pp. 169–194, May 2007.
[14] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp, "Aggregation for Gaussian regression," *Ann. Statist.*, vol. 35, no. 4, pp. 1674–1697, Aug. 2007.
[15] E. J. Candès and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger then n," *Ann. Statist.*, vol. 35, no. 6, pp. 2313–2351, 2007.
[16] S. Chatterjee and A. Barron, "Information theoretic validity of penalized likelihood," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2014, pp. 3027–3031.
[17] S. Chatterjee and A. Barron, "Information theory of penalized likelihoods and its statistical implications," 2014, *arXiv:1401.6714*. [Online]. Available: http://arxiv.org/abs/1401.6714
[18] A. Cichocki and S.-I. Amari, "Families of alpha-beta- and gamma-divergences: Flexible and robust measures of similarities," *Entropy*, vol. 12, no. 6, pp. 1532–1568, Jun. 2010.
[19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley-Interscience, 2006.
[20] I. Csiszar, "Sanov property, generalized I-projection and a conditional limit theorem," *Ann. Probab.*, vol. 12, no. 3, pp. 768–793, Aug. 1984.
[21] T. van Erven and P. Harremoës, "Divergence and Kullback-Leibler divergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.
[22] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical Lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, Jul. 2008.
[23] P. D. Grünwald, *The Minimum Description Length Principle*. Cambridge, MA, USA: MIT Press, 2007.
[24] P. D. Grünwald, I. J. Myung, and M. A. Pitt, *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA, USA: MIT Press, 2005.
[25] P. D. Grünwald and N. A. Mehta, "Fast rates for general unbounded loss functions: From ERM to generalized Bayes," 2016, *arXiv:1605.00252*. [Online]. Available: http://arxiv.org/abs/1605.00252
[26] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer-Verlag, 2001.
[27] M. Ishikawa, "Structural learning with forgetting neural networks," *Neural Netw.*, vol. 9, no. 3, pp. 509–521, 1996.
[28] Q. Li, "Estimation of mixture models," Ph.D. dissertation, Dept. Statistics, Yale Univ., New Haven, CT, USA, 1999.
[29] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Statist. Probab.*, vol. 1, 1961, pp. 547–561.
[30] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, Sep. 1978.
[31] J. R. Schott, *Matrix Analysis for Statistics*, 2nd ed. Hoboken, NJ, USA: Wiley, 2005.
[32] J. Takeuchi, "An introduction to the minimum description length principle," in *A Mathematical Approach to Research Problems of Science and Technology*. Tokyo, Japan: Springer, 2014, pp. 279–296.
[33] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *Ann. Statist.*, vol. 36, no. 3, pp. 1171–1220, 2008.
[34] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Stat. Soc., Ser. B (Methodol.)*, vol. 58, no. 1, pp. 267–288, Jan. 1996.
[35] K. Yamanishi, "A learning criterion for stochastic rules," *Mach. Learn.*, vol. 9, nos. 2–3, pp. 165–203, Jul. 1992.
[36] T. Zhang, "On the convergence of MDL density estimation," in *Learning Theory*, vol. 3120, J. Shawe-Taylor and Y. Singer, Eds. Berlin, Germany: Springer, 2004.
[37] T. Zhang, "From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation," *Ann. Statist.*, vol. 34, no. 5, pp. 2180–2210, Oct. 2006.
[38] T. Zhang, "Some sharp performance bounds for least squares regression with l_1 regularization," *Ann. Statist.*, vol. 37, no. 5, pp. 2109–2144, 2009.

Masanori Kawakita was born in Nagoya, Japan, in 1977. He received the B.E. and M.E. degrees from Keio University in 2001 and 2003, respectively, and the Ph.D. degree in statistical science from the Graduate University for Advanced Studies in 2006. From 2007 to 2018, he was an Assistant Professor with the Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka, Japan. In 2018, he joined Mie Toyopet Cooperation, Mie, Japan, where he is currently the Director. He is currently a Cooperative Researcher of the Graduate School of Informatics, Nagoya University, Nagoya. His research interests include machine learning, statistical science, information geometry, and information theory.

Jun'ichi Takeuchi (Member, IEEE) was born in Tokyo, Japan, in 1964. He received the B.Sc. degree in physics and the Dr.Eng. degree in mathematical engineering from the University of Tokyo in 1989 and 1996, respectively. From 1989 to 2006, he worked with NEC Corporation, Japan. In 2006, he joined Kyushu University, Fukuoka, Japan, where he is currently a Professor of mathematical engineering. From 1996 to 1997, he was a Visiting Research Scholar with the Department of Statistics, Yale University, New Haven, CT, USA. His research interests include mathematical statistics, information geometry, information theory, data science, and machine learning. He is a member of the IEICE, IPSJ, and JSIAM.