Information Theoretic validity of Penalized Likelihood

Sabyasachi Chatterjee
Department of Statistics
Yale University
New Haven, Connecticut 06511
Email: sabyasachi.chatterjee@yale.edu

Andrew Barron
Department of Statistics
Yale University
New Haven, Connecticut 06511
Email: andrew.barron@yale.edu

***Abstract*—Building upon past work, which developed information theoretic notions of when a penalized likelihood procedure can be interpreted as codelengths arising from a two stage code and when the statistical risk of the procedure has a redundancy risk bound, we present new results and risk bounds showing that the l_1 penalty in Gaussian Graphical Models fits the above story. We also show how the traditional l_0 penalty times plus lower order terms which stay bounded on the whole parameter space has a conditional two stage description length interpretation.**

I. Introduction

It is known that the MDL principle motivates viewing a penalized log likelihood procedure as minimizing the codelengths of a two stage code. Traditionally, this has required that the minimizing space be countable. In past works [3], [4] the authors address this issue and develop a notion as to how to interpret a penalized log likelihood as codelengths arising from a two stage code even when the minimization is done over an uncountable parameter space. We describe the framework laid out in these past works briefly.

We denote the sample space by $\mathcal{X}$ and its elements by x. For any integer n, we denote the n fold cross product of $\mathcal{X}$ by $\mathcal{X}^n$. We denote a generic element in $\mathcal{X}^n$ by $\underline{x}$ and a random realization (data) from $\mathcal{X}^n$ by $X = (x_1, \dots, x_n)$. A probabilistic source p is a sequence of probability distributions $p^{(1)}, p^{(2)}, \dots$ on $\mathcal{X}^1, \mathcal{X}^2, \dots$ so that they are consistent. By consistency we mean that the marginal distribution of $p^{(n+1)}$ restricted to the first n coordinates is $p^{(n)}$. We drop the subscript n and write $p(\underline{x})$ instead of $p^{(n)}(\underline{x})$ whenever it is clear from the context. For some probability source p and some element $\underline{x}$, whenever we write $p(\underline{x})$, it refers to the probability mass function corresponding to the source p or the probability density function, corresponding to the source p, with respect to some dominating measure. In this document, the uncountable sample spaces are always euclidean spaces of some dimension and the dominating measure is the Lebesgue measure. We will also distinguish between countable and uncountable parameter spaces in our notations. Generically we denote a countable parameter space by $\mathcal{F}$ and an uncountable parameter space by Θ. We generically denote the elements of $\mathcal{F}$ by $\tilde{\theta}$ and elements in Θ by θ. We also consistently denote a penalty function on Θ by pen and a penalty function on $\mathcal{F}$ by V. We also measure codelengths in nats instead of bits in this manuscript.

The Kraft's inequality gives a correspondence between probability distributions and prefix free codes on a countable sample space. This correspondence may be extended to uncountable sample spaces as well, namely $-\log p(\underline{x})$ may be regarded as Kraft satisfying codelengths on $\mathcal{X}^n$, see [1] for more details. For the countable parameter space $\mathcal{F}$, the two stage codelength on $\mathcal{X}^n$ minimize the negative log likelihood of the data plus a kraft inequality satisfying codelength on $\mathcal{F}$. These two stage codelengths are clearly uniquely decodable and hence correspond to a prefix free code. In the case when the parameter space Θ is uncountable, one of the ways in which a penalized log likelihood expression could still be interpreted as Kraft satisfying codelengths on the sample space is as follows. Assume there exists a countable subset $\mathcal{F} \subset \Theta$ and any Kraft summable penalty $V(\tilde{\theta})$ on $\mathcal{F}$ satisfying firstly,

$$\sum_{\tilde{\theta} \in \mathcal{F}} exp(-V(\theta)) \leq 1 \tag{1}$$

and secondly for all $\underline{x}$

$$\begin{aligned} &\min_{\theta \in \Theta} \{-\log p_{\theta}(\underline{x}) + pen(\theta)\} \geq \\ &\min_{\tilde{\theta} \in \mathcal{F}} \{-\log p_{\tilde{\theta}}(\underline{x}) + V(\tilde{\theta})\}. \end{aligned} \tag{2}$$

Since the right side of (2) is a Kraft summable codelength by virtue of having a codelength associated with a two stage code, the left side also is a Kraft summable codelength. We call a penalty function pen for which one can find a countable subset $\mathcal{F}$ and a penalty V on it satisfying (1) and (2) as *codelength valid*. Arguments in [3] and [4] can be adapted to show the l_1 penalty in linear regression and log density estimation problems is indeed codelength valid. Just showing codelength validity of a penalty is relevant for data compression purposes but not sufficient for generalization guarantees on future data.

The authors in [3] also define a condition for good statistical risk properties to hold. The loss function we consider

978-1-4799-5186-4/14/$31.00 ©2014 IEEE

here is the Bhattacharya divergence between two probability measures. For two probability distributions on $\mathcal{X}^n$, with probability mass functions or densities with respect to some dominating measure given by p and q, the Bhattacharya divergence between p and q is defined as

$$B(p,q) = -\frac{1}{2}\log(E_p\sqrt{\frac{q(X)}{p(X)}}). \tag{3}$$

Let Θ now denote the parameter space which is uncountable. Say our candidate family of distributions with p.m.f's or densities is given by $\{p_\theta : \theta \in \Theta\}$ and the true distribution has p.m.f or density given by $p^\star$. Let pen be a penalty function defined on Θ. The penalized likelihood estimator is now defined as

$$\hat{\theta}(X) = \underset{\theta\in\Theta}{\operatorname{argmin}}\{-\log p_\theta(X) + pen(\theta)\}. \tag{4}$$

Now analogous to (2) if there exists a countable $\mathcal{F} \subset \Theta$ and a codelength $V(\tilde{\theta})$ on $\mathcal{F}$ satisfying $\sum_{\tilde{\theta}\in\mathcal{F}} exp\left(-\frac{V(\tilde{\theta})}{2}\right) \leq 1$ such that the following condition holds:

$$\begin{aligned}&\min_{\tilde{\theta}\in\mathcal{F}}\left(\log\frac{p_{\tilde{\theta}}(\underline{x})}{p^\star(\underline{x})} - B(p^\star, p_\theta) + V(\tilde{\theta})\right) \leq \\ &\min_{\theta\in\Theta}\left(\log\frac{p_\theta(\underline{x})}{p^\star(\underline{x})} - B(p^\star, p_\theta) + pen(\theta)\right)\end{aligned} \tag{5}$$

then, the right side inherits the positive expectation property from the left side in (5) as shown in [3]. Then replacing the minimum over Θ by setting $\theta = \hat{\theta}$ and rearranging, one can conclude, for the estimator given by (4) risk bounds as follows:

$$\mathbb{E}B(p^\star, p_{\hat{\theta}}) \leq \mathbb{E}\min_{\theta\in\Theta}\left(\log\frac{p^\star(\underline{x})}{p_\theta(\underline{x})} + pen(\theta)\right). \tag{6}$$

The above expression is also called the redundancy of the two stage code with respect to the given class of codes or distributions $\{p_\theta : \theta \in \Theta\}$. In i.i.d cases (6) becomes

$$\mathbb{E}B(p^\star, p_{\hat{\theta}}) \leq \mathbb{E}\min_{\theta\in\Theta}\left(\frac{1}{n}\sum_{i=1}^{n}\log\frac{p^\star(\underline{x_i})}{p_\theta(\underline{x_i})} + \frac{pen(\theta)}{n}\right), \tag{7}$$

where n is the sample size and p now refers to probabilty mass function or a density on $\mathcal{X}$ and not $\mathcal{X}^n$. In the i.i.d case, it is clear that the Bhattacharya Divergence between the product distributions on $\mathcal{X}^n$ is n times the divergence between the respective distributions on $\mathcal{X}$.

Remark I.1. *The redundancy which is an expected minimum excess codelength, can be further upper bounded by the minimum expected excess codelength which is called the index of resolvability as in [2].*

$$Res = \min_{\theta\in\Theta}\left(D(p^\star, p_\theta) + pen(\theta)\right).$$

Hence, we have the relation

$$Risk \leq Redundancy \leq Resolvability. \tag{8}$$

In the i.i.d case, by interchanging expectation and minimum in (7) *we get*

$$\mathbb{E}B(p^\star, p_{\hat{\theta}}) \leq \mathbb{E}\min_{\theta\in\Theta}\{D(p^\star, p_\theta) + \frac{pen(\theta)}{n}\}.$$

As we see from the above display, the upper bound of the risk is governed by an ideal tradeoff between Kulback approximation error and complexity. So, in this sense the two stage estimator is adaptive, it looks at the tradeoff between approximation error and complexity at the population size relative to the sample size given.

We call the penalties for which the penalized likelihood procedure gives risk bounds of the form

$$Risk \leq Redundancy$$

as *risk valid* penalties. In particular penalties for which (5) can be verified are risk valid penalties. It is shown in [3] [4] that the l_1 penalty in linear regression and log density estimation problems is indeed risk valid. We add a new example to this story in this paper where we present the risk validity of the l_1 type penalty in Gaussian Graphical Models. Also in this paper, we present a new interpretation of the traditional l_0 penalty in problems such as linear or logistic regression being codelength valid. Traditionally in the MDL literature, the ℓ_0 penalty with a $\log(n)+o(n)$ multiplier has been shown to be codelength valid in linear regression but with the drawback that the $o(n)$ term is unbounded with respect to θ as we go out to the edges of the parameter space. Here we partially resolve this issue as we show that the traditional ℓ_0 penalty plus lower order terms can be indeed thought of as codelength valid in the regime $n > p$ with the lower order term divided by the number of non zero parameters remaining bounded as a function of θ. Also in the linear regression problem, we can leverage the codelength interpretation of the l_0 penalty to derive redundancy risk bounds which is laid out in detail in [8] but is not the focus of the current manuscript.

Let us now make some relevant definitions. We denote the integer lattice in $\mathbb{R}^p$ by Z^p. So Z^p contains all p dimensional vectors z, every coordinate of which are integers. We now define a codelength C on Z_p as follows

$$C(z) = |z|_1\log(4(p+1)) + \log 2. \tag{9}$$

The following lemma says that C defined as in (9) indeed satisfies a Kraft type inequality.

Lemma I.1. *With Z^p being the integer lattice, C as defined in* (9)*, C satisfies the inequality*

$$\sum_{z\in Z^p}\exp(-C(z)) \leq 1. \tag{10}$$

A. Gaussian Graphical Models

Consider the problem of estimating the inverse covariance matrix of a multivariate gaussian random vector. Suppose we observe $\underline{x} = (x_1, \ldots, x_n)$, each of which is a p dimensional vector drawn i.i.d from $N(0, \theta^{-1})$. We denote the true inverse covariance matrix to be $\theta^\star$. Let us denote the $-\log det$ function as ϕ. This ϕ is a convex function on the space of all $p \times p$ matrices with the convention that ϕ takes value $+\infty$ on any matrix that is not positive definite. Inspecting the log likelihood of this model we have

$$\frac{1}{n} \log p_\theta(\underline{x}) = \frac{p}{2} \log(2\pi) + \frac{1}{2} Tr(S\theta) + \frac{\phi(\theta)}{2}$$

Here, $Tr(S\theta)$ is the sum of diagonals of the matrix $S\theta$ and $S = \frac{1}{n} \sum_{i=1}^n x_i^T x_i$ is the sample covariance matrix. In this setting $\theta_{ij} = 0$ means that the ith and jth variables are conditionally independent given the others. We outline the proof of the fact that the penalty $|\theta|_1$, which is just the sum of absolute values of all the entries of the inverse covariance matrix, is a statistical risk valid penalty. The Bhattacharya Divergence between p dimensional multivariate normals with zero means and inverse covariances θ_1 and θ_2 is

$$B(\theta_1, \theta_2) = \frac{1}{2}[\phi(\theta_1) + \phi(\theta_2)] - \phi([\theta_1 + \theta_2]/2).$$

We assume that the truth $\theta^\star$ is sufficiently positive definite in the following way. We assume that for any matrix $\{\Delta : \|\Delta\|_\infty \leq \delta\}$ we have

$$det(\theta^\star + \Delta) > 0. \tag{11}$$

Here $\|\Delta\|_\infty$ means the maximum absolute entry of the matrix Δ and a matrix being $\succ 0$ means it is positive definite. We remark that this is our only assumption on the true inverse covariance and the δ in the assumption is the same δ used in constructing the countable set. The value of δ is specified later. Now we give an idea as to how we verify (5) and establish risk validity of the l_1 penalty times a multiplier. We define $\mathcal{F}$ to be the set of all matrices, which when vectorized in some order, lie in the δ integer lattice, intersected with the space of positive definite symmetric matrices which we denote by S_+^p. So we have

$$\mathcal{F} = \{\delta z \in \mathbb{R}^{p \times p} : vec(z) \in Z^{p^2}, z \in S_+^p\}. \tag{12}$$

Clearly, $\mathcal{F}$ is a countable set. We also define the penalty function V on $\mathcal{F}$ in the following way

$$V(\delta z) = 2C(z). \tag{13}$$

It is clear from lemma (I.1) that V defined as above on $\mathcal{F}$ satisfies the Kraft inequality requirement. To show risk validity of the ℓ_1 penalty we need to show that (5) holds for some multiplier times the $\ell_!$ penalty. An equivalent way to show a penalty pen satisfies (5) is to show for every fixed θ in Θ and data X the penalty pen is not less than the following expression

$$\min_{\tilde{\theta} \in \mathcal{F}} \{\log \frac{p_{\tilde{\theta}}(\underline{x})}{p_\theta(\underline{x})} - B(p_{\theta^\star}, p_{\tilde{\theta}}) + B(p_{\theta^\star}, p_\theta) + V(\tilde{\theta})\}. \tag{14}$$

The expression in (14) is a minimum over the entire δ lattice and hence not lesser than the minimum over the 2^{p^2} vertices of the cube that θ lives in. We further upper bound this minimum by an expectation over these vertices with a particular random choice of vertices in a way such that this random choice is unbiased for θ. Taylor expanding the log likelihoods and the Bhattacharya divergence terms upto the second order permits us to do careful reasoning and obtain that (14) can be upper bounded by

$$16n(\sigma_{max})^2 \delta |\theta|_1 + \frac{2|\theta|_1}{\delta} \log(4(p+1)^2) + 2\log 2$$

where σ_{max} is the maximum diagonal of the true covariance matrix $(\theta^\star)^{-1}$. By setting $\delta^2 = \frac{\log(4(p+1)^2)}{8n(\sigma_{max})^2}$ it follows that by defining the penalty function on Θ defined as follows

$$pen(\theta) = 4\sqrt{\sigma_{max} \log(4(p+1)^2) 2n} |\theta|_1 + 2\log 2 \tag{15}$$

we construct a risk valid penalty. So with the definition of pen above, the estimator defined as follows

$$\hat{\theta} = \underset{\theta \in S_+^p}{\operatorname{argmin}} \left(\frac{1}{2} Tr(S\theta) + \frac{\phi(\theta)}{2} + \frac{pen(\theta)}{n} \right). \tag{16}$$

enjoys the adaptive risk properties we desire. Under the assumption (11) where now δ has been specified, we have the following risk bound

$$\mathbb{E} B(p_{\theta^\star}, p_{\hat{\theta}}) \leq \mathbb{E} \inf_{\theta \in S_+^p} \left(\frac{1}{2} Tr(S(\theta - \theta^\star)) + \right.$$
$$\left. \frac{\phi(\theta) - \phi(\theta^\star)}{2} + \frac{pen(\theta)}{n} \right).$$

By taking the expectation inside the infimum we now present our theorem.

Theorem I.2. *For the estimator $\hat{\theta}$ as in* (16) *with $\hat{\Sigma}^{-1} = \hat{\theta}^{-1}$ and the penalty* (15) *we have the risk bound*

$$\begin{aligned} \mathbb{E} B(p_{\theta^\star}, p_{\hat{\theta}}) \leq \inf_{\theta \in S_+^p} & \left(\frac{1}{2} [Tr(\hat{\theta} \Sigma^\star) - p] + \right. \\ & \left. \frac{1}{2} [\phi(\hat{\theta}) - \phi(\theta^\star))] + \frac{pen(\theta)}{n} \right). \end{aligned} \tag{17}$$

Remark I.2. *By setting $\theta = \theta^\star$ in the right side of the bound, as long as $\theta^\star$ has finite l_1 norm, one has the standard risk bound of the order $\sqrt{\frac{\log(4p^2)}{n}} \|\theta^\star\|_1$. The main purpose of the risk bound is to demonstrate the adaptation properties of the l_1 penalized estimator and to demonstrate redundancy, a coding notion, as the upper bound to the statistical risk which has been championed in [7].*

Remark I.3. *The assumption* (11) *says that the true inverse covariance matrix $\theta^\star$ should be in the interior of the cone of positive definite matrix by a little margin. This assumption may be acceptable even in high dimensions as it does not prohibit collinearity.*

II. Validity of l_0 penalty in Linear Regression

In this section we consider the linear regression setup to show the codelength validity of the l_0 penalty. We consider the known variance σ^2 setup and do our analysis on a fixed design matrix Ψ. Our model is

$$y_{n\times 1} = \Psi_{n\times p}\theta_{p\times 1} + \epsilon_{n\times 1}$$

where $\epsilon \sim N(0, \sigma^2 I_{n\times n})$ and Ψ is the design matrix. Let $X = (y_{n\times 1}, \Psi_{n\times p})$ denote the data. The log likelihood of the model is

$$-\log p_\theta(X) = \frac{1}{2\sigma^2}\|y - \Psi\theta\|_2^2 + \frac{n}{2}\log 2\pi\sigma^2.$$

We assume our model is well specified and there is a true vector of coefficients $\theta^\star$. Our results would be in the regime when the sample size n is larger than the number of explanatory variables p. We divide the data X into $X_{in} = (y_{in}, \Psi_{in})$ consisting of p samples and $X_f = (y_f, \Psi_f)$ consisting of $(n-p)$ samples. Here in is intended to suggest initial and f is intended to mean final. It does not really matter which p samples are chosen to represent the initial sample as long as it is done once and then remains frozen. The purpose of such division of data is to use the initial p samples X_{in} to create a Kraft summable penalty on the countable cover we will choose and then this penalty together with the cover is used to derive codelength interpretation for the ℓ_0 penalized log likelihood. This codelength interpretation can also be leveraged to show risk bounds for the estimator minimizing the ℓ_0 penalized log likelihood. The risk bound calculations can be found in [8].

We now make some relevant definitions and set up some notations. Let $\theta \in \mathbb{R}^p$ be a given vector. We define $k(\theta) = \sum_{i=1}^{p} I\{\theta_i \neq 0\}$. In other words $k(\theta)$ is the number of non zeros of the vector θ. We denote the support of θ or the set of indices where θ is non zero by $S(\theta)$. Clearly $|S(\theta)| = k(\theta)$. Let $S^\star$ be the support of the true vector of coefficients $\theta^\star$. For any subset $S \subset [1:p]$, let $\Psi_{in,S}$ denote the initial part of the design matrix Ψ with column indices in S in natural order. Hence $\Psi_{in,S}$ is a p by $|S|$ matrix. Let us denote the matrix $(\Psi_{in,S}^T\Psi_{in,S})^{-1/2}$ by M_S. For any matrix we use the notation Tr to mean the trace of the matrix. A special role is played by the following quantity $\frac{1}{|S|}Tr\left((\Psi_{in,S}^T\Psi_{in,S})^{-1}(\Psi_{f,S}^T\Psi_{f,S})\right)$ which we denote by Υ_S for any $S \subset [1:p]$.

Let $\mathcal{Z}$ denote the set of integers as before. Also fix some $\delta > 0$. Consider the set $\delta(\mathcal{Z} - \{0\})^m \subset \mathbb{R}^m$ for some positive integer m. It is the set of all m dimensional integer vectors none of whose coordinates are zero. Clearly this set is countable. We denote this set by $\mathcal{G}^m$. For any given subset S we define a countable set

$$C_S = \{M_S v : v \in \mathcal{G}^{|S|}\} \tag{18}$$

As we have defined, C_S is a subset of $\mathbb{R}^{|S|}$ but by appending the coordinates in the complement of S as zeroes, we treat C_S as a subset of $\mathbb{R}^p$. We want to construct Kraft satisfying codelengths and hence subprobabilities on C_S which are proportional to $\left(\frac{(P_\phi(X_{in})}{P_{\theta^\star}(X_{in})}\right)$. For this purpose we want to estimate the normalizer which is the quantity $\sum_{\phi\in C_S}\left(\frac{(P_\phi(X_{in}))}{P_{\theta^\star}(X_{in})}\right)$. The following lemma helps us do exactly that.

Lemma II.1.

$$\sum_{\phi\in C_S}\left(\frac{P_\phi(X_{in})}{P_{\theta^\star}(X_{in})}\right)\delta^{|S|} \leq U(X_{in}, S) \tag{19}$$

where

$$U(X_{in}, S) = \exp\left(\|O_{\Psi_{in,S\cup S^\star}} y_{in} - \Psi_{in,S^\star}\theta^\star\|_2^2\right) \\ (2\pi)^{|S|/2} \tag{20}$$

and $O_{\Psi_{in,S\cup S^\star}}$ denotes the orthogonal projection matrix onto the column space of the matrix $\Psi_{in,S\cup S^\star}$.

We now define the countable set $\mathcal{C} \subset \mathbb{R}^p$ as follows

$$\mathcal{C} = \cup_{k=0}^{p} \cup_{\{S:|S|=k\}} C_S \tag{21}$$

$\mathcal{C}$ is the union of the countable sets $C_{S,\eta}$ over all subsets $S \subset [1:p]$. Hence $\mathcal{C}$ itself is a countable subset of $\mathbb{R}^p$. By definition, $\mathcal{C}$ varies with δ but we dont explicitly write it to minimize notational clutter. We now define penalty functions satisfying Kraft type inequalities on the countable set $\mathcal{C}$. First we define a family of subprobabilities h on $\mathcal{C}$ as follows

$$h(\tilde\theta, X_{in}) = \left(\frac{1}{2}\right)^{k(\tilde\theta)+1}\frac{1}{\binom{p}{k(\tilde\theta)}}\left(\frac{P_{\tilde\theta}(X_{in})}{P_{\theta^\star}(X_{in})}\right)\ \delta^{k(\tilde\theta)} \\ \frac{1}{U(X_{in}, S(\tilde\theta))}. \tag{22}$$

We claim that $h(\tilde\theta)$ is a subprobability on $\mathcal{C}$ for every X_{in}. This can be seen by first summing $h(\tilde\theta)$ over non negative integers k from 0 to p, then summing over all subsets of $[1:p]$ with cardinality k and then summing over C_S. We can now define Kraft satisfying codelengths $l(\tilde\theta, X_{in})$ on $\mathcal{C}$ by defining

$$l(\tilde\theta, X_{in}) = -\log h(\tilde\theta) \tag{23}$$

Then because of h being a subprobability, it is clear that l satisfies (1).

Now we are ready to show that the classical penalty of the order $k(\theta)\log n$ is codelength valid in a certain sense. Let $pen(\theta|X_{in})$ be a penalty function defined on $\Theta = \mathbb{R}^p$ which is a function of X_{in} also. So it is infact a random penalty. The notation is deliberately designed to make the reader think of $pen(\theta|X_{in})$ as a penalty conditional on the initial data X_{in}. Analogous to (2) we intend to show the existence of a countable set $\mathcal{F} \subset \Theta$ and a Kraft valid codelength $V(\tilde\theta|X_{in})$ on $\tilde\Theta$ such that the following inequality holds

$$\min_{\theta\in\Theta}\{-\log P_\theta(X) + pen(\theta|X_{in})\} \geq \\ \min_{\tilde\theta\in\mathcal{F}}\{-\log P_{\tilde\theta}(X_f) + V(\tilde\theta|X_{in})\} \tag{24}$$

where now the right side of (24) gives a two stage codelength interpretation provided we treat it as codelengths on X_f conditional on X_{in} and hence the left side as a function on X_f, being not less than the right side, also has a two stage conditional codelength interpretation. We now proceed to find out a suitable conditional penalty $pen(\theta|X_{in})$ which would satisfy (24).

We declare our countable set $\mathcal{F} = \mathcal{C}$ as defined in (21). We also define our data dependent $V = l$ as defined in (23). Then we have

$$V(\tilde{\theta}|X_{in}) = (k(\tilde{\theta}) + 1)\log(2) + \log\binom{p}{k(\tilde{\theta})} +$$
$$k(\tilde{\theta})\log(\frac{1}{\delta}) + \log(U(X_{in}, S(\theta)) - \log\frac{P_{\tilde{\theta}}(X_{in})}{P_{\theta^\star}(X_{in})}.$$

The task now is to verify (24). An equivalent way to verify (24) is to verify the following for any given $\theta \in \Theta$ and data X,

$$\min_{\tilde{\theta}\in\mathcal{F}}\{-\log\frac{P_{\tilde{\theta}}(X_f)}{P_{\theta^\star}(X_f)} + \log\frac{P_\theta(X)}{P_{\theta^\star}(X)} + V(\tilde{\theta}|X_{in})\} \\ \leq pen(\theta|X_{in}). \tag{25}$$

In the case when X_{in} and X_f are independent, the log likelihood of the full data X is the sum of log likelihoods of X_{in} and X_f and so we can write the left side of the above equation as

$$\min_{\tilde{\theta}\in\mathcal{F}}\{-\log\frac{P_{\tilde{\theta}}(X)}{P_\theta(X)} + \left(V(\tilde{\theta}|X_{in}) + \log\frac{P_{\tilde{\theta}}(X_{in})}{P_{\theta^\star(X_{in})}}\right)\}. \tag{26}$$

Now our strategy to upper bound the minimum of the above expression is to restrict the minimum over $\tilde{\theta} \in C_{S(\theta)}$ where $C_{S(\theta)}$ is as defined in (18). Doing this cannot decrease the overall minimum because $C_{S(\theta)} \subset \mathcal{F}$ by definition of $\mathcal{F}$. Restricted to $\tilde{\theta} \in C_{S(\theta)}$ one can check that the term $V(\tilde{\theta}|X_{in}) + \log\frac{P_{\tilde{\theta}}(X_{in})}{P_{\theta^\star(X_{in})}}$ remains a constant. Now we state a lemma which helps us in upper bounding (26).

Lemma II.2.

$$\min_{\tilde{\theta}\in C_{S(\theta)}}\{-\log\frac{P_{\tilde{\theta}}(X)}{P_\theta(X)}\} \leq 2(1 + \Upsilon_{S(\theta)})\, k(\theta)\delta^2. \tag{27}$$

By the above lemma and the fact that $V(\tilde{\theta}|X_{in}) + \log\frac{P_{\tilde{\theta}}(X_{in})}{P_{\theta^\star(X_{in})}}$ is constant on $C_{S(\theta)}$ we write down the upper bound we get for the left side of (26) which is as follows

$$2(1 + \Upsilon_{S(\theta)})\, k(\theta)\delta^2 + (k(\theta) + 1)\log(2) + \log\binom{p}{k(\theta)} +$$
$$k(\theta)\log(\frac{1}{\delta}) + \log(U(X_{in}, S(\theta)).$$

Setting $\delta^2 = \frac{1}{4(1+\Upsilon_{S(\theta)})}$ we see that a valid penalty satisfying (24) would be

$$pen(\theta|X_{in}) = \frac{k(\theta)}{2} + (k(\theta) + 1)\log(2) +$$
$$\log\binom{p}{k(\theta)} + \frac{k(\theta)}{2}\log(4(1 + \Upsilon_{S(\theta)})) + \log(U(X_{in}, S(\theta)). \tag{28}$$

Rearranging and expanding $U(X_{in}, S(\theta))$ we have

$$pen(\theta|X_{in}) = \frac{k(\theta)}{2}\log(4(1 + \Upsilon_{S(\theta)})) + \log\binom{p}{k(\theta)} +$$
$$k(\theta)\left(\frac{3\log(2)}{2} + \frac{\log(2\pi)}{2}\right)$$
$$+ \frac{1}{2}\|O_{\Psi_{in,S(\theta)\cup S^\star}} y_{in} - \Psi_{in,S^\star}\theta^\star\|_2^2. \tag{29}$$

With a fixed design matrix there is only one term in the above expression which is random. It can be checked that the term $\frac{1}{2}\|O_{\Psi_{in,S\cup S^\star}} y_{in} - \Psi_{in,S^\star}\theta^\star\|_2^2$ is distributed as a χ^2 random variable with degree of freedom at most $k(\theta) + k^\star$. So its expected value is going to be at most $k(\theta) + k^\star$. In the case when the design matrices Ψ_{in} and Ψ_f have orthogonal columns and the ℓ_2 norms of each of the columns of Ψ_{in} and Ψ_f are atmost p and $n - p$ respectively we then have for any subset S, $\Psi_{in,S}^T\Psi_{in,S} = pI_{|S|\times|S|}$ and $\Psi_{f,S}^T\Psi_{f,S} = (n-p)I_{|S|\times|S|}$. In that case it can be checked that $\gamma_S = \frac{n-p}{p}$. Hence in this situation, our codelength valid penalty conditional on X_{in} becomes

$$pen(\theta|X_{in}) = \frac{k(\theta)}{2}\log(\frac{4n}{p}) + \log\binom{p}{k(\theta)} +$$
$$k(\theta)\left(\frac{3\log(2)}{2} + \frac{\log(2\pi)}{2}\right) \tag{30}$$
$$+ \frac{1}{2}\|O_{\Psi_{in,S(\theta)\cup S^\star}} y_{in} - \Psi_{in,S^\star}\theta^\star\|_2^2.$$

Note that the leading term of the expected penalty $pen(\theta|X_{in})$ is indeed going to be the traditional $\frac{\log(n)}{2}k(\theta)$ in case p does not grow with n. In case p grows as n^β for some $0 < \beta < 1$ then the leading term of of the expected penalty $pen(\theta|X_{in})$ is still some constant times $k(\theta)\log(n)$. Complete proofs and detailed discussions are in the full paper [8].

REFERENCES

[1] T. M. Cover and J. A. Thomas, "Elements of Information Theory," Second Edition, John Wiley and Sons, Inc., Hoboken, NJ, USA

[2] A.R. Barron and T.M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inform. Theory*. Vol.37, No.4, pp.1034–1054. 1991.

[3] A.R. Barron, C. Huang, J.Q. Li, X. Luo, "The MDL principle, penalized likelihoods, and statistical risk," In *Festschrift for Jorma Rissanen*. Tampere University press, Tampere, Finland, 2008.

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

[5] A.R. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inform. Theory*. Vol.44, No.6, pp.2743–2760. 1998. Special Commemorative Issue: Information Theory: 1948-1998.

[6] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by probability distributions," *Bull. Calcutta Math. Soc.* Vol.35, pp.99–109. 1943.

[7] P. Grünwald, *The Minimum Description Length Principle*. Cambridge, MA, MIT Press. 2007.

[8] S. Chatterjee A.R. Barron, "Information Theory of Penalized Likelihoods and its Statistical Implications," 2014. Available at www.stat.yale.edu arb4