Information Theory of Penalized Likelihoods and its Statistical Implications.

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*—We extend the correspondence between two-stage coding procedures in data compression and penalized likelihood procedures in statistical estimation. Traditionally, this had required restriction to countable parameter spaces. We show how to extend this correspondence in the uncountable parameter case. Leveraging the description length interpretations of penalized likelihood procedures we devise new techniques to derive adaptive risk bounds of such procedures. We show that the existence of certain countable coverings of the parameter space implies adaptive risk bounds and thus our theory is quite general. We apply our techniques to illustrate risk bounds for ℓ_1 type penalized procedures in canonical high dimensional statistical problems such as linear regression and Gaussian graphical Models. In the linear regression problem, we also demonstrate how the traditional l_0 penalty times $\frac{\log(n)}{2}$ plus lower order terms has a two stage description length interpretation and present risk bounds for this penalized likelihood procedure.**

I. Introduction

There are close connections between good data compression and good estimation in statistical settings. Shannon's recipe for finding the minimum expected codelength when we know the data generating distribution shows the correspondence between probability distributions on data and optimal codelengths on the sample space. Also, Kraft's inequality stating that for every probability mass function there exists a prefix free code with lengths negative log probability gives an operational meaning to probability. The Kraft's inequality allows one to think of prefix free codes and probabilities interchangeably. The MDL principle has further developed this connection by considering the case where we do not necessarily know the data generating distribution. From now on, codes are always meant to be prefix free. In this MDL framework one considers a family of codes or equivalently a set of probability sources, possibly indexed by a parameter space Θ. One codes the observed data by using one of the codes in the family considered. The idea in one shot data compression is to compress the observed data sequence well. But for statistical purposes, we want to devise a coding or estimation strategy based on the observed data that should compress or predict well for future data assumed to be arising from the same generating distribution.

A fundamental concept in the MDL philosophy is that of universal coding or modelling. The aim of universal coding or modelling is find a single code u that allows us to compress data almost as well as the best code in our class of codes Θ either in expectation or high probability with respect to the generation of the data X. This universal distribution can be constructed mainly in four different ways as described in [4]. These four ways can be categorized as Two-stage codes, Bayes mixture codes, Predictive codes and Normalized Maximum Likelihood codes. In the present manuscript we will focus on the Two-stage coding procedure being brought into statistical play.

One of the earliest ways to build a universal code is to build what is called a two stage code **cite rissanen,barron and cover**. The basic idea is to first devise a code or description of all the possible codes in Θ. Also for each possible code one encodes or describes the data using that code. Then one chooses the code which minimizes the sum of the two descriptions, one describing the code and the other describing data given the code. Now one can play the same game in the learning setup where now the codes are replaced by a family pf probability sources and the estimated source from the data is the that minimized the sum of the two descriptions. This is indeed the penalized likelihood estimator where the penalty corresponds to the description lengths of probability distributions in our model. Traditionally, the statistical properties of this penalized likelihood procedure has been studied in the countable parameter space setting. Past work as in **cite li,barron and ??** shows how the expected pointwise redundancy controls the statistical risk in countable parameter spaces. **Describe pointwise redundancy**.

One of the main contributions of this present manuscript is to extend such risk bounds when the parameter space is uncountable maintaining the description length interpretation.

The main idea here is to construct countable subcovers of the parameter space Θ and leverage the results from the countable case. These subcovers are variable in size and are constructed according to the interaction of the negative log likelihood and the penalty as is made clear in section (II-B). In this way these covers are different from traditional metric entropy covers. This is becasue they do not necessarily arise from a metric and instead, they are based on what is essential, the negative log likelihood and the penalty. We show that the loss function we consider, is not much more than the pointwise redundancy, both in expectation and with high probability. As we would see, these risk bounds that we get also reveal the adaptation properties of these penalized likelihood procedures.

The main idea is to propose conditions on the penalty and the negative log likelihood in a general setting to derive adaptive risk bounds of the form as long as the penalized likelihood estimator mirrors the construction of a two stage code. In a preliminary form, this idea has appeared in **cite Barron,Luo conference papers**. This paper lays out this general theory in more detail and then show that our conditions are satisfied and our risk bounds are valid in canonical high dimensional statistical problems such as linear regression and inverse covariance estimation in Gaussian models.

In section (II) we describe the general technique of how to interpret penalized negative log likelihoods as two stage description lengths in the uncountable parameter situation. We also lay out the general strategy for deriving adaptive risk bounds whenever the codelength interpretation holds. In this section, we describe the conditions needed on the penalty and the negative log likelihood which allows us to prove risk bounds. In section (III) we apply our theory to the ℓ_1 penalty in linear regression case and fully illustrate different ways of verifying the conditions we need. We then present a new result on Inverse covariance matrix estimation in a multivariate Normal setting which shows that our theory can handle not just location type problems but scale problems as well. In section (IV) we then turn our attention to the ℓ_0 penalty in the linear regression case. We devise a new way to interpret the ℓ_0 penalty times a $\log(n)/2$ factor as Kraft satisfying codelengths and leverage this interpretation to recover adaptive risk bounds. Thus the penalties we consider in this manuscript are traditionally the two most commonly used in statistics, namely the ℓ_0 penalty or the number of parameters times a suitable multiplier and ℓ_1 type penalties with suitable multipliers.

A. *Notational Conventions*

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 $\underline{X}$. A probabilistic source p is a sequence of probability distributions $p^{(1)}, p^{(2)}, \ldots$ on $\mathcal{X}^1, \mathcal{X}^2, \ldots$ 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 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. Generically we denote a countable parameter space by $\mathcal{F}$ and an uncountable parameter space by Θ. We will also 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.

II. General Technique

In this section, we first propose a way to extend the interpretation of the penalized log likelihood expression as akin to a two-stage codelength in the case when the parameter space is uncountable. Then we show how our proposed extension also helps us derive adaptive risk bounds for the penalized likelihood procedures.

A. *Codelength validity*

First, let us describe the two-stage code in the case when we have a countable parameter space. Let the parameter space $\mathcal{F}$ be countable, and V be a penalty function on $\mathcal{F}$ satisfying Kraft's inequality $\sum_{\tilde{\theta}\in\mathcal{F}} exp(-V(\tilde{\theta})) \leq 1$. Then the total two stage description length l is as follows

$$l(\underline{x}) = \min_{\tilde{\theta}\in\mathcal{F}} \left(-\log p_{\theta}(\underline{x}) + V(\tilde{\theta})\right). \qquad (1)$$

As one can notice, codelengths l are a sum of two description lengths; description of the parameter space by V and the description of the data given the parameter by $-\log p_{\theta}(\underline{x})$. For countable sample space $\mathcal{X}$, negative log being a Kraft satisfying codelength is immediate. The extension to uncountable sample spaces can be carried out and is described in remark in the appendix . Henceforth, even for uncountable sample spaces we regard negative log density as codelengths satisfying Kraft's inequality.

By the above description, the two stage procedure is clearly uniquely decodable as first one can decode the code used and then using that code decode the data. Hence by standard results in Information Theory **cite cover and thomas** there exists a prefix free code on Ω^n with codelengths l. Thus the codelengths l satisfy Kraft's inequality. 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 $\mathcal{F}$ such that the following holds

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

In this case the right side of the above display will satisfy Kraft's inequality by virtue of being a two-stage codelength on the countable set $\mathcal{F}$. Then the left side of the last display being not less than the right side also satisfies Kraft's inequality. So the upshot is, that for an uncountable parameter space Θ and a penalty function pen, as long as one verifies (2), one can assert that the following codelengths on Ω^n

$$l(\underline{x}) = \min_{\theta \in \Theta}\{-\log p_\theta(\underline{x}) + pen(\theta)\} \tag{3}$$

satisfy Kraft's inequality and hence again correspond to a prefix free code. In this way we link the countable and the uncountable cases. For a penalty function pen on Θ if there exists a countable F and Kraft satisfying V defined on $\mathcal{F}$ satisfying (2) then we say pen is a *codelength valid* penalty.

Remark II.1. *The condition* (2) *can also be equivalently restated as the following: There exists a countable subset $\mathcal{F} \subset \Theta$ and any Kraft summable penalty $V(\tilde{\theta})$ on $\mathcal{F}$ such that for every $\theta \in \Theta$ there exists a $\tilde{\theta} \in \mathcal{F}$ satisfying*

$$\{-\log p_\theta(\underline{x}) + pen(\theta)\} \geq \{-\log p_{\tilde{\theta}}(\underline{x}) + V(\tilde{\theta})\}. \tag{4}$$

In other words, for every $\theta \in \Theta$ there exists its representer in $\mathcal{F}$ satisfying the above inequality. In this sense, $\mathcal{F}$ is indeed a cover for Θ. This cover is built out of the interaction of the negative log likelihoods, the penalty pen and the function V which is indeed a theoretical construct but in our applications would be very closely related to pen.

Remark II.2. *Discuss the adaptation issue in universal data compression.*

B. Risk Validity

Now we demonstrate how to derive adaptive risk bounds for penalized likelihood procedures.

1) Countable parameter Space: First we consider the countable parameter space case. Let $\mathcal{F}$ be a countable parameter space and $\{p^{(n)}_{\tilde{\theta}} : \tilde{\theta} \in \mathcal{F}\}$ denoting probability mass functions or densities on $\mathcal{X}^n$ with respect to some dominating measure be our model. Let V be a penalty function on Θ. We want to investigate the statistical risk properties of the following penalized log likelihood estimator

$$\hat{\theta}(\underline{x}) = \operatorname*{argmin}_{\theta \in \mathcal{F}} \Big(-\log(p^{(n)}_\theta(\underline{x}) + V(\theta)\Big) \tag{5}$$

For any $0 < \alpha \leq 1$, we define a family, indexed by α, of loss functions between two probability measures $p^{(n)}$ and $q^{(n)}$ on Ω^n by

$$L_\alpha(p^{(n)}, q^{(n)}) = -\frac{1}{\alpha}\log \mathbb{E}_{p^{(n)}}\left(\frac{q^{(n)}(X)}{p^{(n)}(X)}\right)^\alpha. \tag{6}$$

We note that in the case $p^{(n)}$ is a n fold i.i.d copy of $p^{(1)}$ and same for the probability source q then we have for all $0 < \alpha < 1$,

$$L_\alpha(p^{(n)}, q^{(n)}) = nL_\alpha(p^{(1)}, p^{(1)}). \tag{7}$$

In the literature, these are sometimes known as the Chernoff-Renyi divergences between probability measures. In the remaining part of this subsection, probability densities are understood to be defined on $\mathcal{X}^n$ and hence we drop the subscript n in this section to minimize notational clutter. Also $\mathbb{E}$ would mean taking expectation with respect to the distribution of the data unless sepcified otherwise.

Remark II.3. *L_α is not symmetric in general. However, it is symmetric when $\alpha = \frac{1}{2}$. In that case $L_{\frac{1}{2}}$ turns out to be the familiar Bhattacharya distance between two probability measures.*

Remark II.4. *The Hellinger loss between two probability distributions p and q is given by*

$$H^2(p, q) = \mathbb{E}_p(\sqrt{p(X)} - \sqrt{q(X)})^2.$$

One can check that $L_{1/2}(p, q) = -2\log(1 - \frac{1}{2}H^2(p, q)$. In particular we have that the Bhattacharya distance is a monotonic transformation of the Hellinger distance. Also, by properties of logarithms, we do have $L_{1/2}(p, q) \leq H^2(p, q)$. The familiar Kulback Leibler divergence D between p and q is defined to be

$$D(p, q) = \mathbb{E}_p \log(\frac{p(X)}{q(X)}).$$

By Jensen's inequality one can check that $L_{1/2}(p, q) \leq D(p, q)$. In fact when the log likelihood ratios of p and q are bounded by constants then $L_{1/2}$ is within a constant factor of $D(p, q)$.

Remark II.5. *An upper bound to L_α means an upper bound to $L_{1/2}$.*

Remark II.6. *In case p and q are multivariate normals with mean vectors μ_1 and μ_2 and covariance matrices Σ_1 and Σ_2 respectively, our loss function evaluates to the following expression*

$$L_\alpha(p,q) = \frac{1-\alpha}{2}(\mu_1-\mu_2)^T(\alpha\Sigma_1+(1-\alpha)\Sigma_2) \\ (\mu_1-\mu_2) + \frac{1}{2\alpha}\log\frac{det(\alpha\Sigma_1+(1-\alpha)\Sigma_2)}{det(\Sigma_1)^\alpha det(\Sigma_2)^{1-\alpha}}. \tag{8}$$

In case the covariance matrices are the same and identity then it is proportional to the ℓ_2 squared norm between the mean vectors.

Now we state a lemma:

Lemma II.1. *Let the true distribution generating the data X be denoted by $p^\star$. For the model $\{p_\theta : \theta \in \mathcal{F}\}$ and the penalized likelihood estimator defined as in* (5)*, if the penalty function satisfies a slightly stronger Kraft type inequality as follows,*

$$\sum_{\theta\in\mathcal{F}} \exp(-\alpha\, pen(\theta)) \leq 1 \tag{9}$$

where $0 < \alpha \leq 1$ is any fixed number, we have the following moment generating inequality:

$$\mathbb{E}\exp\left(\alpha \max_{\theta\in\mathcal{F}}(L_\alpha(p^\star,p_\theta) - \log(\frac{p^\star(X)}{p_\theta(X)}) \\ -pen(\theta)\right) \leq 1. \tag{10}$$

Proof: By positivity of the exponential function and the by monotonicity and linearity of expectation we have

$$\mathbb{E}\exp\left(\alpha \max_{\theta\in\mathcal{F}}\{L_\alpha(p^{(n)},p_\theta^{(n)}) - \log(\frac{p(X)}{p_\theta(X)}) \\ -pen(\theta)\}\right) \leq \sum_{\theta\in\mathcal{F}} \mathbb{E}\exp\left(\alpha(L_\alpha(p,p_\theta) - \\ \log(\frac{p(X)}{p_\theta(X)}) - pen(\theta)\right).$$

The right side of the above inequality can be rewritten as

$$\sum_{\theta\in\mathcal{F}} \exp(\alpha L_\alpha(p^\star,p_\theta))\mathbb{E}(\frac{p_\theta(X)}{p^\star(X)})^\alpha \\ \exp(-\alpha\, pen(\theta)). \tag{11}$$

By the definition of the loss function (7) the above simplifies to

$$\sum_{\theta\in\mathcal{F}} \exp(-\alpha\, pen(\theta)) \tag{12}$$

and the summability condition (9) implies that the above display is not greater than 1. This completes the proof of lemma (II.1). ∎

Theorem II.2. *Under the same conditions as in lemma* (II.1) *we have the following risk bound:*

$$\mathbb{E}L_\alpha(p^\star,p_{\hat\theta}) \leq \mathbb{E}\inf_{\theta\in\mathcal{F}}\left(\log\frac{p^\star(X)}{p_\theta(X)} + pen(\theta)\right). \tag{13}$$

Proof: Interchanging $\mathbb{E}_p$ and the exponential cannot increase the left side of equation (10) *so we have the inequality*

$$\exp(\alpha\,\mathbb{E}\max_{\theta\in\mathcal{F}}\{L_\alpha(p^\star,p_\theta) - \log(\frac{p^\star(X)}{p_\theta(X)}) - \\ pen(\theta)\}) \leq 1.$$

Monotonicity of the exponential function and $\alpha > 0$ implies

$$\mathbb{E}\max_{\theta\in\mathcal{F}}\{L_\alpha(p,p_\theta) - \log(\frac{p(X)}{p_\theta(X)}) - pen(\theta)\} \leq 0.$$

Setting $\theta = \hat\theta$ in the left side of the above equation cannot increase it and hence we have

$$\mathbb{E}\{L_\alpha^{(n)}(p^{(n)},p_{\hat\theta}^{(n)}) - \log(\frac{p(X)}{p_{\hat\theta}(X)}) - pen(\theta)\} \leq 0.$$

Taking the loss term on the other side and multiplying by -1, we get the desired risk bound by recalling the definition of $\hat{\hat\theta}$. This completes the proof of theorem (II.2). ∎

2) Extension to Uncountable parameter Spaces: The previous argument only works for countable parameter spaces. This is because we cannot take a sum over uncountable possibilities as in the first step of the proof of lemma (10). In statistical applications, the estimators are optimized over continuous spaces and it is awkward to force an user to construct countable discretizations of the parameter space. In this section we show how to extend the idea of the previous section to obtain risk bounds for estimators minimizing negative log likelihood plus a penalty term over uncountable choices. We identify conditions on the penalty pen and the log likelihood in order to be able to mimic the countable case and derive risk bounds. Let Θ now denote the parameter space which is uncountable. Let pen be a penalty function defined on Θ. The penalized likelihood estimator is now defined as

$$\hat\theta(X) = \operatorname*{argmin}_{\theta\in\Theta}\left(-\log(p_\theta(X) + pen(\theta)\right). \tag{14}$$

Analogous to (2) let us assume the existence a countable subset $\mathcal{F} \subset \Theta$ and a penalty function V on $\mathcal{F}$ such that the following holds for any fixed $0 < \alpha < 1$

$$\max_{\tilde\theta\in\mathcal{F}}\left(L_\alpha(p^\star,p_{\tilde\theta}) - \log\frac{p^\star(X)}{p_{\tilde\theta}(X)} - V(\tilde\theta)\right) \geq \\ \max_{\theta\in\Theta}\left(L_\alpha(p^\star,p_\theta) - \log\frac{p^\star(X)}{p_\theta(X)} - pen(\theta)\right). \tag{15}$$

Also analogous to (9) let us assume V satisfies a similar inequality on F

$$\sum_{\tilde{\theta}\in\mathcal{F}} \exp(-\alpha\, V(\tilde{\theta})) \le 1. \tag{16}$$

We now state the following theorem for the uncountable parameter case.

Theorem II.3. *We again denote the true distribution generating the data X by $p^\star$. For the model $\{p_\theta : \theta \in \Theta\}$, if the assumptions* (15) *and* (16) *are met then we have the desired risk bound for the estimator* (14) *as follows*

$$\mathbb{E}L_\alpha(p^\star, p_{\hat{\theta}}) \le \mathbb{E}\inf_{\theta\in\Theta}\left(\log\frac{p^\star(X)}{p_{\hat{\theta}}(X)} + pen(\theta)\right). \tag{17}$$

Proof: Since V satisfies (9) *on F which is countable by lemma* (II.1) *we obtain*

$$\mathbb{E}\exp(\alpha\max_{\tilde{\theta}\in\mathcal{F}}\{L_\alpha(p^\star, p_{\tilde{\theta}}) - \log(\frac{p(X)}{p_{\tilde{\theta}}(X)}) - V(\tilde{\theta})\}) \le 1.$$

By assumption (15) *and monotonicity of the exponential function and the expectation operator we therefore have our moment generating inequality*

$$\mathbb{E}\exp(\alpha\max_{\theta\in\Theta}\{L_\alpha(p^\star, p_\theta) - \log(\frac{p^\star(X)}{p_\theta(X)}) - pen(\theta)\}) \le 1. \tag{18}$$

Again by interchanging exponential and expectation and then by the monotonicity of the exponential function we have the following

$$E\max_{\theta\in\Theta}(L_\alpha(p^\star, p_\theta) - \log\frac{p(X)}{p_\theta(X)} + pen(\theta)) \le 0.$$

By setting $\theta = \hat{\theta}(X)$ we cannot increase the expectation and hence we have

$$E_p\left(L_\alpha^{(n)}(p, p_{\hat{\theta}}) - \log\frac{p(X)}{p_{\hat{\theta}}(X)} + pen(\hat{\theta})\right) \le 0.$$

Taking the log term and the penalty term on the right side and recalling the definition of $\hat{\theta}$ we obtain the desired risk bound. This completes the proof of theorem (II.3). ∎

For a penalty function pen on Θ if there exists a countable F and a penalty function V defined on $\mathcal{F}$ satisfying (15) and (16) then we say pen is a *risk valid penalty*.

Remark II.7. *The condition* (15) *is very similar to* (2) *with the loss terms added. Condition* (15) *can be interpreted in another way which is going to be sometimes more convenient for us. For a penalty pen defined on Θ to be valid for redundancy risk bounds such as* (17), *condition* (15) *behooves us to find a countable $\mathcal{F} \subset \Theta$ and a penalty V defined on $\mathcal{F}$ satisfying Kraft* (9) *such that for any given $\theta \in \Theta$ and any given data point X, we have the following inequality*

$$\min_{\tilde{\theta}\in\mathcal{F}}(L_\alpha(p^\star, p_\theta) - L_\alpha(p^\star, p_{\tilde{\theta}}) + \log\frac{p^\star(X)}{p_{\tilde{\theta}}(X)} - \log\frac{p^\star(X)}{p_\theta(X)} + V(\tilde{\theta})) \le pen(\theta). \tag{19}$$

Consequently, for every $\theta \in \Theta$ there must exist a representer $\tilde{\theta} \in \mathcal{F}$ such that the left side of the above equation without the minimum is less than $pen(\theta)$. This representer may also depend on the data X. In this sense, again $\mathcal{F}$ is a cover for Θ built out of differences of log likelihoods and the penalty function pen. Existence of such a cover allows us to mimic the countable parameter space situation and lets us prove desired risk bounds.

Remark II.8. *Note that we have a variety of risk bounds with loss functions L_α parametrized by $0 < \alpha \le 1$. Note also that our requirement on the penalty also changes with α. Also $p^\star$ denotes the true data generating probability measure which need not be in the model we consider for our risk bounds to be valid.*

We now again include the subscript n when writing probability mass functions or densities. An important case is when the data is i.i.d, that is when the data generating distribution p^n is the n fold product of a distribution p^1 on $\mathcal{X}$ and the model consists of $\{p_\theta^n : \theta \in \Theta\}$ where p_θ^n refers to the n fold product of $p_\theta^{(1)}$. In this setting, as can be readily checked from (7), we have for two probability sources p and q

$$L_\alpha(p^n, q^n) = nL_\alpha(p^1, q^1). \tag{20}$$

In this case we write our risk bound in the following corollary.

Corollary II.4. *Let the model be consisting of n fold products of $\{p_\theta^{(1)} : \theta \in \Theta\}$ and the true data generating distribution be the n fold product of $p^\star$. In this case, under the same assumptions as in theorem* (II.2) *we have the risk bound for all $0 < \alpha \le 1$,*

$$\mathbb{E}L_\alpha(p^\star, p_{\hat{\theta}}) \le \mathbb{E}\inf_{\theta\in\Theta}(\frac{1}{n}\log\frac{p^n(X)}{p_\theta^n(X)} + \frac{pen(\theta)}{n}). \tag{21}$$

Proof: The proof follows by dividing throughout by n in equation (17) and because we are in the i.i.d setting. ∎

We note that by interchanging expectation and infimum in the right side of the risk bound in the last display we have

$$\mathbb{E}L_\alpha(p^\star, p_{\hat{\theta}}) \le \inf_{\theta\in\Theta}\left(D(p^\star, p_{\hat{\theta}}) + \frac{pen(\theta)}{n}\right). \tag{22}$$

The right side in the last display is called the index of resolvability as in []. As it can be seen, the index of resolvability is an ideal tradeoff between the KL approximation and the penalty or the complexity relative to the sample size. In this way the risk of the penalized likelihood estimator adapts to subclasses in the parameter space with varying complexity or penalty level.

So far we have provided finite sample upper bounds for the expected loss. In case of i.i.d data finite sample high probability upper bounds are also readily available for the loss.

Corollary II.5. *In case of i.i.d data we have the probability of the event that the loss exceeds the redundancy by a positive number $\tau > 0$ is exponentially small in n. We have the following inequality*

$$P(L_\alpha(p, p_{\hat{\theta}}) > \frac{1}{n}\sum_{i=1}^{n}\log(\frac{p^\star(x_i)}{p_{\hat{\theta}}(x_i)}) + \frac{pen(\hat{\theta})}{n\alpha} + \tau) < e^{-n\alpha\tau}. \tag{23}$$

Proof: We take equation (18) as our starting point. In the i.i.d setting we can rewrite it as

$$\mathbb{E}\exp(n\alpha\max_{\theta\in\Theta}\{L_\alpha(p^\star, p_\theta) - \sum_{i=1}^{n}\log(\frac{p^\star(x_i)}{p_{\hat{\theta}}(x_i)}) - \frac{pen(\theta)}{n}\}) \leq 1.$$

By setting $\theta = \hat{\theta}$ the above equation implies

$$\mathbb{E}\exp(n\alpha\{L_\alpha(p^\star, p_{\hat{\theta}}) - \sum_{i=1}^{n}\log(\frac{p^\star(x_i)}{p_{\hat{\theta}}(x_i)}) - \frac{pen(\hat{\theta})}{n}\}) \leq 1.$$

Let τ be any positive number. By applying Markov's inequality and the previous equation we complete the proof of this corollary. ■

Remark II.9. *$p^\star$ denotes the true data generating probability measure which need not be in the model we consider for our risk bounds to be valid.*

Remark II.10. *In order to apply theorem* (II.2) *to particular models, we need to be able to check condition* (15) *which means we have to come up with a choice of a countable subset $\mathcal{F} \subset \Theta$ and a penalty function V defined on $\mathcal{F}$ satisfying* (16). *We will show in the coming sections how to demonstrate that these conditions hold in canonical high dimensional parametric problems such as Linear Models and Gaussian Graphical Models with the penalty being a suitable multiple of the l_1 penalty. We will also show how to use theorem* (II.2) *to obtain adaptive risk bounds for a suitable multiplier times the l_0 penalty in the Linear model case in the regime where $n > p$. Our aim is to demonstrate that the existence condition of countable covers of the parameter space that we have proposed are natural and are satisfied for the canonical problems we consider in high dimensional statistics.*

III. Validity of the l_1 penalty

In this section we show that a certain weighted $\ell_!$ type penalty with a suitable multiplier is codelength valid and risk valid in the linear regression problem. We also show that the ℓ_1 penalty is risk valid in the setting of Gaussian graphical models. We essentially verify conditions (15) and (16) in both these models. Our point is to convince the reader that our conditions are indeed satisfied in these canonical problems.

A. Linear Regression

To illustrate our techniques of obtaining adaptive risk bounds we first choose the setting of linear regression which is one of the canonical location problems in statistics. We have a real valued response variable y and a vector valued predictor vector x. We assume y conditional on x is Gaussian with conditional mean function $f^\star(x)$ and known variance σ^2. We are given n realizations $\{(y_i, x_i)\}_{i=1}^{n}$ from the joint distribution of (y, x). The goal in this setting might be to estimate this unknown $f^\star$ as that completely specifies the conditional density of y given x under the Gaussian assumption. What we would do in this problem is given the predictor variables $\{(x_i)\}_{i=1}^{n}$ treat this as a conditional density estimation problem, define an appropriate ℓ_1 penalized likelihood estimator, apply our theory, check our conditions and get risk bounds, conditional on all the predictor values. At the end we can always take a final expectation with respect to the distribution of $\{(x_i)\}_{i=1}^{n}$ to get our final risk bounds. So we now treat the predictor values $\{(x_i)\}_{i=1}^{n}$ as given and do a conditional analysis. We assume that we have a dictionary $\mathcal{D}$ of fixed functions $\{f_j\}_{j=1}^{p}$ where p could be very large compared to n. The dictionary could have been obtained from a previous training sample or otherwise. We restrict attention to estimators of the conditional mean function, which take the form of a data dependent linear combination of the functions $f \in \mathcal{D}$. In other words, our estimators would be a member of the set

$$\{f : f = \sum_{j=1}^{p}\theta_j f_j\}$$

where $\theta = (\theta_1, \ldots, \theta_p) \in \mathbb{R}^p$. Hence our parameter space Θ could be identified with $\mathbb{R}^p$. For any $\theta \in$
R^p we denote the function $f = \sum_{j=1}^{p}\theta_j f_j$ by f_θ. Now

we proceed to show risk validity of a certain weighted ℓ_1 penalty. We would need to define a countable set $\mathcal{F} \in \Theta$ and a penalty function V satisfying (16) defined on $\mathcal{F}$ such that equation (19) holds. What we essentially do to define our penalty is upper bound the left side of equation (19) and define the upper bound itself as the penalty, thus automatically satisfying (19). Our loss functions between conditional densities with means f_θ and $f_{\theta'}$ are denoted by $L_\alpha(\theta, \theta')$ for notational simplicity. The loss functions L_α, as can be checked from (8) turn out to be

$$L_\alpha(\theta, \theta') = \frac{(1-\alpha)\sigma^2}{2} \sum_{i=1}^{n} (f_\theta(x_i) - f_{\theta'}(x_i))^2. \tag{24}$$

Now we proceed to verify (19). Expanding the left side of (19) in this setting translates to the following

$$\begin{aligned} &\min_{\tilde{\Theta} \in \mathcal{F}} [\frac{(1-\alpha)}{2\sigma^2} \sum_{i=1}^{n} \big((f_\theta(x_i) - f^\star(x_i))^2 \; - \\ &(f_{\tilde{\theta}}(x_i) - f^\star(x_i))^2\big) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} \big((y_i - f_{\tilde{\theta}}(x_i))^2 \\ &- (y_i - f_\theta(x_i))^2\big) + V(\tilde{\theta})] \end{aligned} \tag{25}$$

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

$$C(z) = |z|_1 \log(4p) + \log 2. \tag{26}$$

The following lemma shows that C defined as in (26) indeed satisfies a Kraft type inequality.

Lemma III.1. *With Z^p being the integer lattice, C as defined in* (26)*, C satisfies the inequality*

$$\sum_{z \in Z^p} \exp(-C(z)) \leq 1. \tag{27}$$

The proof of this lemma is given in appendix.

By expanding (25) we see that we have to minimize the following expression $H_\theta(\tilde{\theta})$ over a countable set $\mathcal{F}$ and a penalty function V to be defined momentarily.

$$\begin{aligned} &H_\theta(\tilde{\theta}) = \frac{\alpha}{2\sigma^2} \sum_{i=1}^{n} (f_\theta(x_i) - f_{\tilde{\theta}}(x_i))^2 - \\ &\frac{\alpha}{\sigma^2} \sum_{i=1}^{n} (y_i - f_\theta(x_i))(f_{\tilde{\theta}}(x_i) - f_\theta(x_i)) - \\ &\frac{(1-\alpha)}{\sigma^2} \sum_{i=1}^{n} (f^\star(x_i) - f_\theta(x_i))(f_{\tilde{\theta}}(x_i) - f_\theta(x_i)) \\ &+ V(\tilde{\theta}). \end{aligned} \tag{28}$$

Our strategy is to upper bound the minimum of $H_\theta(\tilde{\theta})$ the above expression by an expectation over a carefully chosen distribution μ on $\mathcal{F}$. Let $\tilde{\theta} \in \mathcal{F}$ be random and distributed according to μ. The minimum is always less than or equal to an average so we have for any distribution μ on $\mathcal{F}$ the following

$$\min_{\tilde{\theta} \in \mathcal{F}} H_\theta(\tilde{\theta}) \leq \mathbb{E}_{\tilde{\theta} \sim \mu} H_\theta(\tilde{\theta})$$

We will now show how to choose this distribution μ. We would arrange for the mean of μ to be θ. Consequently, the cross product terms in would be zero on an average. So the terms we would have to control in (28) are the averages of the quadratic term and the penalty term. We introduce some more notations. If we denote the design matrix by Ψ, where $\Psi_{ij} = f_i(x_j)$, then we define weights $\{w_j\}_{j=1}^p$ as follows

$$w_j = \frac{1}{n} (\Psi^T \Psi)_{jj} \tag{29}$$

The weight vector w is nothing but the empirical ℓ_2 norms of the columns of the design matrix Ψ. For any vector $v \in \mathbb{R}^p$ we denote its weighted ℓ_1 norm as

$$|v|_{1,w} = \sum_{j=1}^{p} w_j |v_j|.$$

Now we illustrate how to sample $\tilde{\theta}$ in $\mathcal{F}$ and hence define a distribution μ on $\mathcal{F}$ for purposes elicited above. We will actually show how to do the above in two other ways in the appendix, which are interesting by themselves. Along the way we will define the countable set $\mathcal{F}$ and the penalty V we would be working with.

1) Sampling method: We now show a way of devising a probability distribution on the countable set $\mathcal{F}$ so that the average of H_θ upper bounds the minimum of it over $\mathcal{F}$ and helps us set a penalty which allows for our desired adaptive risk bounds. We now define our countable set $\mathcal{F}$. Let W denote the diagonal matrix of weights as defined in (29). We define the set $\mathcal{F}$ as follows

$$\mathcal{F} = \delta\{W^{-1}z : z \in Z^p\} \tag{30}$$

Clearly $\mathcal{F}$ is countable since Z^p is so. A careful observation shows that h always takes values in $\mathcal{F}$. We now define a penalty function V on $\mathcal{F}$ derived from C. So we define V in the following manner

$$V(\delta W^{-1} z) = \frac{C(z)}{\alpha}$$

for all vectors z in Z_p. It is clear from (9) that V satisfies the Kraft inequality (16). So the symbol $\mathbb{E}$ would now mean expectation with respect to the distribution of $\bar{h}$ on $\mathcal{F}$. Let $\theta \in \mathbb{R}^p$ be given and $\delta > 0$ be a given number. We can always write θ in the following way

$$\theta = \delta(\frac{m_1}{w_1}, \dots, \frac{m_p}{w_p})$$

for some vector $(m_1, \ldots, m_p)$. We now describe our sampling strategy. For any integer $1 \leq l \leq p$ we define a random variable $\mathbf{h}_l$ in the following way.

$$
\begin{aligned}
\mathbf{h}_l &= \frac{\delta}{w_l} \lceil m_l \rceil \text{ with probability } (\lceil m_l \rceil - m_l) \\
&= \frac{\delta}{w_l} \lfloor m_l \rfloor \text{ with probability } (m_l - \lfloor m_l \rfloor) \qquad (31)\\
&= \frac{\delta}{w_l} m_l \text{ with probability } 1 - (\lceil m_l \rceil - \lfloor m_l \rfloor)
\end{aligned}
$$

We note some facts about the random variable $\mathbf{h}_l$. If θ_l is an integer multiple of δ then $\mathbf{h}_l = \theta_l$ with probability 1. Secondly, $\mathbf{h}_l$ itself, is an integer multiple of δ with probability 1, regardless of what θ_l is. Thirdly, $\mathbf{h}_l$ is unbiased for θ_l, that is

$$E\mathbf{h}_l = \theta_l$$

Now we define the random vector $\mathbf{h} = (\mathbf{h}_1, \ldots, \mathbf{h}_p)$ where the coordinate random variables $\{\mathbf{h}_l\}_{l=1}^{p}$ are jointly independent. Then by the properties of $\mathbf{h}_l$ for each $1 \leq l \leq p$, we have $E\mathbf{h} = \theta$ and $\mathbf{h} \in \mathcal{F}$ with probability 1. Now, we are going to compute the average of the expression in (25) with respect to the distribution of $\mathbf{h}$ on $\mathcal{F}$. Again by unbiasedness of h we have the cross product terms zero on an average. To control the quadratic term we have to control

$$\frac{\alpha}{2\sigma^2} \sum_{i=1}^{n} \mathbb{E}\left(\sum_{j=1}^{p} (\mathbf{h}_j - \theta_j) f_j(x_i)\right)^2 \qquad (32)$$

By unbiasedness of $\mathbf{h}$ and independence of each of its coordinates the expected crossproduct terms in the inner sum are zero. Hence after interchanging the order of summation the last display equals the following

$$n \sum_{j=1}^{p} \mathbb{E}(\mathbf{h}_j - \theta_j) w_j^2.$$

Now after some calculations similar to the calculation of the variance of a bernoulli random variable, it can be shown that for each $1 \leq l \leq p$,

$$E(\mathbf{h}_l - \theta_l)^2 = (\frac{\delta}{w_l})^2 (m_l - \lfloor m_l \rfloor)(\lceil m_l \rceil - m_l).$$

Also it can be checked that for all numbers m_l we have the following inequality

$$(m_l - \lfloor m_l \rfloor)(\lceil m_l \rceil - m_l) \leq |m_l|.$$

So from the arguments above, we obtain an upper bound $\frac{\alpha}{2\sigma^2} n\delta^2 \sum_{l=1}^{p} |m_l|$ for the expected quadratic term. Now by dividing and multiplying by w_l within every term in the sum and recalling the definition of θ we get the upper bound $\frac{\alpha}{2\sigma^2} n\delta |\theta|_{w,1}$.

We note that each coordinate of $\mathbf{h}$ has a fixed sign depending on the signs of the coordinates of θ. Hence, we again have

$$\mathbb{E}V(\mathbf{h}) = \frac{|\theta|_{w,1}}{\alpha\delta} \log(4p) + \frac{\log 2}{\alpha}. \qquad (33)$$

Hence, we have the upper bound for the expectation of H_θ to be

$$\frac{\alpha}{2\sigma^2} n\delta |\theta|_{w,1} + \frac{|\theta|_{w,1}}{\alpha\delta} \log(4p) + \frac{\log 2}{\alpha}.$$

Setting $\delta^2 = \frac{2\sigma^2 \log 4p}{\alpha^2 n}$ we finally obtain the following

$$\min_{\tilde{\theta} \in \mathcal{F}} H_\theta(\tilde{\theta}) \leq \frac{1}{\sigma} \sqrt{2n \log(4p)} |\theta|_{w,1} + \frac{\log 2}{\alpha}.$$

It follows that by defining the penalty function on Θ defined as follows

$$pen(\theta) = \frac{1}{\sigma} \sqrt{2n \log(4p)} |\theta|_{w,1} + \frac{\log 2}{\alpha}. \qquad (34)$$

we have the risk validity of a weighted ℓ_1 penalty given by pen. Since pen is a risk valid penalty, by a direct application of theorem (II.2) and some minor rearranging of terms we obtain for all $0 < \alpha < 1$

$$
\begin{aligned}
&\mathbb{E} \frac{1}{2n\sigma^2} \sum_{i=1}^{n} (f_{\hat{\theta}}(x_i) - f^\star(x_i))^2 \leq \\
&(\frac{1}{1-\alpha}) \mathbb{E} \inf_{\theta \in \mathbb{R}^p} \left(\frac{1}{2n\sigma^2} \sum_{i=1}^{n} [(y_i - f_\theta(x_i))^2 - \right. \qquad (35)\\
&\left. (y_i - f^\star(x_i))^2] + \frac{1}{\sigma} \sqrt{\frac{2\log(4p)}{n}} |\theta|_{w,1} + \frac{\log(2)}{\alpha n} \right).
\end{aligned}
$$

By taking the expectation inside the infimum on the right side of the above display we present a theorem in this linear regression setting

Theorem III.2. *For the penalized likelihood estimator $\hat{\theta}$ defined as in* (5) *and the penalty given by* (73) *we have the following oracle inequality type result*

$$
\begin{aligned}
&\mathbb{E} \frac{1}{2n\sigma^2} \sum_{i=1}^{n} (f_{\hat{\theta}}(x_i) - f^\star(x_i))^2 \leq \\
&(\frac{1}{1-\alpha}) \mathbb{E} \inf_{\theta \in \mathbb{R}^p} \left(\frac{1}{2n\sigma^2} \sum_{i=1}^{n} [(f_\theta(x_i) - f^\star(x_i))^2] \right. \qquad (36)\\
&\left. + \sqrt{\frac{2\log(4p)}{n}} |\theta|_{w,1} + \frac{\log(2)}{\alpha n} \right).
\end{aligned}
$$

Remark III.1. *The leading constant on the right side can be made to be arbitrarily close to* 1 *by choosing α arbitrarily near* 0 *but then we pay for it as we have to divide the penalty term by α in the risk bound.*

Remark III.2. *We do not need any conditions on the design matrix Ψ in order for our risk bound to hold.*

B. Gaussian Graphical Models

A canonical scale problem in statistics is the problem of estimating the inverse covariance matrix of a multivariate Gaussian random vector. We observe $X = \{x_i\}_{i=1}^n$, each of which is drawn i.i.d from $N_p(0, \theta)$. Here $\theta_{p \times p}$ denotes the inverse covariance matrix of the random gaussian vectors. We denote the corresponding covariance matrices by $\Sigma = \theta^{-1}$. We assume that the model is well specified and we denote the true inverse covariance matrix to be $\theta^\star$. In this section we denote the $-\log det$ function on matrices by ϕ. We follow the the convention that ϕ takes value $+\infty$ on any matrix that is not positive definite. Then it follows that ϕ is a convex function on the space of all $p \times p$ matrices. Inspecting the log likelihood of this model we have

$$\frac{1}{n}\log P_\theta(X) = \frac{p}{2}\log(2\pi) + \frac{1}{2}Tr(S\theta) + \frac{\phi(\theta)}{2}$$

Here, $Tr(S\theta)$ is the sum of diagonals of the matrix $S\theta$ and $S = \frac{1}{n}\sum_{i=1}^n \tilde{x_i}^T \tilde{x_i}$. 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 risk valid penalty. We show our risk bounds in the case when 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

$$(\theta^\star + \Delta) \succ 0. \tag{37}$$

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 value of the δ in the assumption is specified later. Now we proceed with our scheme of things. Let us denote the space of $p \times p$ positive definite symmetric matrices by S_+^p. In this setting the parameter space could be identified with a convex cone of $\mathbb{R}^{p^2}$, the convex cone being the cone of positive definite symmetric matrices. We define $\mathcal{F}$ to be the δ integer lattice intersected with S_+^p. So we have

$$\mathcal{F} = \{\delta z \in \mathbb{R}^{p \times p} : vec(z) \in Z^{p^2}, z \in S_+^p\}. \tag{38}$$

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

$$V(\delta z) = \frac{C(z)}{\alpha}. \tag{39}$$

By (27) it is clear that V defined as above on $\mathcal{F}$ satisfies the Kraft type inequality (16). For this i.i.d model, our loss function turns out to be

$$\begin{aligned} L_\alpha(\theta_1, \theta_2) = &\frac{n}{2\alpha}[\alpha\phi(\theta_2) \\ &+ (1-\alpha)\phi(\theta_1) - \phi(\alpha\theta_2 + (1-\alpha)\theta_1)]. \end{aligned} \tag{40}$$

Since ϕ is a convex function, by Jensen' inequality one can see that $L_\alpha \geq 0$. In this setting, for technical reasons, we do not allow α to be too close to 1. So, in the following section, we will present our risk bounds for $0 < \alpha \leq \frac{1}{2}$ although there is nothing special about $\frac{1}{2}$ and in principle it can be replaced by any constant strictly less than 1 with the corresponding change in the factor of 2 in our assumption (37). Now we need to verify (19) in order to set a risk valid penalty. Denote the left side of (19) again by $H_\theta(\tilde{\theta})$. We now expand and simplify $H_\theta(\tilde{\theta})$. We have

$$\begin{aligned} &H_\theta(\tilde{\theta}) = \frac{n}{2}Tr(S(\tilde{\theta} - \theta)) + \frac{n}{2}[\phi(\tilde{\theta}) - \phi(\theta)] + \\ &V(\tilde{\theta}) + \frac{n}{2\alpha}\Big((\alpha)[\phi(\theta) - \phi(\tilde{\theta})] + \\ &\frac{n}{2\alpha}[\phi(\alpha\tilde{\theta} + (1-\alpha)\theta^\star) - \phi(\alpha\theta + (1-\alpha)\theta^\star)] \end{aligned}$$

After some cancellations we are left with

$$\begin{aligned} H_\theta(\tilde{\theta}) = &\frac{n}{2\alpha}[\phi(\alpha\tilde{\theta} + (1-\alpha)\theta^\star) \\ &- \phi(\alpha\theta + (1-\alpha)\theta^\star)] + \frac{n}{2}Tr(S(\tilde{\theta} - \theta)) + V(\tilde{\theta}). \end{aligned}$$

One can check that by treating ϕ as a function of p^2 variables, one has for a given positive definite matrix $M_{p \times p}$ and for any pair of indices i, j we have $\frac{\partial}{\partial M_{i,j}}\phi(X) = -(M^{-1})_{i,j}$. Also for any other pair of indices k, l the second derivatives are given by $\frac{\partial^2}{\partial M_{k,l}\partial M_{i,j}}\phi(X) = (M^{-1}E_{k,l}M^{-1})_{i,j} = (M^{-1})_{i,k}(M^{-1})_{j,l}$. Here $E_{k,l}$ is a $p \times p$ matrix with all zero entries except a 1 at the k, l position. By Taylor expanding ϕ about A upto the second order term we have the following equality for all positive definite symmetric matrices A and $A + B$ where t is some number between 0 and 1

$$\begin{aligned} &\phi(A + B) - \phi(A) = -Tr(BA^{-1}) + \\ &vec(B)^T H\left((A + tB)^{-1}\right) vec(B). \end{aligned} \tag{41}$$

where H evaluated at a positive definite matrix $M_{p \times p}$ is a $p^2 \times p^2$ matrix and is given by

$$H_{(i,j),(k,l)} = M_{i,k}M_{j,l}.$$

Let us now set $A = (1-\alpha)\theta^\star + \alpha\theta$ and $B = \alpha(\tilde{\theta} - \theta)$ in the above Taylor expansion. Then we can write $H_\theta(\tilde{\theta})$ as

$$\begin{aligned} &H_\theta(\tilde{\theta}) = -\frac{n}{2\alpha}Tr(BA^{-1}) + \frac{n}{2}Tr(SB) + V(\tilde{\theta}) + \\ &\frac{n}{2\alpha}vec(B)^T H\left((A + tB)^{-1}\right) vec(B). \end{aligned}$$

We again upper bound the minimum of $H_\theta(\tilde{\theta})$ over $\tilde{\theta} \in \mathcal{F}$ by an expectation over a chosen distribution on $\mathcal{F}$. This distribution is exactly similar to the second sampling method used in the linear regression setting. **Expand on this point?** So our random choice of $\tilde{\theta}$ is unbiased for θ and hence the average of B is zero. Consequently the trace terms are zero on an average. Then we have to control the quadratic form and the penalty term. Since the coordinates of the random choice of $\tilde{\theta}$ are independent the cross terms in the quadratic form are zero on an average. We note that an important property of our sampling strategy is that the ℓ_∞ distance between the random choice $\tilde{\theta}$ and θ is not greater than δ. Hence it follows that $|B|_\infty \leq \alpha\delta$. Now by assumption (37) one can check for all $0 < t < 1$ and all $0 < \alpha \leq \frac{1}{2}$ it follows that $\frac{(1-\alpha)}{2}\theta^\star + tB \succ 0$. Also we have by definition of A and B here,

$$A + tB - \frac{(1-\alpha)}{2}\theta^\star - \alpha\theta = \frac{(1-\alpha)}{2}\theta^\star + tB. \quad (42)$$

The above two equations imply that for all $0 < t < 1$ and all $0 < \alpha \leq \frac{1}{2}$ we have

$$A + tB \succ \frac{(1-\alpha)}{2}\theta^\star \succ 0. \quad (43)$$

In particular we are always inside the region of differentiability of ϕ and hence our taylor expansion is valid. We first consider the following expected quadratic form for any $0 \leq t \leq 1$

$$\mathbb{E}(vec(B)^T H(A+tB)^{-1} vec(B)).$$

Since the cross terms are zero on an average due to independence of the coordinates and the fact that $\mathbb{E}vec(B) = 0$ we have the last display equalling

$$\mathbb{E}\sum_{l=1}^{p^2}(vec(B)_l)^2(H(A+tB)^{-1})_{ll}.$$

Now by definition of H any of the diagonals of $(H(A+tB)^{-1})$ is not greater than the maximum diagonal of $(A+tB)^{-1}$ squared. Now (43) implies that the maximum diagonal of $(A+tB)^{-1}$ is not greater than the maximum diagonal of $\frac{2}{1-\alpha}\Sigma^\star$. Let us denote the maximum diagonal of Σ by σ_{max}. Then we have the following inequality for all $1 \leq l \leq p^2$,

$$((A+tB)^{-1} \otimes (A+tB)^{-1})_{ll} \leq \frac{4(\sigma_{max})^2}{(1-\alpha)^2}. \quad (44)$$

Now, as in the linear regression case, it can be shown that for each coordinate l, the variance of $vec(B)_l$ is upper bounded by $\delta|vec(\theta)_l|$. Hence we can write

$$\mathbb{E}(vec(B)^T H\left((A+tB)^{-1} vec(B)\right) \leq \frac{4(\sigma_{max})^2}{(1-\alpha)^2}$$
$$\delta|vec(\theta)|_1.$$

As for the penalty term, the sampling method ensures that the signs of each of the coordinates of the random choice $\tilde{\theta}$ does not change. Hence the expected penalty term is just the penalty evaluated at θ. So then we have

$$\mathbb{E}H_\theta(\tilde{\theta}) \leq \frac{4n(\sigma_{max})^2}{2\alpha(1-\alpha)^2}\delta|\theta|_1 + \frac{|\theta|_1}{\alpha\delta}\log(4p^2) + \frac{\log 2}{\alpha}.$$

Again by setting $\delta^2 = \frac{\log(4p^2)(1-\alpha)^2}{2n(\sigma_{max})^2}$ it follows that by defining the penalty function on Θ defined as follows

$$pen(\theta) = \frac{\sqrt{\sigma_{max}\log(4p^2)2n}}{\alpha(1-\alpha)}|\theta|_1 + \frac{\log 2}{\alpha} \quad (45)$$

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

$$\hat{\theta} = \underset{\theta\in S_+^p}{\operatorname{argmin}}\left(\frac{1}{2}Tr(S\theta) + \frac{\phi(\theta)}{2} + \frac{pen(\theta)}{n}\right). \quad (46)$$

enjoys the adaptive risk properties we desire. Under the assumption (37) where now δ has been specified, we have the following risk bound for all $0 < \alpha \leq \frac{1}{2}$

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

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

Theorem III.3. *For the estimator $\hat{\theta}$ as in* (46) *with $\hat{\Sigma}^{-1} = \hat{\theta}^{-1}$ and the penalty* (45) *we have the risk bound*

$$\mathbb{E}L_\alpha(\theta^\star, \hat{\theta}) \leq \inf_{\theta\in S_+^p}\left(\frac{1}{2}[Tr(\hat{\theta}\Sigma^\star) - p] + \frac{1}{2}[\phi(\hat{\theta}) - \phi(\theta^\star))] + \frac{pen(\theta)}{n}\right). \quad (47)$$

Remark III.3. *Remark about the factor outside the oracle inequality.*

Remark III.4. *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 $\sqrt{\frac{\log(4p^2)}{n}}\|\theta^\star\|_1$. 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*

Remark III.5. *The assumption* (37) *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.*

IV. Validity of l_0 Penalty in Linear Regression

In this section we return to the linear regression setup to show the codelength and risk validity of the l_0 penalty. We consider the fixed design Ψ and known variance σ^2 setup. 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 or risk bounds for the estimator minimizing the ℓ_0 penalized log likelihood.

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 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. We also denote the quantity $\frac{1}{|S|} Tr\left((\Psi_{in,S}^T \Psi_{in,S})^{-1}(\Psi_{f,S}^T \Psi_{f,S})\right)$ by Υ_S.

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

$$C_S = \{M_S v : v \in \mathcal{G}^{|S|}\} \tag{48}$$

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 is a subset of $\mathbb{R}^p$. We want to construct Kraft satisfying codelengths and hence subprobabilities on C_S which are proportional to $\left(\frac{(P_\phi(X_{in})}{P_{\theta^\star}(X_{in})}\right)^\eta$ for any fixed but arbitrary $0 < \eta \leq 1$. For that purpose we want to estimate the normalizer which is the quantity $\sum_{\phi\in C_S}\left(\frac{(P_\phi(X_{in}))}{P_{\theta^\star}(X_{in})}\right)^\eta$. The following lemma helps us do exactly that.

Lemma IV.1. *For all $0 < \eta \leq 1$ we have*

$$\sum_{\phi\in C_S} \left(\frac{P_\phi(X_{in})}{P_{\theta^\star}(X_{in})}\right)^\eta \delta^{|S|} \leq U_\eta(X_{in}, S) \tag{49}$$

where

$$U_\eta(X_{in}, S) = \exp\left(\frac{\eta}{2}\|O_{\Psi_{in,S\cup S^\star}} y_{in} - \Psi_{in,S^\star}\theta^\star\|_2^2\right) \left(\frac{2\pi}{\eta}\right)^{|S|/2} \tag{50}$$

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

The proof of this lemma is given in the appendix.

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

$$\mathcal{C} = \cup_{k=0}^p \cup_{\{S:|S|=k\}} C_S \tag{51}$$

$\mathcal{C}$ 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 and in applications we will set δ to be something specific. We now define penalty functions satisfying Kraft type inequalities on the countable set $\mathcal{C}$. First we define a family of subprobabilities h_η on $\mathcal{C}$ as follows

$$h_\eta(\tilde{\theta}, X_{in}) = \left(\frac{1}{2}\right)^{k(\tilde{\theta})+1} \frac{1}{\binom{p}{k(\tilde{\theta})}} \left(\frac{P_{\tilde{\theta}}(X_{in})}{P_{\theta^\star}(X_{in})}\right)^\eta \delta^{k(\tilde{\theta})} \frac{1}{U_\eta(X_{in}, S(\tilde{\theta}))}. \tag{52}$$

We claim that $h_\eta(\tilde{\theta})$ is a subprobability on $\tilde{C}_\eta$ for every X_{in}. This can be seen by first summing $h_\eta(\tilde{\theta})$ over non negative integers k from 0 to p, then summing over all subsets of $[1:p]$ with cardinality k and then summing over $C_{S,\eta}$. The inner sum over $C_{S,\eta}$ of $\left(\frac{P_{\tilde{\theta}}(X_{in})}{P_{\theta^\star}(X_{in})}\right)\delta^{|S|}\frac{1}{U_\eta(X_{in},S)}$ is no more than 1 by lemma (IV.1). Then for each k we sum over $\binom{p}{k(\tilde{\theta})}$ subsets and the factor $\frac{1}{\binom{p}{k(\tilde{\theta})}}$ keeps the overall sum still no more than 1. Similarly, the factor $(\frac{1}{2})^{k(\tilde{\theta})+1}$ makes the whole sum less than or equal to 1 when we sum over k from 0 to p, which can be seen by summing up the geometric series. Hence, we prove our claim.

We can now define Kraft satisfying codelengths $l_\eta(\tilde{\theta}, X_{in})$ on $\mathcal{C}$ by defining

$$l_\eta(\tilde{\theta}, X_{in}) = -\frac{1}{\eta}\log h_\eta(\tilde{\theta}) \tag{53}$$

Then because of h_η being a subprobability, it is clear that l_η satisfies the following inequality for all X_{in}

$$\sum_{\tilde{\theta}\in\mathcal{C}} \exp(-\eta l_\eta(\tilde{\theta}, X_{in})) \leq 1. \tag{54}$$

A. Codelength Validity

In this section we 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

$$\min_{\theta\in\Theta}\{-\log P_\theta(X) + pen(\theta|X_{in})\} \geq \min_{\tilde{\theta}\in\mathcal{F}}\{-\log P_{\tilde{\theta}}(X_f) + V(\tilde{\theta}|X_{in})\} \tag{55}$$

where now the right side of (55) 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 (55).

We declare our countable set $\mathcal{F} = \mathcal{C}$ as defined in (51). We also define $V = l_\eta$ with $\eta = 1$ as defined in (53). Then we have

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

The task now is to verify (55). An equivalent way to verify (55) is to verify the following for any given $\theta \in \Theta$ and data X,

$$\min_{\tilde{\theta}\in\mathcal{F}}\{-\log\frac{P_{\tilde{\theta}}(X_f)}{P_{\theta^\star}(X_f)} + \log\frac{P_\theta(X)}{P_{\theta^\star}(X)} + V(\tilde{\theta}|X_{in})\} \leq pen(\theta|X_{in}). \tag{56}$$

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

$$\min_{\tilde{\theta}\in\mathcal{F}}\{-\log\frac{P_{\tilde{\theta}}(X)}{P_{\theta^\star}(X)} + \log\frac{P_\theta(X)}{P_{\theta^\star}(X)} + \left(V(\tilde{\theta}|X_{in}) + \log\frac{P_{\tilde{\theta}}(X_{in})}{P_{\theta^\star(X_{in})}}\right)\}. \tag{57}$$

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 (48). 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 (57).

Lemma IV.2.

$$\min_{\tilde{\theta}\in C_{S(\theta)}}\{-\log\frac{P_{\tilde{\theta}}(X)}{P_{\theta^\star}(X)} + \log\frac{P_\theta(X)}{P_{\theta^\star}(X)}\} \leq 2(1+\Upsilon_{S(\theta)})\,k(\theta)\delta^2. \tag{58}$$

The proof of the above lemma is given in the appendix.

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 $C_{S(\theta),1}$ we write down the upper bound we get for the left side of (57) 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 (55) would be

$$pen(\theta|X_{in}) = \frac{k(\theta)}{2} + (k(\theta)+1)\log(2) + \log\binom{p}{k(\theta)} + \frac{k(\theta)}{2}\log(4(1+\Upsilon_{S(\theta)})) + \log(U(X_{in}, S(\theta)). \tag{59}$$

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

$$pen(\theta|X_{in}) = \frac{k(\theta)}{2}\log(4(1+\Upsilon_{S(\theta)})) + \log\binom{p}{k(\theta)} + k(\theta)\left(\frac{3\log(2)}{2} + \frac{\log(2\pi)}{2}\right) + \frac{1}{2}\|O_{\Psi_{in,S(\theta)\cup S^\star}}\, y_{in} - \Psi_{in,S^\star}\theta^\star\|_2^2. \tag{60}$$

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

$$pen(\theta|X_{in}) = \frac{k(\theta)}{2}\log(\frac{4n}{p}) + \log\binom{p}{k(\theta)} + k(\theta)\left(\frac{3\log(2)}{2} + \frac{\log(2\pi)}{2}\right) + \frac{1}{2}\|O_{\Psi_{in,S(\theta)\cup S^\star}} y_{in} - \Psi_{in,S^\star}\theta^\star\|_2^2. \tag{61}$$

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 $k(\theta)\log(n)$. We remind the reader that $k(\theta)\log(p/k(\theta)) \leq \log(\binom{p}{k(\theta)}) \leq k(\theta)\log(ep/k(\theta))$. So the term $\log(\binom{p}{k(\theta)})$ again contributes a constant times $k(\theta)\log(n)$ term in case p is growing as some power of n.

B. Risk validity

In this section we show the risk validity of the l_0 penalty by leveraging its codelength interpretation as shown in the last subsection. To prove risk bounds by the same reasoning as in section (II-B) we need to adapt the arguments in section (II-B) to the case when we have data split into two parts. We define our family of loss functions between two probability distributions p and q on $\mathcal{X}^n$ in the same way as before except that it only depends on the final part of the data X_f. Let $0 < \alpha \leq 1$ be a fixed, arbitrary number. We define our loss function as follows

$$L_\alpha(p, q) = -\frac{1}{\alpha}\log(\mathbb{E}(\frac{q(X_f)}{p(X_f)})^\alpha). \tag{62}$$

Also for a penalty $pen(\theta|X_{in})$ depending on X_{in} we define our penalized likelihood estimator to be

$$\hat{\theta}(X) = \operatorname*{argmin}_{\theta\in\Theta}\{-\log P_\theta(X) + pen(\theta|X_{in})\}. \tag{63}$$

We now present the theorem which will help us in proving risk bounds for the ℓ_0 penalized likelihood estimator. Fix $0 < \alpha < 1$. For our countable set $\mathcal{F} = \mathcal{C}$ and codelengths $V = l_\alpha$ as defined in (53), clearly the following is true by (54).

$$\sum_{\tilde{\theta}\in\mathcal{F}} \exp(-\alpha V(\tilde{\theta}|X_{in})) \leq 1. \tag{64}$$

We expand V to get

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

We would like to now verify the following

$$\min_{\tilde{\theta}\in\tilde{\Theta}}\{-\log\frac{P_{\tilde{\theta}}(X_f)}{P_{\theta^\star}(X_f)} + \log\frac{P_\theta(X)}{P_{\theta^\star}(X)} + L_\alpha(P, P_\theta) - L_\alpha(P, P_{\tilde{\theta}}) + V(\tilde{\theta}|X_{in})\} \leq pen(\theta). \tag{65}$$

Verifying the above gives us risk bounds as is shown in the following lemma.

Lemma IV.3. *Assuming the existence of a countable subset $\mathcal{F} \subset \Theta$ and a penalty function $V(.|X_{in})$ defined on F satisfying* (64) *and* (65)*, we have the following risk bound*

$$\mathbb{E}L(P_{\theta^\star}, P_{\hat{\theta}}) \leq \mathbb{E}\min_{\theta\in\Theta}\left(\log\frac{P_{\theta^\star}(X)}{P_\theta(X)} + pen(\theta|X_{in})\right).$$

The proof of this lemma parallels the proof of theorem (II.3) and is given in the appendix.

Now we proceed to verify (64) in order to derive the risk bound in theorem (IV.3) for the l_0 penalized estimator in the linear regression setting. We can write the left side in (65) as

$$\min_{\tilde{\theta}\in\tilde{\Theta}}\{-\log\frac{P_{\tilde{\theta}}(X)}{P_{\theta^\star}(X)} + \log\frac{P_\theta(X)}{P_{\theta^\star}(X)} + L_\alpha(P, P_\theta) - L_\alpha(P, P_{\tilde{\theta}}) + \left(V_\alpha(\tilde{\theta}|X_{in}) + \log\frac{P_{\tilde{\theta}}(X_{in})}{P_{\theta^\star(X_{in})}}\right)\}.$$

Again we upper bound the minimum of the above expression by restricting to $\tilde{\theta} \in C_{S(\theta)}$ which cannot decrease the overall minimum. Restricted to $\tilde{\theta} \in C_{S(\theta)}$ it turns out that the term $V_\alpha(\tilde{\theta}|X_{in}) + \log\frac{P_{\tilde{\theta}}(X_{in})}{P_{\theta^\star(X_{in})}}$ remains a constant. The following lemma now helps us.

Lemma IV.4. *Given any θ and data X we have the following inequality*

$$\min_{\tilde{\theta}\in C_{S(\theta)}}\left(\log(\frac{P_\theta(X)}{P_{\tilde{\theta}}(X)}) + L_\alpha(\tilde{\theta}, \theta^\star) - L_\alpha(\theta, \theta^\star)\right) \leq 2k(\theta)\delta^2(1 + \alpha\Upsilon_{S(\theta)})$$

Hence we get an upper bound for the left side of (65) which is as follows

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

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

$$pen_\alpha(\theta|X_{in}) = \frac{k(\theta)}{2\alpha}\log(4\alpha(1 + \alpha\Upsilon_{S(\theta)})) + \frac{k(\theta)}{2\alpha} + \frac{(k(\theta)+1)\log(2)}{\alpha} + \frac{\log\binom{p}{k(\theta)}}{\alpha} + \frac{\log(U_\alpha(X_{in}, S(\theta))}{\alpha}.$$

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

$$pen_{\alpha}(\theta|X_{in}) = \frac{k(\theta)}{2\alpha}\log(4\alpha(1+\alpha\Upsilon_{S(\theta)})) + \alpha\Psi_f^T\Psi_f)) + \frac{(k(\theta)+1)\log(2)}{\alpha} + \frac{\log\binom{p}{k(\theta)}}{\alpha} + \frac{k(\theta)}{2}\log(\frac{2\pi}{\alpha}) + \frac{1}{2}\|O_{\Psi_{in,S\cup S^\star}}y_{in} - \Psi_{in,S^\star}\theta^\star\|_2^2 \tag{66}$$

By taking the expectation inside the minimum in the right side and then doing some algebraic manipulations of the statement of theorem (IV.3), we get the resolvability risk bound which we write down below as a theorem.

Theorem IV.5. *With the estimator being defined as in* (63) *and* $pen_{\alpha}(\theta|X_{in})$ *as defined in* (66) *we have the risk bound for all* $0 < \alpha \leq 1$,

$$\mathbb{E}\frac{1}{2n}\|X_f(\hat{\theta}-\theta^\star)\|_2^2 \leq \frac{1-\alpha}{\sigma^2}\inf_{\theta\in\mathbb{R}^p}\left(\frac{1}{2n}\|X_f(\hat{\theta}-\theta^\star)\|_2^2 + \frac{1}{n}\mathbb{E}_{in}\, pen(\theta|X_{in})\right).$$

Remark IV.1. *As we can see, as* α *is taken to be near zero, the constant outside the right side in theorem* (IV.5) *approaches the desired value* 1. *But then we have to pay for the fact that the penalty contains terms divided by* α *which blow up when* α *is brought near zero.*

By setting $\theta = \theta^\star$ inside the infimum in the above theorem we obtain

$$\mathbb{E}\frac{1}{2n}\|X_f(\hat{\theta}-\theta^\star)\|_2^2 \leq \frac{1-\alpha}{\sigma^2}\mathbb{E}_{in}\frac{pen(\theta^\star|X_{in})}{n}. \tag{67}$$

Remark IV.2. *The random part depending on* y_{in} *in* $pen((\theta^\star|X_{in}))$ *is* $\|O_{\Psi_{in,S\cup S^\star}}y_{in} - \Psi_{in,S^\star}\theta^\star\|_2^2$ *which is distributed as a* χ^2 *random variable with degrees of freedom at most* $k(\theta)+k(\theta^\star)$. *In the case when the design matrices* Ψ_{in} *and* Ψ_f *are orthogonal and the* ℓ_2 *norms of the columns of* Ψ_{in} *and* Ψ_f *are atmost* p *and* $n-p$ *respectively we then have* $\Upsilon_S = \frac{n-p}{p}$. *Then the leading term of the expected penalty is of the order* $k(\theta)\log(n)$ *and hence we atleast have a* $k(\theta^\star)\log(n)/n$ *rate of convergence of the left side in* (67).

Remark IV.3. *Our risk bounds are useful even when* p *grows like a constant fraction of* n.

V. Appendix

We hereby prove lemma (III.1).

Proof: We first define a subprobability measure π on $\mathcal{Z}^p$. We define π in two stages. For each non negative integer k, let $\pi(C_k) = 1/2^{k+1}$ where $C_k \subset \mathcal{Z}^p$ denotes all integer p tuples with l_1 norm equalling k. We claim that

$$|C_k| \leq (2p)^k.$$

We prove the claim in the end. Now let the conditional distribution of π given the set C_k be the uniform distribution on C_k for all non negative k. Then, for any $z \in \mathcal{Z}^p$, if we define π to be

$$\pi(z) = \frac{1}{(2)^{|z|_1+1}}\frac{1}{(2p)^{|z|_1}}$$

then π is clearly a subprobability distribution on $\mathcal{Z}^p$. Now by taking negative log of the subprobability π we have

$$-\log(\pi(z)) = |z|_1\log(4p) + \log 2.$$

Clearly $-\log(\pi(z))$ satisfies the desired Kraft type inequality. Now we prove our claim. Let us first give an upper bound for the set of length p positive integer sequences which sums to k. This is exactly equal to the number of ways of assigning k unlabelled balls to p labelled cells. It is certainly upper bounded by the number of ways of assigning k labelled balls to p labelled cells. The latter is easy to count and it is precisely p^k. Now given a positive integer sequence one can flip the signs of non zero coordinates and the l_1 norm would remain k. There would be atmost k non zero coordinates as the sum is k and hence the final upper bound $(2p)^k$. ■

1) Sampling method 1: Let θ be any given vector in $\mathbb{R}^p$ and let us first consider the quadratic term which is the following

$$\frac{\alpha}{2\sigma^2}\sum_{i=1}^{n}(f_\theta(x_i) - f_{\tilde{\theta}}(x_i))^2.$$

By expanding out f_θ and $f_{\tilde{\theta}}$ in terms of the dictionary functions the last display becomes

$$\frac{\alpha}{2\sigma^2}\sum_{i=1}^{n}(\sum_{j=1}^{p}(\tilde{\theta}_j - \theta_j)f_j(x_i))^2. \tag{68}$$

Let δ be a positive real number. Let $K(\theta) = \lceil\frac{|\theta|_{w,1}}{\delta}\rceil$. K is the least integer larger than or equal to $|\theta|_{w,1}$ divided by δ. We will write $K = K(\theta)$ to minimize notational clutter. In order to explain our sampling strategy, we first define a random variable h. Let $\{\tilde{e_j}\}_{j=1}^p$ denote the canonical basis of $\mathbb{R}^p$. The random vector h takes value $K\delta sign(\theta_j)\frac{\tilde{e_j}}{w_j}$ with probability $\frac{w_j\theta_j}{K\delta}$ for all $j = 1,\ldots,p$. With the remaining probability, h takes the form of the zero vector. One can check that h defined this way is unbiased for θ, that is $\mathbb{E}(h_1) = \theta$. Say we sample K i.i.d copies $h_1,\ldots,h_k$. of h. We now consider the mean of these random vectors $\overline{h} = \frac{1}{K}\sum_{i=1}^{K}h_i$. Clearly, $\overline{h}$ is also unbiased for θ. $\overline{h}$ would be our random choice of $\tilde{\theta} \in \mathcal{F}$. We are now in a position to define our countable set $\mathcal{F}$. Let W denote the diagonal matrix of weights as defined in (29). We define the set $\mathcal{F}$ as follows

$$\mathcal{F} = \delta\{W^{-1}z : z \in Z^p\} \tag{69}$$

Clearly $\mathcal{F}$ is countable since Z^p is so. A careful observation shows that h always takes values in $\mathcal{F}$. We now define a penalty function V on $\mathcal{F}$ derived from C. So we define V in the following manner

$$V(\delta W^{-1}z) = \frac{C(z)}{\alpha}$$

for all vectors z in Z_p. It is clear from (9) that V satisfies the Kraft inequality (16). So the symbol $\mathbb{E}$ would now mean expectation with respect to the distribution of $\overline{h}$ on $\mathcal{F}$. After making the definitions of $\mathcal{F}$ and V clear, let us now compute the expectation of the expression in (32) over our random choice of $\overline{h} \in \mathcal{F}$. We first note that

$$\mathbb{E}(\sum_{j=1}^{p}((\overline{h}_j - \theta_j)f_j(x_i))^2 = \frac{1}{K}\mathbb{E}(\sum_{j=1}^{p}((h_j - \theta_j)f_j(x_i))^2.$$

This is because $\overline{h}$ is the sum of K i.i.d copies of h. Now we can upper bound the expectation over $\{h_l\}_{l=1}^{K}$ of the above term as follows

$$\mathbb{E}(\sum_{j=1}^{p}((h_j - \theta_j)f_j(x_i))^2 \leq \mathbb{E}(\sum_{j=1}^{p} h_j f_j(x_i))^2.$$

The above inequality follows due to unbiasedness of h_1 and by the simple fact that the variance of any random variable is at most the expected square of that random variable. We note the fact that for any $j \neq l$ the cross product terms $h_j h_l = 0$ pointwise by definition of h. Now summing over i and combining the previous two inequalities we obtain the following result

$$\mathbb{E}\sum_{i=1}^{n}(\sum_{j=1}^{p}(\overline{h}_j - \theta_j)f_j(x_i))^2 \leq \frac{n}{K}\sum_{j=1}^{p} w_j^2 E(h_j)^2. \tag{70}$$

For any j we also have by definition of the random variable h

$$\mathbb{E}h_j^2 = \frac{K\delta\theta_j}{w_j}. \tag{71}$$

So combining the last two equations we have

$$\mathbb{E}\sum_{i=1}^{n}(\sum_{j=1}^{p}(\overline{h}_j - \theta_j)f_j(x_i))^2 \leq n\delta|\theta|_{w,1}. \tag{72}$$

Now we consider the penalty term. We note that each coordinate of h has a fixed sign depending on the signs of the coordinates of θ. Therefore, linearity of expectation extends to absolute values also, in other words it can be checked that the following holds for any $j \in [1:p]$

$$\mathbb{E}|\overline{h}_j| = \mathbb{E}|h_j|$$

The above equation and the definition of V implies the following fact

$$\mathbb{E}V(\overline{h}) = \mathbb{E}V(h).$$

It is clear now from the definition of h and (26) the following

$$\mathbb{E}V(\overline{h}) = \frac{K\log(4p) + \log 2}{\alpha}.$$

So by the above arguments we can conclude that H_θ on an average is upper bounded by the following expression

$$\frac{\alpha}{2\sigma^2}n\delta|\theta|_{w,1} + \frac{K\log(4p) + \log 2}{\alpha}.$$

Using the fact that $K \leq \frac{|\theta|_{w,1}}{\delta} + 1$ we have the expression in the last display can be further upper bounded by the following expression

$$\frac{\alpha}{2\sigma^2}n\delta|\theta|_{w,1} + \frac{|\theta|_{w,1}}{\delta}\frac{\log(4p)}{\alpha} + \frac{\log(4p) + \log 2}{\alpha}.$$

Setting $\delta^2 = \frac{2\sigma^2 \log 4p}{\alpha^2 n}$ we finally obtain the following

$$\min_{\tilde{\theta} \in \mathcal{F}} H_\theta(\tilde{\theta}) \leq \frac{1}{\sigma}\sqrt{2n\log(4p)}|\theta|_{w,1} + \frac{\log(4p) + \log 2}{\alpha}.$$

It follows that by defining the penalty function on Θ defined as follows

$$pen(\theta) = \frac{1}{\sigma}\sqrt{2n\log(4p)}|\theta|_{w,1} + \frac{\log(4p) + \log 2}{\alpha}. \tag{73}$$

we define a risk valid penalty. **Check whether it is** $\log(2p)$ **or** $\log(4p)$.

In this case the random matrix $\Psi_{in,S}^T\Psi_{in,S}$ has a Wishart distribution denoted by $W_k(\Sigma_S, n)$. By Bartlett's decomposition for Wishart matrices we have the following lemma:

Lemma V.1.

$$\mathbb{E}\log det\Psi_{in,S}^T\Psi_{in,S} = \sum_{i=1}^{k}\psi(\frac{p+1-i}{2}) + k\log 2$$

where ψ is known in literature as the digamma function.

We also have the following well known upper and lower bound on the digamma function which we do not prove here. **Give reference.**

Lemma V.2.

$$\log(x-1) \leq \psi(x) \leq \log(x) \quad \forall x > 0.$$

We now control the behaviour of $\mathbb{E}\log det\Psi_{in,S}^T\Psi_{in,S}$ in the following lemma

Lemma V.3. *If there exists a constant C such that $det(\Sigma_S) > |S|C$ for all S then there exists a universal constant C_0 such that*

$$\log\binom{p}{k} - \mathbb{E}\log det\Psi_{in,S}^T\Psi_{in,S} \leq C_0 k \tag{74}$$

Proof: Using the lower bound of ψ in lemma (V.2) we obtain

$$\sum_{i=1}^{k}\psi(\frac{p+1-i}{2})+k\log 2\geq\sum_{i=1}^{k}\log(p-i-1).$$

We can further lower bound by replacing each of the log terms by $\log(n-k-1)$ to obtain

$$\sum_{i=1}^{k}\psi(\frac{p+1-i}{2})+k\log 2\geq k\log(p-k-1) \tag{75}$$

Also we have the inequality

$$\log\binom{p}{k}\leq k\log(\frac{ep}{k}). \tag{76}$$

By equations (75) and (76) we have

$$\log\binom{p}{k}-\mathbb{E}\log det\Psi_{in,S}^{T}\Psi_{in,S}\leq k\log(\frac{ep}{k(p-k-1)})-\log det(\Sigma_S) \tag{77}$$

Now $k(p-k-1)$ is minimized at the extremes $k=1$ and $k=p-2$. Hence $log(\frac{ep}{k(p-k-1)})\leq\log(\frac{ep}{p-2})$. Now we can definitely find a constant C_0 such that $\log(\frac{ep}{p-2})\leq C_0$. Hence we prove the lemma. ∎

for obtaining a redundancy risk bound. It is very similar to the arguments in section (II-B1). In the following, $0\leq\alpha<1$ is any fixed but arbitrary number. We define our family of loss functions between two probability measures P and Q in the same way as before except that it only depends on the final part of the data X_f. They are defined as follows

$$L_\alpha(P,Q)=-\frac{1}{\alpha}\log(\mathbb{E}(\frac{Q(X_f)}{P(X_f)})^\alpha). \tag{78}$$

For a penalty $pen(\theta|X_{in})$ depending on X_{in} we define our penalized likelihood estimator to be

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

By the definition of $\hat{\theta}$ we have

$$L_\alpha(P_{\theta^\star},P_{\hat{\theta}})=\left(L_\alpha(P_{\theta^\star},P_{\hat{\theta}})-\log\frac{P_{\theta^\star}(X)}{P_{\hat{\theta}}(X)}-pen(\hat{\theta})\right)+\min_{\theta\in\Theta}\left(\log\frac{P_{\theta^\star}(X)}{P_\theta(X)}+pen(\theta)\right) \tag{79}$$

If we focus on the first term in the brackets in the above display we can write it as

$$L_\alpha(P_{\theta^\star},P_{\hat{\theta}})-\log\frac{P_{\theta^\star}(X)}{P_{\hat{\theta}}(X)}-pen(\hat{\theta}|X_{in})=\frac{1}{\alpha}\log\left(\exp(\alpha L_\alpha(P_{\theta^\star},P_{\hat{\theta}}))(\frac{P_{\hat{\theta}}(X)}{P(X)})^\alpha\exp(-\alpha pen(\hat{\theta}|X_{in}))\right). \tag{80}$$

Assume we are able to define a countable set $\tilde{\theta}$ and a penalty $V(\tilde{\theta}|X_{in})$ on it so that for any fixed data X and θ, there would exist a $\tilde{\theta}\in\tilde{\Theta}$ satisfying the following inequality

$$\log\left(\alpha\exp(\alpha L_\alpha(P_{\theta^\star},P_\theta))(\frac{P_\theta(X)}{P(X)})^\alpha\exp(-\alpha pen(\theta|x_{in}))\right)\leq\log\left(\exp(\alpha L_\alpha(P_{\theta^\star},P_{\tilde{\theta}}))(\frac{P_{\tilde{\theta}}(X_f)}{P(X_f)})^\alpha\exp(-\alpha V(\tilde{\theta}|X_{in}))\right) \tag{81}$$

where $V(\tilde{\theta},X_{in})$ also satisfies for all X_{in},

$$\sum_{\tilde{\theta}\in\Theta}\exp(-\alpha V(\tilde{\theta}|X_{in}))\leq 1. \tag{82}$$

Then we note that (81) implies a further upper bound

$$\log\left(\exp(\alpha L_\alpha(P_{\theta^\star},P_\theta))(\frac{P_\theta(X)}{P(X)})^\alpha\exp(-\alpha pen(\theta|X_{in}))\right)\leq\log\left(\sum_{\tilde{\theta}\in\tilde{\Theta}}\exp(\alpha L_\alpha(P_{\theta^\star},P_{\tilde{\theta}}))(\frac{P_{\tilde{\theta}}(X_f)}{P(X_f)})^\alpha\exp(-\alpha V(\tilde{\theta}|X_{in}))\right).$$

We can now take expectation with respect to X_f conditional on X_{in} and move the expectation inside the log by Jensen's inequality. Now by equations (82), (83) and by definition of L_α

$$\mathbb{E}_f\log\left(\sum_{\tilde{\theta}\in\tilde{\Theta}}\exp(\alpha L(P_{\theta^\star},P_{\hat{\theta}}))(\frac{P_{\hat{\theta}}(X_f)}{P(X_f)})^\alpha\exp(-\alpha V(\tilde{\theta}|X_{in}))\right)\leq 0.$$

From the above and (79) we get an upper bound of the expected loss function conditional on X_{in} where expectation is taken over X_f,

$$\mathbb{E}_f L(P_{\theta^\star},P_{\hat{\theta}})\leq\mathbb{E}_f\min_{\theta\in\Theta}\left(\log\frac{P(\Psi)}{P_\theta(\Psi)}+pen(\theta|X_{in})\right).$$

Now by taking expectation with respect to X_{in} we obtain the following risk bound, where $\mathbb{E}$ refers to expectation taken over the whole data,

$$\mathbb{E}L_\alpha(P_{\theta^\star},P_{\hat{\theta}})\leq\mathbb{E}\min_{\theta\in\Theta}\left(\log\frac{P_{\theta^\star}(X)}{P_\theta(X)}+pen(\theta|X_{in})\right).$$

Lemma V.4. *Given θ and data X we have the following inequality*

$$\min_{\tilde{\theta}\in C_{S(\theta),\alpha}}\left(\log(\frac{P_\theta(X)}{P_{\tilde{\theta}}(X)})+L_\alpha(\tilde{\theta},\theta^\star)-L_\alpha(\theta,\theta^\star)\right)\leq 2\delta^2(Tr(\Psi_{in}^T\Psi_{in}+\alpha\Psi_f^T\Psi_f)).$$

Proof: For all $\tilde{\theta} \in C_{S(\theta),\alpha}$ clearly $k(\tilde{\theta}) = k(\theta)$. Now if we expand the expression we have to minimize we obtain

$$\frac{1}{2}\|y - \Psi\tilde{\theta}\|_2^2 - \frac{1}{2}\|y - \Psi\theta\|_2^2 + \frac{1-\alpha}{2}\|\Psi_f\theta - \Psi_f\theta^\star\|_2^2 - \frac{1-\alpha}{2}\|\Psi_f\tilde{\theta} - \Psi_f\theta^\star\|_2^2. \tag{83}$$

After simplifications and noting $\Psi^T\Psi = \Psi_{in}^T\Psi_{in} + \Psi_f^T\Psi_f$ the above term can be written as

$$\frac{1}{2}\|\Psi_{in}\tilde{\theta} - \Psi_{in}\theta\|_2^2 + \frac{\alpha}{2}\|\Psi_f\tilde{\theta} - \Psi_f\theta\|_2^2 + l(\tilde{\theta} - \theta)$$

where l is a affine function. Again, our strategy is to upper bound the minimum by an expectation with respect to a carefully chosen distribution. We now describe the choice of the distribution. Consider θ and its position in the δ integer lattice in $\mathbb{R}^{|S(\theta)|}$. It is in one of the cubes in the lattice. Consider this cube and all the other neighbouring cubes. Each of these cubes have one and only one point belonging to $C_{S(\theta)}$. It can be checked that the set of these points contain the point θ in its convex hull. Hence one can devise a distribution on this set of points with the property that the each coordinate of the random vector drawn from this distribution is independent and the average of this distribution is θ. Hence l, being an affine function of $\tilde{\theta} - \theta$ is zero on an average. Now for any quadratic form $(\tilde{\theta}-\theta)^T A(\tilde{\theta}-\theta)$ where A is some non negative definite matrix its expectation boils down to the expectation of the diagonal terms. We mean that

$$\mathbb{E}(\tilde{\theta} - \theta)^T A(\tilde{\theta} - \theta) = \sum_{i=1}^{k(\theta)} A_{ii}\mathbb{E}((\tilde{\theta} - \theta))^2.$$

Now since we are only choosing points from neighbouring cubes any coordinate of $\tilde{\theta} - \theta$ is atmost 2δ. So the total ℓ_2 distance of $\mathbb{E}((\tilde{\theta} - \theta)_i)^2$ is for all random choices atmost $k(\theta)4\delta^2$. So we actually have the bound

$$\mathbb{E}(\tilde{\theta} - \theta)^T A(\tilde{\theta} - \theta) \leq 4Tr(A)\delta^2.$$

So applying the above fact to the quadratic forms in the next display we have the following inequality

$$\mathbb{E}\frac{1}{2}\|\Psi_{in}\tilde{\theta} - \Psi_{in}\theta\|_2^2 + \mathbb{E}\frac{\alpha}{2}\|\Psi_f\tilde{\theta} - \Psi_f\theta\|_2^2 + \mathbb{E}l(\tilde{\theta} - \theta) \leq 2\delta^2(Tr(\Psi_{in}^T\Psi_{in} + \alpha\Psi_f^T\Psi_f)). \tag{84}$$

Now by the column normalization condition, we have that the trace terms are upper bounded by n. ■

Now we prove lemma

REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] S. Chatterjee A.R. Barron, "The information theoretic validity of penalized likelihood," 2014. Available at www.stat.yale.edu arb4