General Inequalities for Gibbs Posterior with Nonadditive Empirical Risk

Cheng Li*, Wenxin Jiang† and Martin A. Tanner‡

Department of Statistics, Northwestern University

Abstract

The Gibbs posterior is a useful tool for risk minimization, which adopts a Bayesian framework and can incorporate convenient computational algorithms such as Markov chain Monte Carlo. We derive risk bounds for the Gibbs posterior using some general nonasymptotic inequalities, which can be used to derive nearly optimal convergence rates and select models to optimally balance the approximation errors and the stochastic errors. These inequalities are formulated in a very general way that does not require the empirical risk to be a usual sample average over independent observations. We apply this framework to studying the convergence rate of the GMM risk (generalized method of moments) and deriving an oracle inequality for the ranking risk, where models are selected based on the Gibbs posterior with a nonadditive empirical risk.

1 Introduction

The Gibbs posterior is a random method of empirical risk minimization obtained from an analogy of statistical physics, where the empirical risk is identified with the

*Address correspondence to Cheng Li, Department of Statistics, Northwestern University, 2006 Sheridan Road, Evanston 60208, Illinois, U.S.A.; e-mail: chengli2014@u.northwestern.edu.

†e-mail: wjiang@northwestern.edu.

‡e-mail: mat132@northwestern.edu.

energy and low probabilities are assigned to high energy configurations. The method has recently been recognized by researchers from various fields, for example, information theorists (Zhang 1999, 2006), econometricians (Chernozhukov and Hong 2003), and statisticians (Jiang and Tanner 2008). Due to its Bayesian flavor, the Gibbs posterior allows application of convenient computational algorithms such as Markov chain Monto Carlo (Chernozhukov and Hong 2003, Belloni and Chernozhukov 2009, Chen, Jiang and Tanner 2010). Given observed data $\boldsymbol{D} = \{D_i, i = 1, 2, ..., n\}$, the general form of Gibbs posterior Q is defined as a probability measure constructed from an empirical risk R_n:

$$Q(\mathrm{d}\theta) = \frac{e^{-\lambda R_n(\theta)}\pi(\mathrm{d}\theta)}{\int_\Theta e^{-\lambda R_n(\theta')}\pi(\mathrm{d}\theta')}, \tag{1}$$

where θ is the parameter of interest, Θ is the space of θ, $R_n(\theta)$ is an empirical risk function that depends on both θ and the sample $\boldsymbol{D}$, λ is a positive scalar and π is a prior distribution over Θ. We will sometimes use $\tilde{\theta} \sim Q$ to denote a random variable that is generated from the Gibbs posterior Q, when it is intended to be distinguished from a dummy argument θ, or θ', of a risk function or of a probability density.

Compared to the posterior distribution derived from a likelihood based procedure, the Gibbs posterior may no longer have the usual interpretation of conditional probability given observed data unless $\lambda R_n(\theta)$ is exactly the negative log-likelihood. However, it can achieve better risk performance under model misspecification compared to the likelihood based Bayesian method, since the Gibbs posterior is directly associated with the risk function of interest (Jiang and Tanner 2008, Yao, Jiang and Tanner 2011).

Although the Gibbs posterior has been studied in different fields, the emphases placed and the terminologies used have been differed. Information theorists have used the term "Gibbs posterior" from the analogy to statistical physics, and have mostly considered the "additive empirical risk" (such as the classification error in machine learning), which is proportional to a sum over n terms from n independent subjects and corresponds to the additive energy of n noninteracting subsystems in statistical physics. Econometricians used the term "quasi-Bayesian" or "Laplace-type" posterior, and without referring to the statistical physics analogy, have been able to include more general nonadditive empirical risks such as the GMM (generalized method of moment) criterion function.

The emphasis of econometric research has been on using the Gibbs posterior to do

parametric inference (e.g., Chernozhukov and Hong 2003), which often studies the asymptotic distribution of the posterior based parameter estimates that typically involves regularity assumptions such as identifiability. The emphasis of information theorists, on the other hand, has been on the risk performance of the Gibbs posterior. While the study of risk performance can be regarded as an intermediate step for the parametric inference under identifiability conditions, it is sometimes an important problem itself, for example, when the risk is the probability of misclassification in machine learning. We note that although the approach used by information theorists usually involves fewer assumptions compared to the approach used by econometricians, they also get weaker results with these assumptions – that is, they only obtain bounds on the risk, but not a finer, distributional characterization of the risk.

In a recent work of the same style as the information theoretic literature, Jiang and Tanner (2008, Proposition 6) have established a nonasymptotic and assumption-free relation between the risk performance of Gibbs posterior and a probability of uniform large deviation, which allows for a theoretical study of the risk performance of Gibbs posterior in very general situations, including dependent data (e.g., Jiang and Tanner 2010) and panel data (e.g., Yao et al. 2011). In principle this general relation is applicable to accommodate nonadditive empirical risks such as the GMM criterion function. However, the relation in Jiang and Tanner (2008) does not always lead to sharp risk convergence rates. The connection they made to the large deviation rate $|R_n(\theta) - \mathrm{E}\, R_n(\theta)| = O_p(n^{-1/2})$ entails a similar risk convergence rate, which is not optimal for many situations when the optimal rate is of order $O_p(n^{-1})$.

The contributions of the current paper will include the following:

1. We extend the approach of Jiang and Tanner (2008) to derive a more general assumption-free inequality, which can be used to derive sharper risk convergence rates that are nearly of the optimal order $O(n^{-1})$ in many situations. We will show how this inequality can be applied to nonadditive empirical risks such as the GMM criterion function.

2. In addition, we derive an assumption-free oracle inequality for a model selection framework, so that nearly optimal risk performance will be achieved across a range

of models under consideration. We will provide an example on how this can be applied to the ranking problem, which involves an empirical risk in the form of a U-statistic that is nonadditive and is analogous to the energy of pairwise interactions in statistical physics.

3. Through these efforts, we demonstrate that the information theoretic approach based on our nonasymptotic inequalities can be successfully extended to the more general nonadditive risks beyond the additive risks studied in most of the current literature.

4. The inequalities that we derive are assumption-free. They reveal some simple yet fundamental relationship behind the construction of Gibbs posterior that are sometimes obscured among the regularity conditions used in more elaborate approaches. In our inequalities, the risk performance reflects how the empirical risk and the theoretical risk differ, and how restrictive the model is in approximating the optimal risk. Since we focus on the risk bounds similar to the machine learning literature, our method involves fewer assumptions compared to the econometric work on posterior asymptotic normality, and has potential application to partially identified situations.

Before we proceed, we provide some examples of empirical risks where data are assumed to be iid (independent and identically distributed). Example 0 involves an additive empirical risk and Examples 1 and 2 involve nonadditive empirical risks.

0. *Classification.* In a classification problem, $Y \in \{0, 1\}$ and $X \in \Re^p$ are random variable / vectors. Let $(Y_i, X_i^\top)_{i=1,\ldots,n}^\top = (D_i)_{i=1}^n = \boldsymbol{D}$ be iid copies of $(Y, X^\top)^\top = D$. Define $\ell(D, \theta) = |Y - I(X^\top\theta > 0)|$, where θ is a parameter in $\Re^p$, $I(\cdot)$ is the indicator function, and ℓ represents the classification loss of a linear rule $I(X^\top\theta > 0)$. Then the sample classification error

$$R_n(\theta) = n^{-1} \sum_{i=1}^{n} \ell(D_i, \theta)$$

is an *additive empirical risk* since it is proportional to a sum over the contributions from n independent individuals. The theoretical risk $R(\theta) = \mathrm{E}\, \ell(D, \theta)$ is the large sample limit of the empirical risk $R_n(\theta)$.

Most existing work in the machine learning literature has focused on this example of classification problem (e.g., Zhang 1999, 2006). In econometrics, recently Jun,

Pinske and Wan (2013) has proposed $n^{1/3}$-consistent Laplace estimators for various regression problems, including an analog to the maximum score estimator for the iid classification problem, and obtained very fine distributional results. (See also a related paper Jun, Pinske and Wan 2011.) In our paper, however, we will instead focus on the *nonadditive empirical risks* as given in the following examples.

1. *GMM (generalized method of moments).* GMM is related to the generalized estimating equation method or Z-estimation in statistics literature and we will consider the case when the weight matrix is the identity for simplicity. The empirical risk function is defined as

$$R_n(\theta) = \Big\| \frac{1}{n} \sum_{i=1}^{n} g(D_i, \theta) \Big\|^2 \tag{2}$$

where $g(D, \theta)$ is the p-dimensional moment vector satisfying the moment condition $\mathrm{E}\, g(D, \theta^*) = 0$ for some θ^*, and $\| \cdot \|$ denotes the L_2 norm. Here, $R_n(\theta)$ is the L_2 norm square of empirical moments and does not have the additive form of a sample average. The asymptotic normality of the Gibbs posterior with the GMM risk has been studied in Chernozhukov and Hong (2003) with a finite dimension of θ, and in Belloni and Chernozhukov (2009) with an increasing dimension of θ. By applying our general results in Section 2, we will show that the nonasymptotic inequalities can be used to derive a nearly optimal convergence rate for the theoretical risk $R(\tilde{\theta}) = \| \mathrm{E}\, g(D, \tilde{\theta}) \|^2$, where $\tilde{\theta} \sim Q$ is generated according to the Gibbs posterior. Such a risk convergence result does not require parametric identifiability conditions and may have potential application to the partially identified situation. In addition, we allow $\dim(\theta)$ to increase with n with order $o(n^{1/2}/\log n)$, which can be much faster than the approximate $o(n^{1/4})$ growth rate in Belloni and Chernozhukov (2009). On the other hand, the stronger assumption $\dim(\theta) = o(n^{1/4})$ in Belloni and Chernozukov (2009) leads to the stronger asymptotic normality result, compared to our convergence rate result for the GMM risk.

2. *Ranking estimation.* In ranking estimation, where the empirical risk function is defined as

$$R_n(\theta) = \frac{1}{n(n-1)} \sum_{i \neq j} I[(Y_i - Y_j) r(X_i, X_j; \theta) < 0], \tag{3}$$

where Y is a scalar random variable, $X \in \mathcal{X}$ is a random vector in $\Re^p$ and the ranking rule $r : \mathcal{X} \times \mathcal{X} \to \Re$ follows $r(x, x'; \theta) > 0$ if x ranks higher than x'

and $r(x, x'; \theta) \leq 0$ otherwise. Since $R_n(\theta)$ involves averaging over paired data, it is no longer an additive empirical risk. One special case of (3) is the maximum rank correlation estimator (MRC) in Han (1987) and Sherman (1993), where the ranking rule is the linear difference $r(x, x'; \theta) = (x - x')^\top \theta$ for $\theta \in \Re^p$. Han (1987) showed the strong consistency of MRC estimator and Sherman (1993) proved the $\sqrt{n}$ convergence rate and asymptotic normality. In this paper, our goal is not to estimate the parameter θ but instead to minimize the theoretical risk of mismatch $R(\theta) = P[(Y - Y')r(X, X'; \theta) < 0]$. The consistency and fast convergence rate of general ranking estimator that minimizes (3) and its convex upper bounds have been studied in recent frequentist papers such as Clémencon, Lugosi and Vayatis (2008) and Rejchel (2012). In this paper, we will introduce a method of random model selection using the Gibbs posterior, where we choose among linear ranking rules with a varying number of selected X-variables. By applying our general results in Section 4 to this method, we will show that an oracle performance of the ranking risk can be achieved to optimally balance between the model complexity and the approximation error.

The structure of the paper will be the following. In Section 2, we derive some general bounds on the risk performance of the Gibbs posterior. In Section 3, we apply the bounds to the GMM example and derive the nearly optimal risk convergence result. In Section 4, we consider the framework of model selection with Gibbs posterior and derive some general oracle inequalities for the risk performance. In Section 5, we demonstrate an application of the oracle inequality to the ranking risk. Section 6 includes some discussion. Technical proofs are included in the appendix.

2 General inequalities for risk convergence

We will first attempt to make no assumptions and derive an inequality for some theoretical risk of interest $R(\theta)$, which will be related to the empirical risk $R_n(\theta)$ and the prior $\pi(\theta)$ used to construct the Gibbs posterior (1). Due to the assumption-free nature of this approach, the relations here can (at least in principle) apply to a wide variety of cases, with either additive or nonadditive empirical risk, with iid data, time series, panel data, or spatial data.

We study, for $a \in \Re$, the *expected posterior probability* that a theoretic risk R exceeds a, i.e., $P_{\boldsymbol{D}} P_{\tilde{\theta}\sim Q}(R(\tilde{\theta}) > a)$, which will be abbreviated as $PQ(R(\tilde{\theta}) > a)$. Here P corresponds to the underlying true distribution of data $P_{\boldsymbol{D}}$, Q corresponds to the Gibbs posterior distribution of $\tilde{\theta}$ conditional on the data $\boldsymbol{D}$, and PQ is defined to be the joint measure $P_{\boldsymbol{D}} P_{\tilde{\theta}\sim Q}$. Different from the prior π used in the Gibbs posterior, PQ can be understood as a mixture distribution that measures the random outcome of the following sampling process: (i) sampling a data set $\boldsymbol{D}$ from the underlying true distribution P, (ii) sampling a parameter $\tilde{\theta}$ from the resulting Gibbs posterior Q conditional on the data $\boldsymbol{D}$ sampled from step (i).

To bound the probability $PQ(R(\tilde{\theta}) > a)$, we construct a *simultaneous coverage interval* for the empirical risk $R_n(\theta)$ appearing in the Gibbs posterior, using the theoretic risk $R(\theta)$. Let $0 < s_1 \leq 1 \leq s_2$ and $\Delta \equiv \Delta(\theta) \geq 0$ be some nonstochastic quantities, possibly dependent on sample size n. Define an event $A = [\forall \theta, s_1 R(\theta) - \Delta(\theta) \leq R_n(\theta) \leq s_2 R(\theta) + \Delta(\theta)]$ and its complement A^c. Then $P(A)$ is the uniform coverage probability of $[s_1 R(\theta) - \Delta(\theta), s_2 R(\theta) + \Delta(\theta)]$ for $R_n(\theta)$. [1] Note that Δ is related to the radius of the coverage interval and is analogous to the standard deviation of $R_n(\theta)$ which characterizes its stochastic error. In applications, the quantity Δ usually decreases with n at a certain rate. Although it is ideal to have $s_1 = s_2 = 1$ (which was the choice of Jiang and Tanner 2008, Proposition 6), we will later see that sometimes it is better to take $s_1 = 1 - \delta$ and $s_2 = 1 + \delta$ for some small positive δ, to allow a smaller radius Δ for a given coverage probability.

We then have the following proposition for the excess probability $PQ(R(\tilde{\theta}) > a)$, where $\tilde{\theta}$ is randomly drawn from the Gibbs posterior Q.

Proposition 1. *(i) When Δ is possibly dependent on θ, we have the following: for any $u \in \Re$:*

$$PQ\Big(R(\tilde{\theta}) > \bar{R} + \frac{u}{s_1 \lambda}\Big) \leq e^{-u} + P\Big(\exists \theta : R_n(\theta) \notin [s_1 R(\theta) - \Delta(\theta), s_2 R(\theta) + \Delta(\theta)]\Big), \quad (4)$$

where

$$\bar{R} \equiv -\frac{1}{s_1 \lambda} \log \left[\frac{\int e^{-\lambda(s_2 R(\theta) + \Delta)} \pi(\mathrm{d}\theta)}{\int e^{\lambda \Delta} \pi(\mathrm{d}\theta)} \right].$$

(ii) When Δ is chosen to be common for all θ, we have the following: let Θ be the support of the prior π. For any $u \in \Re$ and any $v > 0$:

$$PQ\Big(R(\tilde{\theta}) > \tilde{R}(v) + \frac{u}{s_1 \lambda}\Big) \leq e^{-u} + P\Big(\exists \theta : R_n(\theta) \notin [s_1 R(\theta) - \Delta, s_2 R(\theta) + \Delta]\Big), \quad (5)$$

where

$$\tilde{R}(v) \equiv \frac{s_2}{s_1}\big(\inf_{\theta\in\Theta} R(\theta) + v\big) + \frac{1}{s_1\lambda}\log \pi(\theta : R(\theta) < \inf_{\theta'\in\Theta} R(\theta') + v)^{-1} + \frac{2\Delta}{s_1}. \quad (6)$$

Remark 1. If we use Proposition 1 to bound the probability of a large excess risk $R(\tilde{\theta}) - \inf_\theta R(\theta)$, then a fundamental relation is revealed by these inequalities: the performance of the excess risk $R(\tilde{\theta}) - \inf_\theta R(\theta)$ is mainly influenced by two factors: the excess of a *nonstochastic* bound $\bar{R} - \inf_\theta R(\theta)$ or $\tilde{R}(v) - \inf_\theta R(\theta)$, as well as a *stochastic* difference between $R_n(\theta)$ and $R(\theta)$, as reflected in $P\big(\exists\theta : R_n(\theta) \notin [s_1R(\theta) - \Delta(\theta), s_2R(\theta) + \Delta(\theta)]\big)$. Now we will discuss our choices of the tuning parameters s_1, s_2, λ, u, v, Δ in the proposition.

- We can choose $s_1 = 1 - \delta$ and $s_2 = 1 + \delta$ for some small positive δ. This can lead to a better risk convergence rate compared to the choice $s_1 = s_2 = 1$ made in Jiang and Tanner (2008), Proposition 6.

- The scalar λ is usually set to be $\lambda = n\psi$, where $\psi > 0$ is a constant sometimes called "inverse temperature" in statistical mechanics.

- We will let $u = 2\log n$ so that $e^{-u} = n^{-2}$.

- We will choose $v = \Delta$ (or about the same order).

- We will choose Δ such that $P\big(\exists\theta : R_n(\theta) \notin [s_1R(\theta) - \Delta, s_2R(\theta) + \Delta]\big)$ can be controlled by $e^{-u} = n^{-2}$.

Therefore given these choices, the right-hand side of the inequalities can be bounded by $2n^{-2}$, implying that

$$R(\tilde{\theta}) \le \tilde{R}(\Delta) + \frac{2\log n}{ns_1\psi} \quad (7)$$

holds for all large n, almost surely in the measure PQ, by Borel-Cantelli lemma, where $\tilde{\theta}$ on the left-hand side is drawn from the Gibbs posterior Q given the data $\boldsymbol{D}$. Here "almost surely in the measure PQ" can be understood in the following way. First, we can rewrite P as P_n and Q as Q_n because both the true probability P and the Gibbs posterior Q depend on the sample size n. Now we want to study the probability about the *joint* event "$A_n = \{R(\tilde{\theta}) \le \tilde{R}(\Delta) + 2\log n/(ns_1\psi)\}$ happens for *all* large enough n". Consider a setup where the data $\boldsymbol{D}$ are drawn independently across different n and

define the product measure $\widetilde{PQ} = P_1Q_1 \times P_2Q_2 \times ....$ Then the event A_n happens for all large enough n, almost surely with respect to this $\widetilde{PQ}$ measure by Borel-Cantelli lemma. Without confusion, we will still write "A_n happens for all large n almost surely with respect to the measure PQ".

Remark 2. Now we discuss the nonstochastic bounds $\bar{R}$ and $\tilde{R}(v)$. The bound $\bar{R}$ in result (i) can be applied to the case of model selection in Section 4, where Δ is specified to depend on the model complexity. The bound $\tilde{R}(v)$ in result (ii) is useful when one chooses Δ to be common for all θ. For example, we will choose $\Delta = p \log n/n$ in the GMM example with $p = \dim(\theta)$. Together with the choices of other tuning parameters in Remark 1, we can show that $\tilde{R}(\Delta) = (s_2/s_1) \inf_{\theta\in\Theta} R(\theta) + O(\Delta)$. Therefore according to (7), we have obtained that

$$R(\tilde{\theta}) \leq \frac{s_2}{s_1} \inf_{\theta\in\Theta} R(\theta) + O(\Delta),$$

for all large n, almost surely in PQ, where $\tilde{\theta}$ on the left-hand side is drawn from the Gibbs posterior Q. When the term $\inf_{\theta\in\Theta} R(\theta)$ is either zero or standardized to be zero by a translation (by redefining the risk to be relative to the best achievable over $\theta \in \Theta$), the risk convergence rate is simply $O(\Delta)$.

Proof of Proposition 1:

For (i): Hereafter the notation PQf denotes the expectation of a random function f under the measure PQ. For any $a \in \Re$, $PQ[R(\tilde{\theta}) > a] \leq PQI[R(\tilde{\theta}) > a]I[A] + P[A^c]$, where $A = [\forall\theta, s_1R(\theta) - \Delta(\theta) \leq R_n(\theta) \leq s_2R(\theta) + \Delta(\theta)]$.

The first term can be bounded by

$$\begin{aligned} & PQI[R(\tilde{\theta}) > a]I[A] = P\left\{ I[A] \cdot \frac{\int I[R(\theta) > a]e^{-\lambda R_n(\theta)}\pi(\mathrm{d}\theta)}{\int e^{-\lambda R_n(\theta)}\pi(d\theta)} \right\} \\ \leq \;& P\left\{ \frac{\int I[R(\theta) > a]e^{-\lambda(s_1R(\theta)-\Delta)}\pi(\mathrm{d}\theta)}{\int e^{-\lambda(s_2R(\theta)+\Delta)}\pi(\mathrm{d}\theta)} \right\} \leq P\left\{ \frac{\int e^{-\lambda(s_1a-\Delta)}\pi(\mathrm{d}\theta)}{\int e^{-\lambda(s_2R(\theta)+\Delta)}\pi(\mathrm{d}\theta)} \right\} \\ \leq \;& e^{-\lambda s_1 a}\left\{ \frac{\int e^{-\lambda(s_2R(\theta)+\Delta)}\pi(\mathrm{d}\theta)}{\int e^{\lambda\Delta}\pi(\mathrm{d}\theta)} \right\}^{-1}. \end{aligned}$$

Then take $a = \bar{R} + u/(s_1\lambda)$ and (4) follows.

For (ii): We show that $\bar{R} \leq \tilde{R}(v)$ for all $v > 0$ and then apply (i). When Δ is a constant in θ, it is obvious that $\bar{R} = -(s_1\lambda)^{-1} \log \int e^{-s_2\lambda R(\theta)}\pi(\mathrm{d}\theta) + 2\Delta/s_1$. Now we lower bound

the integral in the first term, by restricting it to the region where $R(\theta) < \inf_{\theta' \in \Theta} R(\theta') + v$ as

$$\begin{aligned}\int_\Theta e^{-s_2 \lambda R(\theta)} \pi(\mathrm{d}\theta) &\geq \int_{R(\theta) < \inf_{\theta' \in \Theta} R(\theta') + v} e^{-s_2 \lambda (\inf_{\theta \in \Theta} R(\theta) + v)} \pi(\mathrm{d}\theta) \\ &= e^{-s_2 \lambda (\inf_{\theta' \in \Theta} R(\theta') + v)} \pi(\theta : R(\theta) < \inf_{\theta' \in \Theta} R(\theta') + v),\end{aligned}$$

which concludes the proof. ■

3 Example: Convergence of GMM risk

We now apply Proposition 1 to the GMM risk defined by (2). Our theoretical risk of interest is $R(\theta) = \| \mathrm{E}\, g(D, \theta)\|^2$, the limit of $R_n(\theta)$ in probability for each fixed $\theta \in \Re^p$. Let $|\cdot|_q$ be the L_q norm for $q \in [1, \infty]$. Then we have the following theorem for $R(\tilde{\theta})$ with $\tilde{\theta}$ sampled from the Gibbs posterior Q. The proofs are given in the appendix.

Theorem 1. *Suppose the following regularity conditions G1-G5 on the moments and the prior hold.*

(G1) $\pi(\theta)$ is a continuous distribution restricted on $\Theta_n = \big\{\theta : \|\theta\| \leq \sqrt{p} \log n\big\}$. For any $\theta_0 \in \Theta_n$, any small enough $\delta > 0$, there exists a constant $\zeta > 0$ such that $\pi(\{\theta : \|\theta - \theta_0\| < \delta\}) \geq (\delta n^{-\zeta})^p$ for all sufficiently large n.

(G2) Let $S^p = \{\eta \in \Re^p : \|\eta\| = 1\}$. Assume both $\sup_{\eta \in S^p} \mathrm{E}[(\eta^\top g(D, \theta^*))^2]$ and $\sup_{\eta \in S^p} (\mathrm{E}[(\eta^\top (g(D, \theta) - g(D, \theta^*)))^2])$ are uniformly bounded on Θ_n for some $\theta^* \in \Theta_n$ that satisfies $\mathrm{E}\, g(D, \theta^*) = 0$.

(G3) $|g(D, \theta)|_\infty \leq c_g \sqrt{\log n}$ for some constant $c_g > 0$, for all D almost surely and all $\theta \in \Theta_n$ when n is sufficiently large.

(G4) $\mathrm{E}\, g(D, \theta)$ is continuously differentiable in θ, and $|\partial \mathrm{E}\, g_j(D, \theta)/\partial \theta_k| \leq n^\xi$ with some constant $\xi > 0$ uniformly for all $\theta \in \Theta_n$, for all $j = 1, 2, ..., p$, $k = 1, 2, ..., p$ and large n.

(G5) Let $H_{[]}(\varepsilon, \mathcal{F})$ be the L_1 bracketing entropy of the class of functions $\mathcal{F} = \{\eta \in S^p, \theta \in \Theta_n : \eta^\top [g(D, \theta) - \mathrm{E}\, g(D, \theta)]\}$ for any $\varepsilon > 0$. [2] Then $H_{[]}(\varepsilon, \mathcal{F}) \leq c_1 p \log(c_2 n/\varepsilon)$ for some positive constants c_1, c_2.

Then for $p = o(n^{1/2}/\log n)$,

(i) there exists a constant $C_1 > 0$, such that $PQ\big(R(\tilde{\theta}) > C_1 p \log n/n\big) \leq 2n^{-2}$ for each sufficiently large n;

(ii) there exists a constant $C_2 > 0$, such that $PQR(\tilde{\theta}) \leq C_2 p \log n/n$ for each sufficiently large n;

(iii) $R(\tilde{\theta})$ converges to zero almost surely with respect to the measure PQ at the rate no slower than $p \log n/n$ for all sufficiently large n, where $\tilde{\theta}$ is randomly drawn from the Gibbs posterior (1).

Remark 3. Part (i) can be derived from the inequality (5) and setting s_1, s_2, λ, u, v as in Remark 1 and $\Delta = O(p \log n/n)$. Part (iii) immediately follows by applying Borel-Cantelli lemma as in Remark 1. Part (ii) says the convergence rate of the posterior mean of $R(\tilde{\theta})$ is of order $O(p \log n/n)$, close to the optimal rate of $O(p/n)$ in the parametric settings. See for example, Ghosal (2000) for generalized linear models, Wang (2011) for generalized estimating equations. Belloni and Chernozhukov (2009) considered general Z-estimation and proved the asymptotic normality of Gibbs posterior with a uniform prior, which immediately implies the $O(p/n)$ convergence rate for the risk $R(\tilde{\theta})$. The Bayesian central limit theorems (or Bernstein-von Mises theorems) in Ghosal (2000) and Belloni and Chernozhukov (2009) are stronger distributional results than our result on risk bounds. On the other hand, since only the risk is of interest in this paper, our assumptions are also weaker than theirs. We do not assume point identification, and the dimension of parameters p in Theorem 1 can grow as fast as $n^{1/2}$ up to some logarithm factors, which is less restrictive than the growth rate about $o(n^{1/4})$ in the condition ZE3 of Belloni and Chernozhukov (2009). It is not clear to us if the extra logarithm factor in the rate $p \log n/n$ can be removed under our weaker assumptions.

Remark 4. Our regularity conditions are mild and general enough to cover the commonly used moment conditions for linear regression and quantile regression. The assumption G1 says that the prior does not vanish too fast in n on any small neighborhood in Θ_n, which is satisfied by many commonly used priors, such as a uniform prior or a normal prior truncated on Θ_n. Note that since the radius of Θ_n is growing with n, such a prior is not restrictive. Assumptions G2-G5 require that the variance, the L_∞ norm, the derivative and the entropy of the moments are bounded in certain order. They are high level conditions comparable with the condition ZE1 in Belloni and Chernozhukov (2009). We can see immediately that they work well, for example, for the quantile regression in the next proposition. The proof is given in the appendix.

Proposition 2. *Suppose $D_i = (Y_i, X_i^\top)^\top, i = 1, 2, ..., n$ are an iid sample where Y is a scalar random variable and X is a p-dimensional random vector including a constant component. The conditional distribution satisfies $F^{-1}_{Y|X}(\tau) = X^\top\theta^*$ for some θ^*, with $\tau \in (0, 1)$ being a fixed quantile. Define $g(D, \theta) = X[I(Y \le X^\top\theta) - \tau]$. Assume that*

1. $|X_j| \le M$ almost surely for all $1 \le j \le p$ for some constant $M > 0$, where X_j is the jth component of X.

2. The conditional distribution $F_{Y|X}$ is continuous with a density $f_{Y|X}$ that is bounded above by $\bar{f} < \infty$.

3. Eigenvalues of $E[XX^\top]$ are bounded above by a constant.

Then it follows that the assumptions G2-G5 are satisfied.

4 General oracle inequalities for model selection

In this section we present a general oracle inequality for random model selection with the Gibbs posterior. In the definition of the Gibbs posterior (1), we let the parameter $\theta = (b, m)$, where m is a model index ($m = 1, 2, ...$) with corresponding model space B_m and b is a parameter in B_m. Sometimes without confusion, we also use m to denote the dimension of B_m. In the model selection framework, the prior distribution can be usually decomposed into $\pi(\mathrm{d}b, m) = \pi(\mathrm{d}b|m)\pi_m$ where π_m is a discrete prior distribution over all models considered, and $\pi(\mathrm{d}b|m)$ is the prior of b on model space B_m. Then (1) can be equivalently written as

$$Q(\mathrm{d}b, m|\boldsymbol{D}) = \frac{e^{-\lambda R_n(b,m)}\pi(\mathrm{d}b|m)\pi_m}{\sum_{m'}\int_{B_{m'}} e^{-\lambda R_n(b',m')}\pi(\mathrm{d}b'|m')\pi_{m'}} \tag{8}$$

In general, we are still interested in a theoretical risk $R(\theta)$, for which $R_n(\theta)$ is the corresponding empirical risk. For example, for additive empirical risk $R_n(\theta) = n^{-1}\sum_{i=1}^n \ell(D_i, \theta)$, we have $R(\theta) = \mathrm{E}_D\,\ell(D, \theta)$. For a non-adidtive empirical risk such as the ranking risk (3), $R(\theta) = P_{X,X',Y,Y'}[(Y - Y')r(X, X'; \theta) < 0]$. Hereafter without confusion, we sometimes omit the dependence on the model index m and write $R_n(b) \equiv R_n(\theta)$ and $R(b) \equiv R(\theta)$ when a model m is given and $b \in B_m$.

The goal here is to let the Gibbs posterior propose good parameters $\tilde{\theta} = (\tilde{b}, \tilde{m})$ with small theoretical risk $R(\tilde{\theta})$. The current framework of model selection can therefore be regarded as special case of the general setup of Section 2. The main difference, however,

is that the dimension of the parameter b can change with the model index m. A constant choice of the margin parameter Δ in Proposition 1 (ii) (common to all models) would not be able to lead to sharp results for all models. Instead, we use the more general setup of Proposition 1(i), and assume from now on that $\Delta(\theta) = \Delta_m$ which depends on model index m (but not b).

Define for any model m and any $v > 0$,

$$\tilde{R}_m(v) \equiv \inf_{b \in B_m} R(b) + v + \frac{1}{s_2 \lambda} \log \pi(b : R(b) < \inf_{b' \in B_m} R(b') + v | m)^{-1}$$

and also

$$\tilde{R}(v) \equiv \inf_m (s_2 \tilde{R}_m(v) + \Delta_m + \lambda^{-1} \log \pi_m^{-1})/s_1 + \tilde{\Delta}/s_1, \tag{9}$$

where

$$\tilde{\Delta} \equiv \lambda^{-1} \log(\sum_m \pi_m e^{\lambda \Delta_m}).$$

The following oracle inequality can be derived from Proposition 1.

Proposition 3. *(Oracle Inequality) For any $u \in \Re$ and $v > 0$,*

$$PQ\Big(R(\tilde{\theta}) > \tilde{R}(v) + \frac{u}{s_1 \lambda}\Big) \leq e^{-u} + P\Big(\exists (b, m) : R_n(b) \notin [s_1 R(b) - \Delta_m, s_2 R(b) + \Delta_m]\Big) \tag{10}$$

where $\tilde{R}(v)$ is defined in (9).

Remark 5. Similar to Remark 2 after Proposition 1, the inequality here typically implies the following oracle relation:

$$R(\tilde{\theta}) \leq \frac{s_2}{s_1} \inf_m \Big\{ \inf_{b \in B_m} R(b) + O(\Delta_m) \Big\}$$

for all large n, almost surely in PQ, where on the left-hand side $\tilde{\theta} = (\tilde{b}, \tilde{m})$ is drawn from the Gibbs posterior (8). The term $\inf_{b \in B_m} R(b)$ measures the accuracy of model m, which typically improves (decreases) when the dimension or complexity m increases, and the radius of the coverage interval Δ_m measures the size of the stochastic error of model m, which typically increases with the model dimension m and decreases with the sample size n. In our ranking risk example, we will choose $\Delta_m = m(\log n)^3/n$. The constant ratio s_2/s_1 can be arbitrarily close to 1.

Remark 6. The prior on models π_m is usually chosen to decrease with the model dimension m to reflect our preference for more parsimonious models. To make the term $\tilde{\Delta}$

in (9) controllable, we will set, for example, $\pi_m \propto e^{-2\lambda\Delta_m}$ to offset the effect from the weighted average of $e^{\lambda\Delta_m}$. This introduces a BIC-type penalization on the model size when Δ_m is linear in m, as in our ranking risk example later.

Proof of Proposition 3:

The result is proved by first showing that (*) $\bar{R} \leq \tilde{R}(v)$ for all $v > 0$ and $\bar{R}$ defined by Proposition 1, and then applying Proposition 1(i). In the expression of $\bar{R}$ in Proposition 1(i), the denominator is exactly $e^{\lambda\tilde{\Delta}}$, and the numerator is lower-bounded as

$$\begin{aligned} & \int e^{-\lambda(s_2 R(\theta)+\Delta)}\pi(\mathrm{d}\theta) = \sum_m \int_{B_m} e^{-\lambda(s_2 R(b)+\Delta_m)}\pi_m \pi(\mathrm{d}b|m) \\ \geq & \sup_m \int_{B_m} e^{-\lambda(s_2 R(b)+\Delta_m+\lambda^{-1}\log\pi_m^{-1})}\pi(\mathrm{d}b|m) = \sup_m e^{-\lambda(s_2\hat{R}_m+\Delta_m+\lambda^{-1}\log\pi_m^{-1})}, \end{aligned}$$

where $\hat{R}_m \equiv -(s_2\lambda)^{-1}\log\int e^{-s_2\lambda R(b)}\pi(\mathrm{d}b|m)$. These lead to (**) $\bar{R} \leq \inf_m(s_2\hat{R}_m + \Delta_m + \lambda^{-1}\log\pi_m^{-1})/s_1 + \tilde{\Delta}/s_1$. Now we bound the integral in $\hat{R}_m$ using the same technique as in the proof of Proposition 1(ii) and obtain: $\hat{R}_m \leq \inf_{b'\in B_m} R(b') + v + (s_2\lambda)^{-1}\log\pi(b : R(b) < \inf_{b'\in B_m} R(b') + v|m)^{-1} \equiv \tilde{R}_m(v)$ for all $v > 0$. Therefore (**) implies (*). ■

5 Example: Oracle performance of ranking risk with model selection

We now apply Proposition 3 to the model selection problem of the ranking risk R_n defined by (3), to select the best linear rule in which only part of the components in X are active. The targeted theoretical risk is the probability of mismatch $R = P[(Y - Y')r(X, X') < 0]$. Proposition 1 in Clémencon et al. (2008) indicates that the best rule possible in theory, namely the Bayes rule, is $r^*(X, X') = P(Y - Y' > 0|X, X') - P(Y - Y' < 0|X, X')$, (or any sign-preserving equivalent). Define the corresponding theoretical risk $R^* = P[(Y - Y')r^*(X, X') < 0]$ as the *optimal Bayes risk*. In general, r^* may depend on X, X' nonparametrically. In the following, we focus on the case where Y is a binary variable taking values in $\{-1, 1\}$, X is a p-dimensional random vector with p growing with n, and consider the set of linear rules $\mathscr{R} = \{b \in \Re^p : r(x, x'; b) = (x - x')^\top b\}$. We assume that the constant component $X_1 = 1$ is always present in the model, and restrict $b_1 = \pm 1$ as a normalization for identification purpose. The parameter is then $\theta = (b, m)$ with $m = 1, 2, ..., p$, where $b \in B_m$ and B_m is the union of all m-dimensional coordinate

subspaces B_{mj} for $j = 1, 2, ..., \binom{p}{m}$. We then have the following theorem for $R(\tilde{\theta})$ with $\tilde{\theta} = (\tilde{b}, \tilde{m})$ sampled from the Gibbs posterior. The proofs are given in the appendix.

Theorem 2. *Suppose the following regularity conditions R1-R4 hold:*

(R1) For any $m = 0, 1, 2, ..., p$, $\pi_m \propto e^{-2\psi m(\log n)^3}$. The priors of all submodels B_{mj} with size m are the same $\binom{p}{m}^{-1}\pi_m$.

(R2) $\pi(b|m, j)$ is a continuous distribution restricted on $\Theta_n = \{b : \|b\| \leq \sqrt{p}\log n\}$, for $1 \leq m \leq p$ and $1 \leq j \leq \binom{p}{m}$. For any $b_0 \in \Theta_n$, any small enough $\delta > 0$, there exists a constant $\zeta > 0$ such that $\pi(\{b : \|b - b_0\| \leq \delta\}|m, j) \geq (\delta n^{-\zeta})^m$ uniformly for all m, j and all sufficiently large n.

(R3) $\mathrm{E}_{X',Y'} I[(y - Y')(x - X')^\top b < 0]$ is continuously differentiable in b for all x, y, and the partial derivatives are bounded as $|\partial \mathrm{E}_{X',Y'} I[(y - Y')(x - X')^\top b < 0]/\partial b_k| \leq n^\xi$ for some constant $\xi > 0$ uniformly for $k = 1, 2, ..., p$, all x, y, and all sufficiently large n.

(R4) The conditional expectation $\eta(X) = \mathrm{E}[Y|X] = P(Y = 1|X)$ has an absolute continuous distribution on $[0, 1]$ with density bounded above by constant $\bar{f}_\eta$.

Then for $p = o(n/(\log n)^3)$, for all $m = 1, 2, ..., p$ and $j = 1, 2, ..., \binom{p}{m}$,

(i) for any $\delta > 0$, for each sufficiently large n,

$$PQ\Big\{R(\tilde{\theta}) - R^* > (1 + \delta)\inf_{m,j}\Big[\inf_{b \in B_{mj}}(R(b) - R^*) + \frac{6m(\log n)^3}{n}\Big]\Big\} \leq 2n^{-2};$$

(ii) for any $\delta > 0$, for each sufficiently large n,

$$PQR(\tilde{\theta}) \leq R^* + (1 + \delta)\inf_{m,j}\Big[\inf_{b \in B_{mj}}(R(b) - R^*) + \frac{7m(\log n)^3}{n}\Big];$$

(iii) for any $\delta > 0$, almost surely in the measure PQ for all sufficiently large n,

$$R(\tilde{\theta}) - R^* \leq (1 + \delta)\inf_{m,j}\Big[\inf_{b \in B_{mj}}(R(b) - R^*) + \frac{6m(\log n)^3}{n}\Big]$$

where $\tilde{\theta} = (\tilde{b}, \tilde{m})$ is randomly drawn from the Gibbs posterior (8).

Remark 7. Consider the special case of dimensional reduction, where the optimal Bayes risk R^* is achievable in a lower dimensional model with unknown dimension $m \leq p$, so that $\inf_{b \in B_{mj}}(R(b) - R^*) = 0$. Then the current theorem implies a near optimal rate $O(m(\log n)^3/n)$ which is *oracle* in the sense that it depends on the *unknown* dimension m of the best linear rule. As a result, the excess risk $R(\tilde{\theta}) - R^*$ converges at this oracle rate, both in the posterior mean and in the sense of almost sure convergence. In this special

case of dimensional reduction, our work has extended the bipartite ranking example (Example 5.1) in Clémencon et al. (2008) in two ways:

1. We allow a framework of adaptive model selection;

2. We achieve a fast oracle rate of about $O(m/n)$, which does not depend on the smoothness parameter α in their Assumption 4.

In fact, our condition R4 guarantees that the α parameter in their paper can take a value arbitrarily close to 1, which allows us to make it depend on n and derive an improved convergence rate, as compared to Corollary 8 of Clémencon et al. (2008).

Remark 8. The condition R1 assigns a prior on models that decreases exponentially fast with the model size, which favors parsimonious models, while all models of the same size has equal prior mass. R2 is a condition about the prior on parameter b similar to G1 in the GMM example. We require a uniform lower bound of the prior probability in any small neighborhood over all submodels. This condition is satisfied, for example, by a uniform prior on each B_{mj}. R3 and R4 impose mild bounds on the partial derivative of the conditional expectation and the density of $\eta(X)$.

Remark 9. In general, we do not require that the Bayes rule $r^*(x, x')$ belongs to the linear family $\mathscr{R}$, nor do we make any model assumptions on the relation between Y and X. Han (1987) proposed a generalized regression model $Y = F_2 \circ F_1(X^\top b^*, \epsilon)$, with b^* being the unknown true parameter, ϵ independent of X, F_1 strictly increasing in both arguments, and F_2 monotonely increasing. By taking $F_1(x_1, x_2) = x_1 + x_2$ and $F_2(x) = I(x \geq 0)$, this becomes a binary choice model of $Y = I(X^\top b^* + \epsilon \geq 0)$ and b^* can be estimated by minimizing (3). This MRC estimator is shown to be $\sqrt{n}$ consistent for b^* and asymptotically normal in Sherman (1993). In our general setup, the true parameter b^* may not exist since we do not assume the existence of such a single index model. Instead, the Bayes rule always exists and we are interested in the performance of the excess risk $R(\tilde{\theta}) - R^*$ with $\tilde{\theta}$ sampled from the Gibbs posterior.

6 Discussion

In this paper, we have introduced some assumption-free inequalities that are useful for studying the performance of Gibbs posterior as a random method of risk minimization. We now discuss several directions of extensions and possible future work:

1. In our examples, we have only considered prior distributions with compact support Θ_n. However, our general inequalities can be directly extended to accommodate prior distributions with thin tails on a noncompact support (such as a normal prior), similar to Jiang and Tanner (2008).

2. Although we have assumed iid data for both GMM and ranking examples, in principle our general inequalities can be applied to dependent data or panel data and improve the convergence results of, e.g., Jiang and Tanner (2010) and Yao et al. (2011).

3. In this paper we have mainly focused on nonadditive empirical risks that have not been commonly studied in the information theoretic literature. It is worth noting that our method certainly encompasses the common additive risk such as the classification risk and the mean square risk for linear regression, and can provide nearly optimal and oracle convergence rates.

4. In the formula of the Gibbs posterior (1), the scaling parameter λ has been taken to be $n\psi$ in this paper, where ψ can be any positive constant, without affecting the risk performance results derived in this paper. We note that ψ corresponds to the inverse temperature in statistical mechanics, and in the classification literature, researchers have considered choosing ψ using data-driven methods such as cross validation (see, e.g., Zhang 1999, Audibert 2004, Catoni 2007). It is an interesting future problem to explore how to choose ψ based on data in our more general setup with possibly nonadditive empirical risk.

Notes

[1]This involves a joint probability over an uncountable space of θ. One can use the outer probability P^* if the measurability problem is concerned. See for example, Section 1.2 of van der Vaart and Wellner (1996).

[2]The L_1 bracketing entropy $H_{[]}(\varepsilon, \mathcal{F})$ is the logarithm of the number of paired functions $[f(D;\eta_1,\theta_1), f(D;\eta_2,\theta_2)]$, such that for any $f(D;\eta,\theta) \in \mathcal{F}$, there exists a pair satisfying $f(D;\eta_1,\theta_1) \le f(D;\eta,\theta) \le f(D;\eta_2,\theta_2)$ for all values of D and $|f(D;\eta_1,\theta_1) - f(D;\eta_2,\theta_2)|_1 \le \varepsilon$.

Acknowledgments

We thank Professor Elie Tamer for useful discussions. The second author thanks Qilu

Securities Institute for Financial Studies, Shandong University, for the hospitality during his visit, when part of this work was done. We thank the Co-Editor Professor Victor Chernozhukov and two anonymous referees for useful comments that have improved the presentation of this paper.

References

[1] Audibert, J. Y. (2004). Classification using Gibbs estimators under complexity and margin assumptions. Technical report, Laboratoire de Probabilités et Modelès Aléatoires, http://www.proba.jussieu.fr/mathdoc/textes/PMA-908.pdf.

[2] Belloni, A. & V. Chernozhukov (2009). On the computational complexity of MCMC-based estimators in large samples. *The Annals of Statistics* 37, 2011-2055.

[3] Catoni, O. (2007). *PAC-Bayesian Supervised Classification (The Thermodynamics of Statistical Learning)*. Lecture Notes-Monograph Series, Vol 56, IMS.

[4] Chen, K., W. Jiang & M. A. Tanner (2010). A note on some algorithms for the Gibbs posterior. *Statistics and Probability Letters* 80, 1234-1241.

[5] Chernozhukov, V. & H. Hong (2003). An MCMC approach to classical estimation. *Journal of Econometrics* 115, 293-346.

[6] Clémencon, S., G. Lugosi & N. Vayatis (2008). Ranking and empirical minimization of U-statistics. *The Annals of Statistics* 36, 844-874.

[7] Ghosal, S. (2000). Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *Journal of Multivariate Analysis* 73, 49-68.

[8] Han, A. K. (1987). Non-parametric analysis of a generalized regression model - the maximum rank correlation estimator. *Journal of Econometrics* 35, 303-316.

[9] Jiang, W. & M. A. Tanner (2008). Gibbs posterior for variable selection in high dimensional classification and data mining. *The Annals of Statistics* 36, 2207-2231.

[10] Jiang, W. & M. A. Tanner (2010). Risk minimization for time series binary choice with variable selection. *Econometric Theory* 26, 1437-1452.

[11] Jun, S. J., J. Pinske & Y. Wan (2011). $\sqrt{n}$-consistent robust integration-based estimation. *Journal of Multivariate Analysis* 102, 828-846.

[12] Jun, S. J., J. Pinske & Y. Wan (2013). Classical Laplace estimation for $\sqrt[3]{n}$-consistent estimators: improved convergence rates and rate-adaptive inference. Technical Report, http://joris.econ.psu.edu/papers/Jun_Pinkse_Wan_cuberoot.pdf.

[13] Massart, P. (2003). Concentration inequalities and model selection. Springer, Berlin.

[14] Rejchel, W. (2012). On ranking and generalization bounds. *Journal of Mahine Learning Research* 13, 1373-1392.

[15] Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica* 61, 123-137.

[16] van der Vaart, A. W. & J. A. Wellner (1996). *Weak convergence and empirical process*. Springer, New York.

[17] Wang, L. (2011). GEE analysis of clustered binary data with diverging number of covariates. *The Annals of Statistics* 39, 389-417.

[18] Yao, L., W. Jiang & M. A. Tanner (2011). Predicting panel data binary choice with the Gibbs posterior. *Neural Compututation* 23, 2683-2712.

[19] Zhang, T. (1999). Theoretical analysis of a class of randomized regularization methods. *COLT '99 Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, 156-163.

[20] Zhang, T. (2006). Information theoretical upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory* 52, 1307-1321.

Appendix

In the appendix, we include the proofs of Theorem 1, Proposition 2 and Theorem 2. To prove Theorem 1, we first prove the following lemma about the coverage probability on the right-hand side of the inequality (5).

Lemma 1. *Suppose the assumptions G2-G5 in Theorem 1 hold. For any $u > 0$, uniformly for all $\theta \in \Theta_n$, for some constant $K > 1$, for all sufficiently large n, there exists a constant $C > 0$ such that with probability at least $1 - e^{-u}$,*

$$R_n(\theta) \in \Big[(1 - K^{-1})R(\theta) - CK\Big(\frac{p \log n}{n} + \frac{u}{n}\Big), (1 + K^{-1})R(\theta) + CK\Big(\frac{p \log n}{n} + \frac{u}{n}\Big)\Big].$$

In the following, we use c to denote a generic positive constant whose value may change in different places. We also denote the maximum of two numbers x_1, x_2 as $x_1 \vee x_2$.

Proof of Lemma 1:

The main tool we use here is Talagrand's inequality. See for example, (5.50) in Massart (2003). Define $S^p = \{\eta \in \Re^p : \|\eta\| = 1\}$, $W_n(\boldsymbol{D}, \theta) = \bar{g}(\boldsymbol{D}, \theta) - \bar{g}(\boldsymbol{D}, \theta^*) - (\mathrm{E}\, g(D, \theta) - \mathrm{E}\, g(D, \theta^*))$ where $\bar{g}(\boldsymbol{D}, \theta) = n^{-1} \sum_{i=1}^n g(D_i, \theta)$ and θ^* is the parameter that satisfies $\mathrm{E}\, g(D, \theta^*) = 0$ in G2. Also define the class of functions $\mathcal{G} = \big\{\eta \in S^p, \theta \in \Theta_n : f(D; \eta, \theta) = \eta^\top[(g(D, \theta) - g(D, \theta^*)) - (\mathrm{E}\, g(D, \theta) - \mathrm{E}\, g(D, \theta^*))]\big\}$. We note that under our assumptions G3 on the L_∞ norm and G5 on the L_1 bracketing entropy, the class $\mathcal{G}$ satisfies the assumption (M) before Theorem 8.3 of Massart (2003), i.e. the pointwise measurability condition (see for example, Section 2.3 of van der Vaart and Wellner 1996). This guarantees that we can generalize the original Talagrand's inequality in Massart (2003) (and also Lemma 4.23 and Lemma 6.5 we will use later) from a countable collection of functions to the uncountable class $\mathcal{G}$, because the supremum over $\mathcal{G}$ can be approximated by the supremum over its dense subset at an arbitrarily fast rate. As a result, we apply Talagrand's inequality to $\mathcal{G}$ and obtain that for any $u' \in \Re$,

$$P\Big[\sup_{\theta \in \Theta_n} \big\|W_n(\boldsymbol{D}, \theta)\big\| \geq 2\,\mathrm{E} \sup_{\theta \in \Theta_n} \big\|W_n(\boldsymbol{D}, \theta)\big\| + \sqrt{2V_f u'} + \frac{4}{3} B_f u'\Big] \leq e^{-u'}, \tag{A.1}$$

where V_f is defined and bounded by

$$\begin{aligned} V_f &\equiv \sup_{\eta \in S^p, \theta \in \Theta_n} \frac{1}{n} \mathrm{Var}\Big[f(D; \eta, \theta)\Big] \\ &= \sup_{\eta \in S^p, b \in B_{mj}} \frac{1}{n} \mathrm{E}\Big[\eta^\top\big[g(D_i, \theta) - \mathrm{E}\, g(D, \theta) - (g(D_i, \theta^*) - \mathrm{E}\, g(D, \theta^*))\big]\Big]^2 \\ &\leq \frac{2}{n} \sup_{\eta \in S^p, \theta \in \Theta_n} \mathrm{E}[(\eta^\top(g(D, \theta) - g(D, \theta^*)))^2] \leq \frac{2c}{n} \end{aligned}$$

given G2, and B_f is defined and bounded by

$$B_f \equiv \sup_{\eta \in S^p, \theta \in \Theta_n} \frac{1}{n} \Big|f(D; \eta, \theta)\Big|_\infty \leq \frac{4c_g \sqrt{p \log n}}{n}$$

given G3.

Next we bound $\mathrm{E}\sup_{\theta\in\Theta_n}\left\|W_n(\boldsymbol{D},\theta)\right\|$ by using Lemma 6.5 in Massart (2003). Note that for $f\in\mathcal{G}$, we will normalize it by dividing $F=2\sup_{D,\eta,\theta}|f(D;\eta,\theta)|_\infty$. Using the assumption G5, the right-hand side of the inequality in Lemma 6.5 is

$$\begin{aligned}12\varphi(\sigma) &\equiv 12\int_0^\sigma\sqrt{H_{[]}(u^2,\mathcal{G})}\mathrm{d}u\le 12\int_0^\sigma\sqrt{c_1p\log(c_2n/u^2)}\mathrm{d}u\\ &\le c\Big(\sqrt{p\log n}\sigma+\sqrt{p\log(\sigma^{-1})}\sigma\Big)\le c\sqrt{p\log n}\sigma\end{aligned}$$

for some $c>0$ and $1\ge\sigma\ge\sigma^*\equiv c\sqrt{p\log n/n}>0$. We choose σ^* such that $\varphi(\sigma)\le\sqrt{n}\sigma^2/4$ in Lemma 6.5 is satisfied. Note that $\sigma^*\to0$ as $n\to\infty$ since $p=o(n^{1/2}/\log n)$. Thus Lemma 6.5 implies that

$$\sqrt{n}\,\mathrm{E}\Big[\sup_{\eta\in S^p,\theta\in\Theta_n,\mathrm{E}\left[f(D;\eta,\theta)^2\right]\le F^2\sigma^2}\|F^{-1}W_n(\boldsymbol{D},\theta)\|\Big]\le c\sqrt{p\log n}\sigma. \tag{A.2}$$

Then based on (A.2), we can apply the pealing lemma (Lemma 4.23) of Massart (2003) and obtain that for any $1\ge\sigma\ge\sigma^*>0$,

$$\mathrm{E}\Big[\sup_{\eta\in S^p,\theta\in\Theta_n}\frac{\|F^{-1}W_n(\boldsymbol{D},\theta)\|}{\mathrm{E}\left[F^{-2}f(D;\eta,\theta)^2\right]+\sigma^2}\Big]\le c\sqrt{\frac{p\log n}{n}}\sigma^{-1},$$

which further implies that

$$\begin{aligned}&\mathrm{E}\sup_{\theta\in\Theta_n}\left\|W_n(\boldsymbol{D},\theta)\right\|\\ \le\ & F\cdot\mathrm{E}\Big[\sup_{\eta\in S^p,\theta\in\Theta_n}\frac{\|F^{-1}W_n(\boldsymbol{D},\theta)\|}{\mathrm{E}\left[F^{-2}f(D;\eta,\theta)^2\right]+\sigma^2}\Big]\cdot\Big[\sup_{\eta\in S^p,\theta\in\Theta_n}\mathrm{E}\left[F^{-2}f(D;\eta,\theta)^2\right]+\sigma^2\Big]\\ \le\ & c\sqrt{\frac{p\log n}{n}}\cdot\Big[(F\sigma)^{-1}\sup_{\eta\in S^p,\theta\in\Theta_n}\mathrm{E}\left[f(D;\eta,\theta)^2\right]+F\sigma\Big]\equiv c\sqrt{\frac{p\log n}{n}}\cdot h(\sigma)\end{aligned}$$

We can minimize $h(\sigma)$ by taking $\sigma=F^{-1}\sqrt{\sup_{\eta\in S^p,\theta\in\Theta_n}\mathrm{E}\left[f(D;\eta,\theta)^2\right]}\vee\sigma^*$. When σ^* is smaller, we have $h(\sigma)\le2\sqrt{\sup_{\eta\in S^p,\theta\in\Theta_n}\mathrm{E}\left[f(D;\eta,\theta)^2\right]}\le2c$ for some c since $\sup_{\eta\in S^p,\theta\in\Theta_n}\mathrm{E}[f(D;\eta,\theta)^2]$ is bounded according to the assumption G2. When σ^* is larger, we have $h(\sigma)\le2F\sigma^*\le16c_gp^{1/2}\cdot c\sqrt{p\log n/n}\le c$ because $F\le8\sqrt{p}|g|_\infty\le8c_g\sqrt{p\log n}$ according to G3 and $p=o(n^{1/2}/\log n)$. In either case, $h(\sigma)$ is bounded above by a constant. Therefore

$$\mathrm{E}\sup_{\theta\in\Theta_n}\left\|W_n(\boldsymbol{D},\theta)\right\|\le c\sqrt{\frac{p\log n}{n}}$$

for some constant c.

We now get back to the inequality (A.1) and plug in the bounds of $\mathrm{E}\sup\|W_n(\boldsymbol{D},\theta)\|$, V_f and B_f. We have that for some constant $c>0$, for any $u'\in\Re$ and each sufficiently large n,

$$P\Big[\sup_{\theta\in\Theta_n}\big\|W_n(\boldsymbol{D},\theta)\big\|\geq c\sqrt{\frac{p\log n}{n}}+\sqrt{\frac{c}{n}u'}+\frac{c\sqrt{p\log n}}{n}u'\Big]\leq e^{-u'}. \tag{A.3}$$

Applying similar technique to the class $\{\eta\in S^p: f(D;\eta)=\eta^\top[g(D,\theta^*)-\mathrm{E}\,g(D,\theta^*)]\}$ yields

$$P\Big[\big\|\bar{g}(\boldsymbol{D},\theta^*)-\mathrm{E}\,g(D,\theta^*)\big\|\geq c\sqrt{\frac{p\log n}{n}}+\sqrt{\frac{c}{n}u'}+\frac{c\sqrt{p\log n}}{n}u'\Big]\leq e^{-u'} \tag{A.4}$$

for some constant $c>0$,

We add (A.3) and (A.4) and obtain that there exists $c>0$, such that for any $u'\in\Re$ and each sufficiently large n, with probability at least $1-2e^{-u'}$,

$$\sup_{\theta\in\Theta_n}\big\|\bar{g}(\boldsymbol{D},\theta)-\mathrm{E}\,g(D,\theta)\big\|\leq c\Big(\sqrt{\frac{p\log n}{n}}+\sqrt{\frac{u'}{n}}\Big),$$

where the simplified order on the right-hand side is due to $p=o(n^{1/2}/\log n)$ and the inequality $\sqrt{x_1+x_2}\leq\sqrt{x_1}+\sqrt{x_2}$ for $x_1,x_2>0$. This can be further rewritten in the interval form ($u=u'-\log 2$)

$$\begin{aligned}&\|\bar{g}(\boldsymbol{D},\theta)\|^2\in\\&\Big[\Big(0\vee\|\mathrm{E}\,g(D,\theta)\|-c\Big(\sqrt{\frac{p\log n}{n}}+\sqrt{\frac{u}{n}}\Big)\Big)^2,\Big(\|\mathrm{E}\,g(D,\theta)\|+c\Big(\sqrt{\frac{p\log n}{n}}+\sqrt{\frac{u}{n}}\Big)\Big)^2\Big],\end{aligned}$$

with probability at least $1-e^{-u}$.

Enlarging the interval will not change our conclusion since it makes the probability even larger. Eventually, we have that for some constant $C>0$ (which is only related to the "c"s in our previous inequalities), for large constant $K>1$, for any $u\in\Re$, for all sufficiently large n, with probability at least $1-e^{-u}$,

$$R_n(\theta)\in\Big[(1-K^{-1})R(\theta)-CK\Big(\frac{p\log n}{n}+\frac{u}{n}\Big),(1+K^{-1})R(\theta)+CK\Big(\frac{p\log n}{n}+\frac{u}{n}\Big)\Big],$$

where we use the fact $2x_1x_2\leq K^{-1}x_1^2+Kx_2^2$. ■

Proof of Theorem 1:

For (i), we set $s_1 = 1 - K^{-1}$, $s_2 = 1 + K^{-1}$, $\lambda = n\psi$, $u = 2\log n$, $v = p\log n/n$. Since $\inf_{\theta\in\Theta} R(\theta) = 0$, we can fix $K = 2$ and $\Delta = 4Cp\log n/n$ on the right-hand side of (5), where C is from Lemma 1. In (6),

$$\begin{aligned}
&\pi(\theta : R(\theta) < \inf_{\theta'\in\Theta} R(\theta') + v) = \pi(\theta : R(\theta) < v) = \pi(\theta : \|\operatorname{E} g(D,\theta) - \operatorname{E} g(D,\theta^*)\| \le \sqrt{v}) \\
\ge\ & \pi\big(\theta : n^{\xi} p\|\theta - \theta^*\| \le \sqrt{v}\big) \ge (p^{-1/2} n^{-\xi-\zeta}\sqrt{\log n})^p \ge n^{-p(\xi+\zeta+1)},
\end{aligned}$$

where we use G1, G4 and $p = o(n^{1/2}/\log n)$. Hence

$$\tilde{R} = 3v + \frac{2(\xi+\zeta+1)p\log n}{n\psi} + 4\Delta \le \frac{cp\log n}{n},$$

where $c = 16C + 3 + 2(\xi+\zeta+1)/\psi$. This together with Lemma 1 implies that in (5), we have

$$PQ\Big[R(\tilde{\theta}) > \frac{cp\log n}{n} + \frac{4\log n}{n\psi}\Big] \le 2n^{-2}.$$

Take $C_1 = 16C+3+2(\xi+\zeta+3)/\psi$ and we have proved (i). Part (iii) immediately follows by Borel-Cantelli lemma.

For (ii), if we set $\Delta = 2C(p\log n + u)/n$ and keep the same values of s_1, s_2, λ, u, v as before, then from Lemma 1 and Proposition 1 we have

$$PQ\Big[R(\tilde{\theta}) > \frac{cp\log n}{n} + \frac{cu}{n}\Big] \le 2e^{-u},$$

with $c = (8C + 3 + 2(\xi+\zeta+1)/\psi) \vee (8C + 2/\psi) \vee 1$. We integrate with respect to $u \in [0, +\infty)$ and obtain

$$PQR(\tilde{\theta}) = \frac{cp\log n}{n} + \int_0^\infty PQ\Big[R(\tilde{\theta}) > \frac{cp\log n}{n} + t\Big]\mathrm{d}t = \frac{c(p\log n + 2)}{n} \le \frac{2cp\log n}{n}$$

when n is large. Set $C_2 = 2c$ and (ii) follows. ■

Proof of Proposition 2:

Since $F^{-1}_{Y|X}(\tau) = X^\top\theta^*$, it immediately follows that $\operatorname{E} g(D,\theta^*) = 0$. Let $\lambda_{\max}(A)$ be the largest eigenvalues of matrix A. Then

$$\begin{aligned}
&\sup_{\eta\in S^p} \operatorname{E}[(\eta^\top g(D,\theta^*))^2] = \sup_{\eta\in S^p} \operatorname{E}[(\eta^\top X)^2 (I(Y \le X^\top\theta^*) - \tau)^2] \\
\le\ & \sup_{\eta\in S^p} \eta^\top \operatorname{E}[XX^\top]\eta = \lambda_{\max}(\operatorname{E}[XX^\top])
\end{aligned}$$

and also

$$\sup_{\eta\in S^p, \theta\in\Theta_n} \operatorname{E}[(\eta^\top(g(D,\theta) - g(D,\theta^*)))^2]$$

$$
\begin{aligned}
&= \sup_{\eta\in S^p,\theta\in\Theta_n} \mathrm{E}[(\eta^\top X)^2\cdot|1(Y\le X^\top\theta)-1(Y\le X^\top\theta^*)|]\\
&\le \sup_{\eta\in S^p}\eta^\top \mathrm{E}(XX^\top)\eta=\lambda_{\max}(\mathrm{E}[XX^\top]).
\end{aligned}
$$

Since $\lambda_{\max}(\mathrm{E}[XX^\top])$ is bounded above by constant, G2 is satisfied. $|g(D,\theta)|_\infty = |X[I(Y\le X^\top\theta)-\tau]|_\infty \le M$ given that $|X_j|\le M$ for all $j=1,2,...,p$, so G3 holds.

The expected moment is $\mathrm{E}\, g(D,\theta)=\mathrm{E}_X\left\{X[F_{Y|X}(X^\top\theta)-\tau]\right\}$. The partial derivative is bounded as

$$
\left|\frac{\mathrm{E}\, g_j(D,\theta)}{\partial\theta_k}\right| \le \left|\,\mathrm{E}[X_kX_jf_{Y|X}(X^\top\theta)]\right| \le M^2\bar f
$$

for any j and k. Thus G4 follows.

For G5, functions in $\mathcal{F}$ are the product of functions in $\mathcal{F}_1=\{\eta\in S^p:\eta^\top X\}$ and $\mathcal{F}_2=\{\theta\in\Theta_n: I[Y\le X^\top\theta]-\tau\}$. For any $\eta,\eta'\in S^p$, $|\eta^\top X-\eta'^\top X|\le M|\eta-\eta'|_1\le Mp|\eta-\eta'|_\infty$. Therefore the L_1 bracketing number of $\mathcal{F}_1$ is bounded by the L_∞ covering number as $N_{[]}(\varepsilon,\mathcal{F}_1)\le N_\infty(\varepsilon/2,\mathcal{F}_1)=(8Mp/\varepsilon)^p$ given that $\|\eta\|=1$. For $\mathcal{F}_2$, we construct the L_1 bracket as follows. Suppose the first component $X_1=1$ is the constant term and rewrite $I[Y\le X^\top\theta]-\tau = I[Y\le X_{-1}^\top\theta_{-1}+\theta_1]-\tau$, where the subscript -1 denotes the remaining vector without the first component. We first pick a $\varepsilon/(8\bar f)$ covering net in L_1 norm, say $\mathcal{N}_1$, for $\{\|\theta_{-1}\|\le\sqrt{p}\log n: X_{-1}^\top\theta_{-1}\}$. For any θ_{-1},θ'_{-1} in this set, it suffices to require $|X_{-1}^\top\theta_{-1}-X_{-1}^\top\theta'_{-1}|\le M(p-1)|\theta_{-1}-\theta'_{-1}|_\infty\le\varepsilon/(8\bar f)$, which implies that the cardinality of $\mathcal{N}_1$ is bounded by $[2\sqrt{p}\log n/(\varepsilon/(8M\bar f(p-1)))+1]^{p-1}\le(32M\bar fp^{3/2}\log n/\varepsilon)^{p-1}$. Then we pick a $\varepsilon/(8\bar f)$ net on the interval $|\theta_1|\le\sqrt{p}\log n$, say $\mathcal{N}_2$, which has at most $2\sqrt{p}\log n/(\varepsilon/(8\bar f))+1\le 32\bar fp^{1/2}\log n/\varepsilon$ points. Now consider the net of θ generated by the Cartesian product $\mathcal{N}_1\times\mathcal{N}_2$. For any $\theta=(\theta_1,\theta_{-1})\in\Theta_n$, we can pick some θ'_{-1} from $\mathcal{N}_1$ such that $|X_{-1}^\top(\theta_{-1}-\theta'_{-1})|\le\varepsilon/(8\bar f)$, and some θ'_1 from $\mathcal{N}_2$ such that $\theta'_1-\theta_1\in[\varepsilon/(4\bar f),3\varepsilon/(8\bar f)]$. Therefore,

$$
\begin{aligned}
& I[Y\le X^\top\theta']-I[Y\le X^\top\theta]\\
=\;& I[Y-X_{-1}^\top\theta_{-1}\le\theta_1+X_{-1}^\top(\theta'_{-1}-\theta_{-1})+(\theta'_1-\theta_1)]-I[Y-X_{-1}^\top\theta_{-1}\le\theta_1]\\
\ge\;& I\Big[Y-X_{-1}^\top\theta_{-1}\le\theta_1-\frac{\varepsilon}{8\bar f}+\frac{1\varepsilon}{4\bar f}\Big]-I[Y-X_{-1}^\top\theta_{-1}\le\theta_1]\\
=\;& I\Big[\theta_1<Y-X_{-1}^\top\theta_{-1}\le\theta_1+\frac{\varepsilon}{8\bar f}\Big]\ge 0,
\end{aligned}
$$

which implies that $I[Y\le X^\top\theta']-\tau$ is an upper bound for $I[Y\le X^\top\theta]-\tau$. Similarly we can find a lower bound indexed by θ'' in $\mathcal{N}_1\times\mathcal{N}_2$ by choosing $|X_{-1}^\top(\theta_{-1}-\theta''_{-1})|\le\varepsilon/(8\bar f)$

and $\theta_1 - \theta_1'' \in [\varepsilon/(4\bar{f}), 3\varepsilon/(8\bar{f})]$. Moreover, since

$$|X^\top\theta' - X^\top\theta''| \le |X_{-1}^\top(\theta'_{-1} - \theta''_{-1})| + |\theta'_1 - \theta''_1| \le 2\Big(\frac{\varepsilon}{8\bar{f}} + \frac{3\varepsilon}{8\bar{f}}\Big) = \frac{\varepsilon}{\bar{f}},$$

the L_1 distance between the upper and lower bounds is controlled by

$$\mathrm{E}\left|I[Y \le X^\top\theta'] - I[Y \le X^\top\theta'']\right| \le \mathrm{E}\, I[Y \text{ between } X^\top\theta' \text{ and } X^\top\theta''] \le \bar{f} \cdot \frac{\varepsilon}{\bar{f}} = \varepsilon.$$

Hence $\mathcal{N}_1 \times \mathcal{N}_2$ is a L_1 bracket of $\mathcal{F}_2$ and the L_1 bracketing number of $\mathcal{F}_2$ is bounded above by the number of pairs (θ', θ'') with $\theta', \theta'' \in \mathcal{N}_1 \times \mathcal{N}_2$, which is at most $N_{[]}(\varepsilon, \mathcal{F}_2) \le \big[(32M\bar{f}p^{3/2}\log n/\varepsilon)^{p-1} \cdot 32\bar{f}p^{1/2}\log n/\varepsilon\big]^2 \le (32M\bar{f}p^{3/2}\log n/\varepsilon)^{2p}$. Therefore, the L_1 bracketing number of $\mathcal{F}$ is bounded above by $N_{[]}(\varepsilon, \mathcal{F}) \le N_{[]}(\varepsilon/2, \mathcal{F}_1) \cdot N_{[]}(\varepsilon/2, \mathcal{F}_2) \le (1024M^2\bar{f}p^{5/2}\log n/\varepsilon)^{2p}$, and its L_1 bracketing entropy is bounded above by $H_{[]}(\varepsilon, \mathcal{F}) \le \log N_{[]}(\varepsilon, \mathcal{F}) \le 2p\log(1024M^2\bar{f}n^{5/4}/\varepsilon) \le c_1 p\log(c_2 n/\varepsilon)$ for some constants c_1, c_2 given that $p = o(n^{1/2}/\log n)$. ■

Before we prove Theorem 2 for the ranking risk, we first introduce the Gibbs posterior with translated risk and the Hoeffding's decomposition.

We set $\lambda = n\psi$ and the Gibbs posterior (8) can be written equivalently as

$$Q(\mathrm{d}b, m|\boldsymbol{D}) = \frac{\exp\{-n\psi R'_n(b)\}\pi(\mathrm{d}b|m)\pi_m}{\sum_{m'}\int_{B_{m'}}\exp\{-n\psi R'_n(b)\}\pi(\mathrm{d}b|m')\pi_{m'}},$$

where $R'_n(b) = \frac{1}{n(n-1)}\sum_{i\ne j}\big\{I[(Y_i - Y_j)r(X_i, X_j; b) < 0] - I[(Y_i - Y_j)r^*(X_i, X_j) < 0]\big\}$, the translated version of $R_n(b)$ by subtracting the risk of Bayes rule. Accordingly, let $R'(b) = R(b) - R^*$ where $R^* = P[(Y - Y')r^*(X, X') < 0]$.

The following Hoeffding's decomposition of U-statistics will be the key tool for the ranking risk. See for example, Appendix A of Clémencon et al. (2008).

$$\begin{aligned}
q(x, y, x', y'; b) &= I[(y - y)r(x, x'; b) < 0)] - I[(y - y')r^*(x, x') < 0] \\
R'_n(b) &= \frac{1}{n(n-1)}\sum_{i\ne j} q(X_i, Y_i, X_j, Y_j; b) \\
R'(b) &= \mathrm{E}\, q(X, Y, X', Y'; b) \\
h(x, y; b) &= \mathrm{E}_{X',Y'} q(x, y, X', Y'; b) - R(b) \\
\hat{h}(x, y, x', y'; b) &= q(x, y, x', y'; b) - R(b) - h(x, y; b) - h(x', y'; b) \\
T_n(b) &= \frac{1}{n}\sum_{i=1}^n h(X_i, Y_i; b)
\end{aligned}$$

$$W_n(b) \;=\; \frac{1}{n(n-1)}\sum_{i\neq j}\hat{h}(X_i,Y_i,X_j,Y_j;b).$$

According to these definitions, it is obvious that $\mathrm{E}\,h(X,Y;b)=0$, $\mathrm{E}_{X',Y'}\,\hat{h}(x,y,X',Y';b)=0$ for all x and y, and most importantly

$$R'_n(b)-R'(b)=2T_n(b)+W_n(b), \tag{A.5}$$

which means that the deviation of risks $R'_n(b)-R'(b)$ can be decomposed into an iid sample average part of $T_n(b)$ and a second-order degenerate U-statistic $W_n(b)$.

We first state two lemmas that will be useful for the proof of Theorem 2.

Lemma 2. *Suppose assumption R4 holds. Then for $n>1$ and any $b\in\Theta_n$,*

$$\mathrm{Var}[h(X,Y;b)]\le 2\bar{f}_\eta\log n\cdot R'(b)^{1-(\log n)^{-1}}$$

Proof of Lemma 2:

In Corollary 7 and Corollary 8 of Clémencon et al. (2008), take $\epsilon=(\log n)^{-1}$. ■

Lemma 3. *Suppose R3-R4 in Theorem 2 hold. For any $u>0$, uniformly for all $b\in B_{mj}$, all $m=1,2,...,p$ and all $j=1,2,...,\binom{p}{m}$, for some constant $K>1$, for all sufficiently large n, there exists a constant $C>0$ such that with probability at least $1-e^{-u}$,*

$$R'_n(b)\in\Big[(1-K^{-1})R'(b)-CK\Big(\frac{m(\log n)^2}{n}+\frac{u}{n}\Big),(1+K^{-1})R'(b)+CK\Big(\frac{m(\log n)^2}{n}+\frac{u}{n}\Big)\Big].$$

Proof of Lemma 3:

Theorem 5 and Corollary 6 of Clémencon et al. (2008) have proved that for some constant $c>0$ and any $u'\in\Re$,

$$P\Big[\sup_{b\in B_{mj}}|W_n(b)|>c\frac{m+u'}{n}\Big]\le e^{-u'}, \tag{A.6}$$

since the VC-dimension of the class of functions $\mathcal{F}_{mj}=\{b\in B_{mj}: r(x,x')=(x-x')^\top b\}$ is at most $m+1$.

Next we bound the term $T_n(b)$ using similar technique to the proof of Theorem 8.3 in Massart (2003). The assumption (M) in Massart (2003) is satisfied by $\{b\in B_{mj}: h(X,Y;b)\}$ since $|h(X,Y;b)|\le 2$ and is continuous in b given the assumption R3. Let $b_{mj}=\arg\min_{b\in B_{mj}}R'(b)$, and define the pseudo distance d to be

$$d^2(b,b')$$

$$\equiv \mathrm{E}[h(X,Y;b) - h(X,Y;b')]^2$$
$$= \mathrm{E}_{X,Y}\big\{\mathrm{E}_{X',Y'|X,Y}(I[(Y-Y')(X-X')^\top b < 0] - I[(Y-Y')(X-X')^\top b' < 0])\big\}^2.$$

Then according to Lemma 2, the condition (8.17) of Massart (2003) in our case becomes

$$\begin{aligned} d^2(b, b_{mj}) &\le 2(\mathrm{Var}[h(X,Y;b)] + \mathrm{Var}[h(X,Y;b_{mj})]) \\ &\le 4\bar{f}_\eta \log n \cdot [R'(b)^{1-(\log n)^{-1}} + R'(b_{mj})^{1-(\log n)^{-1}}] \\ &\le 8\bar{f}_\eta \log n \cdot R'(b)^{1-(\log n)^{-1}} \equiv w^2(\sqrt{R'(b)}), \end{aligned} \tag{A.7}$$

where we define the function $w(x) = \sqrt{8\bar{f}_\eta \log n} \cdot x^{1-(\log n)^{-1}}$.

On the other hand, within the class $\mathcal{F}_{mj}$, we apply Lemma 6.5 of Massart (2003) ($|h(x,y;b)| \le 2$), and obtain

$$\sqrt{n}\,\mathrm{E} \sup_{b \in B_{mj}, d^2(b,b_{mj}) \le \sigma^2} \Big| T_n(b) - T_n(b_{mj}) \Big| \le 48\varphi(\sigma), \tag{A.8}$$

where $0 < \sigma \le 1$ and $\varphi(\sigma)$ is defined as

$$\varphi(\sigma) \equiv \int_0^\sigma \sqrt{H_{[]}(u^2, \mathcal{F}_{mj})}\mathrm{d}u \le \int_0^\sigma \sqrt{H_\infty(u^2/2, \mathcal{F}_{mj})}\mathrm{d}u.$$

Here we used the relation between the L_1 bracketing entropy and the L_∞ entropy. To further bound $H_\infty(u^2/2, \mathcal{F}_{mj})$, we only need that for any x, y and $b, b' \in \mathcal{F}_{mj}$, $|h(x,y;b) - h(x,y;b')| \le n^\xi m|b-b'|_\infty \le u^2/2$, where the derivative is bounded by n^ξ according to the assumption R3. Note that $|b-b'|_\infty \le 2p^{1/2}\log n$ due to the restriction $\|b\| \le p^{1/2}\log n$. The L_∞ covering number of $\mathcal{F}_{mj}$ is at most $N_\infty(u^2/2, \mathcal{F}_{mj}) \le [4mp^{1/2}n^\xi/u^2 + 1]^m < [5n^{\xi+2}/u^2]^m$, and $\varphi(\sigma)$ is bounded by

$$\begin{aligned} \varphi(\sigma) &\le \int_0^\sigma \sqrt{m \log\Big(\frac{5n^{\xi+2}}{u^2}\Big)}\mathrm{d}u \\ &\le 2\sqrt{(\xi+3)m\log n}\sigma + \sqrt{2m\log(\sigma^{-1})}\sigma \\ &\le 3\sqrt{(\xi+3)m\log n}\sigma \end{aligned}$$

if $\sigma > n^{-1}$ and n is large. Since $m \le p = o(n/(\log n)^3)$, as long as $1 \ge \sigma \ge c\sqrt{m/n}\log n \to 0$, the condition $\varphi(\sigma) \le \sqrt{n}\sigma^2/4$ is satisfied by all sufficiently large n.

Now based on (A.7) and (A.8), we can apply the inequality (8.26) in the proof of Theorem 8.3 in Massart (2003) (derived from a two-sided Talagrand's inequality and the pealing lemma), which gives

$$P\left[\sup_{b \in B_{mj}} \frac{\big|T_n(b) - T_n(b_{mj})\big|}{R'(b) + a^2} \ge \frac{1}{4K}\right] < e^{-u'}, \tag{A.9}$$

where $u' \in \Re$, $K > 1$ is a large positive constant that does not depend on n, and

$$a = cK\sqrt{\epsilon^{*2} + cu'/n} \tag{A.10}$$

with some constant $c > 1$. ϵ^* here is the solution to equation $\sqrt{n}\epsilon^{*2} = 48\varphi(w(\epsilon^*))$. This together with $m = o(n/(\log n)^2)$ implies that for some constant $c > 0$,

$$\epsilon^{*2} = \left[\frac{cm(\log n)^2}{n}\right]^{1-(\log n+1)^{-1}} \approx \frac{\mathrm{e} \cdot cm(\log n)^2}{n} \tag{A.11}$$

as n becomes large, where $\mathrm{e} = 2.71828....$

Now we combine (A.9), (A.10) and (A.11), replace u' with $u' + m\log p + m\log 2$ and obtain that for some $c > 0$, some constant $K > 1$, uniformly over all B_{mj} with $1 \le m \le p$ and $1 \le j \le \binom{p}{m}$,

$$\begin{aligned} & P\left[\forall b \in B_{mj}, \left|T_n(b) - T_n(b_{mj})\right| > (4K)^{-1}R'(b) + cK\frac{m(\log n)^2 + m\log p + m\log 2 + u'}{n}\right] \\ \le & \sum_{m=1}^{p}\sum_{j=1}^{\binom{p}{m}} e^{-m\log p - m\log 2 - u'} \le e^{-u'} \end{aligned}$$

where we use the fact $\binom{p}{m} \le p^m$. This is equivalent to

$$P\left[\forall b \in B_{mj}, \left|T_n(b) - T_n(b_{mj})\right| > (4K)^{-1}R'(b) + cK\frac{m(\log n)^2 + u'}{n}\right] \le e^{-u'}. \tag{A.12}$$

We can then apply a similar technique to the class of functions $\{1 \le m \le p, 1 \le j \le \binom{p}{m} : r(x, x'; b_{mj}) = (x - x')^\top b_{mj}\} \cup \{r^*(x, x')\}$, and obtain that for any $u' \in \Re$, some $c > 0$, some large constant $K > 0$, uniformly over all m, j,

$$P\left[\left|T_n(b_{mj})\right| > (4K)^{-1}R'(b_{mj}) + cK\frac{m\log 2 + u'}{n}\right] \le e^{-u'}.$$

Since for any $b \in B_{mj}$, $R'(b) \ge R'(b_{mj})$, it follows that

$$P\left[\forall b \in B_{mj}, \left|T_n(b_{mj})\right| > (4K)^{-1}R'(b) + cK\frac{m\log 2 + u'}{n}\right] \le e^{-u'}. \tag{A.13}$$

We finally add inequalities (A.6), (A.12), (A.13) according to the decomposition (A.5) and have that uniformly for all B_{mj}, for any $u' \in \Re$, and for sufficiently large n, there exists some constant $c > 0$ and some large constant $K > 0$,

$$P\left[\forall b \in B_{mj}, \left|R'_n(b) - R'(b)\right| > K^{-1}R'(b) + cK\frac{m(\log n)^2 + u'}{n}\right] \le 3e^{-u'},$$

which implies the conclusion of Lemma 3 after we set $u = u' - \log 3$ and $C = 2c$. ■

Proof of Theorem 2:

We will apply the oracle inequality (10) in Proposition 3. For (i), we set $s_1 = 1 - K^{-1}$, $s_2 = 1 + K^{-1}$, $\lambda = n\psi$, $u = 2\log n$, $v = m/n$, $\Delta_m = m(\log n)^3/n$. Note that for sufficiently large n, $\Delta_m > 2CKm(\log n)^2/n$ with C, K defined in Lemma 3. We define $\tilde{R}'(v), \tilde{R}'_{mj}(v)$ to be the translated version of $\tilde{R}(v), \tilde{R}_{mj}(v)$ and rewrite $\tilde{R}'_m(v)$ in Proposition 3 as $\tilde{R}'_{mj}(v)$, since we consider the subspace B_{mj}. In the definition of $\tilde{R}'_{mj}(v)$, using the assumptions R2 and R3 and the choice of v, we have

$$\begin{aligned} & \pi(b : R'(b) - \inf_{b' \in B_{mj}} R'(b') < v | m, j) = \pi\Big(b : R'(b) - R'(b_{mj}) < \frac{m}{n}\Big| m, j\Big) \\ \geq \; & \pi\Big(b : n^{\xi} m \|b - b_{mj}\| < \frac{m}{n}\Big| m, j\Big) \geq n^{-m(\xi+\zeta+1)}, \end{aligned}$$

which implies

$$\tilde{R}'_{mj}(v) \leq \inf_{b \in B_{mj}} R'(b) + \frac{m}{n} + \frac{(\xi + \zeta + 1)m \log n}{n s_2 \psi} \leq \inf_{b \in B_{mj}} R'(b) + \frac{cm \log n}{n}$$

with $c = 1 + (\xi + \zeta + 1)/\psi$.

Now we bound the right-hand side of (9). Δ_m is the same for all $j = 1, 2, ..., \binom{p}{m}$. According to R1, the prior mass on B_m is $\pi_m = e^{-2\psi m(\log n)^3} / \sum_{m'=1}^{p} e^{-2\psi m'(\log n)^3} = e^{-2\psi(m-1)(\log n)^3}(1 - e^{-2\psi(\log n)^3})/(1 - e^{-2\psi p(\log n)^3})$. Hence $\lambda^{-1} \log \pi_{mj}^{-1} = (n\psi)^{-1}\big[\log \binom{p}{m} + 2\psi(m-1)(\log n)^3 + \log(1 - e^{-2\psi p(\log n)^3}) - \log(1 - e^{-2\psi(\log n)^3})\big] \leq 3m(\log n)^3/n$ for sufficiently large n. The term $\tilde{\Delta}$ can be bounded by

$$\begin{aligned} \tilde{\Delta} &= (n\psi)^{-1} \log\Big(\sum_{m=1}^{p} \pi_m e^{\psi m(\log n)^3}\Big) = (n\psi)^{-1} \log\left[\frac{\sum_{m=1}^{p} e^{-\psi m(\log n)^3}}{\sum_{m=1}^{p} e^{-2\psi m(\log n)^3}}\right] \\ &= (n\psi)^{-1} \log\left[\frac{e^{-\psi(\log n)^3}(1 - e^{-\psi p(\log n)^3})(1 - e^{-2\psi(\log n)^3})}{e^{-2\psi(\log n)^3}(1 - e^{-2\psi p(\log n)^3})(1 - e^{-\psi(\log n)^3})}\right] \leq \frac{2(\log n)^3}{n}. \end{aligned}$$

when n is sufficiently large. Therefore, (9) becomes

$$\begin{aligned} \tilde{R}'(v) &\leq \inf_{m,j}\Big\{\frac{1 + K^{-1}}{1 - K^{-1}} \inf_{b \in B_{mj}} R'(b) + \frac{4m(\log n)^3}{(1 - K^{-1})n}\Big\} + \frac{2(\log n)^3}{(1 - K^{-1})n} \\ &\leq \inf_{m,j}\Big\{\frac{1 + K^{-1}}{1 - K^{-1}} \inf_{b \in B_{mj}} R'(b) + \frac{5m(\log n)^3}{n}\Big\}, \end{aligned}$$

when n is sufficiently large and $K > 9$. If we replace $u = 2\log n$ in the oracle inequality (5), we get for fixed large K and each sufficiently large n,

$$PQ\Big[R'(\tilde{\theta}) > \frac{1 + K^{-1}}{1 - K^{-1}} \inf_{m,j}\Big(\inf_{b \in B_{mj}} R'(b) + \frac{6m(\log n)^3}{n}\Big)\Big] \leq 2n^{-2}.$$

Set $\delta = 2/(K-1)$ and (i) is proved since $R'(\tilde{\theta}) = R(\tilde{\theta}) - R^*$. (iii) follows by Borel-Cantelli Lemma.

For (ii), we notice that $|R'(b)| \leq 2$ for all b. Let the event $A = \{R'(\tilde{\theta}) \leq (1+\delta)\inf_{m,j}[\inf_{b\in B_{mj}} R'(b) + 6m(\log n)^3/n]\}$. Therefore when n is sufficiently large, using the result of (i) we have

$$\begin{aligned} PQR'(\tilde{\theta}) &= PQ[R'(\tilde{\theta})I(A)] + PQ[R'(\tilde{\theta})I(A^c)] \\ &\leq (1+\delta)\inf_{m,j}\Big(\inf_{b\in B_{mj}} R'(b) + \frac{6m(\log n)^3}{n}\Big) + 4n^{-2} \\ &\leq (1+\delta)\inf_{m,j}\Big(\inf_{b\in B_{mj}} R'(b) + \frac{7m(\log n)^3}{n}\Big), \end{aligned}$$

which concludes the proof. ■