JMLR: Workshop and Conference Proceedings vol 30 (2013) 1–10

General Oracle Inequalities for Gibbs Posterior with Application to Ranking

Cheng Li CHENGLI2014@U.NORTHWESTERN.EDU
Wenxin Jiang WJIANG@NORTHWESTERN.EDU
Martin A. Tanner MAT132@NORTHWESTERN.EDU
Department of Statistics, Northwestern University, 2006 Sheridan Road, Evanston 60208, Illinois, U.S.A.

Abstract

In this paper, we summarize some recent results in Li et al. (2012), which can be used to extend an important PAC-Bayesian approach, namely the Gibbs posterior, to study the nonadditive ranking risk. The methodology is based on assumption-free risk bounds and nonasymptotic oracle inequalities, which leads to nearly optimal convergence rates and optimal model selection to balance the approximation errors and the stochastic errors.

Keywords: Gibbs posterior, model selection, oracle inequalities, ranking, risk minimization.

1. Introduction

As summarized in an authoritative text Catoni (2007), the PAC-Bayesian approach has found great success in supervised classification. There has been a lot of interest over the past few years in using the Gibbs posterior, and more generally the exponential weights, to derive finite sample oracle inequalities (e.g. Alquier and Lounici 2011, Lecué 2007, Rigollet and Tsybakov 2011). Although these studies have mostly focused on iid (independent and identically distributed) data, recent works have extended to the study of weakly dependent time series in, for example, Alquier and Wintenberger (2012).

The current paper summarizes some recent results we obtained (Li et al. 2012) on the risk performance of the Gibbs posterior. Our results are different in nature from those of Catoni (2007), in the sense that the relation we revealed does not depend on the specific form of risk functions, nor on the data generation process. The relation is the same whether we have additive empirical risk or not, whether we have independent or dependent data. Therefore it is more general than the ones appeared in existing literature. It is both general and simple, and therefore fundamental. The flip side of being very general is that our approach may not always lead to optimal oracle inequalities. As we will discuss later, in some situations, the Gibbs posterior could produce less sharp results compared to the state-of-art PAC-Bayesian methods.

Below, we will first introduce the framework of the Gibbs posterior, and summarize the general relations derived in Li et al. (2012). Then we will apply these relations to the ranking problem, which involves a nonadditive empirical risk that has not been addressed

in the PAC-Bayesian literature before. Li et al. (2012) also considered another application to the generalized method of moments (GMM), which is a popular research direction in econometrics and is not included in this paper.

2. The Gibbs posterior

The Gibbs posterior is a randomization method of empirical risk minimization obtained from an analogy to statistical physics, where the empirical risk is identified with the 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 (e.g. Zhang 1999, 2006), econometricians (e.g. Chernozhukov and Hong 2003), and statisticians (e.g. Jiang and Tanner 2008). Due to its Bayesian flavor, the Gibbs posterior allows application of convenient computational algorithms such as Markov chain Monto Carlo (e.g. Chernozhukov and Hong 2003, Belloni and Chernozhukov 2009, Chen et al. 2010). Given the 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.

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, 2010).

The goal of the current paper is to derive oracle inequalities for model selection using the Gibbs posterior, so that nearly optimal risk performance will be achieved across a range of models under consideration. 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

$$Q(\mathrm{d}b, m) = \frac{e^{-\lambda R_n(b,m)}\pi(\mathrm{d}b|m)\pi_m}{\sum_{m'}\int_{B_{m'}} e^{-\lambda R_n(b,m')}\pi(\mathrm{d}b|m')\pi_{m'}} \tag{2}$$

In general, we are interested in some theoretical risk $R(\theta)$, for which $R_n(\theta)$ is the corresponding empirical risk. With slight abuse of notation, we sometimes also write $R(b)$ and $R_n(b)$ since b is the primary parameter to be estimated in the risk functions.

3. General oracle inequalities for risk convergence and model selection

In this section, we will make no assumptions and describe an inequality for some theoretical risk of interest $R(\theta)$, which will be related to the empirical risk R_n and the prior π 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 (nonstochastic) theoretic risk R exceeds a, i.e. $PQ(R(\theta) > a)$, where P corresponds to the underlying true distribution of data $\boldsymbol{D}$, and Q corresponds to the Gibbs posterior conditional on data $\boldsymbol{D}$. The probability distribution PQ, which is different from the prior π used in the Gibbs posterior, 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 θ from the resulting Gibbs posterior Q conditional on the data $\boldsymbol{D}$ sampled from step (i).

To bound the probability $PQ(R(\theta) > a)$, we construct a *simultaneous coverage interval* for the empirical risk R_n appearing in the Gibbs posterior, using the theoretic risk R. Let $0 < s_1 \leq s_2$ and $\Delta(\theta) \geq 0$ 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 $[s_1 R - \Delta, s_2 R + \Delta]$ for R_n. [1] Note that Δ is related to the radius of the coverage interval and is analogous to the standard deviation of R_n which characterizes its stochastic error. In applications, the quantity Δ usually decreases with n and increases with the model complexity. Although it is ideal to have $s_1 = s_2 = 1$ for strict oracle inequalities, 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 define the quantity

$$\bar{R} \equiv -(s_1\lambda)^{-1} \log \left[\frac{\int e^{-\lambda(s_2 R + \Delta)} \pi(\mathrm{d}\theta)}{\int e^{\lambda\Delta} \pi(\mathrm{d}\theta)} \right]. \tag{3}$$

for general θ. For model selection framework with $\theta = (b, m)$, we let the radius of the coverage interval Δ depend on m but not on b, and we write Δ_m instead for this dependence. We define for any model m and any $v > 0$,

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

and also

$$\tilde{R}(v) \equiv \inf_m (s_2 \tilde{R}_m(v) + \Delta_m + \lambda^{-1} \log \pi_m^{-1})/s_1 + \tilde{\Delta}/s_1, \tag{4}$$

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

where

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

Then we have the following proposition for the excess probability $PQ(R(\theta) > a)$ and oracle inequality for $R(\theta)$. The proofs are given in Li et al. (2012).

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

$$PQ\left[R > \bar{R} + \frac{u}{s_1 \lambda}\right] \leq e^{-u} + P\Big\{\exists \theta : R_n \notin [s_1 R - \Delta, s_2 R + \Delta]\Big\}, \tag{5}$$

where $\bar{R}$ is defined in (3).

(ii) (Oracle Inequality) In model selection case, for any $u \in \Re$ and $v > 0$,

$$PQ\left[R > \tilde{R}(v) + \frac{u}{s_1 \lambda}\right] \leq e^{-u} + P\Big\{\exists (b, m) : R_n \notin [s_1 R - \Delta_m, s_2 R + \Delta_m]\Big\} \tag{6}$$

where $\tilde{R}(v)$ is defined in (4).

Here $\tilde{R}(v)$ is chosen to be slightly larger than $\bar{R}$, so inequality (5) is tighter than (6) but the latter will be more useful when applied to model selection problems. If we use Proposition 1 to bound the probability of a large excess risk $R - \inf_\theta R$, then these inequalities reveal the fundamental relation that the performance of excess risk $R - \inf_\theta R$ is mainly determined by two factors: the excess of a *nonstochastic* bound $\bar{R} - \inf_\theta R$ or $\tilde{R}(v) - \inf_\theta R$, as well as a *stochastic* difference between R_n and R, as reflected in $P\{\exists \theta : R_n \notin [s_1 R - \Delta, s_2 R + \Delta]\}$.

Now we will discuss our choices of the tuning parameters s_1, s_2, λ, u, Δ, v and π in the proposition.

- Strict oracle inequalities are supposed to have $s_1 = s_2 = 1$. But sometimes 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 Δ_m such that $P\{\exists (b, m) : R_n \notin [s_1 R - \Delta_m, s_2 R + \Delta_m]\}$ can be controlled by $e^{-u} = n^{-2}$. To achieve this, we make use of the concentration tools in our proofs (e.g. Massart 2003). In later examples, typically $\Delta_m = m/n^c$ up to some logarithm factors, where $c = 1/2$ if $s_1 = s_1 = 1$ and $c = 1$ if $s_1 = 1 - \delta$ and $s_2 = 1 + \delta$.
- We will choose $v = \Delta_m$ (or about the same order up to some logarithm factors).
- We will choose $\pi_m \propto e^{-2\lambda \Delta_m}$. This induces a BIC-type penalty on the model complexity, which reflects our preference for more parsimonious models.

Therefore, given the choices of all tuning parameters, by setting $u = 2\log n$, $\lambda = n\psi$, the right-hand side of both inequalities can be bounded by $2n^{-2}$, implying that

$$R \leq \tilde{R}(\Delta) + \frac{2\log n}{ns_1\psi}, \tag{7}$$

for all large n, almost surely in the measure PQ, by Borel-Cantelli lemma. 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 that the *joint* event "$A_n = \{R \leq \tilde{R}(\Delta) + 2\log n/(ns_1\psi)\}$ 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 the Borel-Cantelli lemma. Without confusion, we will still write "A_n happens for all large n almost surely with respect to the measure PQ". [2]

To derive an oracle inequality, we still need to simplify the nonstochastic bound $\tilde{R}(v)$ in (3). Given the choice of π and Δ_m, it can be shown that $\tilde{\Delta}$ will approximately have the same order as Δ_1. Furthermore, since only linear ranking rules will be considered in our application, we can prove that

$$\inf_{v>0}[(v + (s_2\lambda)^{-1}\log\pi(R - \inf_{b\in B_m} R < v|m)^{-1}] = O(\Delta_m),$$

as long as the prior $\pi(\cdot|m)$ on each model space B_m does not vanish exponentially fast in sample size n for any small neighborhood contained in the subspace B_m. Therefore, in the expression (4), we have

$$\tilde{R}(v) = (s_2/s_1)\inf_m\Big\{\inf_{b\in B_m} R + O(\Delta_m) + O(\Delta_m)\Big\} + O(\Delta_1) = (s_2/s_1)\inf_m\Big\{\inf_{b\in B_m} R + O(\Delta_m)\Big\},$$

since Δ_m increases with m and $O(\Delta_1)$ can be absorbed into $O(\Delta_m)$. Combining this with (7) leads to

$$R \leq \frac{s_2}{s_1}\inf_m\Big\{\inf_{b\in B_m} R + O(\Delta_m)\Big\} \tag{8}$$

for all large n, almost surely in PQ, which is the oracle inequality for model selection.

Based on this "almost surely" statement, in our ranking example, we can further obtain the oracle inequality of Gibbs posterior mean

$$PQR \leq \frac{s_2}{s_1}\inf_m\Big\{\inf_{b\in B_m} R + O(\Delta_m)\Big\}. \tag{9}$$

The oracle inequalities (8) and (9) imply that even if the underlying best candidate model B_m is unknown, the theoretical risk $R(b)$ with b sampled from the Gibbs posterior

2. Here we used the PQ measure instead of the more common P measure. This is a consequence related to the generality of the relation we obtained, which is not dependent on the risk function or the data generation process. For example, using the P measure together with the posterior mean, instead of PQ, often would require the loss function to be convex, while our result holds in PQ measure regardless of the form of the loss function.

will always have the best possible convergence rate among all candidate models, both in the posterior mean and in the sense of almost surely convergence. In (8) and (9), $\inf_{b\in B_m} R$ 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. The leading constant s_2/s_1 is either 1 or $\frac{1+\delta}{1-\delta}$ for small $\delta > 0$ such that the ratio is close to 1.

4. Oracle performance of ranking risk with model selection

In ranking estimation, the empirical risk function is defined as

$$R_n(b) = \frac{1}{n(n-1)} \sum_{i\neq j} I[(Y_i - Y_j) r(X_i, X_j; b) < 0], \tag{10}$$

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 $r(x, x'; b) > 0$ if x ranks higher than x' and $r(x, x'; b) \leq 0$ otherwise. Since R_n involves averaging over paired data, it is a *nonadditive empirical risk* and is analogous to the energy of pairwise interactions in statistical physics. In this paper, our goal is to minimize the theoretical risk of mismatch $R(b) = P[(Y - Y')r(X, X'; b) < 0]$. The consistency and fast convergence rate of general ranking estimator that minimizes (10) and its convex upper bounds have been studied in recent frequentist papers such as Clémencon et al. (2008) and Rejchel (2012).

We now apply Proposition 1 to the model selection problem of the ranking risk R_n, to select the best linear rule ranking rules of the form $r(x, x'; b) = (x - x')^\top b$ for $b \in \Re^p$ in which only part of the components in X are active. 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 $R(\theta)$ with $\theta = (b, m)$ sampled from the Gibbs posterior. The proofs are given in Li et al. (2012).

Theorem 1 *Suppose the following regularity conditions R1-R4 hold:*

(R1) For any $m = 0, 1, 2, ..., p$, $\pi_m \propto e^{-2\psi m(\log n)^3}$ with $\psi > 0$ a constant. 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 $B_{mj} \cap \Theta_n$ with $\Theta_n = \left\{b : \|b\| \leq \sqrt{p}\log n\right\}$, for $1 \leq m \leq p$ and $1 \leq j \leq \binom{p}{m}$. For any $b_0 \in B_{mj} \cap \Theta_n$, any small enough $\delta > 0$,

there exists a constant $\zeta > 0$ *such that* $\pi(\{b : \|b - b_0\| \le \delta\}|m, j) \ge (\delta n^{-\zeta})^m$ *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* $|\partial \mathrm{E}_{X',Y'} I[(y - Y')(x - X')^\top b < 0]/\partial b_k| \le n^\xi$ *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 $\delta > 0$, *almost surely in the measure* PQ *for all sufficiently large* n, *there exists a constant* $C_1 > 0$, *such that*

$$R - R^* \le (1 + \delta) \inf_{m,j} \Big[\inf_{b \in B_{mj}} (R - R^*) + \frac{C_1 m (\log n)^3}{n} \Big]$$

where (b, m) *is randomly drawn from the Gibbs posterior* (2);

(ii) for any $\delta > 0$, *for each sufficiently large* n, *there exists a constant* $C_2 > 0$, *such that*

$$PQR \le R^* + (1 + \delta) \inf_{m,j} \Big[\inf_{b \in B_{mj}} (R - R^*) + \frac{C_2 m (\log n)^3}{n} \Big].$$

This theorem has extended the "bipartite" ranking example (Example 5.1) in Clémencon et al. (2008) in two ways. First, we allow a framework of adaptive model selection without knowing the best model dimension m. Second, 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).

Our regularity conditions R1 and R2 on the prior are mild and general. The assumption R1 assigns a prior on models that decreases exponentially fast with the model size, which favors parsimonious models. The assumption R2 says that the prior on B_{mj} does not vanish too fast in n on any small neighborhood and has a uniform lower bound over all submodels, which is satisfied by many commonly used priors, such as a uniform prior or a normal prior truncated on the L_2 bounded set Θ_n. Note that since the radius of Θ_n is growing with n, such a prior is not restrictive.

R3 and R4 impose mild upper bounds on the partial derivatives of the conditional expectation $\mathrm{E}_{X',Y'} I[(y - Y')(x - X')^\top b < 0]$ and the density of random variable $\eta(X)$. 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 $F_2(x) = I(x \ge 0)$, this becomes a binary choice model of $Y = I[X^\top b^* + \epsilon > 0]$ and b^* can be estimated by minimizing (10). This maximum rank correlation 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 $R(\theta) - R^*$ with θ sampled from the Gibbs posterior.

5. Discussion

In this paper, we have introduced assumption-free oracle inequalities that are useful for studying the performance of Gibbs posterior as a random method of risk minimization and model selection. Our method can be generally applied to nonadditive risk functions, and can give adaptive and nearly optimal oracle rate.

The generality of our inequalities could potentially lead to less sharp results compared to PAC-Bayesian methods. Specifically, when considering model selection in classification with large margin for iid data, we will get a nonstrict oracle inequality with an additional $(1+\delta)$ factor, while the techniques in Catoni (2007) are more refined for this situation and can achieve the strict oracle inequality with a better leading constant 1. [3]

We have only considered prior distributions with compact support Θ_n in our ranking example. However, our general inequalities can be directly extended to accommodate prior distributions with thin tails on a noncompact support, such as a normal prior. It is also possible to include high dimensional predictors with $p \gg n$ in the ranking example, with some additional adjustment of the concentration tools used in the proofs. For example, we can impose an additional model size cap in order to keep the number of candidate models from growing too fast.

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 dependent data and nonadditive empirical risk.

Acknowledgments

We are grateful to the anonymous referees for providing additional references, and for their valuable comments telling us which should be emphasized in our paper.

3. We thank a referee who pointed out that in Catoni (2007), it is proven that if one uses a Gibbs estimator in each model and then selects one of them following a Lepski-type procedure, then one can obtain a leading constant 1 (Theorem 2.2.11 of Catoni 2007). It is also known that such a result cannot be obtained for the Gibbs estimator (Audibert 2010).

References

P. Alquier and K. Lounici. PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electronic Journal of Statistics*, 5:127–145, 2011.

P. Alquier and O. Wintenberger. Model selection for weakly dependent time series forecasting. *Bernoulli*, 18:883–913, 2012.

J. Y. Audibert. Classification using gibbs estimators under complexity and margin assumptions. preprint, laboratoire de probabilités et modelès aléatoires, http://www.proba.jussieu.fr/mathdoc/textes/pma-908.pdf. 2004.

J. Y. Audibert. PAC-Bayesian aggregation and multi-armed bandits. habilitation thesis. université paris est. 2010.

A. Belloni and V. Chernozhukov. On the computational complexity of MCMC-based estimators in large samples. *The Annals of Statistics*, 37:2011–2055, 2009.

O. Catoni. *PAC-Bayesian Supervised Classification (The Thermodynamics of Statistical Learning)*, volume 37. IMS, 2007.

K. Chen, W. Jiang, and M. A. Tanner. A note on some algorithms for the Gibbs posterior. *Statistics and Probability Letters*, 80:1234–1241, 2010.

V. Chernozhukov and H. Hong. An MCMC approach to classical estimation. *Journal of Econometrics*, 115:293–346, 2003.

S. Clémencon, G. Lugosi, and N. Vayatis. Ranking and empirical minimization of u-statistics. *The Annals of Statistics*, 36:844–874, 2008.

A. K. Han. Non-parametric analysis of a generalized regression model - the maximum rank correlation estimator. *Journal of Econometrics*, 35:303–316, 1987.

W. Jiang and M. A. Tanner. Gibbs posterior for variable selection in high dimensional classification and data mining. *The Annals of Statistics*, 36:2207–2231, 2008.

W. Jiang and M. A. Tanner. Risk minimization for time series binary choice with variable selection. *Econometric Theory*, 26:1437–1452, 2010.

G Lecué. Suboptimality of penalized empirical risk minimization in classification. *COLT'07 Proceedings of the 20th Annual Conference on Learning Theory*, pages 142–156, 2007.

C. Li, W. Jiang, and M. A. Tanner. General inequalities for Gibbs posterior with nonadditive empirical risk. *Manuscript Submitted*, 2012.

P. Massart. *Concentration inequalities and model selection.* Spriner, Berlin, 2003.

W. Rejchel. On ranking and generalization bounds. *Journal of Machine Learning Research*, 13:1373–1392, 2012.

P. Rigollet and A. Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39:731–771, 2011.

R. P. Sherman. The limiting distribution of the maximum rank correlation estimator. *Econometrica*, 61:123–137, 1993.

A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical process*. Spriner, New York, 1996.

T. Zhang. Theoretical analysis of a class of randomized regularization methods. *COLT '99 Proceedings of the 12th Annual Conference on Computational Learning Theory*, pages 156–163, 1999.

T. Zhang. Information theoretical upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52:1307–1321, 2006.