General Oracle Inequalities for Gibbs Posterior with Application to Ranking

Cheng Li, Wenxin Jiang and Martin A. Tanner

Department of Statistics, Northwestern University

June 2013, Conference on Learning Theory

Gibbs Posterior and Oracle Inequalities

Our goal is to minimize some theoretical risk of interest R. For example, in bipartite ranking, the probability of misranking

$$R(\theta) = P[(Y - Y')r(X, X'; \theta) < 0]$$

Y is binary in $\{-1, 1\}$, X is a p-dimensional predictor vector, $r(x, x'; \theta) = 1$ if x ranks higher than x' and $r(x, x'; \theta) = -1$ otherwise. r is parameterized by θ.

Gibbs posterior is a randomization method of empirical risk minimization

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

- θ is the parameter vector; π is the prior; λ is the inverse temperature parameter.
- R_n is the empirical risk based on a sample of size n. In the ranking example,

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

We expect that θ sampled from the Gibbs posterior has good theoretical risk performance $R(\theta)$.

Under model selection, $\theta = (b, m)$, m is the model index, b is the parameter of interest in model space B_m.

The Gibbs posterior can be written as

$$Q(\mathrm{d}b, m) = \frac{e^{-\lambda R_n(b)}\pi(\mathrm{d}b|m)\pi_m}{\sum_m \int_{B_m} e^{-\lambda R_n(b)}\pi(\mathrm{d}b|m)\pi_m}$$

Oracle inequality (general form): Δ_m is the risk convergence rate on model m.

$$ER(b) \le (1+\delta)\inf_m \left\{ \inf_{b \in B_m} R(b) + O(\Delta_m) \right\}$$

with $\theta = (b, m)$ sampled from the Gibbs posterior Q, and small $\delta \ge 0$.

• PAC-Bayesian model selection $\Rightarrow$ Oracle inequalities for Gibbs posterior

Catoni (2007), Lecué (2007) COLT, Audibert (2010), Alquier and Lounici (2011), Rigollet and Tsybakov (2011), etc.

• Most of the PAC-Bayesian literature has focused on the **additive risk** $R_n(\theta)$, with the **iid summation** form. For example, the iid classification risk with classifier g, $R_n(\theta) = \frac{1}{n}\sum_{i=1}^n \left| Y_i - I(g(X_i, \theta) > 0) \right|$.

• Our work extends to **nonadditive risk** $R_n(\theta)$. For example, the ranking risk with ranking rule r,$R_n(\theta) = \frac{1}{n(n-1)}\sum_{i \neq j} I[(Y_i - Y_j) r(X_i, X_j; \theta) < 0]$.

Main Result: Oracle Inequalities for Gibbs Posterior

P is the true probability measure of the data. Q is the Gibbs posterior measure.

(1) Almost surely in PQ-measure, for small $\delta \geq 0$ and all large n,

$$R(b) \leq (1+\delta) \inf_m \left\{ \inf_{b \in B_m} R(b) + O(\Delta_m) \right\};$$

with (b, m) sampled from the Gibbs posterior.

(2) The posterior mean risk can be bounded as

$$E_{PQ} R \leq (1+\delta) \inf_m \left\{ \inf_{b \in B_m} R(b) + O(\Delta_m) \right\}.$$

Pros and cons of our approach:

+ No assumption on the form of R_n;

+ Applicable to the ranking risk and dependent data;

+ Adaptive to the unknown best candidate model;

+ Optimal or near optimal convergence rate;

– Not always strict oracle inequality (leading constant $1+\delta$ instead of 1).

Application to Bipartite Ranking

Minimal risk R^* is achieved by the $r^*(x, x') = 2I[\eta(x) - \eta(x') > 0] - 1$ where $\eta(x) = P(Y = 1|X = x)$.

We consider linear rules $r(x, x'; b) = 2I[(x - x')^\top b > 0] - 1$ with b in all coordinate subspaces $B_{m,j}$ of $\Re^p$, $m = 1, 2, ..., p$ and $j = 0, 1, ..., \binom{p}{m}$.

(1) Almost surely in PQ-measure for any $\delta > 0$ and all large n, there exists a constant $C_1 > 0$, such that

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

(2) For any $\delta > 0$ and all large n, there exists a constant $C_2 > 0$, such that

$$E_{PQ} R \le R^* + (1 + \delta) \inf_{m,j} \left[\inf_{b \in B_{m,j}} (R(b) - R^*) + \frac{C_2 m (\log n)^3}{n} \right]$$

+ Adaptive to the unknown best candidate model;

+ The fast oracle rate of about $O(m/n)$ instead of $O(\sqrt{m/n})$, without assuming the smoothness condition in Clémencon et al. (2008);

+ Dimension $p = o(n/(\log n)^3)$ can increase with the sample size n.

Welcome to my poster for details

Thank You!