An Analysis of Gibbs Posterior Concentration in Terms of the Separation α-Entropy

Olivier Binette & Yu Luo

Université du Québec à Montréal – olivier.binette@duke.edu & McGill University – yu.t.luo@mail.mcgill.ca

Key Points

- We introduce the *separation α-entropy*, a local measure of prior complexity, as a theoretical tool for the study of Gibbs posterior concentration in nonparametric models.
- It allows us to state probable posterior concentration bounds around risk-minimizing parameters in simple terms of separation entropy, prior concentration and sample size.
- It has computationally convenient properties, and it generalizes metric entropies and prior summability conditions.
- This is work in early progress and we are exploring particular applications.

Background

Renewed interest in behaviour of posterior distributions under misspecification. While it is well-known that posterior distributions tend to concentrate around Kullback-Leibler (KL) minimizing parameters under regularity assumptions, there are **some problems**:

- examples of inconsistency in natural situations;
- KL minimizer is heavily influenced by distributional tails and does not always exist.

Two related solutions:

- Raise the likelihood to a fractional power (fractional posterior distributions);
- use pseudo-likelihoods to target learning about a general risk-minimizing parameter (Gibbs posterior distributions).

Advantages of Gibbs:

- full data model not always required;
- also suited to MCMC computation, which is useful when the irregularity of the loss function hampers optimization-based procedures;
- uncertainty quantification, e.g. with calibrated credible sets.

Theoretical studies:

- Zhang (2006) provides PAC-Bayes theory for Gibbs posteriors (see also Grunwald et Mehta (2016) and work by Bhattacharya et al. (2019) in the context of fractional posteriors).

Framework

Let $\mathcal{X}$ be a sample space, X be a random variable on $\mathcal{X}$ and let Θ be a model associated with loss functions $\ell_\theta : \mathcal{X} \to \mathbb{R}$. That is, $\ell_\theta(X)$ represents the loss in using θ to fit the data X. The goal is use X to learn about a risk-minimizing parameter

$$\theta_0 \in \operatorname*{argmin}_{\theta \in \Theta} \mathbb{E}\left[\ell_\theta(X)\right].$$

Given a prior π on Θ used to regularize learning, the Gibbs posterior distribution on Θ is defined as the probability measure

$$\begin{aligned}\pi(\cdot \mid X) &= \operatorname*{argmin}_{\hat\pi} \left\{\mathbb{E}_{\theta\sim\hat\pi}\left[\ell_\theta(X)\right] + D(\hat\pi\|\pi)\right\}\\ &= \int_{\cdot} e^{-\ell_\theta(X)}\,\pi(d\theta) \Big/ \int_\Theta e^{-\ell_\theta(X)}\,\pi(d\theta)\end{aligned}$$

for $D(\hat\pi\|\pi)$ the Kullback-Leibler divergence (Zhang, 2006).

Density estimation. Let Θ parametrize a set of density functions $\{p_\theta \mid \theta \in \Theta\}$ and consider the loss $\ell_\theta(X) = -\eta \log p_\theta(X)$ for $\eta \in (0, 1]$. Then we recover posterior distributions $\pi(A \mid X) \propto \int_A p_\theta(X)^\eta\,\pi(d\theta)$ with $\eta = 1$ being usual. The case $\eta < 1$ corresponds to fractional posteriors.

Classification. Suppose $X = (U, Y)$ where $Y \in \{0, 1\}$ is a binary response and U is a predictor. Let Θ be a collection of classifiers and consider the loss $\ell_\theta(X) = \mathbb{I}(Y \neq \theta(U))$. The risk is then the missclassification rate and it is minimized at the oracle Bayes classifier.

Some Results

Let θ_0 be any fixed parameter, typically risk-minimizing from the excess risk support of the prior. We make use of the Rényi-type divergence

$$d_\alpha(\theta, \theta_0) = -\alpha^{-1} \log \mathbb{E}\left[e^{\alpha(\ell_{\theta_0}(X) - \ell_\theta(X))}\right]$$

and the excess risk $d_0(\theta, \theta_0) = \mathbb{E}[\ell_\theta(X) - \ell_{\theta_0}(X)]$. In the context of standard posterior distributions in well-specified models, d_0 is the Kullback-Leibler divergence and d_α is the Rényi divergence (comparable to the Hellinger distance).

We can then consider concentration and convergence in neighborhoods of the form $A = \{\theta \mid d_\alpha(\theta, \theta_0) < \varepsilon\}$ for some $\varepsilon > 0$ and $\alpha \in (0, 1)$. Our main tool is the separation α-entropy of A with separation parameter $\delta > 0$, which we denote by $\mathcal{S}_\alpha(A^c, \delta)$. Our results are stated in the i.i.d. setup, where $X^{(n)} = (X_1, X_2, \ldots, X_n)$ and we consider associated additive losses $\ell_\theta(X^{(n)}) = \sum_i \ell_\theta(X_i)$.

Theorem 1

Let $\alpha \in (0, 1]$, $\delta > 0$ and let

$$B(\delta) = \left\{\theta \mid d_{-1/2}(\theta, \theta_0) \le \delta\right\}.$$

With probability at least $1 - 2e^{-\alpha n\delta/2}$, we have that

$$\log \pi(A^c \mid X^{(n)}) \le \mathcal{S}_\alpha(A^c, 2\delta) - \log B(\delta) - n\delta.$$

Remarks.

- For the upper bound to be finite, it is necessary that $A \supset \{\theta \mid d_\alpha(\theta, \theta_0) < 2\delta\}$.
- The set A can depend on n to provide rates.

Theorem 2

Suppose there is a $\delta > 0$ such that

$$\pi\left(\{\theta \mid d_0(\theta, \theta_0) < \delta\}\right) > 0.$$

If $A \subset \Theta$ is such that $\mathcal{S}_\alpha(A^c, \delta) < \infty$ for some $\alpha \in (0, 1]$, then

$$\pi(A^c \mid X^{(n)}) \to 0$$

almost surely as $n \to \infty$.

Definition of $\mathcal{S}_\alpha$

Given a subset $A \subset \Theta$, denote by $\langle A\rangle$ the convexification of the set of pseudo-likelihoods $e^{-A} = \{x \mapsto e^{-\ell_\theta(x)} \mid \theta \in \Theta\}$. We say that A is δ-separated from θ_0 with respect to d_α if, for every $f \in \langle A\rangle$,

$$d_\alpha(f, \theta_0) := -\alpha^{-1} \log \mathbb{E}\left[\left(f(X)e^{\ell_{\theta_0}(X)}\right)^\alpha\right] \ge \delta.$$

Now given $\alpha \in (0, 1)$, π the prior on Θ and the fixed target θ_0, the separation α-entropy of a set $A \subset \Theta$ with separation parameter $\delta > 0$ is defined as

$$\mathcal{S}_\alpha(A, \delta) = \inf \alpha^{-1} \log \sum_{i=1}^\infty \pi(A_i)^\alpha$$

where the infimum is taken over all coverings $\{A_i\}$ of A such that each A_i is δ-separated from θ_0 with respect to d_α. When no such covering exists, we let $\mathcal{S}_\alpha(A, \delta) := \infty$.

This is inspired by the Hausdorff α-entropy of Xing (2009), from the notion of δ-separation discussed in Choi et al. (2008) and from the prior summability conditions of Barron (1986) and Walker (2004).

Computation

- In the case where $\alpha = 1$ and $A = \{\theta \mid d_1(\theta, \theta_0) < \delta\}$, automatically
$$\mathcal{S}_1(A^c, \delta) = \log \pi(A^c).$$
- Let $N(B, \delta)$ denote the minimal cardinality of a δ-separated covering of a set B. Similarly as for the Hausdorff α-entropy, for any partition $\{B_i\}$ of A and $\alpha \in (0, 1)$, we have
$$\begin{aligned}\mathcal{S}_\alpha(A, \delta) &\le \alpha^{-1} \log \sum_i e^{\alpha \mathcal{S}_\alpha(B_n, \delta)}\\ &\le \alpha^{-1} \log \sum_i \pi(B_n)^\alpha N(B_n, \delta)^{1-\alpha},\end{aligned}$$

an upper bound on $\mathcal{S}_\alpha$ in terms of covering numbers.

References

See appended page.

Acknowledgements

We aknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) and of the Fonds de Recherche Nature et Technologies du Québec.