PAC-Bayes Meets Variational Inference: Theory and Generalizations

Badr-Eddine Chérief-Abdellatif

CNRS, LPSM, Sorbonne Université

Post-Bayes Seminar Series

November 2025

Outline

Generalized Bayes in a nutshell (Chapter 1)

- Dataset $\mathcal{S} = \{x_1, \dots, x_n\}$.
- Statistical model $\{p_\theta : \theta \in \Theta\}$.
- Prior distribution $\pi(\theta)$ over Θ.

The Bayesian Posterior:

$$\pi(\mathrm{d}\theta \mid \mathcal{S}) \propto \left[\prod_{i=1}^{n} p_\theta(x_i)\right] \pi(\mathrm{d}\theta) .$$

- In learning settings, no statistical model.
- Objects of inference $\theta \in \Theta$.
- Loss function $\ell(\theta, x)$ measuring the quality of θ on x.

The Gibbs Posterior:

$$\pi(\mathrm{d}\theta \mid \mathcal{S}) \propto \exp\left(-\lambda_n \cdot \sum_{i=1}^{n} \ell(\theta, x_i)\right) \pi(\mathrm{d}\theta) .$$

Reasonable-ness of Generalized Bayes (Chapter 1)

The Gibbs Posterior:

$$\pi(\mathrm{d}\theta \mid \mathcal{S}) \propto \exp\left(-\lambda_n \cdot \sum_{i=1}^{n} \ell(\theta, x_i)\right) \pi(\mathrm{d}\theta) \,.$$

Is $\pi(\theta \mid \mathcal{S})$ a reasonable set of beliefs?

- Are inferences based on are *reliable/reasonable/useful*?
- *Reasonable-ness* measured here via the *large sample* behavior.
- In particular via *Posterior Concentration*: Does $\pi(\theta \mid \mathcal{S})$ assign high mass to regions where loss is small?

Reasonable-ness of Generalized Bayes (Chapter 1)

The Gibbs Posterior:

$$\pi(\mathrm{d}\theta \mid \mathcal{S}) \propto \exp\left(-\lambda_n \cdot \sum_{i=1}^{n} \ell(\theta, x_i)\right) \pi(\mathrm{d}\theta)\,.$$

Is $\pi(\theta \mid \mathcal{S})$ a reasonable set of beliefs?

- Are inferences based on are *reliable/reasonable/useful*?
- *Reasonable-ness* measured here via the *large sample* behavior.
- In particular via *Posterior Concentration*: Does $\pi(\theta \mid \mathcal{S})$ assign high mass to regions where loss is small?

The *Old School* theory was presented by **David F.** (in his own words).

Let's shortly investigate the *New School* theory based on **PAC-Bayes**.

The *Old School* theory in a nutshell (Chapter 1)

Assume that $\mathcal{S} = \{x_1, \ldots, x_n\}$ is i.i.d. from $P_\star$. Then define:

$$L(\theta) = \mathbb{E}_{X \sim P_\star}\left[\ell(\theta, X)\right] \quad , \quad \widehat{L}(\theta, \mathcal{S}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, x_i)\,.$$

Hope is that $\pi(\theta \mid \mathcal{S})$ *concentrates* onto the population loss minimizer:

$$\theta_\star := \arg\min_{\theta \in \Theta} L(\theta)\,.$$

The *Old School* theory in a nutshell (Chapter 1)

Assume that $\mathcal{S} = \{x_1, \ldots, x_n\}$ is i.i.d. from $P_\star$. Then define:

$$L(\theta) = \mathbb{E}_{X \sim P_\star}\left[\ell(\theta, X)\right] \quad , \quad \widehat{L}(\theta, \mathcal{S}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, x_i)\,.$$

Hope is that $\pi(\theta \mid \mathcal{S})$ *concentrates* onto the population loss minimizer:

$$\theta_\star := \arg\min_{\theta \in \Theta} L(\theta)\,.$$

Definition: The Gibbs Posterior is said to concentrate toward $\theta_\star$ at rate (at least) ε_n with respect to a metric $d(\theta, \theta')$ if

$$\mathbb{E}_{\mathcal{S}}\left[\pi\Big(\theta : d(\theta, \theta_\star) > M_n \varepsilon_n \mid \mathcal{S}\Big)\right] \xrightarrow[n \to +\infty]{} 0$$

where $M_n \to +\infty$ arbitrarily slowly or is a sufficiently large constant.

The *Old School* theory in a nutshell (Chapter 1)

Assume that $\mathcal{S} = \{x_1, \dots, x_n\}$ is i.i.d. from $P_\star$. Then define:

$$L(\theta) = \mathbb{E}_{X \sim P_\star}\left[\ell(\theta, X)\right] \quad , \quad \widehat{L}(\theta, \mathcal{S}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, x_i)\,.$$

Hope is that $\pi(\theta \mid \mathcal{S})$ *concentrates* onto the population loss minimizer:

$$\theta_\star := \arg\min_{\theta \in \Theta} L(\theta)\,.$$

Definition: The Gibbs Posterior is said to concentrate toward $\theta_\star$ at rate (at least) ε_n with respect to a metric $d(\theta, \theta')$ if

$$\mathbb{E}_{\mathcal{S}}\left[\pi\Big(\theta : d(\theta, \theta_\star) > M_n \varepsilon_n \mid \mathcal{S}\Big)\right] \xrightarrow[n \to +\infty]{} 0$$

where $M_n \to +\infty$ arbitrarily slowly or is a sufficiently large constant.

Informal (watch David F.'s talk for more details): The Gibbs Posterior concentrates toward $\theta_\star$ provided two key conditions: a *Prior mass* condition and a *well-behaved* loss

The *Old School* theory in a nutshell (Chapter 1)

Assume that $\mathcal{S} = \{x_1, \ldots, x_n\}$ is i.i.d. from $P_\star$. Then define:

$$L(\theta) = \mathbb{E}_{X \sim P_\star}\left[\ell(\theta, X)\right] \quad , \quad \widehat{L}(\theta, \mathcal{S}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, x_i)\,.$$

Hope is that $\pi(\theta \mid \mathcal{S})$ *concentrates* onto the population loss minimizer:

$$\theta_\star := \arg\min_{\theta \in \Theta} L(\theta)\,.$$

Definition: The Gibbs Posterior is said to concentrate toward $\theta_\star$ at rate (at least) ε_n with respect to a metric $d(\theta, \theta')$ if

$$\mathbb{E}_{\mathcal{S}}\left[\pi\Big(\theta : d(\theta, \theta_\star) > M_n \varepsilon_n \mid \mathcal{S}\Big)\right] \xrightarrow[n \to +\infty]{} 0$$

where $M_n \to +\infty$ arbitrarily slowly or is a sufficiently large constant.

Informal (watch David F.'s talk for more details): The Gibbs Posterior concentrates toward $\theta_\star$ provided two key conditions: a *Prior mass* condition and a *well-behaved* loss (+ a right choice of $d(\theta, \theta')$ and λ_n).

The *New School* theory in a nutshell (Chapter 3)

$$\mathbb{E}_{\mathcal{S}}\left[\pi\Big(\theta : d(\theta,\theta_\star) > M_n\varepsilon_n \mid \mathcal{S}\Big)\right] \xrightarrow[n\to+\infty]{??} 0 \quad \text{as} \quad M_n \to +\infty$$

The *New School* theory in a nutshell (Chapter 3)

$$\mathbb{E}_{\mathcal{S}}\left[\pi\left(\theta : d(\theta,\theta_\star) > M_n\varepsilon_n \mid \mathcal{S}\right)\right] \xrightarrow[n\to+\infty]{??} 0 \quad \text{as} \quad M_n \to +\infty$$

Several ingredients:

1. Markov's inequality:

$$\mathbb{E}_{\mathcal{S}}\left[\pi\left(\theta : d(\theta,\theta_\star) > M_n\varepsilon_n \mid \mathcal{S}\right)\right] \leq \frac{\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\pi(\theta|\mathcal{S})}\left[d(\theta,\theta_\star)\right]}{M_n\varepsilon_n} \xrightarrow[n\to+\infty]{??} 0\,.$$

   It is then enough to show that $\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\pi(\theta|\mathcal{S})}\left[d(\theta,\theta_\star)\right] \leq \varepsilon_n$.

The *New School* theory in a nutshell (Chapter 3)

$$\mathbb{E}_{\mathcal{S}}\left[\pi\Big(\theta : d(\theta,\theta_\star) > M_n\varepsilon_n \mid \mathcal{S}\Big)\right] \xrightarrow[n\to+\infty]{??} 0 \quad \text{as} \quad M_n \to +\infty$$

Several ingredients:

1. Markov's inequality:

$$\mathbb{E}_{\mathcal{S}}\left[\pi\Big(\theta : d(\theta,\theta_\star) > M_n\varepsilon_n \mid \mathcal{S}\Big)\right] \leq \frac{\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\pi(\theta|\mathcal{S})}\big[d(\theta,\theta_\star)\big]}{M_n\varepsilon_n} \xrightarrow[n\to+\infty]{??} 0\,.$$

   It is then enough to show that $\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\pi(\theta|\mathcal{S})}\big[d(\theta,\theta_\star)\big] \leq \varepsilon_n$.

2. Use the following PAC-Bayes result (proof to be detailed later):

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\pi(\theta|\mathcal{S})}\big[L(\theta)\big] \leq L(\theta_\star) + \widetilde{\mathcal{O}}\left(\lambda_n\right) + \widetilde{\mathcal{O}}\left(\frac{1}{\lambda_n\, n}\right)$$

   as soon as $\ell(\theta, x)$ is bounded and a *prior mass* condition is satisfied.

The *New School* theory in a nutshell (Chapter 3)

$$\mathbb{E}_{\mathcal{S}}\left[\pi\Big(\theta : d(\theta,\theta_\star) > M_n\varepsilon_n \mid \mathcal{S}\Big)\right] \xrightarrow[n\to+\infty]{??} 0 \quad \text{as} \quad M_n \to +\infty$$

Several ingredients:

1. Markov's inequality:

$$\mathbb{E}_{\mathcal{S}}\left[\pi\Big(\theta : d(\theta,\theta_\star) > M_n\varepsilon_n \mid \mathcal{S}\Big)\right] \leq \frac{\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\pi(\theta|\mathcal{S})}\big[d(\theta,\theta_\star)\big]}{M_n\varepsilon_n} \xrightarrow[n\to+\infty]{??} 0\,.$$

   It is then enough to show that $\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\pi(\theta|\mathcal{S})}\big[d(\theta,\theta_\star)\big] \leq \varepsilon_n$.

2. Use the following PAC-Bayes result (proof to be detailed later):

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\pi(\theta|\mathcal{S})}\big[L(\theta)\big] \leq L(\theta_\star) + \widetilde{\mathcal{O}}\left(\lambda_n\right) + \widetilde{\mathcal{O}}\left(\frac{1}{\lambda_n\, n}\right)$$

   as soon as $\ell(\theta, x)$ is bounded and a *prior mass* condition is satisfied.

3. Choose the *excess risk* metric $d(\theta,\theta_\star) = L(\theta) - L(\theta_\star)$ to achieve concentration at rate (up to a log)

$$\varepsilon_n = \lambda_n + \frac{1}{\lambda_n\, n}\,.$$

Beyond the Gibbs posterior

The Gibbs posterior concentrates w.r.t. $d(\theta, \theta_\star) = L(\theta) - L(\theta_\star)$ at rate:

$$\varepsilon_n = \lambda_n + \frac{1}{\lambda_n n}\,.$$

Beyond the Gibbs posterior

The Gibbs posterior concentrates w.r.t. $d(\theta, \theta_\star) = L(\theta) - L(\theta_\star)$ at rate:

$$\varepsilon_n = \lambda_n + \frac{1}{\lambda_n \, n} \, .$$

Two quick remarks:

1. The boundedness assumption is relevant in the learning framework, which we focus on for the moment.
2. The temperature parameter has to be tuned, with an optimal scaling in $\lambda_n \propto n^{-1/2}$ leading to concentration in $n^{-1/2}$. This makes sense.

So we're fine!

Beyond the Gibbs posterior

The Gibbs posterior concentrates w.r.t. $d(\theta, \theta_\star) = L(\theta) - L(\theta_\star)$ at rate:

$$\varepsilon_n = \lambda_n + \frac{1}{\lambda_n\, n}\,.$$

Two quick remarks:

1. The boundedness assumption is relevant in the learning framework, which we focus on for the moment.
2. The temperature parameter has to be tuned, with an optimal scaling in $\lambda_n \propto n^{-1/2}$ leading to concentration in $n^{-1/2}$. This makes sense.

So we're fine! Is that the end of the story?

Beyond the Gibbs posterior

The Gibbs posterior concentrates w.r.t. $d(\theta, \theta_\star) = L(\theta) - L(\theta_\star)$ at rate:

$$\varepsilon_n = \lambda_n + \frac{1}{\lambda_n \, n} \, .$$

Two quick remarks:

1. The boundedness assumption is relevant in the learning framework, which we focus on for the moment.
2. The temperature parameter has to be tuned, with an optimal scaling in $\lambda_n \propto n^{-1/2}$ leading to concentration in $n^{-1/2}$. This makes sense.

So we're fine! Is that the end of the story?

- No, because the Gibbs posterior is rarely available in practice...

Beyond the Gibbs posterior

The Gibbs posterior concentrates w.r.t. $d(\theta, \theta_\star) = L(\theta) - L(\theta_\star)$ at rate:

$$\varepsilon_n = \lambda_n + \frac{1}{\lambda_n \, n} \, .$$

Two quick remarks:

1. The boundedness assumption is relevant in the learning framework, which we focus on for the moment.
2. The temperature parameter has to be tuned, with an optimal scaling in $\lambda_n \propto n^{-1/2}$ leading to concentration in $n^{-1/2}$. This makes sense.

So we're fine! Is that the end of the story?

- No, because the Gibbs posterior is rarely available in practice...

Can we do something to solve this problem?

Beyond the Gibbs posterior

The Gibbs posterior concentrates w.r.t. $d(\theta, \theta_\star) = L(\theta) - L(\theta_\star)$ at rate:

$$\varepsilon_n = \lambda_n + \frac{1}{\lambda_n \, n} \, .$$

Two quick remarks:

1. The boundedness assumption is relevant in the learning framework, which we focus on for the moment.
2. The temperature parameter has to be tuned, with an optimal scaling in $\lambda_n \propto n^{-1/2}$ leading to concentration in $n^{-1/2}$. This makes sense.

So we're fine! Is that the end of the story?

- No, because the Gibbs posterior is rarely available in practice...

Can we do something to solve this problem?

- Yes, approximate the Gibbs posterior via **Variational Inference**!

Beyond the Gibbs posterior

The Gibbs posterior concentrates w.r.t. $d(\theta, \theta_\star) = L(\theta) - L(\theta_\star)$ at rate:

$$\varepsilon_n = \lambda_n + \frac{1}{\lambda_n\, n}\,.$$

Two quick remarks:

1. The boundedness assumption is relevant in the learning framework, which we focus on for the moment.
2. The temperature parameter has to be tuned, with an optimal scaling in $\lambda_n \propto n^{-1/2}$ leading to concentration in $n^{-1/2}$. This makes sense.

So we're fine! Is that the end of the story?

- No, because the Gibbs posterior is rarely available in practice...

Can we do something to solve this problem?

- Yes, approximate the Gibbs posterior via **Variational Inference**!
- But does the approximation retain the nice reasonable-ness properties of the posterior it approximates?

Gibbs Posteriors and PAC-Bayes

Approximate Bayes and Variational Inference

Variational Inference and PAC-Bayes

Discussion and generalization

Approximate Bayes: VI original definition

Computing the normalizing constant is often challenging in complex models:

$$Z = \mathbb{E}_{\vartheta \sim \pi}\left[\exp\left(-\lambda_n \cdot \sum_{i=1}^{n} \ell(\vartheta, x_i)\right)\right].$$

Approximate Bayes: VI original definition

Computing the normalizing constant is often challenging in complex models:

$$Z = \mathbb{E}_{\vartheta \sim \pi} \left[\exp \left(-\lambda_n \cdot \sum_{i=1}^{n} \ell(\vartheta, x_i) \right) \right] .$$

Idea of VI: choose a family $\mathcal{Q}$ of probability distributions on Θ and approximate $\pi(\cdot \mid \mathcal{S})$ by the closest distribution in the variational set $\mathcal{Q}$, i.e.

$$\widetilde{\pi}(\cdot \mid \mathcal{S}) := \arg\min_{q \in \mathcal{Q}} \mathrm{KL}\left(q \,\middle\|\, \pi(\cdot \mid \mathcal{S}) \right) .$$

Approximate Bayes: VI original definition

Computing the normalizing constant is often challenging in complex models:

$$Z = \mathbb{E}_{\vartheta \sim \pi}\left[\exp\left(-\lambda_n \cdot \sum_{i=1}^{n} \ell(\vartheta, x_i)\right)\right] .$$

Idea of VI: choose a family $\mathcal{Q}$ of probability distributions on Θ and approximate $\pi(\cdot \mid \mathcal{S})$ by the closest distribution in the variational set $\mathcal{Q}$, i.e.

$$\widetilde{\pi}(\cdot \mid \mathcal{S}) := \arg\min_{q \in \mathcal{Q}} \mathrm{KL}\left(q \,\middle\|\, \pi(\cdot \mid \mathcal{S})\right) .$$

Examples of sets $\mathcal{Q}$:

- parametric ($\Theta \subset \mathbb{R}^d$):

$$\left\{\mathcal{N}(\mu, \Sigma) \colon \mu \in \mathbb{R}^d, \Sigma \in \mathcal{S}_d^+\right\} .$$

- mean-field ($\Theta = \Theta_1 \times \Theta_2$):

$$q(\mathrm{d}\theta) = q_1(\mathrm{d}\theta_1) \times q_2(\mathrm{d}\theta_2).$$

From Approximate Bayes to Variational Inference

Seems sound, but why the exclusive KL?

From Approximate Bayes to Variational Inference

Seems sound, but why the exclusive KL? To remove the normalizing constant Z in the optimization objective, thanks to the following straightforward derivation:

$$
\begin{aligned}
\mathsf{KL}\Big(q \,\Big\|\, \pi(\cdot \mid \mathcal{S})\Big) &= \mathbb{E}_{\theta\sim q}\left[\log \frac{\mathrm{d}q}{\mathrm{d}\pi(\cdot \mid \mathcal{S})}(\theta)\right] \\
&= \mathbb{E}_{\theta\sim q}\left[\log\left(\frac{Z}{\exp\left(-\lambda_n \cdot \sum_{i=1}^{n} \ell(\theta, x_i)\right)} \cdot \frac{\mathrm{d}q}{\mathrm{d}\pi}(\theta)\right)\right] \\
&= \log Z + \mathbb{E}_{\theta\sim q}\left[\lambda_n \cdot \sum_{i=1}^{n} \ell(\theta, x_i)\right] + \mathsf{KL}\big(q \,\big\|\, \pi\big)\,.
\end{aligned}
$$

From Approximate Bayes to Variational Inference

Seems sound, but why the exclusive KL? To remove the normalizing constant Z in the optimization objective, thanks to the following straightforward derivation:

$$\begin{aligned}
\mathsf{KL}\Big(q \,\Big\|\, \pi(\cdot \mid \mathcal{S})\Big) &= \mathbb{E}_{\theta\sim q}\left[\log \frac{\mathrm{d}q}{\mathrm{d}\pi(\cdot \mid \mathcal{S})}(\theta)\right] \\
&= \mathbb{E}_{\theta\sim q}\left[\log\left(\frac{Z}{\exp\left(-\lambda_n \cdot \sum_{i=1}^n \ell(\theta, x_i)\right)} \cdot \frac{\mathrm{d}q}{\mathrm{d}\pi}(\theta)\right)\right] \\
&= \log Z + \mathbb{E}_{\theta\sim q}\left[\lambda_n \cdot \sum_{i=1}^n \ell(\theta, x_i)\right] + \mathsf{KL}\Big(q \,\Big\|\, \pi\Big) .
\end{aligned}$$

So the normalizing constant does not appear in the optimization objective:

$$\begin{aligned}
\widetilde{\pi}(\cdot \mid \mathcal{S}) &= \arg\min_{q\in\mathcal{Q}} \mathsf{KL}\Big(q \,\Big\|\, \pi(\cdot \mid \mathcal{S})\Big) \\
&= \arg\min_{q\in\mathcal{Q}} \left\{\mathbb{E}_{\theta\sim q}\left[\widehat{L}(\theta, \mathcal{S})\right] + \frac{\mathsf{KL}(q\|\pi)}{\lambda_n\, n}\right\} .
\end{aligned}$$

From Approximate Bayes to Variational Inference

Seems sound, but why the exclusive KL? To remove the normalizing constant Z in the optimization objective, thanks to the following straightforward derivation:

$$\begin{aligned}
\mathsf{KL}\Big(q \,\Big\|\, \pi(\cdot \mid \mathcal{S})\Big) &= \mathbb{E}_{\theta\sim q}\left[\log \frac{\mathrm{d}q}{\mathrm{d}\pi(\cdot \mid \mathcal{S})}(\theta)\right] \\
&= \mathbb{E}_{\theta\sim q}\left[\log\left(\frac{Z}{\exp\left(-\lambda_n \cdot \sum_{i=1}^n \ell(\theta, x_i)\right)} \cdot \frac{\mathrm{d}q}{\mathrm{d}\pi}(\theta)\right)\right] \\
&= \log Z + \mathbb{E}_{\theta\sim q}\left[\lambda_n \cdot \sum_{i=1}^n \ell(\theta, x_i)\right] + \mathsf{KL}\Big(q \,\Big\|\, \pi\Big)\,.
\end{aligned}$$

So the normalizing constant does not appear in the optimization objective:

$$\begin{aligned}
\widetilde{\pi}(\cdot \mid \mathcal{S}) &= \arg\min_{q\in\mathcal{Q}} \mathsf{KL}\Big(q \,\Big\|\, \pi(\cdot \mid \mathcal{S})\Big) \\
&= \arg\min_{q\in\mathcal{Q}} \left\{\mathbb{E}_{\theta\sim q}\left[\widehat{L}(\theta, \mathcal{S})\right] + \frac{\mathsf{KL}(q\|\pi)}{\lambda_n\, n}\right\}\,.
\end{aligned}$$

There are two different perspectives on VI: an **approximate Bayes** perspective (update-then-project) and a **variational** perspective (constrain-then-optimize).

The variational perspective and PAC-Bayes

Same objective for both the Gibbs posterior and its variational approximation:

$$\widetilde{\pi}(\cdot \mid \mathcal{S}) = \arg \min_{q \in \mathcal{Q}} \left\{ \mathbb{E}_{\theta \sim q} \left[\widehat{L}(\theta, \mathcal{S}) \right] + \frac{\mathrm{KL}(q \| \pi)}{\lambda_n \, n} \right\}$$

$$\pi(\cdot \mid \mathcal{S}) = \arg \min_{q \in \mathcal{P}(\Theta)} \left\{ \mathbb{E}_{\theta \sim q} \left[\widehat{L}(\theta, \mathcal{S}) \right] + \frac{\mathrm{KL}(q \| \pi)}{\lambda_n \, n} \right\} .$$

The variational perspective and PAC-Bayes

Same objective for both the Gibbs posterior and its variational approximation:

$$\widetilde{\pi}(\cdot \mid \mathcal{S}) = \arg \min_{q \in \mathcal{Q}} \left\{ \mathbb{E}_{\theta \sim q} \left[\widehat{L}(\theta, \mathcal{S}) \right] + \frac{\mathrm{KL}(q \| \pi)}{\lambda_n \, n} \right\}$$
$$\pi(\cdot \mid \mathcal{S}) = \arg \min_{q \in \mathcal{P}(\Theta)} \left\{ \mathbb{E}_{\theta \sim q} \left[\widehat{L}(\theta, \mathcal{S}) \right] + \frac{\mathrm{KL}(q \| \pi)}{\lambda_n \, n} \right\} .$$

Furthermore, the optimization objective = Catoni's PAC-Bayes bound (2003): for any sample size n, any $\lambda_n > 0$, any (data-free) prior π, any bounded loss, and any (possibly data-dependent) posterior q, we have

$$\mathbb{E}_{\mathcal{S}} \mathbb{E}_{\theta \sim q} \left[L(\theta) \right] \leq \mathbb{E}_{\mathcal{S}} \left[\mathbb{E}_{\theta \sim q} \left[\widehat{L}(\theta, \mathcal{S}) \right] + \frac{\mathrm{KL}(q \| \pi)}{\lambda_n \, n} \right] + \frac{\lambda_n}{8} .$$

Question: can we exploit this result to derive concentration rates?

From PAC-Bayes bounds to concentration rates

Reminder: ε_n is a concentration rate (w.r.t. the excess risk metric) if

$$\mathbb{E}_{\mathcal{S}}\left[\pi\Big(\theta : L(\theta) - L(\theta_\star) > M_n\varepsilon_n \mid \mathcal{S}\Big)\right] \xrightarrow[n\to+\infty]{} 0\,,$$

From PAC-Bayes bounds to concentration rates

Reminder: ε_n is a concentration rate (w.r.t. the excess risk metric) if

$$\mathbb{E}_{\mathcal{S}}\left[\pi\Big(\theta : L(\theta) - L(\theta_\star) > M_n\varepsilon_n \mid \mathcal{S}\Big)\right] \xrightarrow[n\to+\infty]{} 0\,,$$

Any sequence ε_n satisfying

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\pi(\theta|\mathcal{S})}\big[L(\theta)\big] \leq L(\theta_\star) + \varepsilon_n$$

is a concentration rate, since it implies

$$\mathbb{E}_{\mathcal{S}}\left[\pi\Big(\theta : L(\theta) - L(\theta_\star) > M_n\varepsilon_n \mid \mathcal{S}\Big)\right] \leq \frac{\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\pi(\theta|\mathcal{S})}\big[L(\theta) - L(\theta_\star)\big]}{M_n\varepsilon_n} \leq \frac{1}{M_n}\,.$$

From PAC-Bayes bounds to concentration rates

Reminder: ε_n is a concentration rate (w.r.t. the excess risk metric) if

$$\mathbb{E}_{\mathcal{S}}\left[\pi\Big(\theta : L(\theta) - L(\theta_\star) > M_n\varepsilon_n \mid \mathcal{S}\Big)\right] \xrightarrow[n\to+\infty]{} 0\,,$$

Any sequence ε_n satisfying

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\pi(\theta|\mathcal{S})}\big[L(\theta)\big] \leq L(\theta_\star) + \varepsilon_n$$

is a concentration rate, since it implies

$$\mathbb{E}_{\mathcal{S}}\left[\pi\Big(\theta : L(\theta) - L(\theta_\star) > M_n\varepsilon_n \mid \mathcal{S}\Big)\right] \leq \frac{\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\pi(\theta|\mathcal{S})}\big[L(\theta) - L(\theta_\star)\big]}{M_n\varepsilon_n} \leq \frac{1}{M_n}\,.$$

Question: can we exploit Catoni's bound to derive such excess risk bounds?

PAC-Bayes derivation of rates for the Gibbs posterior (1)

Objective: find ε_n such that $\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\pi(\theta|\mathcal{S})}\left[L(\theta)\right] \leq L(\theta_\star)+\varepsilon_n$

PAC-Bayes derivation of rates for the Gibbs posterior (1)

Objective: find ε_n such that $\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\pi(\theta|\mathcal{S})}\left[L(\theta)\right] \leq L(\theta_\star)+\varepsilon_n$

- Applying Catoni's bound to the Gibbs posterior: for any $\lambda_n > 0$, any π,

$$
\begin{aligned}
\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi(\cdot|\mathcal{S})}\left[L(\theta)\right] &\leq \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\pi(\cdot|\mathcal{S})}\left[\widehat{L}(\theta,\mathcal{S})\right] + \frac{\mathrm{KL}(\pi(\cdot \mid \mathcal{S})\|\pi)}{\lambda_n\, n}\right] + \frac{\lambda_n}{8} \\
&= \mathbb{E}_{\mathcal{S}}\left[\inf_{q\in\mathcal{P}(\Theta)}\left\{\mathbb{E}_{\theta\sim q}\left[\widehat{L}(\theta,\mathcal{S})\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n\, n}\right\}\right] + \frac{\lambda_n}{8} \\
&\leq \inf_{q\in\mathcal{P}(\Theta)}\left\{\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim q}\left[\widehat{L}(\theta,\mathcal{S})\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n\, n}\right]\right\} + \frac{\lambda_n}{8} \\
&= \inf_{q\in\mathcal{P}(\Theta)}\left\{\mathbb{E}_{\theta\sim q}\left[L(\theta)\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n\, n}\right\} + \frac{\lambda_n}{8}\,.
\end{aligned}
$$

PAC-Bayes derivation of rates for the Gibbs posterior (1)

Objective: find ε_n such that $\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\pi(\theta|\mathcal{S})}\left[L(\theta)\right] \leq L(\theta_\star) + \varepsilon_n$

- Applying Catoni's bound to the Gibbs posterior: for any $\lambda_n > 0$, any π,

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi(\cdot|\mathcal{S})}\left[L(\theta)\right] &\leq \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\pi(\cdot|\mathcal{S})}\left[\widehat{L}(\theta,\mathcal{S})\right] + \frac{\mathrm{KL}(\pi(\cdot \mid \mathcal{S})\|\pi)}{\lambda_n\, n}\right] + \frac{\lambda_n}{8} \\
&= \mathbb{E}_{\mathcal{S}}\left[\inf_{q\in\mathcal{P}(\Theta)}\left\{\mathbb{E}_{\theta\sim q}\left[\widehat{L}(\theta,\mathcal{S})\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n\, n}\right\}\right] + \frac{\lambda_n}{8} \\
&\leq \inf_{q\in\mathcal{P}(\Theta)}\left\{\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim q}\left[\widehat{L}(\theta,\mathcal{S})\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n\, n}\right]\right\} + \frac{\lambda_n}{8} \\
&= \inf_{q\in\mathcal{P}(\Theta)}\left\{\mathbb{E}_{\theta\sim q}\left[L(\theta)\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n\, n}\right\} + \frac{\lambda_n}{8}\,.
\end{aligned}$$

- Restrict minimization to the subset $\pi_r(\mathrm{d}\theta) \propto \mathbb{1}(\theta\in\mathcal{B}_r)\cdot\pi(\mathrm{d}\theta)$ where $\mathcal{B}_r = \{\theta : L(\theta) \leq L(\theta_\star) + r\}$ are the loss minimizer neighborhoods:

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi(\cdot|\mathcal{S})}\left[L(\theta)\right] &\leq \inf_{q\in\mathcal{P}(\Theta)}\left\{\mathbb{E}_{\theta\sim q}\left[L(\theta)\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n\, n}\right\} + \frac{\lambda_n}{8} \\
&\leq \inf_{r>0}\left\{\mathbb{E}_{\theta\sim\pi_r}\left[L(\theta)\right] + \frac{\mathrm{KL}(\pi_r\|\pi)}{\lambda_n\, n}\right\} + \frac{\lambda_n}{8}\,.
\end{aligned}$$

PAC-Bayes derivation of rates for the Gibbs posterior (2)

Objective: find ε_n such that $\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\pi(\theta|\mathcal{S})}\left[L(\theta)\right] \le L(\theta_\star)+\varepsilon_n$

- Restrict minimization to the subset $\pi_r(\mathrm{d}\theta) \propto \mathbb{1}(\theta \in \mathcal{B}_r) \cdot \pi(\mathrm{d}\theta)$ where $\mathcal{B}_r = \{\theta : L(\theta) \le L(\theta_\star) + r\}$ are the loss minimizer neighborhoods:

$$
\begin{aligned}
\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi(\cdot|\mathcal{S})}\left[L(\theta)\right] &\le \inf_{r>0}\left\{\mathbb{E}_{\theta\sim\pi_r}\left[L(\theta)\right] + \frac{\mathrm{KL}(\pi_r\|\pi)}{\lambda_n\, n}\right\} + \frac{\lambda_n}{8} \\
&\le \inf_{r>0}\left\{L(\theta_\star) + r + \frac{\mathrm{KL}(\pi_r\|\pi)}{\lambda_n\, n}\right\} + \frac{\lambda_n}{8} \\
&= L(\theta_\star) + \inf_{r>0}\left\{r + \frac{-\log \pi(\mathcal{B}_r)}{\lambda_n\, n}\right\} + \frac{\lambda_n}{8}\,.
\end{aligned}
$$

PAC-Bayes derivation of rates for the Gibbs posterior (2)

Objective: find ε_n such that $\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\pi(\theta|\mathcal{S})}\left[L(\theta)\right] \leq L(\theta_\star) + \varepsilon_n$

- Restrict minimization to the subset $\pi_r(\mathrm{d}\theta) \propto \mathbb{1}(\theta \in \mathcal{B}_r) \cdot \pi(\mathrm{d}\theta)$ where $\mathcal{B}_r = \{\theta : L(\theta) \leq L(\theta_\star) + r\}$ are the loss minimizer neighborhoods:

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi(\cdot|\mathcal{S})}\left[L(\theta)\right] &\leq \inf_{r>0}\left\{\mathbb{E}_{\theta\sim\pi_r}\left[L(\theta)\right] + \frac{\mathrm{KL}(\pi_r\|\pi)}{\lambda_n\, n}\right\} + \frac{\lambda_n}{8} \\
&\leq \inf_{r>0}\left\{L(\theta_\star) + r + \frac{\mathrm{KL}(\pi_r\|\pi)}{\lambda_n\, n}\right\} + \frac{\lambda_n}{8} \\
&= L(\theta_\star) + \inf_{r>0}\left\{r + \frac{-\log \pi(\mathcal{B}_r)}{\lambda_n\, n}\right\} + \frac{\lambda_n}{8}\,.
\end{aligned}$$

- Under the prior mass condition: $\pi(\mathcal{B}_r) \geq \left(\frac{r}{c}\right)^d$ for some $c > 0$, $d > 0$:

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi(\cdot|\mathcal{S})}\left[L(\theta)\right] &\leq L(\theta_\star) + \inf_{r>0}\left\{r + \frac{d\,\log(c/r)}{\lambda_n\, n}\right\} + \frac{\lambda_n}{8} \\
&\leq L(\theta_\star) + \frac{d\,\log(ce\lambda_n n/d)}{\lambda_n\, n} + \frac{\lambda_n}{8}\,.
\end{aligned}$$

PAC-Bayes derivation of rates for the Gibbs posterior (2)

Objective: find ε_n such that $\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\pi(\theta|\mathcal{S})}\left[L(\theta)\right] \leq L(\theta_\star)+\varepsilon_n$

- Restrict minimization to the subset $\pi_r(\mathrm{d}\theta) \propto \mathbb{1}(\theta \in \mathcal{B}_r) \cdot \pi(\mathrm{d}\theta)$ where $\mathcal{B}_r = \{\theta : L(\theta) \leq L(\theta_\star) + r\}$ are the loss minimizer neighborhoods:

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi(\cdot|\mathcal{S})}\left[L(\theta)\right] &\leq \inf_{r>0}\left\{\mathbb{E}_{\theta\sim\pi_r}\left[L(\theta)\right] + \frac{\mathrm{KL}(\pi_r\|\pi)}{\lambda_n\, n}\right\} + \frac{\lambda_n}{8} \\
&\leq \inf_{r>0}\left\{L(\theta_\star) + r + \frac{\mathrm{KL}(\pi_r\|\pi)}{\lambda_n\, n}\right\} + \frac{\lambda_n}{8} \\
&= L(\theta_\star) + \inf_{r>0}\left\{r + \frac{-\log \pi(\mathcal{B}_r)}{\lambda_n\, n}\right\} + \frac{\lambda_n}{8}\,.
\end{aligned}$$

- Under the prior mass condition: $\pi(\mathcal{B}_r) \geq \left(\frac{r}{c}\right)^d$ for some $c > 0$, $d > 0$:

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi(\cdot|\mathcal{S})}\left[L(\theta)\right] &\leq L(\theta_\star) + \inf_{r>0}\left\{r + \frac{d\,\log(c/r)}{\lambda_n\, n}\right\} + \frac{\lambda_n}{8} \\
&\leq L(\theta_\star) + \frac{d\,\log(ce\lambda_n n/d)}{\lambda_n\, n} + \frac{\lambda_n}{8}\,.
\end{aligned}$$

We finally have the rate (up to a log) $\varepsilon_n = \lambda_n + \dfrac{1}{\lambda_n\, n}$.

PAC-Bayes derivation of rates for the approximations (1)

Objective: find ε_n such that $\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\widetilde{\pi}(\theta|\mathcal{S})}\left[L(\theta)\right] \leq L(\theta_\star) + \varepsilon_n$

- Same route as before: apply Catoni's bound to the approximations to get for any $\lambda_n > 0$, any π,

$$
\begin{aligned}
\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\widetilde{\pi}(\cdot|\mathcal{S})}[L(\theta)] &\leq \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\widetilde{\pi}(\cdot|\mathcal{S})}\left[\widehat{L}(\theta,\mathcal{S})\right] + \frac{\mathrm{KL}(\widetilde{\pi}(\cdot\mid\mathcal{S})\|\pi)}{\lambda_n n}\right] + \frac{\lambda_n}{8} \\
&= \mathbb{E}_{\mathcal{S}}\left[\inf_{q\in\mathcal{Q}}\left\{\mathbb{E}_{\theta\sim q}\left[\widehat{L}(\theta,\mathcal{S})\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n n}\right\}\right] + \frac{\lambda_n}{8} \\
&\leq \inf_{q\in\mathcal{Q}}\left\{\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim q}\left[\widehat{L}(\theta,\mathcal{S})\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n n}\right]\right\} + \frac{\lambda_n}{8} \\
&= \inf_{q\in\mathcal{Q}}\left\{\mathbb{E}_{\theta\sim q}[L(\theta)] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n n}\right\} + \frac{\lambda_n}{8}.
\end{aligned}
$$

PAC-Bayes derivation of rates for the approximations (1)

Objective: find ε_n such that $\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\widetilde{\pi}(\theta|\mathcal{S})}\left[L(\theta)\right] \leq L(\theta_\star)+\varepsilon_n$

- Same route as before: apply Catoni's bound to the approximations to get for any $\lambda_n > 0$, any π,

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\widetilde{\pi}(\cdot|\mathcal{S})}\left[L(\theta)\right] &\leq \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\widetilde{\pi}(\cdot|\mathcal{S})}\left[\widehat{L}(\theta,\mathcal{S})\right] + \frac{\mathrm{KL}(\widetilde{\pi}(\cdot\mid\mathcal{S})\|\pi)}{\lambda_n\, n}\right] + \frac{\lambda_n}{8} \\
&= \mathbb{E}_{\mathcal{S}}\left[\inf_{q\in\mathcal{Q}}\left\{\mathbb{E}_{\theta\sim q}\left[\widehat{L}(\theta,\mathcal{S})\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n\, n}\right\}\right] + \frac{\lambda_n}{8} \\
&\leq \inf_{q\in\mathcal{Q}}\left\{\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim q}\left[\widehat{L}(\theta,\mathcal{S})\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n\, n}\right]\right\} + \frac{\lambda_n}{8} \\
&= \inf_{q\in\mathcal{Q}}\left\{\mathbb{E}_{\theta\sim q}\left[L(\theta)\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n\, n}\right\} + \frac{\lambda_n}{8}\,.
\end{aligned}$$

- Problem: we do not necessarily have $\pi_r(\mathrm{d}\theta) \propto \mathbb{1}(\theta\in\mathcal{B}_r)\cdot\pi(\mathrm{d}\theta) \subset \mathcal{Q}$...

PAC-Bayes derivation of rates for the approximations (1)

Objective: find ε_n such that $\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\widetilde{\pi}(\theta|\mathcal{S})}\left[L(\theta)\right] \leq L(\theta_\star) + \varepsilon_n$

- Same route as before: apply Catoni's bound to the approximations to get for any $\lambda_n > 0$, any π,

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\widetilde{\pi}(\cdot|\mathcal{S})}\left[L(\theta)\right] &\leq \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\widetilde{\pi}(\cdot|\mathcal{S})}\left[\widehat{L}(\theta,\mathcal{S})\right] + \frac{\mathrm{KL}(\widetilde{\pi}(\cdot\mid\mathcal{S})\|\pi)}{\lambda_n\, n}\right] + \frac{\lambda_n}{8} \\
&= \mathbb{E}_{\mathcal{S}}\left[\inf_{q\in\mathcal{Q}}\left\{\mathbb{E}_{\theta\sim q}\left[\widehat{L}(\theta,\mathcal{S})\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n\, n}\right\}\right] + \frac{\lambda_n}{8} \\
&\leq \inf_{q\in\mathcal{Q}}\left\{\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim q}\left[\widehat{L}(\theta,\mathcal{S})\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n\, n}\right]\right\} + \frac{\lambda_n}{8} \\
&= \inf_{q\in\mathcal{Q}}\left\{\mathbb{E}_{\theta\sim q}\left[L(\theta)\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n\, n}\right\} + \frac{\lambda_n}{8}.
\end{aligned}$$

- Problem: we do not necessarily have $\pi_r(\mathrm{d}\theta) \propto \mathbb{1}(\theta\in\mathcal{B}_r)\cdot\pi(\mathrm{d}\theta) \subset \mathcal{Q}$...
- Question: how can we make

$$\inf_{q\in\mathcal{Q}}\left\{\mathbb{E}_{\theta\sim q}\left[L(\theta)\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n\, n}\right\} \leq L(\theta_\star) + \widetilde{\mathcal{O}}\left(\frac{1}{\lambda_n\, n}\right) \quad ??$$

PAC-Bayes derivation of rates for the approximations (1)

Objective: find ε_n such that $\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\widetilde{\pi}(\theta|\mathcal{S})}\left[L(\theta)\right] \leq L(\theta_\star)+\varepsilon_n$

- Same route as before: apply Catoni's bound to the approximations to get for any $\lambda_n > 0$, any π,

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\widetilde{\pi}(\cdot|\mathcal{S})}[L(\theta)] &\leq \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\widetilde{\pi}(\cdot|\mathcal{S})}\left[\widehat{L}(\theta,\mathcal{S})\right] + \frac{\mathrm{KL}(\widetilde{\pi}(\cdot\mid\mathcal{S})\|\pi)}{\lambda_n n}\right] + \frac{\lambda_n}{8} \\
&= \mathbb{E}_{\mathcal{S}}\left[\inf_{q\in\mathcal{Q}}\left\{\mathbb{E}_{\theta\sim q}\left[\widehat{L}(\theta,\mathcal{S})\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n n}\right\}\right] + \frac{\lambda_n}{8} \\
&\leq \inf_{q\in\mathcal{Q}}\left\{\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim q}\left[\widehat{L}(\theta,\mathcal{S})\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n n}\right]\right\} + \frac{\lambda_n}{8} \\
&= \inf_{q\in\mathcal{Q}}\left\{\mathbb{E}_{\theta\sim q}[L(\theta)] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n n}\right\} + \frac{\lambda_n}{8}.
\end{aligned}$$

- Problem: we do not necessarily have $\pi_r(\mathrm{d}\theta) \propto \mathbb{1}(\theta\in\mathcal{B}_r)\cdot\pi(\mathrm{d}\theta) \subset \mathcal{Q}$...
- Question: how can we make

$$\inf_{q\in\mathcal{Q}}\left\{\mathbb{E}_{\theta\sim q}[L(\theta)] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n n}\right\} \leq L(\theta_\star) + \widetilde{\mathcal{O}}\left(\frac{1}{\lambda_n n}\right) \quad ??$$

Answer: assume it explicitly!

PAC-Bayes derivation of rates for the approximations (2)

Central requirement: $$\inf_{q \in \mathcal{Q}} \left\{ \mathbb{E}_{\theta \sim q} \left[L(\theta) \right] + \frac{\mathrm{KL}(q \| \pi)}{\lambda_n \, n} \right\} \leq L(\theta_\star) + \widetilde{\mathcal{O}} \left(\frac{1}{\lambda_n \, n} \right) \quad ??$$

PAC-Bayes derivation of rates for the approximations (2)

Central requirement: $\inf_{q\in\mathcal{Q}}\left\{\mathbb{E}_{\theta\sim q}\left[L(\theta)\right]+\frac{\mathrm{KL}(q\|\pi)}{\lambda_n\, n}\right\}\leq L(\theta_\star)+\widetilde{\mathcal{O}}\left(\frac{1}{\lambda_n\, n}\right)$??

The *extended* prior mass condition: there exists a sequence of distributions $q_n\in\mathcal{Q}$ such that

$$\mathbb{E}_{\theta\sim q_n}\left[L(\theta)\right]\leq L(\theta_\star)+\widetilde{\mathcal{O}}\left(\frac{1}{\lambda_n\, n}\right)\quad\text{and}\quad \mathrm{KL}(q_n\|\pi)\leq\widetilde{\mathcal{O}}\left(1\right).$$

PAC-Bayes derivation of rates for the approximations (2)

Central requirement: $$\inf_{q\in\mathcal{Q}}\left\{\mathbb{E}_{\theta\sim q}\left[L(\theta)\right]+\frac{\mathrm{KL}(q\|\pi)}{\lambda_n\, n}\right\}\leq L(\theta_\star)+\widetilde{\mathcal{O}}\left(\frac{1}{\lambda_n\, n}\right)\quad ??$$

The *extended* prior mass condition: there exists a sequence of distributions $q_n\in\mathcal{Q}$ such that

$$\mathbb{E}_{\theta\sim q_n}\left[L(\theta)\right]\leq L(\theta_\star)+\widetilde{\mathcal{O}}\left(\frac{1}{\lambda_n\, n}\right)\quad\text{and}\quad \mathrm{KL}(q_n\|\pi)\leq\widetilde{\mathcal{O}}\left(1\right).$$

Informal: The variational approximation of the Gibbs Posterior concentrates toward $\theta_\star$ (w.r.t. the excess loss) at the exact same rate as the Gibbs

$$\varepsilon_n=\lambda_n+\frac{1}{\lambda_n\, n}$$

for a bounded loss as soon as the extended prior mass condition is satisfied.

PAC-Bayes derivation of rates for the approximations (2)

Central requirement: $\inf_{q \in \mathcal{Q}} \left\{ \mathbb{E}_{\theta \sim q}[L(\theta)] + \frac{\mathrm{KL}(q \| \pi)}{\lambda_n n} \right\} \leq L(\theta_\star) + \widetilde{\mathcal{O}}\left(\frac{1}{\lambda_n n}\right)$??

The *extended* prior mass condition: there exists a sequence of distributions $q_n \in \mathcal{Q}$ such that

$$\mathbb{E}_{\theta \sim q_n}[L(\theta)] \leq L(\theta_\star) + \widetilde{\mathcal{O}}\left(\frac{1}{\lambda_n n}\right) \quad \text{and} \quad \mathrm{KL}(q_n \| \pi) \leq \widetilde{\mathcal{O}}(1) .$$

Informal: The variational approximation of the Gibbs Posterior concentrates toward $\theta_\star$ (w.r.t. the excess loss) at the exact same rate as the Gibbs

$$\varepsilon_n = \lambda_n + \frac{1}{\lambda_n n}$$

for a bounded loss as soon as the extended prior mass condition is satisfied.

Main question: is the extended prior mass condition realistic?

PAC-Bayes derivation of rates for the approximations (2)

Central requirement: $\inf_{q \in \mathcal{Q}} \left\{ \mathbb{E}_{\theta \sim q} [L(\theta)] + \frac{\mathrm{KL}(q \| \pi)}{\lambda_n n} \right\} \leq L(\theta_\star) + \widetilde{\mathcal{O}}\left(\frac{1}{\lambda_n n}\right)$??

The *extended* prior mass condition: there exists a sequence of distributions $q_n \in \mathcal{Q}$ such that

$$\mathbb{E}_{\theta \sim q_n} [L(\theta)] \leq L(\theta_\star) + \widetilde{\mathcal{O}}\left(\frac{1}{\lambda_n n}\right) \quad \text{and} \quad \mathrm{KL}(q_n \| \pi) \leq \widetilde{\mathcal{O}}(1) \ .$$

Informal: The variational approximation of the Gibbs Posterior concentrates toward $\theta_\star$ (w.r.t. the excess loss) at the exact same rate as the Gibbs

$$\varepsilon_n = \lambda_n + \frac{1}{\lambda_n n}$$

for a bounded loss as soon as the extended prior mass condition is satisfied.

Main question: is the extended prior mass condition realistic? Yes, quite often!

The extended prior mass condition in practice

The *extended* prior mass condition: there exists a sequence of distributions $q_n \in \mathcal{Q}$ such that

$$\mathbb{E}_{\theta \sim q_n}[L(\theta)] \leq L(\theta_\star) + \widetilde{\mathcal{O}}\left(\frac{1}{\lambda_n n}\right) \quad \text{and} \quad \mathsf{KL}(q_n \| \pi) \leq \widetilde{\mathcal{O}}(1) .$$

Does the standard prior mass condition imply the extended one?

The extended prior mass condition in practice

The *extended* prior mass condition: there exists a sequence of distributions $q_n \in \mathcal{Q}$ such that

$$\mathbb{E}_{\theta \sim q_n}[L(\theta)] \leq L(\theta_\star) + \widetilde{\mathcal{O}}\left(\frac{1}{\lambda_n n}\right) \quad \text{and} \quad \mathrm{KL}(q_n \| \pi) \leq \widetilde{\mathcal{O}}(1) .$$

Does the standard prior mass condition imply the extended one?

- When $\mathcal{Q} = \mathcal{P}(\Theta)$, yes: Simply take $q_n = \pi_r$ with $r = d/\lambda_n n$.

The extended prior mass condition in practice

The *extended* prior mass condition: there exists a sequence of distributions $q_n \in \mathcal{Q}$ such that

$$\mathbb{E}_{\theta \sim q_n}[L(\theta)] \leq L(\theta_\star) + \widetilde{\mathcal{O}}\left(\frac{1}{\lambda_n n}\right) \quad \text{and} \quad \mathrm{KL}(q_n \| \pi) \leq \widetilde{\mathcal{O}}(1) \ .$$

Does the standard prior mass condition imply the extended one?

- When $\mathcal{Q} = \mathcal{P}(\Theta)$, yes: Simply take $q_n = \pi_r$ with $r = d/\lambda_n n$.
- When $\mathcal{Q} = \{\mathcal{N}(m, \sigma^2 I_p) : m \in \mathbb{R}^p, \sigma^2 > 0\}$ and the loss is Lipschitz in θ, yes: Simply take $q_n = \mathcal{N}(\theta_\star, \sigma_n^2 I_p)$ with $\sigma_n = 1/\lambda_n n$.

The extended prior mass condition in practice

The *extended* prior mass condition: there exists a sequence of distributions $q_n \in \mathcal{Q}$ such that

$$\mathbb{E}_{\theta \sim q_n}[L(\theta)] \leq L(\theta_\star) + \widetilde{\mathcal{O}}\left(\frac{1}{\lambda_n n}\right) \quad \text{and} \quad \mathrm{KL}(q_n \| \pi) \leq \widetilde{\mathcal{O}}(1) .$$

Does the standard prior mass condition imply the extended one?

- When $\mathcal{Q} = \mathcal{P}(\Theta)$, yes: Simply take $q_n = \pi_r$ with $r = d/\lambda_n n$.
- When $\mathcal{Q} = \{\mathcal{N}(m, \sigma^2 I_p) : m \in \mathbb{R}^p, \sigma^2 > 0\}$ and the loss is Lipschitz in θ, yes: Simply take $q_n = \mathcal{N}(\theta_\star, \sigma_n^2 I_p)$ with $\sigma_n = 1/\lambda_n n$.

In some sense, when $\mathcal{Q} \subsetneq \mathcal{P}(\Theta)$, just approximate the choice

$$q_n \propto \mathbb{1}\left(L(\theta) \leq L(\theta_\star) + \frac{1}{\lambda_n n}\right) \cdot \pi(\mathrm{d}\theta) \quad \text{by} \quad q_n = \mathcal{N}\left(\theta_\star, \frac{1}{(\lambda_n n)^2} I_p\right)$$

given some additional smoothness structure.

The extended prior mass condition in practice

The *extended* prior mass condition: there exists a sequence of distributions $q_n \in \mathcal{Q}$ such that

$$\mathbb{E}_{\theta \sim q_n}[L(\theta)] \leq L(\theta_\star) + \widetilde{\mathcal{O}}\left(\frac{1}{\lambda_n n}\right) \quad \text{and} \quad \mathrm{KL}(q_n \| \pi) \leq \widetilde{\mathcal{O}}(1) .$$

Does the standard prior mass condition imply the extended one?

- When $\mathcal{Q} = \mathcal{P}(\Theta)$, yes: Simply take $q_n = \pi_r$ with $r = d/\lambda_n n$.
- When $\mathcal{Q} = \{\mathcal{N}(m, \sigma^2 I_p) : m \in \mathbb{R}^p, \sigma^2 > 0\}$ and the loss is Lipschitz in θ, yes: Simply take $q_n = \mathcal{N}(\theta_\star, \sigma_n^2 I_p)$ with $\sigma_n = 1/\lambda_n n$.

In some sense, when $\mathcal{Q} \subsetneq \mathcal{P}(\Theta)$, just approximate the choice

$$q_n \propto \mathbb{1}\left(L(\theta) \leq L(\theta_\star) + \frac{1}{\lambda_n n}\right) \cdot \pi(\mathrm{d}\theta) \quad \text{by} \quad q_n = \mathcal{N}\left(\theta_\star, \frac{1}{(\lambda_n n)^2} I_p\right)$$

given some additional smoothness structure.

Takeaway: concentration rates of Gibbs posteriors are usually still valid for their variational approximations, provided structural additional conditions.

Failure in statistical modeling $\ell(x, \theta) = -\log p_\theta(x)$

The Gibbs posterior concentrates w.r.t. $d(\theta, \theta_\star) = \mathrm{KL}(P_{\theta_\star} \| P_\theta)$ at rate:

$$\varepsilon_n = \lambda_n + \frac{1}{\lambda_n \, n}$$

provided the prior mass condition when $p_\theta(x)$ is lower bounded.

Failure in statistical modeling $\ell(x,\theta) = -\log p_\theta(x)$

The Gibbs posterior concentrates w.r.t. $d(\theta,\theta_\star) = \mathrm{KL}(P_{\theta_\star}\|P_\theta)$ at rate:

$$\varepsilon_n = \lambda_n + \frac{1}{\lambda_n\, n}$$

provided the prior mass condition when $p_\theta(x)$ is lower bounded. But:

1. The boundedness assumption is relevant only in the learning framework.
2. No concentration guarantee for the Bayes posterior for which $\lambda_n = 1$.
3. The optimal choice of $\lambda_n \propto n^{-1/2}$ prevents from achieving rate n^{-1} with respect to the squared Euclidean distance in regular parametric models.

Failure in statistical modeling $\ell(x,\theta) = -\log p_\theta(x)$

The Gibbs posterior concentrates w.r.t. $d(\theta,\theta_\star) = \mathrm{KL}(P_{\theta_\star}\|P_\theta)$ at rate:

$$\varepsilon_n = \lambda_n + \frac{1}{\lambda_n\, n}$$

provided the prior mass condition when $p_\theta(x)$ is lower bounded. But:

1. The boundedness assumption is relevant only in the learning framework.
2. No concentration guarantee for the Bayes posterior for which $\lambda_n = 1$.
3. The optimal choice of $\lambda_n \propto n^{-1/2}$ prevents from achieving rate n^{-1} with respect to the squared Euclidean distance in regular parametric models.

However, the Bayes posterior concentrates! What's known from the literature:

1. The Bayes posterior ($\lambda_n = 1$) concentrates w.r.t. $\mathcal{H}^2(P_{\theta_\star}\|P_\theta)$ at rate

   $$\varepsilon_n = \frac{1}{n}$$

   provided the prior mass condition + test conditions (GGV, AoS 2000).
2. The tempered posterior ($\lambda_n < 1$) concentrates w.r.t. $\mathrm{R}_{\lambda_n}(P_\theta\|P_{\theta_\star})$ at rate

   $$\varepsilon_n = \frac{\lambda_n}{(1-\lambda_n)n}$$

   provided the prior mass condition alone (BPY, AoS 2019).

A PAC-Bayes bound for model-based posteriors

The previous analysis on the Gibbs posterior fails for the tempered and Bayes posteriors is that they are based on Catoni's bound, which is vacuous if $\lambda_n \nrightarrow 0$:

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim q}\left[\mathsf{KL}(P_{\theta_\star}\|P_\theta)\right] \leq \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim q}\left[\widehat{\mathsf{KL}}(P_{\theta_\star}\|P_\theta)\right] + \frac{\mathsf{KL}(q\|\pi)}{\lambda_n\, n}\right] + \frac{\lambda_n}{8}\,.$$

A PAC-Bayes bound for model-based posteriors

The previous analysis on the Gibbs posterior fails for the tempered and Bayes posteriors is that they are based on Catoni's bound, which is vacuous if $\lambda_n \nrightarrow 0$:

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim q}\left[\mathrm{KL}(P_{\theta_\star}\|P_\theta)\right] \leq \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim q}\left[\widehat{\mathrm{KL}}(P_{\theta_\star}\|P_\theta)\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n\, n}\right] + \frac{\lambda_n}{8}\,.$$

Question: are the tempered and Bayes posteriors minimizers of non-vacuous PAC-Bayes bounds?

1. For the tempered posterior ($\lambda_n < 1$): Yes! Based on:

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim q}\left[\mathrm{R}_{\lambda_n}(P_\theta\|P_{\theta_\star})\right] \leq \frac{\lambda_n}{1-\lambda_n}\cdot\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim q}\left[\widehat{\mathrm{KL}}(P_{\theta_\star}\|P_\theta)\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n\, n}\right]\,.$$

(BPY, AoS 2019) derived concentration based on this bound, extended to variational approximations by (YPB, AoS 2020) and (AR, AoS 2020).

A PAC-Bayes bound for model-based posteriors

The previous analysis on the Gibbs posterior fails for the tempered and Bayes posteriors is that they are based on Catoni's bound, which is vacuous if $\lambda_n \nrightarrow 0$:

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim q}\left[\mathrm{KL}(P_{\theta_\star}\|P_\theta)\right] \leq \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim q}\left[\widehat{\mathrm{KL}}(P_{\theta_\star}\|P_\theta)\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n\, n}\right] + \frac{\lambda_n}{8}\,.$$

Question: are the tempered and Bayes posteriors minimizers of non-vacuous PAC-Bayes bounds?

1. For the tempered posterior ($\lambda_n < 1$): Yes! Based on:

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim q}\left[\mathrm{R}_{\lambda_n}(P_\theta\|P_{\theta_\star})\right] \leq \frac{\lambda_n}{1-\lambda_n}\cdot\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim q}\left[\widehat{\mathrm{KL}}(P_{\theta_\star}\|P_\theta)\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n\, n}\right]\,.$$

   (BPY, AoS 2019) derived concentration based on this bound, extended to variational approximations by (YPB, AoS 2020) and (AR, AoS 2020).

2. For the Bayes posterior ($\lambda_n = 1$): No...

A PAC-Bayes bound for model-based posteriors

The previous analysis on the Gibbs posterior fails for the tempered and Bayes posteriors is that they are based on Catoni's bound, which is vacuous if $\lambda_n \nrightarrow 0$:

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim q}\left[\mathrm{KL}(P_{\theta_\star}\|P_\theta)\right] \le \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim q}\left[\widehat{\mathrm{KL}}(P_{\theta_\star}\|P_\theta)\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n\, n}\right] + \frac{\lambda_n}{8}\,.$$

Question: are the tempered and Bayes posteriors minimizers of non-vacuous PAC-Bayes bounds?

1. For the tempered posterior ($\lambda_n < 1$): Yes! Based on:

   $$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim q}\left[\mathrm{R}_{\lambda_n}(P_\theta\|P_{\theta_\star})\right] \le \frac{\lambda_n}{1-\lambda_n}\cdot\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim q}\left[\widehat{\mathrm{KL}}(P_{\theta_\star}\|P_\theta)\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n\, n}\right]\,.$$

   (BPY, AoS 2019) derived concentration based on this bound, extended to variational approximations by (YPB, AoS 2020) and (AR, AoS 2020).

2. For the Bayes posterior ($\lambda_n = 1$): No...

Question: if concentration of the Bayes posterior cannot be obtained from our route, is it possible to derive concentration for its variational approximation?

A PAC-Bayes bound for model-based posteriors

The previous analysis on the Gibbs posterior fails for the tempered and Bayes posteriors is that they are based on Catoni's bound, which is vacuous if $\lambda_n \nrightarrow 0$:

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim q}\left[\mathrm{KL}(P_{\theta_\star}\|P_\theta)\right] \leq \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim q}\left[\widehat{\mathrm{KL}}(P_{\theta_\star}\|P_\theta)\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n\, n}\right] + \frac{\lambda_n}{8}\,.$$

Question: are the tempered and Bayes posteriors minimizers of non-vacuous PAC-Bayes bounds?

1. For the tempered posterior ($\lambda_n < 1$): Yes! Based on:

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim q}\left[\mathrm{R}_{\lambda_n}(P_\theta\|P_{\theta_\star})\right] \leq \frac{\lambda_n}{1-\lambda_n}\cdot\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim q}\left[\widehat{\mathrm{KL}}(P_{\theta_\star}\|P_\theta)\right] + \frac{\mathrm{KL}(q\|\pi)}{\lambda_n\, n}\right]\,.$$

   (BPY, AoS 2019) derived concentration based on this bound, extended to variational approximations by (YPB, AoS 2020) and (AR, AoS 2020).

2. For the Bayes posterior ($\lambda_n = 1$): No...

Question: if concentration of the Bayes posterior cannot be obtained from our route, is it possible to derive concentration for its variational approximation?

Yes, see (ZG, AoS 2020) who relied on the **approximate Bayes** nature of VI. Different proof, but PAC-Bayes change-of-measure inequalities remain the key!

Generalize vanilla VI in two directions

Vanilla Variational inference

$$\widetilde{\pi}(\cdot \mid \mathcal{S}) = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}\Big(q \,\Big\|\, \pi(\cdot \mid \mathcal{S})\Big)$$
$$= \arg\min_{q \in \mathcal{Q}} \left\{ \mathbb{E}_{\theta \sim q}\left[\widehat{L}(\theta, \mathcal{S})\right] + \frac{\mathrm{KL}(q \| \pi)}{\lambda_n\, n} \right\}.$$

can be extended in two directions:

1. Generalized Variational Inference:

$$\widetilde{\pi}(\cdot \mid \mathcal{S}) = \arg\min_{q \in \mathcal{Q}} \left\{ \mathbb{E}_{\theta \sim q}\left[\widehat{L}(\theta, \mathcal{S})\right] + \frac{\mathrm{D}(q \| \pi)}{\lambda_n\, n} \right\}.$$

2. Discrepancy Variational Inference:

$$\widetilde{\pi}(\cdot \mid \mathcal{S}) = \arg\min_{q \in \mathcal{Q}} \mathrm{D}\Big(q \,\Big\|\, \pi(\cdot \mid \mathcal{S})\Big).$$

Generalize vanilla VI in two directions

Vanilla Variational inference

$$\widetilde{\pi}(\cdot \mid \mathcal{S}) = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}\Big(q \,\Big\|\, \pi(\cdot \mid \mathcal{S})\Big)$$

$$= \arg\min_{q \in \mathcal{Q}} \left\{ \mathbb{E}_{\theta \sim q} \left[\widehat{L}(\theta, \mathcal{S}) \right] + \frac{\mathrm{KL}(q \| \pi)}{\lambda_n \, n} \right\} .$$

can be extended in two directions:

1. Generalized Variational Inference:

$$\widetilde{\pi}(\cdot \mid \mathcal{S}) = \arg\min_{q \in \mathcal{Q}} \left\{ \mathbb{E}_{\theta \sim q} \left[\widehat{L}(\theta, \mathcal{S}) \right] + \frac{\mathrm{D}(q \| \pi)}{\lambda_n \, n} \right\} .$$

2. Discrepancy Variational Inference:

$$\widetilde{\pi}(\cdot \mid \mathcal{S}) = \arg\min_{q \in \mathcal{Q}} \mathrm{D}\Big(q \,\Big\|\, \pi(\cdot \mid \mathcal{S})\Big) .$$

There is a need to develop PAC-Bayes theory to understand alternative choices of divergences, see e.g. (AG, ML 2018; A, ICML 2021).

Generalize vanilla VI in two directions

Vanilla Variational inference

$$\widetilde{\pi}(\cdot \mid \mathcal{S}) = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}\Big(q \,\Big\|\, \pi(\cdot \mid \mathcal{S})\Big)$$

$$= \arg\min_{q \in \mathcal{Q}} \left\{ \mathbb{E}_{\theta \sim q} \left[\widehat{L}(\theta, \mathcal{S}) \right] + \frac{\mathrm{KL}(q \| \pi)}{\lambda_n \, n} \right\} .$$

can be extended in two directions:

1. Generalized Variational Inference:

$$\widetilde{\pi}(\cdot \mid \mathcal{S}) = \arg\min_{q \in \mathcal{Q}} \left\{ \mathbb{E}_{\theta \sim q} \left[\widehat{L}(\theta, \mathcal{S}) \right] + \frac{\mathrm{D}(q \| \pi)}{\lambda_n \, n} \right\} .$$

2. Discrepancy Variational Inference:

$$\widetilde{\pi}(\cdot \mid \mathcal{S}) = \arg\min_{q \in \mathcal{Q}} \mathrm{D}\Big(q \,\Big\|\, \pi(\cdot \mid \mathcal{S})\Big) .$$

There is a need to develop PAC-Bayes theory to understand alternative choices of divergences, see e.g. (AG, ML 2018; A, ICML 2021).

We can also discuss the role of the empirical loss $\widehat{L}(\theta, \mathcal{S})$...

Discussion

Still many things to be discussed/discovered:

- On the tightness of the bounds: in fact, VI can act as regularization and even lead to faster rates, which cannot be established using PAC-Bayes.
- A unifying picture of VI is still missing: is there a bound to rule them all?
- PAC-Bayes theory is a very active field, but the connection between empirical bounds and theoretical guarantees is still overlooked.
- What is the impact of the discrepancy in generalized/discrepancy variational inference on the concentration rate of the variational posterior?
- Is it possible to improve the rates by using a localization argument?
- Does there exist a finer analysis of each (of the possibly many) minimizer?
- How can PAC-Bayes be used to analyze gradient algorithms?
- How about the role of PAC-Bayes in uncertainty quantification?
- Beyond the large-sample theory: is it possible to evaluate the behavior of such objects in overparameterized regimes?
- ...

Main references

- (GGV, AoS 2000): S. Ghosal, J.K. Ghosh, A.W. Van Der Vaart. *Convergence rates of posterior distributions*. **The Annals of Statistics** 2000.
- (C, 2003): O. Catoni. *A PAC-Bayesian approach to adaptive classification*. Preprint LPMA 2003.
- (Z, 2006): T. Zhang. *Information-theoretic upper and lower bounds for statistical estimation*. **IEEE Transactions on Information Theory** 2006.
- (ARC, JMLR 2016): P. Alquier, J. Ridgway & N. Chopin, *On the Properties of Variational Approximations of Gibbs Posteriors*. **The Journal of Machine Learning Research** 2016.
- (AG, ML 2018): P. Alquier, B. Guedj. *Simpler PAC-Bayesian Bounds for Hostile Data*. **Machine Learning** 2018.
- (BPY, AoS 2019): A. Bhattacharya, D. Pati & Y. Yang. *Bayesian fractional posteriors*. **The Annals of Statistics** 2019.
- (AR, AoS 2020): P. Alquier & J. Ridgway. *Concentration of tempered posteriors and of their variational approximations*. **The Annals of Statistics** 2020.
- (YPB, AoS 2020): Y. Yang, D. Pati & A. Bhattacharya. *α-variational inference with statistical guarantees*. **The Annals of Statistics** 2020.
- (ZG, AoS 2020): F. Zhang & C. Gao. *Convergence rates of variational posteriors*. **The Annals of Statistics** 2020.
- (A, AoS 2021): P. Alquier. *Non-exponentially Weighted Aggregation: Regret Bounds for Unbounded Loss Functions*. **The International Conference on Machine Learning** 2021.