Rates of convergence in Bayesian meta-learning

Pierre Alquier

ESSEC BUSINESS SCHOOL

12th Workshop on High Dimensional Data Analysis
ESSEC and University Mohamed V
Rabat, Morocco

Charles RIOU

University of Tokyo,

RIKEN AIP

Badr-Eddine CHÉRIEF-ABDELLATIF

CNRS,

Sorbonne Université

Riou, C., Alquier, P. and Chérief-Abdellatif, B.-E. (2023). Bayes meets Bernstein at the Meta Level : an Analysis of Fast Rates in Meta-Learning with PAC-Bayes. Preprint arXiv :2302.11709.

$p(x|\theta)$

$\pi(\theta)$ $\quad \mathcal{S} = (x_1, \dots, x_n)$

$\pi(\theta|x_1, \dots, x_n)$

$\propto \pi(\theta) \prod_{i=1}^n p(x_i|\theta)$

$\ell(x, \theta)$

$\pi(\theta)$ $\quad \mathcal{S} = (x_1, \dots, x_n)$

α

$\rho(\theta)$

$\propto \pi(\theta) \mathrm{e}^{-\alpha \sum_{i=1}^n \ell(x_i, \theta)}$

To keep the results as simple as possible :

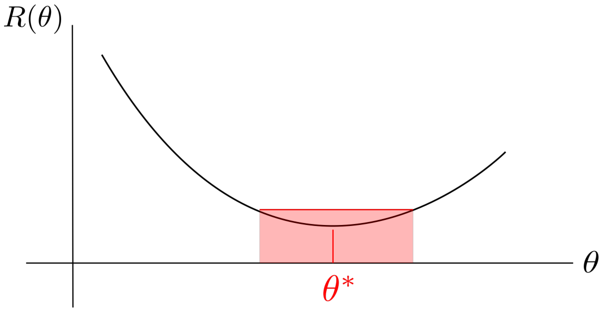
- $\mathcal{S} = (X_1, \dots, X_n)$ i.i.d. from P,
- $\ell(x, \theta)$ bounded by 1.

- Generalization risk : $R(\theta) = \mathbb{E}_{X \sim P}[\ell(X, \theta)]$.
- Objective : $\theta^* = \arg\min_{\theta \in \Theta} R(\theta)$.
- Risk of the "Bayes" procedure ρ : $\mathbb{E}_{\theta \sim \rho}[R(\theta)]$.

Theorem (stated informally)

$$\mathbb{E}_{\mathcal{S}}\left\{\mathbb{E}_{\theta \sim \rho}[R(\theta)]\right\} \leq R(\theta^*) + c\sqrt{\frac{d \log(n)}{n}}$$

where $d = d(P, \pi)$ defined in the next slide, for α well chosen.



$$N(\theta^*, s) := \{\theta \in \Theta : R(\theta) - R(\theta^*) \leq s\}.$$

$d = d(P, \pi)$ is the smallest number such that, for any s small enough :

$$\pi(N(\theta^*, s)) \geq s^d.$$

How do we prove the theorem ?

$$\rho(\theta) \propto \pi(\theta) \mathrm{e}^{-\alpha \sum_{i=1}^{n} \ell(x_i,\theta)}$$
$$= \arg\min_{p \in \mathcal{P}rob(\Theta)} \left\{ \mathbb{E}_{\theta \sim p} \left[\frac{1}{n} \sum_{i=1}^{n} \ell(x_i, \theta) \right] + \frac{KL(p,\pi)}{\alpha n} \right\}.$$

PAC-Bayes / Information bounds

$$\mathbb{E}_{\mathcal{S}} \left\{ \mathbb{E}_{\theta\sim\rho}[R(\theta)] \right\} \leq \inf_{p} \left\{ \mathbb{E}_{\theta\sim p}[R(\theta)] + \alpha + \frac{KL(p,\pi)}{\alpha n} \right\}.$$

In particular, for p as the restriction of π to $N(\theta^*, s)$,

$$\mathbb{E}_{\mathcal{S}} \left\{ \mathbb{E}_{\theta\sim\rho}[R(\theta)] \right\} \leq \inf_{s>0} \left\{ R(\theta^*) + s + \alpha + \frac{d \log \frac{1}{s}}{\alpha n} \right\}.$$

- Old result : in a "noiseless setting", when there is a θ such that $\ell(x,\theta)=0$ almost surely for $x\sim P$,

$$\mathbb{E}_{\mathcal{S}}\left\{\mathbb{E}_{\theta\sim\rho}[R(\theta)]\right\} \leq \underbrace{R(\theta^*)}_{=0} + c\frac{d\log(n)}{n}.$$

- Similar fast rates obtained in classification under Mammen and Tsybakov margin assumption (1999).
- Also with Lipschitz and strongly convex losses $\ell(x,\cdot)$ by Bartlett and Mendelson (2006).

All these assumptions turned out to be a special case of :

Bernstein condition

$$\mathbb{E}_{x\sim P}\left\{\left[\ell(x,\theta)-\ell(x,\theta^*)\right]^2\right\} \leq C\left[R(\theta)-R(\theta^*)\right].$$

What about variational Bayes ?

Let $\mathcal{W}$ be a subset of $\mathcal{P}rob(\Theta)$, and put :

$$\rho^{\mathcal{W}}(\theta) = \arg\min_{p \in \cancel{\mathcal{P}rob(\Theta)}\,\mathcal{W}} \left\{ \mathbb{E}_{\theta\sim p}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(x_i,\theta)\right] + \frac{KL(p,\pi)}{\alpha n} \right\}.$$

P. Alquier, J. Ridgway , N. Chopin (2016). On the Properties of Variational Approximations of Gibbs Posteriors. JMLR.

provides minimal assumptions on $\mathcal{W}$ ensuring

$$\mathbb{E}_{\mathcal{S}}\left\{\mathbb{E}_{\theta\sim\rho^{\mathcal{W}}}[R(\theta)]\right\} \leq R(\theta^*) + c\left(\frac{d(P,\pi)\log(n)}{n}\right)^{\beta}$$

where $\beta = 1$ under Bernstein condition, $\beta = 1/2$ otherwise.

Recap

$$\rho(\theta) \propto \pi(\theta) \mathrm{e}^{-\alpha \sum_{i=1}^{n} \ell(x_i,\theta)}.$$

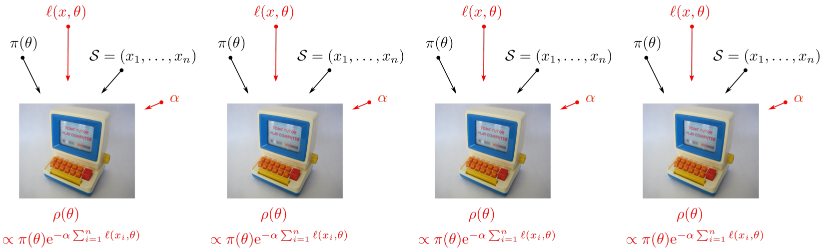
We have :

$$\mathbb{E}_{\mathcal{S}} \left\{ \mathbb{E}_{\theta\sim\rho}[R(\theta)] \right\} \leq R(\theta^*) + c \left(\frac{d \log(n)}{n} \right)^{\beta}$$

where $\beta = 1$ under Bernstein condition, $\beta = 1/2$ otherwise.

- The generalization error is driven by $d = d(P, \pi)$ that depends on π.
- Tempting to learn a better π, but π is not allowed to depend on the data...

Idea of Bayesian meta-learning :



- We solve many related tasks (say T) using Bayesian learning.
- By related, we mean that the same prior could be used in all tasks.
- Based on past tasks, can we define a π that would work better for future tasks ?

Notations :

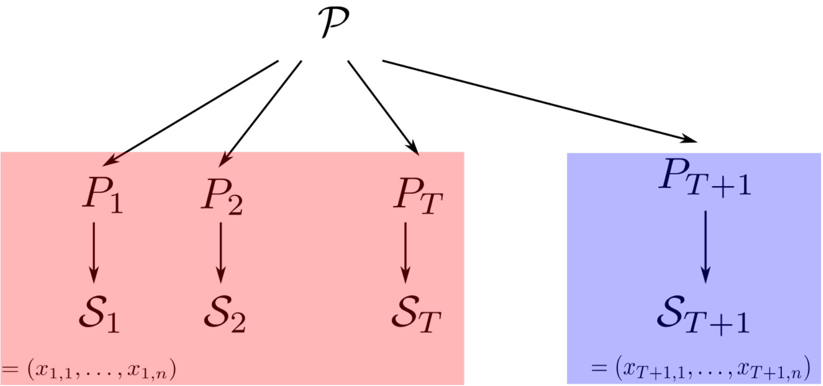
- Tasks : $t = 1, \dots, T$.
- $P_1, \dots, P_T$ are i.i.d from $\mathcal{P}$.
- Task t : $\mathcal{S}_t = (x_{t,1}, \dots, x_{t,n})$ i.i.d from P_t.
- Generalization error in task t : $R_t(\theta) = \mathbb{E}_{x \sim P_t}[\ell(x, \theta)]$.
- Best error in task t : $R_t(\theta_t^*) = \min_\theta R_t(\theta)$.
- $\rho_t(\pi, \alpha)(\theta) \propto \pi(\theta) \exp[-\alpha \sum_{i=1}^n \ell(\theta, x_{t,i})]$.

Objective

- Learn $\hat{\pi} = \hat{\pi}(\mathcal{S}_1, \dots, \mathcal{S}_T)$.
- For a new task $P_{T+1} \sim \mathcal{P}$, $\mathcal{S}_{T+1} = (x_{T+1,1}, \dots, x_{T+1,n})$ i.i.d. from P_{T+1}, we want :

$$\mathbb{E}_{\theta \sim \rho_{T+1}(\hat{\pi}, \alpha)} \left[R_{T+1}(\theta) \right] \leq \mathbb{E}_{\theta \sim \rho_{T+1}(\pi, \alpha)} \left[R_{T+1}(\theta) \right].$$

$$
\begin{tikzcd}
 & & \mathcal{P} \arrow[dll] \arrow[dl] \arrow[dr] \arrow[drr] & & \\
P_1 \arrow[d] & P_2 \arrow[d] & & P_T \arrow[d] & P_{T+1} \arrow[d] \\
\mathcal{S}_1 & \mathcal{S}_2 & & \mathcal{S}_T & \mathcal{S}_{T+1} \\
= (x_{1,1}, \dots, x_{1,n}) & & & & = (x_{T+1,1}, \dots, x_{T+1,n})
\end{tikzcd}
$$



- Past tasks, used to learn a better prior. Expectation with respect to $P_1, \dots, P_T, \mathcal{S}_1, \dots, \mathcal{S}_T$ denoted by $\mathbb{E}_{\rm data}$.
- New task. Expectation with respect to P_{T+1} and $\mathcal{S}_{T+1}$ will be denoted by $\mathbb{E}_{\rm new}$.

Ultimate (non-achievable) performance :

$$\mathcal{E}^* = \mathbb{E}_{\rm new}[R_{T+1}(\theta^*_{T+1})] = \mathbb{E}_{P_{T+1}\sim\mathcal{P}}[R_{T+1}(\theta^*_{T+1})].$$

With a fixed prior :

$$\mathcal{E}(\pi) = \mathbb{E}_{\rm new}\left\{\mathbb{E}_{\theta\sim\rho_{T+1}(\pi,\alpha)}[R_{T+1}(\theta)]\right\}$$
$$\leq \mathcal{E}^* + c\,\mathbb{E}_{P_{T+1}\sim\mathcal{P}}\left[\left(\frac{d(P_{T+1},\pi)\log(n)}{n}\right)^{\beta}\right].$$

To give an overview of our results, let us consider first an easy situation : we want to find the best of K priors, say

$$\pi_1,\dots,\pi_K.$$

Recall :

$$\rho_t(\pi,\alpha) = \arg\min_p \Big\{ \underbrace{\mathbb{E}_{\theta\sim p}\left[\frac{1}{n}\sum_{i=1}^{n} \ell(x_{t,i},\theta)\right] + \frac{KL(p,\pi)}{\alpha n}}_{\hat{\mathcal{R}}_t(p,\pi)} \Big\}.$$

In this case, our procedure boils down to :

$$\hat{\pi} = \arg\min_{\pi\in\{\pi_1,\dots,\pi_K\}} \left\{ \frac{1}{T}\sum_{t=1}^{T} \hat{\mathcal{R}}_t\Big[\rho_t(\pi,\alpha),\pi\Big] \right\}.$$

Theorem

$$\mathbb{E}_{\text{data}}[\mathcal{E}(\hat{\pi})] \leq \min_{k=1,\dots,K} \mathcal{E}(\pi_k) + \frac{\log K}{T} \leq \mathcal{E}^* + c \min_{k=1,\dots,K} \mathbb{E}_{P_{T+1}\sim\mathcal{P}}\left[\left(\frac{d(P_{T+1},\pi_k)\log(n)}{n}\right)^{\beta}\right] + \frac{\log K}{T}.$$

Important observations :

- gain expected only if $T \gg n$.
- the rate for learning the prior is in $1/T$ regardless or the rate within tasks ($\beta = 1$ or $\beta = 1/2$).

- More generally, we can learn the best prior in an infinite set $\mathcal{Q}$ (for example, all Gaussian priors, etc).
- The definition of $\hat{\pi}$ gets a little more convoluted.
- We will recover similar results

$$\mathbb{E}_{\rm data}[\mathcal{E}(\hat{\pi})] \leq \min_{\pi \in \mathcal{Q}} \mathcal{E}(\pi) + \frac{\mathcal{C}(\mathcal{Q})}{T}$$

where $\mathcal{C}(\mathcal{Q})$ is a complexity measure of $\mathcal{Q}$.

Example 1 : Gaussian priors.

- $\theta \in \mathbb{R}^p$.
- $\mathcal{Q} = \{\mathcal{N}(\mu, \Sigma), \mu \in \mathbb{R}^p, \Sigma \in \mathcal{S}_+^p\}$.
- fix some m and put $V = \mathbb{E}_{\text{new}}\left[\|\theta^*_{T+1} - m\|^2\right]$.

Very approximatevely,

$$\mathbb{E}_{\text{data}}[\mathcal{E}(\hat{\pi})] \leq \mathcal{E}^* + \frac{p}{T} + \begin{cases} \frac{V + p\log n}{n} & \text{if } V > \frac{n}{T} \\ 0 & \text{otherwise} \end{cases}.$$

Example 2 : mixture of Gaussian priors.

- $\theta \in \mathbb{R}^p$.
- $\mathcal{Q} = \left\{ \sum_{k=1}^{K} p_k \mathcal{N}(\mu_k, \Sigma_k) \right\}$.
- fix $m_1, \dots, m_K$ and put $V = \mathbb{E}_{\text{new}} \left[\min_k \left\| \theta^*_{T+1} - m_k \right\|^2 \right]$.

$$\mathbb{E}_{\text{data}}[\mathcal{E}(\hat{\pi})] \leq \mathcal{E}^* + \frac{pK}{T} + \frac{\log K}{n} + \begin{cases} \frac{K(V + p \log n)}{n} & \text{if } V > \frac{n}{T} \\ 0 & \text{otherwise} . \end{cases}$$

The general procedure $\hat{\pi} = \hat{\pi}(\mathcal{S}_1, \dots, \mathcal{S}_T)$ is a little more convoluted, it is actually a Bayesian procedure :

- fix a prior Π on the set of priors $\mathcal{Q}$: $\Pi \in \mathcal{P}rob(\mathcal{Q})$,
- define :

$$\hat{\Lambda} = \underset{\Lambda \in \mathcal{P}rob(\mathcal{Q})}{\arg\min} \left\{ \mathbb{E}_{\pi\sim\Lambda} \left[\frac{1}{T} \sum_{t=1}^{T} \hat{\mathcal{R}}_t\Big(\rho_t(\pi,\alpha),\pi\Big) \right] + \frac{KL(\Lambda,\Pi)}{\gamma T} \right\},$$

- draw $\hat{\pi} \sim \hat{\Lambda}$.

Define

$$\pi^* = \arg\min_{\pi} \mathbb{E}_{\text{new}} \left[\hat{\mathcal{R}}_{T+1}\Big(\rho_{T+1}(\pi,\alpha), \pi \Big) \right].$$

Lemma – Bernstein condition at the meta-level

For any $\pi \in \mathcal{Q}$,

$$\mathbb{E}_{\text{new}} \left[\left(\hat{\mathcal{R}}_{T+1}\Big(\rho_{T+1}(\pi,\alpha), \pi \Big) - \hat{\mathcal{R}}_{T+1}\Big(\rho_{T+1}(\pi^*,\alpha), \pi^* \Big) \right)^2 \right]$$
$$\leq C \, \mathbb{E}_{\text{new}} \left[\hat{\mathcal{R}}_{T+1}\Big(\rho_{T+1}(\pi,\alpha), \pi \Big) - \hat{\mathcal{R}}_{T+1}\Big(\rho_{T+1}(\pi^*,\alpha), \pi^* \Big) \right].$$

Theorem

$$\mathbb{E}_{\text{data}}\left\{\mathbb{E}_{\hat{\pi}\sim\hat{\Lambda}}[\mathcal{E}(\hat{\pi})]\right\} \leq \mathcal{E}^{*} + \min_{\Lambda\in\mathcal{P}rob(\mathcal{Q})} \mathbb{E}_{\pi\sim\Lambda}\left\{\mathbb{E}_{P_{T+1}\sim\mathcal{P}}\left[\left(\frac{d(P_{T+1},\pi)\log(n)}{n}\right)^{\beta}\right] + \frac{\mathcal{K}(\Lambda,\Pi)}{\gamma T}\right\}.$$

The aforementioned examples are obtained by specification of Π, and taking an explicit Λ above.

Remark :

$$\hat{\Lambda} = \underset{\Lambda \in \mathcal{P}rob(\mathcal{Q})}{\arg\min} \left\{ \mathbb{E}_{\pi \sim \Lambda} \left[\frac{1}{T} \sum_{t=1}^{T} \hat{\mathcal{R}}_t \Big(\rho_t(\pi, \alpha), \pi \Big) \right] + \frac{KL(\Lambda, \Pi)}{\gamma T} \right\}.$$

What happens if we minimize over a smaller set $\mathcal{V} \subset \mathcal{P}rob(\mathcal{Q})$?

$$\hat{\Lambda}_{\mathcal{V}} = \underset{\Lambda \in \mathcal{V}}{\arg\min} \left\{ \mathbb{E}_{\pi \sim \Lambda} \left[\frac{1}{T} \sum_{t=1}^{T} \hat{\mathcal{R}}_t \Big(\rho_t(\pi, \alpha), \pi \Big) \right] + \frac{KL(\Lambda, \Pi)}{\gamma T} \right\}.$$

Note : can be seen as a variational Bayes version of $\hat{\Lambda}$.

Theorem

$$\mathbb{E}_{\text{data}}\left\{\mathbb{E}_{\hat{\pi}\sim\hat{\Lambda}_{\mathcal{V}}}[\mathcal{E}(\hat{\pi})]\right\} \leq \mathcal{E}^* + \min_{\Lambda\in\mathcal{V}} \mathbb{E}_{\pi\sim\Lambda}\left\{\mathbb{E}_{P_{T+1}\sim\mathcal{P}}\left[\left(\frac{d(P_{T+1},\pi)\log(n)}{n}\right)^{\beta}\right] + \frac{\mathcal{K}(\Lambda,\Pi)}{\gamma T}\right\}.$$

For example, in the case $\mathcal{Q} = \{\pi_1, \dots, \pi_K\}$, taking $\mathcal{V}$ as the set of Dirac masses allows to define $\hat{\pi}$ by a minimization rather than by randomisation.

Note however that our result require to use "exact" Bayes within tasks.

Lemma – Bernstein condition at the meta-level

For any $\pi \in \mathcal{Q}$,

$$\mathbb{E}_{\rm new}\left[\left(\hat{\mathcal{R}}_{T+1}\Big(\rho_{T+1}(\pi,\alpha),\pi\Big) - \hat{\mathcal{R}}_{T+1}\Big(\rho_{T+1}(\pi^*,\alpha),\pi^*\Big)\right)^2\right]$$
$$\leq C\, \mathbb{E}_{\rm new}\left[\hat{\mathcal{R}}_{T+1}\Big(\rho_{T+1}(\pi,\alpha),\pi\Big) - \hat{\mathcal{R}}_{T+1}\Big(\rho_{T+1}(\pi^*,\alpha),\pi^*\Big)\right].$$

We don't know how to extend this lemma if we replace $\rho_{T+1}(\pi,\alpha)$ by a variational approximation.

Some important open questions :

- extending the Lemma to allow variational approximations.
- lower bounds.