On adaptation for the posterior distribution under local and sup-norm

Judith Rousseau, Marc Hoffman and Johannes Schmidt - Hieber

ENSAE - CREST et CEREMADE, Université Paris-Dauphine

Brown

Outline

Outline

Generalities

- **Model :** $Y_1^n|\theta \sim p_\theta^n$ (density wrt μ), $\theta \in \Theta$

A priori : $\theta \sim \Pi$: prior distribution

$\longrightarrow$ posterior distribution

$$d\Pi(\theta|X^n) = \frac{d\Pi(\theta)p_\theta^n(Y_1^n)}{m(Y_1^n)}, \quad Y_1^n = (Y_1, \ldots, Y_n)$$

- **Posterior concentration** $d(.,.) =$ loss on Θ & $\theta_0 \in \Theta$ = True

$$E_{\theta_0}\left(\Pi\left[U_{\epsilon_n}|Y_1^n\right]\right) = 1 + o(1), \quad U_{\epsilon_n} = \{\theta; d(\theta, \theta_0) \leq \epsilon_n\} \quad \epsilon_n \downarrow 0$$

- **Minimax concentration rates** on a Class $\Theta_\alpha(L)$,

$$\sup_{\theta_0 \in \Theta_\alpha(L)} E_{\theta_0}\left(\Pi\left[U^c_{M\epsilon_n(\alpha)}|Y_1^n\right]\right) = o(1),$$

where $\epsilon_n(\alpha) =$ minimax rate under $d(.,.)$ & over $\Theta_\alpha(L)$.

Examples of Models-losses for which nice results exist

- **Density estimation** $Y_i \sim p_\theta$ i.i.d.

$$d(p_\theta, p_{\theta'})^2 = \int (\sqrt{p_\theta} - \sqrt{p_{\theta'}})^2(x)dx, \quad d(p_\theta, p_{\theta'}) = \int |p_\theta - p_{\theta'}|(x)dx$$

- **Regression function**

$$Y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \quad \theta = (f, \sigma)$$

$$d(p_\theta, p_{\theta'}) = \|f - f'\|_2, \quad d(p_\theta, p_{\theta'}) = n^{-1} \sum_{i=1}^{n} H^2(p_\theta(y|X_i), p_{\theta'}(y|X_i))$$

H = Hellinger

- **White noise**

$$dY(t) = f(t)dt + n^{-1/2}dW(t) \quad \Leftrightarrow Y_i = \theta_i + n^{-1/2}\epsilon_i, \quad i \in \mathbb{N}$$

$$d(p_\theta, p_{\theta'}) = \|f - f'\|_2$$

Examples : functional classes

$\Theta_\alpha(L) =$ Hölder $(\mathcal{H}(\alpha, L))$

$$\epsilon_n(\alpha) = n^{-\alpha/(2\alpha+1)} \quad \text{minimax rate over } \mathcal{H}(\alpha, L)$$

- **Density example : Hellinger loss**

Prior = DPM

$$f(x) = f_{P,\sigma}(x) = \int \phi_\sigma(x-\mu) dP(\mu), \quad \sigma \sim \text{I}\Gamma(a,b) \quad P \sim \text{DP}(A, G_0)$$

$$\sup_{f_0 \in \Theta_\alpha(L)} E_{f_0} \left(\Pi \left[U^c_{M(n/\log n)^{-\alpha/(2\alpha+1)}}(f_0) | Y_1^n \right] \right) = o(1),$$

$U_\epsilon(f_0) = \{f, h(f_0, f) \leq \epsilon\}$ [$\log n$ term necessary ?]

$$\Rightarrow \quad E_{f_0} \left[h(\hat{f}, f_0)^2 \right] \lesssim (n/\log n)^{-\alpha/(2\alpha+1)}, \quad \hat{f}(x) = E^\pi [f(x) | Y^n]$$

Outline

Adaptation

For such $d(.,.)$ Adaptation is easy : The prior does not depend on α :

$$\sup_{\alpha_1 \leq \alpha \leq \alpha_2} \sup_{\theta_0 \in \Theta_\alpha(L)} E_{\theta_0} \left(\Pi \left[U^c_{M(n/\log n)^{-\alpha/(2\alpha+1)}} | Y_1^n \right] \right) = o(1),$$

- **why ?**

Outline

Outline

$U_n = U_{M(n/\log n)^{-\alpha/(2\alpha+1)}}$ and $l_n(\theta) = \log p_\theta^n(Y_1^n)$
$\bar{\epsilon}_n = (n/\log n)^{-\alpha/(2\alpha+1)}$

$$\Pi\left[U_n^c | Y_1^n\right] = \frac{\int_{U_n^c} e^{l_n(\theta)-l_n(\theta_0)} d\Pi(\theta)}{\int_\Theta e^{l_n(\theta)-l_n(\theta_0)} d\Pi(\theta)} := \frac{N_n}{D_n}$$

$\phi_n = \phi_n(Y_1^n) \in [0,1]$

$$P_{\theta_0}\left(\Pi\left[U_n^c | Y_1^n\right] > e^{-\tau n\epsilon_n^2}\right) \leq E_{\theta_0}^n[\phi_n] + P_{\theta_0}\left[D_n < e^{-cn\bar{\epsilon}_n^2}\right] + e^{(c+\tau)n\epsilon_n^2} \int_{U_n^c} E_\theta\left[1-\phi_n\right] d\pi(\theta)$$

Constraints

$$E_{\theta_0}^n[\phi_n] = o(1) \quad \& \quad \sup_{d(\theta,\theta_0)>M\bar{\epsilon}_n} E_\theta[1-\phi_n] = o(e^{-cn\bar{\epsilon}_n^2}) \to d(.,.)$$

$$P_{\theta_0}\left[D_n < e^{-cn\bar{\epsilon}_n^2}\right] = o(1) \quad \text{We need :}$$

$$\begin{aligned} D_n &\geq \int_{S_n} e^{l_n(\theta)-l_n(\theta_0)} d\Pi(\theta) \\ &\geq e^{-2n\epsilon_n^2}\Pi\left[S_n \cap \{l_n(\theta) - l_n(\theta_0) > -2n\bar{\epsilon}_n^2\}\right] \end{aligned}$$

Ok if $S_n = \{KL(p_{\theta_0}^n, p_\theta^n) \leq n\bar{\epsilon}_n^2; V(p_{\theta_0}^n, p_\theta^n) \leq n\bar{\epsilon}_n^2\}$ and

$$\Pi(S_n) \geq e^{-cn\bar{\epsilon}_n^2} \to \text{links } d(.,.) \quad \text{with } KL(.,.)$$

example : white noise model + L_2 loss

$$Y_{ik} = \theta_{ik} + n^{-1/2}\epsilon_{ik} \quad \epsilon_{ik} \sim \mathcal{N}(0,1), \quad i \in \mathbb{N}, k \leq 2^{i-1}$$

$$(dY(t) = f(t)dt + n^{-1/2}dW(t))$$

- **Hölder class** (α)

$$\theta_0 \in \{\theta; |\theta_{ik}| \leq Li^{-\alpha-1/2}, \quad \forall i,k\}$$

- **Prior : spike and slab**

$$\theta_{ik} \sim (1-p_n)\delta_{(0)} + p_n g, \quad e.g. \quad g = \mathcal{N}(0,v), \quad p_n = 1/n$$

- **Concentration**

$$S_n \approx \{\|\theta-\theta_0\|^2 \leq (n/\log n)^{-2\alpha/(2\alpha+1)}\} \rightarrow \forall j \geq J_{n,\alpha}, k \leq 2^j; \theta_{j,k} = 0$$

$$2^{J_{n,\alpha}} = (n/\log n)^{1/(2\alpha+1)} := R_n, \quad \Pi(S_n) \gtrsim e^{-CR_n \log n} := e^{-Cn\epsilon_n^2}$$

$$E_{\theta_0}[\phi_n] = o(1), \& \quad \sup_{\theta \in \Theta_n; \|\theta-\theta_0\| \gtrsim \epsilon_n} E_\theta[1-\phi_n] \leq e^{-cn\epsilon_n^2}$$

What about $f(x_0)$? or $\|f - f_0\|_\infty$?

$$Y_{ik} = \theta_{ik} + n^{-1/2}\epsilon_{ik} \quad \epsilon_{ik} \sim \mathcal{N}(0,1), \quad \theta_0 \in \{\theta; |\theta_{ik}| \leq L i^{-\alpha-1/2}, \quad \forall i,k\}$$

- **Prior : spike and slab** $\theta_{ik} = (1-p_n)\delta_{(0)} + p_n g$, $p_n = 1/n$
- **losses** :

$$l(\theta,\theta_0) = (\sum_{ik}(\theta_{ik} - \theta^o_{ik})\psi_{ik}(x_0)2^{i/2})^2 \quad \text{(local)}$$

$$l(\theta,\theta_0) = \|\theta - \theta_0\|_\infty = \sum_i \max_k |\theta_{ik} - \theta^o_{ik}|2^{i/2} \quad \text{(sup)}$$

- **Bayesian concentration** $\forall \alpha > 0, \exists \theta_0 \in \Theta_\alpha(L)$ s.t.

$$E_{\theta_0}\left(\Pi\left[l(\theta,\theta_0) \leq n^{-(\alpha-1/2)/(2\alpha+1)} \log n^q | Y_1^n\right]\right) = o(1)$$

Sub-optimal $\theta^o_{i0} = \rho_n 2^{-i/2}$ and $\theta^o_{ik} = 0$, $i \leq l_n : \forall J > 0$

$$\sum_{i>J}\sum_k (\theta^o_{ik})^2 \leq n^{-2\alpha/(2\alpha+1)}, \quad \sum_{i>J}\max_k |\theta^o_{ik}| > n^{-(\alpha-1/2)/(2\alpha+1)} \log n^q$$

Risk ?

$$Y_{ik} = \theta_{ik} + n^{-1/2}\epsilon_{ik} \quad \epsilon_{ik} \sim \mathcal{N}(0,1), \quad \theta_0 \in \{\theta; |\theta_{ik}| \leq L i^{-\alpha-1/2}, \quad \forall i,k\}$$

- Prior : $\theta_{ik} = (1-p_n)\delta_{(0)} + p_n g$, $p_n = 1/n$
- Suboptimal concentration BUT $\hat{\theta} = E^{\pi}[\theta|Y^n]$

$$\limsup_n \sup_{\alpha_1 \leq \alpha \leq \alpha_2} (n/\log n)^{2\alpha/(2\alpha+1)} \sup_{\theta_0 \in \Theta_\alpha} E^n_{\theta_0}\left[l(\hat{\theta},\theta_0)\right] < +\infty$$

  - Questions

Risk ?

$$Y_{ik} = \theta_{ik} + n^{-1/2}\epsilon_{ik} \quad \epsilon_{ik} \sim \mathcal{N}(0,1), \quad \theta_0 \in \{\theta; |\theta_{ik}| \leq L i^{-\alpha-1/2}, \quad \forall i,k\}$$

- Prior : $\theta_{ik} = (1-p_n)\delta_{(0)} + p_n g$, $p_n = 1/n$
- Suboptimal concentration BUT $\hat{\theta} = E^{\pi}[\theta|Y^n]$

$$\limsup_n \sup_{\alpha_1 \leq \alpha \leq \alpha_2} (n/\log n)^{2\alpha/(2\alpha+1)} \sup_{\theta_0 \in \Theta_\alpha} E^n_{\theta_0}\left[l(\hat{\theta},\theta_0)\right] < +\infty$$

- Questions
- ▶ **Question 1** How general is this (negative) result ?

Risk ?

$$Y_{ik} = \theta_{ik} + n^{-1/2}\epsilon_{ik} \quad \epsilon_{ik} \sim \mathcal{N}(0,1), \quad \theta_0 \in \{\theta; |\theta_{ik}| \leq L i^{-\alpha-1/2}, \quad \forall i,k\}$$

- Prior : $\theta_{ik} = (1-p_n)\delta_{(0)} + p_n g$, $p_n = 1/n$
- Suboptimal concentration BUT $\hat{\theta} = E^{\pi}[\theta|Y^n]$

$$\limsup_n \sup_{\alpha_1 \leq \alpha \leq \alpha_2} (n/\log n)^{2\alpha/(2\alpha+1)} \sup_{\theta_0 \in \Theta_\alpha} E^n_{\theta_0}\left[l(\hat{\theta},\theta_0)\right] < +\infty$$

- Questions
- ▶ **Question 1** How general is this (negative) result ?
- ▶ **Question 2** What does it tell us about posterior concentration ?

A first general result

$$\mathcal{H}(\alpha_1, L) \cup \mathcal{H}(\alpha_2, L) \subset \Theta, \quad \alpha_1 < \alpha_2$$

- **Local loss** $l(\theta, \theta_0) = (\theta(x) - \theta_0(x))^2$

Result : *There exist no prior that leads to adaptive minimax concentration over any collection of Hölder balls :*

$\forall \pi$ prior on Θ, $\forall M > 0$

$$\max_j \sup_{\theta_0 \in \mathcal{H}(\alpha_j, L)} E_{\theta_0} \left(\Pi \left[l(\theta, \theta_0) > M n^{-2\alpha_j/(2\alpha_j+1)} | Y^n \right] \right) = 1$$

- What do we loose ?
- L_∞ **and local loss**

If $\exists \theta_0 \in \Theta$,

$$P_{\theta_0} \left(\Pi \left[l(\theta, \theta_0) > M n^{-2\alpha_2/(2\alpha_2+1)} | Y^n \right] > e^{-n^\tau} \right) = o(1), \quad \tau > 0$$

Then worse

$$\max_j \sup_{\theta_0 \in \mathcal{H}(\alpha_j, L)} E_{\theta_0} \left(\Pi \left[l(\theta, \theta_0) > n^{-(2\alpha_j - \tau)/(2\alpha_j+1)} | Y^n \right] \right) = 1$$

Still not completely satisfying

- For local loss : If we could find a prior with only $\log n$ loss then who cares !
- L_∞ loss : Smaller than e^{-n^τ} to be expected because of test

can we be more precise ?

Slightly

Another negative result

$$\mathcal{H}(\alpha_1, L) \cup \mathcal{H}(\alpha_2, L) \subset \Theta, \alpha_1 < \alpha_2$$
$$\epsilon_n(\alpha) = (n/\log n)^{-\alpha/(2\alpha+1)}$$

If there exists $\theta_0 \in \mathcal{H}(\alpha_2, L)$ and $2^{J_{n,\alpha_2}} = (n/\log n)^{1/(2\alpha_2+1)}$.

- $$\pi\left(\|\theta - \theta_0\|_2 \leq c\epsilon_n(\alpha_2)\right) \gtrsim e^{-n\epsilon_n^2(\alpha_2)}$$

Then there $\exists \theta_1 \in \mathcal{H}(\alpha_1, L)$

$$E_{\theta_1}\left(\Pi\left[I(\theta, \theta_0) >> \epsilon_n(\alpha_1) | Y^n\right]\right) = 1$$

Another negative result

$$\mathcal{H}(\alpha_1, L) \cup \mathcal{H}(\alpha_2, L) \subset \Theta, \alpha_1 < \alpha_2$$
$$\epsilon_n(\alpha) = (n/\log n)^{-\alpha/(2\alpha+1)}$$

If there exists $\theta_0 \in \mathcal{H}(\alpha_2, L)$ and $2^{J_{n,\alpha_2}} = (n/\log n)^{1/(2\alpha_2+1)}$.

- $$\pi\left(\|\theta - \theta_0\|_2 \leq c\epsilon_n(\alpha_2)\right) \gtrsim e^{-n\epsilon_n^2(\alpha_2)}$$
- $$\pi\left(\sum_{j \geq J_{n,\alpha_2}} \sum_k \theta_{jk}^2 > A\epsilon_n(\alpha_2)^2\right) \leq e^{-Bn\epsilon_n^2(\alpha_2)}$$

Then there $\exists \theta_1 \in \mathcal{H}(\alpha_1, L)$

$$E_{\theta_1}\left(\Pi\left[I(\theta, \theta_0) >> \epsilon_n(\alpha_1) | Y^n\right]\right) = 1$$

Another negative result

$$\mathcal{H}(\alpha_1, L) \cup \mathcal{H}(\alpha_2, L) \subset \Theta, \alpha_1 < \alpha_2$$
$$\epsilon_n(\alpha) = (n/\log n)^{-\alpha/(2\alpha+1)}$$

If there exists $\theta_0 \in \mathcal{H}(\alpha_2, L)$ and $2^{J_{n,\alpha_2}} = (n/\log n)^{1/(2\alpha_2+1)}$.

- $$\pi\left(\|\theta - \theta_0\|_2 \leq c\epsilon_n(\alpha_2)\right) \gtrsim e^{-n\epsilon_n^2(\alpha_2)}$$
- $$\pi\left(\sum_{j \geq J_{n,\alpha_2}} \sum_k \theta_{jk}^2 > A\epsilon_n(\alpha_2)^2\right) \leq e^{-Bn\epsilon_n^2(\alpha_2)}$$
- $\exists \rho_n \downarrow 0$ s.t.

$$\pi\left(\sum_{j \geq J_{n,\alpha_2}} 2^{j/2} \max_k |\theta_{jk}| > \rho_n \epsilon_n(\alpha_1)\right) \leq e^{-Bn\epsilon_n^2(\alpha_2)}$$

Then there $\exists \theta_1 \in \mathcal{H}(\alpha_1, L)$

$$E_{\theta_1}\left(\Pi\left[I(\theta, \theta_0) >> \epsilon_n(\alpha_1) | Y^n\right]\right) = 1$$

Conclusion

• Bayesian is great for risks that are related to Kullback : L_2 in regression, hellinger or L_1 in density etc.
• How to understand some specific features in these big models ?

More tricky

• Can we prove that $\forall \pi$: No adaptation in L_∞ for concentration rates ?
• Why should we care ? $\rightarrow$ interpretation of credible bands ! ?
• Are these negative results related to the non existence of adaptive confidence bands in L_∞ ?
• If no adaptive prior : Important to understand the types of θ_0 that won't work. e.g. . . .

THANK YOU