When can a Bayesian divide and conquer?
General optimality guarantees for distributed nonparametric Bayesian inference

Lasse Vuursteen

TU Delft

This talk is based on joint works with Botond Szabó and Harry van Zanten.

Università Bocconi
MILANO

VU UNIVERSITY AMSTERDAM

The (computational) cost of doing Bayes

Many traditional methods of inference do not scale to very large datasets, due to e.g. computational or memory bottlenecks.

The (computational) cost of doing Bayes

Many traditional methods of inference do not scale to very large datasets, due to e.g. computational or memory bottlenecks.

- Even for iid data $X = (X_1, \ldots, X_n)$, just evaluating the likelihood

$$\prod_{i=1}^{n} p_{\theta}(X_i)$$

  has computational complexity $O(n)$. This makes many optimization or MCMC based methods infeasible when n is very large.

The (computational) cost of doing Bayes

Many traditional methods of inference do not scale to very large datasets, due to e.g. computational or memory bottlenecks.

- Even for iid data $X = (X_1, \ldots, X_n)$, just evaluating the likelihood

$$\prod_{i=1}^{n} p_{\theta}(X_i)$$

  has computational complexity $O(n)$. This makes many optimization or MCMC based methods infeasible when n is very large.
- Popular methods like Gaussian process regression typically have computational complexity $O(n^3)$ and memory complexity $O(n^2)$ even in conjugate models.

Divide and conquer

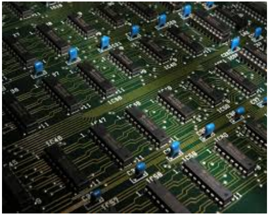
- A simple method to scale Bayesian computation is to divide the n observations equally over m servers/processors which we will call *machines*.

Divide and conquer

- A simple method to scale Bayesian computation is to divide the n observations equally over m servers/processors which we will call *machines*.
- Each machine computes a posterior based on only its local data, after which draws from each of the local posteriors are averaged across the machines.

Divide and conquer

- A simple method to scale Bayesian computation is to divide the n observations equally over m servers/processors which we will call *machines*.
- Each machine computes a posterior based on only its local data, after which draws from each of the local posteriors are averaged across the machines.



Divide and conquer - formally

Formally:

1. Observations $X_1, \ldots, X_n$ are divided over n machines, machine $j = 1, \ldots, n$ receiving n/m observations $X^{(j)} = (X_1^{(j)}, \ldots, X_{n/m}^{(j)})$

Divide and conquer - formally

Formally:

1. Observations $X_1, \ldots, X_n$ are divided over n machines, machine $j = 1, \ldots, n$ receiving n/m observations $X^{(j)} = (X_1^{(j)}, \ldots, X_{n/m}^{(j)})$
2. Each machine j runs its own routine to compute the posterior $\Pi(\cdot|X^{(j)})$ based on some prior $\Pi^{(j)}$ on Θ.

Divide and conquer - formally

Formally:

1. Observations $X_1, \dots, X_n$ are divided over n machines, machine $j = 1, \dots, n$ receiving n/m observations $X^{(j)} = (X_1^{(j)}, \dots, X_{n/m}^{(j)})$
2. Each machine j runs its own routine to compute the posterior $\Pi(\cdot|X^{(j)})$ based on some prior $\Pi^{(j)}$ on Θ.
3. Inference is based on the average of the draws of the local posteriors:

$$\frac{1}{m}\sum_{j=1}^{m}\theta^{(j)} \text{ for } \theta^{(j)} \sim \Pi(\cdot|X^{(j)}).$$

Divide and conquer - formally

Formally:

1. Observations $X_1, \dots, X_n$ are divided over n machines, machine $j = 1, \dots, n$ receiving n/m observations $X^{(j)} = (X_1^{(j)}, \dots, X_{n/m}^{(j)})$
2. Each machine j runs its own routine to compute the posterior $\Pi(\cdot|X^{(j)})$ based on some prior $\Pi^{(j)}$ on Θ.
3. Inference is based on the average of the draws of the local posteriors:

$$\frac{1}{m}\sum_{j=1}^{m}\theta^{(j)} \text{ for } \theta^{(j)} \sim \Pi(\cdot|X^{(j)}).$$

The distribution of the average of the draws of the local posteriors give us a generalized "posterior": the m-times convolution of the local posteriors rescaled with m:

$$\tilde{\Pi}(A|X^{(1:m)}) := \mathop{\circledast}_{j=1}^{m} \Pi(m \cdot |X^{(j)})(A).$$

We perform inference on θ using $\tilde{\Pi}(\cdot|X^{(1:m)})$.

Divide and conquer - nomenclature

- *"Bayesian divide and conquer"* for example in Rihui Ou, Deborshee Sen, and David Dunson(2021), "Scalable Bayesian inference for time series via divide-and-conquer"

Divide and conquer - nomenclature

- *"Bayesian divide and conquer"* for example in Rihui Ou, Deborshee Sen, and David Dunson(2021), "Scalable Bayesian inference for time series via divide-and-conquer"
- *"Embarrassingly parallel Bayes"* for example in Michael Minyi Zhang and Sinead A Williamson(2019), "Embarrassingly parallel inference for Gaussian processes"

Divide and conquer - nomenclature

- *"Bayesian divide and conquer"* for example in Rihui Ou, Deborshee Sen, and David Dunson(2021), "Scalable Bayesian inference for time series via divide-and-conquer"
- *"Embarrassingly parallel Bayes"* for example in Michael Minyi Zhang and Sinead A Williamson(2019), "Embarrassingly parallel inference for Gaussian processes"
- *"Product of experts"* for example in Botond Szabó and Harry Van Zanten(2019), "An asymptotic analysis of distributed nonparametric methods."

Divide and conquer - nomenclature

- *"Bayesian divide and conquer"* for example in Rihui Ou, Deborshee Sen, and David Dunson(2021), "Scalable Bayesian inference for time series via divide-and-conquer"
- *"Embarrassingly parallel Bayes"* for example in Michael Minyi Zhang and Sinead A Williamson(2019), "Embarrassingly parallel inference for Gaussian processes"
- *"Product of experts"* for example in Botond Szabó and Harry Van Zanten(2019), "An asymptotic analysis of distributed nonparametric methods."
- *"Consensus Monte Carlo"* in Steven L Scott et al.(2016), "Bayes and big data: The consensus Monte Carlo algorithm"

Divide and conquer - benefits

- It allows you to scale any method which computational/memory complexity is dominated in the number of observations.

Divide and conquer - benefits

- It allows you to scale any method which computational/memory complexity is dominated in the number of observations.

$$O(n^3) \text{ becomes } \begin{cases} O\left(\frac{n^3}{m^3}\right) & \text{if ran in parallel,} \\ O\left(\frac{n^3}{m^2}\right) & \text{otherwise.} \end{cases}$$

  So we can even obtain linear or constant complexity in n if $m \simeq n$.
- It is super simple to implement (even parallely): no need to rewrite code for existing Bayesian methods (it is “plug-and-play”).

So when can a Bayesian divide and conquer?

Short answer: *(almost) always, but it is not always good. Inference might become highly suboptimal.*

So when can a Bayesian divide and conquer?

Short answer: *(almost) always, but it is not always good. Inference might become highly suboptimal.*

In this work, we provide general sufficiency conditions that guarantee optimal frequentist convergence rates for the divide and conquer strategy.

- We demonstrate that the guarantees hold for a variety of nonparametric models and priors.

So when can a Bayesian divide and conquer?

Short answer: *(almost) always, but it is not always good. Inference might become highly suboptimal.*

In this work, we provide general sufficiency conditions that guarantee optimal frequentist convergence rates for the divide and conquer strategy.

- We demonstrate that the guarantees hold for a variety of nonparametric models and priors.
- Similar guarantees are provided for methods that are based on general loss functions instead of likelihoods for G-Bayes and M-estimation.

Setting and notation

In this talk, I will assume iid $X_i \sim P_{\theta^\circ}$ for some $\theta^\circ \in \Theta$ and that the model $\{P_\theta : \theta \in \Theta\}$ is dominated.

As shorthand notation, let

$$\ell(\theta_1, \theta_2; X_i) := \log\left(\frac{p_{\theta_1}}{p_{\theta_2}}(X_i)\right)$$

where p_θ is the density of P_θ.

Setting and notation

In this talk, I will assume iid $X_i \sim P_{\theta^\circ}$ for some $\theta^\circ \in \Theta$ and that the model $\{P_\theta : \theta \in \Theta\}$ is dominated.

As shorthand notation, let

$$\ell(\theta_1, \theta_2; X_i) := \log\left(\frac{p_{\theta_1}}{p_{\theta_2}}(X_i)\right)$$

where p_θ is the density of P_θ.

I will also write

$$\bar{\theta} = \frac{1}{m}\sum_{j=1}^{m} \theta^{(j)}.$$

Setting and notation

In this talk, I will assume iid $X_i \sim P_{\theta^\circ}$ for some $\theta^\circ \in \Theta$ and that the model $\{P_\theta : \theta \in \Theta\}$ is dominated.

As shorthand notation, let

$$\ell(\theta_1, \theta_2; X_i) := \log\left(\frac{p_{\theta_1}}{p_{\theta_2}}(X_i)\right)$$

where p_θ is the density of P_θ.

I will also write

$$\bar{\theta} = \frac{1}{m}\sum_{j=1}^{m} \theta^{(j)}.$$

Recall:

$$\bar{\theta} \sim \tilde{\Pi}(\cdot | X^{(1:m)}) := \circledast_{j=1}^{m} \Pi(m \cdot | X^{(j)}).$$

Divide and conquer convergence rate

Definition: The posterior distribution $\Pi(\cdot|X^{(n)})$ is said to *contract at rate* $\varepsilon_n \to 0$ in d at θ° if

$$\Pi\left(\theta \in \Theta : d(\theta, \theta^\circ) \leq M_n \varepsilon_n | X^{(n)}\right) \to 1 \text{ in } P_{\theta^\circ}\text{-probability},$$

for any $M_n \to \infty$ as $n \to \infty$.

Divide and conquer convergence rate

Definition: The posterior distribution $\Pi(\cdot|X^{(n)})$ is said to *contract at rate* $\varepsilon_n \to 0$ in d at θ° if

$$\Pi\left(\theta \in \Theta : d(\theta, \theta^\circ) \leq M_n \varepsilon_n | X^{(n)}\right) \to 1 \text{ in } P_{\theta^\circ}\text{-probability},$$

for any $M_n \to \infty$ as $n \to \infty$.

Suppose the full data posterior contracts at rate ε_n, does

$$\tilde{\Pi}\left(\bar{\theta} : d(\bar{\theta}, \theta^\circ) \leq M_n \varepsilon_n \middle| X^{(1:m)}\right) \overset{P_{\theta^\circ}}{\to} 1$$

also for all $M_n \to \infty$?

Divide and conquer convergence rate Theorem

Theorem ((Szabó, V., van Zanten))

*Let ε_n be a sequence of positive numbers for which the **likelihood geometry condition** and **testing condition** hold. Assume furthermore that $\Pi^{(j)}$'s are such that the **prior mass** and **entropy conditions** are satisfied.*

Divide and conquer convergence rate Theorem

Theorem ((Szabó, V., van Zanten))

Let ε_n be a sequence of positive numbers for which the ***likelihood geometry condition*** *and* ***testing condition*** *hold. Assume furthermore that $\Pi^{(j)}$'s are such that the* ***prior mass*** *and* ***entropy conditions*** *are satisfied.*

Then, ε_n is a contraction rate for the generalized posterior. That is, it holds that

$$\tilde{\Pi}\left(\bar{\theta} : d(\bar{\theta}, \theta^\circ) \leq M_n \varepsilon_n \middle| X^{(1:m)}\right) \stackrel{P_{\theta^\circ}}{\to} 1 \tag{1}$$

for any $M_n \to \infty$.

Likelihood geometry condition

The likelihood geometry condition is satisfied if

$$\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{\theta^\circ}\ell\left(\theta^{(j)},\theta^\circ;X_i^{(j)}\right) \leq c_\ell \mathbb{E}_{\theta^\circ}\ell\left(m^{-1}\sum_{j=1}^{m}\theta^{(j)},\theta^\circ;X_i^{(j)}\right) \tag{2}$$

whenever $-\mathbb{E}_{\theta^\circ}\ell\left(\bar{\theta},\theta^\circ;X_i^{(j)}\right) \leq C_\ell \varepsilon_n{}^2$ for some constants $c_\ell, C_\ell > 0$.

Likelihood geometry condition

The likelihood geometry condition is satisfied if

$$\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{\theta^\circ}\ell\left(\theta^{(j)},\theta^\circ;X_i^{(j)}\right)\le c_\ell\mathbb{E}_{\theta^\circ}\ell\left(m^{-1}\sum_{j=1}^{m}\theta^{(j)},\theta^\circ;X_i^{(j)}\right) \tag{2}$$

whenever $-\mathbb{E}_{\theta^\circ}\ell\left(\bar{\theta},\theta^\circ;X_i^{(j)}\right)\le C_\ell\varepsilon_n{}^2$ for some constants $c_\ell, C_\ell > 0$.

NB1: $\mathbb{E}_{\theta^\circ}\ell\left(\bar{\theta},\theta^\circ;X_i^{(j)}\right) = -D_{KL}(P_{\bar{\theta}};P_{\theta^\circ})$.

Likelihood geometry condition

The likelihood geometry condition is satisfied if

$$\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{\theta^\circ}\ell\left(\theta^{(j)},\theta^\circ;X_i^{(j)}\right)\leq c_\ell\mathbb{E}_{\theta^\circ}\ell\left(m^{-1}\sum_{j=1}^{m}\theta^{(j)},\theta^\circ;X_i^{(j)}\right)\tag{2}$$

whenever $-\mathbb{E}_{\theta^\circ}\ell\left(\bar{\theta},\theta^\circ;X_i^{(j)}\right)\leq C_\ell{\varepsilon_n}^2$ for some constants $c_\ell, C_\ell > 0$.

NB1: $\mathbb{E}_{\theta^\circ}\ell\left(\bar{\theta},\theta^\circ;X_i^{(j)}\right) = -D_{KL}(P_{\bar{\theta}};P_{\theta^\circ})$.

NB2: The assumption certainly holds if the likelihood is log-concave.

Likelihood geometry condition

The likelihood geometry condition is satisfied if

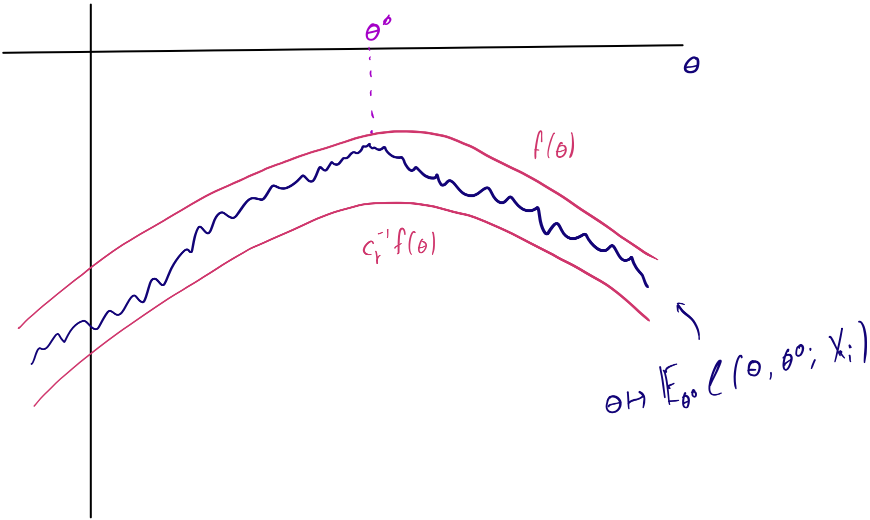
$$\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{\theta^\circ}\ell\left(\theta^{(j)},\theta^\circ;X_i^{(j)}\right)\le c_\ell\mathbb{E}_{\theta^\circ}\ell\left(m^{-1}\sum_{j=1}^{m}\theta^{(j)},\theta^\circ;X_i^{(j)}\right) \tag{2}$$

whenever $-\mathbb{E}_{\theta^\circ}\ell\left(\bar{\theta},\theta^\circ;X_i^{(j)}\right)\le C_\ell\varepsilon_n{}^2$ for some constants $c_\ell, C_\ell > 0$.

NB1: $\mathbb{E}_{\theta^\circ}\ell\left(\bar{\theta},\theta^\circ;X_i^{(j)}\right) = -D_{KL}(P_{\bar{\theta}};P_{\theta^\circ})$.

NB2: The assumption certainly holds if the likelihood is log-concave.

Likelihood geometry condition – a cartoon



Divide and conquer Ghoshal & van der Vaart conditions

The following **prior mass conditions** and **entropy** and **testing conditions** respectively:

$$\bigotimes_{j=1}^{m} \Pi^{(j)} \left(\theta^{(1:m)} \in \Theta^m : -\mathbb{E}_{\theta^\circ} \ell \left(\bar{\theta}, \theta^\circ ; X_i^{(j)} \right) \leq {\varepsilon_n}^2 \right) \geq e^{-C_{KL} n {\varepsilon_n}^2}, \tag{3}$$

and that there exists a $C_0 > 0$ such that for all $C \geq C_0$ there exists $\mathscr{P}_n \subset \mathscr{P}$ such that

$$\bigotimes_{j=1}^{m} \Pi^{(j)} (\theta^{(1:m)} \in \Theta^m \, : \, \bar{\theta} \in {\mathscr{P}_n}^c) \leq e^{-C n {\varepsilon_n}^2} \tag{4}$$

$$\log N(\varepsilon_n, d, \mathscr{P}_n) \leq C n {\varepsilon_n}^2. \tag{5}$$

Furthermore, writing $\bar{\theta} = m^{-1} \sum_{j=1}^{m} \theta^{(j)}$, assume that there exists a test T such that

$$P_{\theta^\circ} T \overset{n \to \infty}{\to} 0 \quad \text{and} \quad \bigotimes_{j=1}^{m} P_{\theta^{(j)}} (1 - T) \leq e^{-C_{test} M_n n {\varepsilon_n}^2} \tag{6}$$

whenever $\bar{\theta} \in \mathscr{P}_n$ and $d(\bar{\theta}, \theta^\circ) \geq M_n \varepsilon_n$.

Divide and conquer - Gaussian processes

Let $\Theta = (\Theta, \|\cdot\|)$ be a separable Banach space and let W be a centered Gaussian process on it.
Suppose that $-\mathbb{E}_{\theta}\ell(\theta', \theta; X_i) \leq \|\theta - \theta'\|^2$ and that W's *concentration function* satisfies[1]

$$\phi_{\theta^\circ}^{W}(\varepsilon_n) = \inf_{h \in \mathbb{H}: \|\theta^\circ - h\| < \varepsilon_n} \|h\|_{\mathbb{H}}^2 - \log \Pr(\|W\| \leq \varepsilon_n) \lesssim n\varepsilon_n{}^2.$$

[1] If the full data posterior corresponding to W is to contract at rate ε_n, this should be the case; see e.g. Ismaël Castillo(2008), "Lower bounds for posterior rates with Gaussian process priors" and A. W. van der Vaart and J. H. van Zanten(June 2008), "Rates of contraction of posterior distributions based on Gaussian process priors".

Divide and conquer - Gaussian processes

Let $\Theta = (\Theta, \|\cdot\|)$ be a separable Banach space and let W be a centered Gaussian process on it.
Suppose that $-\mathbb{E}_{\theta}\ell(\theta', \theta; X_i) \leq \|\theta - \theta'\|^2$ and that W's *concentration function* satisfies[1]

$$\phi_{\theta^\circ}^{W}(\varepsilon_n) = \inf_{h \in \mathbb{H}: \|\theta^\circ - h\| < \varepsilon_n} \|h\|_{\mathbb{H}}^2 - \log \Pr(\|W\| \leq \varepsilon_n) \lesssim n{\varepsilon_n}^2.$$

Take each of the $\Pi^{(j)}$ equal to the distribution of $\sqrt{m}W$. Then, the **prior mass** and **entropy** conditions are satisfied.

[1] If the full data posterior corresponding to W is to contract at rate ε_n, this should be the case; see e.g. Ismaël Castillo(2008), "Lower bounds for posterior rates with Gaussian process priors" and A. W. van der Vaart and J. H. van Zanten(June 2008), "Rates of contraction of posterior distributions based on Gaussian process priors".

Divide and conquer - Gaussian processes

Let $\Theta = (\Theta, \|\cdot\|)$ be a separable Banach space and let W be a centered Gaussian process on it.
Suppose that $-\mathbb{E}_{\theta}\ell(\theta', \theta; X_i) \leq \|\theta - \theta'\|^2$ and that W's *concentration function* satisfies[1]

$$\phi^{W}_{\theta^\circ}(\varepsilon_n) = \inf_{h \in \mathbb{H}: \|\theta^\circ - h\| < \varepsilon_n} \|h\|^2_{\mathbb{H}} - \log \Pr(\|W\| \leq \varepsilon_n) \lesssim n{\varepsilon_n}^2.$$

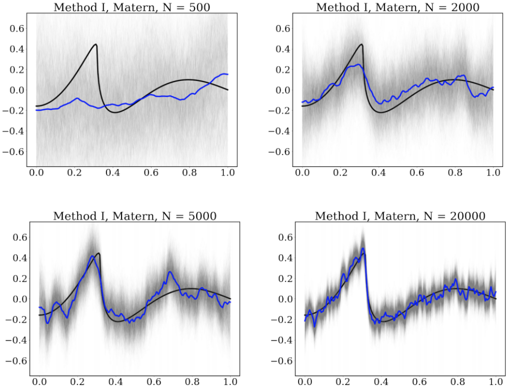
Take each of the $\Pi^{(j)}$ equal to the distribution of $\sqrt{m}W$. Then, the **prior mass** and **entropy** conditions are satisfied.

The prior blow-up of $\sqrt{m}$ turns out to be strictly necessary. We have a theorem that stating that taking $\Pi^{(j)}$ equal to the distribution of W results in a convergence rate that corresponds to throwing away $\frac{n(m-1)}{m}$ observations in typical settings.

[1] If the full data posterior corresponding to W is to contract at rate ε_n, this should be the case; see e.g. Ismaël Castillo(2008), "Lower bounds for posterior rates with Gaussian process priors" and A. W. van der Vaart and J. H. van Zanten(June 2008), "Rates of contraction of posterior distributions based on Gaussian process priors".

Divide and conquer - Gaussian processes

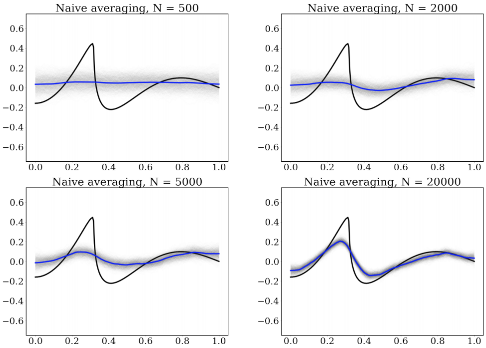
W centered Gaussian with squared exponential kernel. Local prior is with $\sqrt{m}$ blow-up.



Picture is courtesy of Amine Hadji, Tammo Hesslink, and Botond Szabó(2022), "Optimal recovery and uncertainty quantification for distributed Gaussian process regression".

Divide and conquer - Gaussian processes

W centered Gaussian with squared exponential kernel. Local prior is without $\sqrt{m}$ blow-up.



Picture is courtesy of Amine Hadji, Tammo Hesslink, and Botond Szabó(2022), "Optimal recovery and uncertainty quantification for distributed Gaussian process regression".

Divide and conquer convergence rate Theorem

Theorem ((Szabó, V., van Zanten))

Let ε_n be a sequence of nonnegative numbers and let W be a Gaussian process such that $\phi^W_{\theta^\circ}(\varepsilon_n) \lesssim n{\varepsilon_n}^2$ holds.

For $j = 1, \ldots, m$, let the local priors $\Pi^{(j)}$ be equal to the distribution of $\sqrt{m}W$. Then,

$$\tilde{\Pi}(\bar{\theta} : \|\bar{\theta} - \theta^\circ\| \leq M_n\varepsilon_n | X^{(1:m)}) \overset{P_{\theta^\circ}}{\to} 1 \tag{7}$$

as long as the ***testing condition*** *and* ***likelihood geometry condition*** *are satisfied.*

Examples: for which models does all of this hold?

- **Density estimation**: Suppose we observe iid X_i's in $[0,1]^d$ from some Hölder smooth density p°. Take as local priors

$$p_{\sqrt{m}W}(X_i) = \frac{e^{\sqrt{m}W(X_i)}}{\int e^{\sqrt{m}W(x)} d\mu(x)}$$

  where W is a zero mean Gaussian process in $\ell_\infty([0,1]^d)$ satisfying (**??**) with ε_n the optimal contraction rate and suppose that p° lies in the support of W. Then, the generalized posterior contracts at rate ε_n also.

Examples: for which models does all of this hold?

- **Density estimation**: Suppose we observe iid X_i's in $[0,1]^d$ from some Hölder smooth density p°. Take as local priors

$$p_{\sqrt{m}W}(X_i) = \frac{e^{\sqrt{m}W(X_i)}}{\int e^{\sqrt{m}W(x)} d\mu(x)}$$

  where W is a zero mean Gaussian process in $\ell_\infty([0,1]^d)$ satisfying (**??**) with ε_n the optimal contraction rate and suppose that p° lies in the support of W. Then, the generalized posterior contracts at rate ε_n also.

- **Regression:** A similar statement holds for iid observations $Y_i|P_\theta$ satisfying

$$Y_i = \theta(i/n) + Z_i$$

  with $Z_i \sim N(0,\sigma^2)$ and θ belonging to e.g. a Sobolev space.

Examples: for which models does all of this hold?

- **Density estimation**: Suppose we observe iid X_i's in $[0,1]^d$ from some Hölder smooth density p°. Take as local priors

$$p_{\sqrt{m}W}(X_i) = \frac{e^{\sqrt{m}W(X_i)}}{\int e^{\sqrt{m}W(x)} d\mu(x)}$$

  where W is a zero mean Gaussian process in $\ell_\infty([0,1]^d)$ satisfying (**??**) with ε_n the optimal contraction rate and suppose that p° lies in the support of W. Then, the generalized posterior contracts at rate ε_n also.
- **Regression:** A similar statement holds for iid observations $Y_i|P_\theta$ satisfying

$$Y_i = \theta(i/n) + Z_i$$

  with $Z_i \sim N(0,\sigma^2)$ and θ belonging to e.g. a Sobolev space.
- **Classification:** Consider iid observations $(Y_1, X_1), \ldots, (Y_n, X_n)$ satisfying

$$\Psi(\theta(x)) = \Pr(Y_i = 1 | X_i = x),$$

  conditionally on P_θ for some link function Ψ and some other assumptions, a divide and conquer approach (using e.g. Gaussian process priors) attains the optimal contraction rate for θ°.

Thank you for listening!

References:

- Botond Szabó and Harry Van Zanten(2019), "An asymptotic analysis of distributed nonparametric methods."
- Amine Hadji, Tammo Hesslink, and Botond Szabó(2022), "Optimal recovery and uncertainty quantification for distributed Gaussian process regression"