An asymptotic analysis of nonparametric divide-and-conquer methods

Botond Szabó and Harry van Zanten

van Dantzig seminar, Delft, 06. 04. 2017.

Table of contents

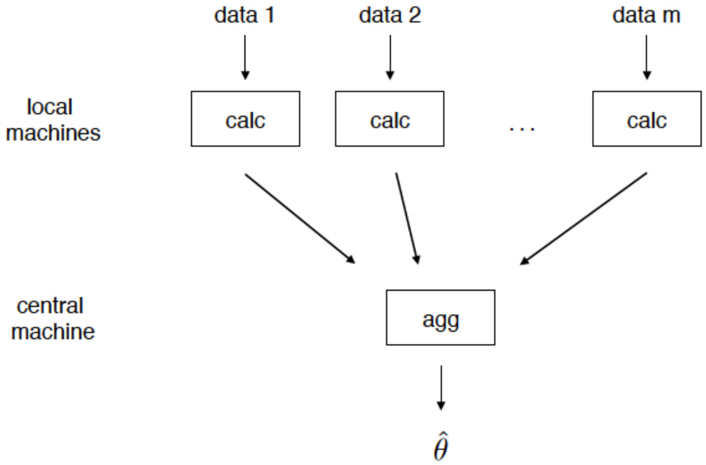
Universiteit Leiden

Distributed methods

Applications

- Volunteer computing (NASA, CERN, SETI,... projects)
- Massive multiplayer online games (peer network)
- Aircraft control systems
- Meteorology, Astronomy
- Medical data from different hospitals

Distributed setting



Universiteit Leiden

Distributed setting II

Interested in high-dimensional and nonparametric models.

- Methods have tunning-, regularity-, sparsity-, bandwidth-hyperparameters to adjust for optimal bias-variance trade-off. How does it work in distributed settings?

Universiteit Leiden

Distributed setting II

Interested in high-dimensional and nonparametric models.

- Methods have tunning-, regularity-, sparsity-, bandwidth-hyperparameters to adjust for optimal bias-variance trade-off. How does it work in distributed settings?
- Several approach in the literature (Consensus MC, WASP, Fast-KRR, Distributed GP,...)
- Limited theoretical underpinning
- No unified framework to compare methods
- Statistical models for illustration:
  - Kernel density estimation,
  - Gaussian white noise model,
  - Random design nonparametric regression.

Universiteit Leiden

Kernel density estimation I

- **Model:** Observe $X_1, ..., X_n \overset{iid}{\sim} f_0$ with $f_0 \in H^{\beta}(L)$.
- **Distributed setting:** distribute data randomly over m machines.
- **Method:**
  - **Local machines:** Kernel density estimation in each

$$\hat{f}_h^{(i)}(x) = \frac{1}{hn/m} \sum_{j=1}^{n/m} K\Big(\frac{x - X_j^{(i)}}{h}\Big).$$

  - **Central machine:** average local estimators

$$\hat{f}_h(x) = \frac{1}{m} \sum_{i=1}^{m} \hat{f}_h^{(i)}(x).$$

Kernel density estimation II

Problem: The choice of the bandwidth parameter h:

- Local bias-variance trade-off:

$$|f_0(x) - E_{f_0}\hat{f}_h^{(i)}(x)| \lesssim h^\beta, \quad \text{and} \quad Var_{f_0}\hat{f}_h^{(i)}(x) \asymp \frac{m}{hn},$$

optimal bandwidth: $h = (n/m)^{-1/(1+2\beta)}$.

Kernel density estimation II

Problem: The choice of the bandwidth parameter h:

- Local bias-variance trade-off:

$$|f_0(x) - E_{f_0}\hat{f}_h^{(i)}(x)| \lesssim h^\beta, \quad \text{and} \quad Var_{f_0}\hat{f}_h^{(i)}(x) \asymp \frac{m}{hn},$$

  optimal bandwidth: $h = (n/m)^{-1/(1+2\beta)}$.

- Global bias-variance trade-off:

$$|f_0(x) - E_{f_0}\hat{f}_h(x)| \lesssim h^\beta, \quad \text{and} \quad Var_{f_0}\hat{f}_h(x) \asymp \frac{1}{hn},$$

  optimal bandwidth: $h = n^{-1/(1+2\beta)}$.

Kernel density estimation II

Problem: The choice of the bandwidth parameter h:

- Local bias-variance trade-off:

$$|f_0(x) - E_{f_0}\hat{f}_h^{(i)}(x)| \lesssim h^\beta, \quad \text{and} \quad Var_{f_0}\hat{f}_h^{(i)}(x) \asymp \frac{m}{hn},$$

  optimal bandwidth: $h = (n/m)^{-1/(1+2\beta)}$.

- Global bias-variance trade-off:

$$|f_0(x) - E_{f_0}\hat{f}_h(x)| \lesssim h^\beta, \quad \text{and} \quad Var_{f_0}\hat{f}_h(x) \asymp \frac{1}{hn},$$

  optimal bandwidth: $h = n^{-1/(1+2\beta)}$.

- Local bias-variance trade-off results too big bias for $\hat{f}_h$: oversmoothing.

Universiteit Leiden

Kernel density estimation II

Problem: The choice of the bandwidth parameter h:

- Local bias-variance trade-off:

$$|f_0(x) - E_{f_0} \hat{f}_h^{(i)}(x)| \lesssim h^{\beta}, \quad \text{and} \quad Var_{f_0} \hat{f}_h^{(i)}(x) \asymp \frac{m}{hn},$$

  optimal bandwidth: $h = (n/m)^{-1/(1+2\beta)}$.

- Global bias-variance trade-off:

$$|f_0(x) - E_{f_0} \hat{f}_h(x)| \lesssim h^{\beta}, \quad \text{and} \quad Var_{f_0} \hat{f}_h(x) \asymp \frac{1}{hn},$$

  optimal bandwidth: $h = n^{-1/(1+2\beta)}$.

- Local bias-variance trade-off results too big bias for $\hat{f}_h$: oversmoothing.
- In practice β is unknown: distributed data-driven methods?

Gaussian white noise model

Single observer:

$$dY_t = f_0(t) + \frac{1}{\sqrt{n}} dW_t, \quad t \in [0, 1].$$

Gaussian white noise model

Single observer:

$$dY_t = f_0(t) + \frac{1}{\sqrt{n}} dW_t, \quad t \in [0,1].$$

Distributed case: m **observer**

$$dY_t^{(i)} = f_0(t) + \sqrt{\frac{m}{n}} dW_t^{(i)}, \quad t \in [0,1], i \in \{1, ..., m\},$$

$W_t^{(i)}$ are independent Brownian motions.

Assumption: $f_0 \in S^{\beta}(L)$, for $\beta > 0$.

Distributed Bayesian approach

- Endow f_0 in each local problem with GP prior of the form

$$f|\alpha \sim \sum_{j=1}^{\infty} j^{-1/2-\alpha} Z_j \phi_j,$$

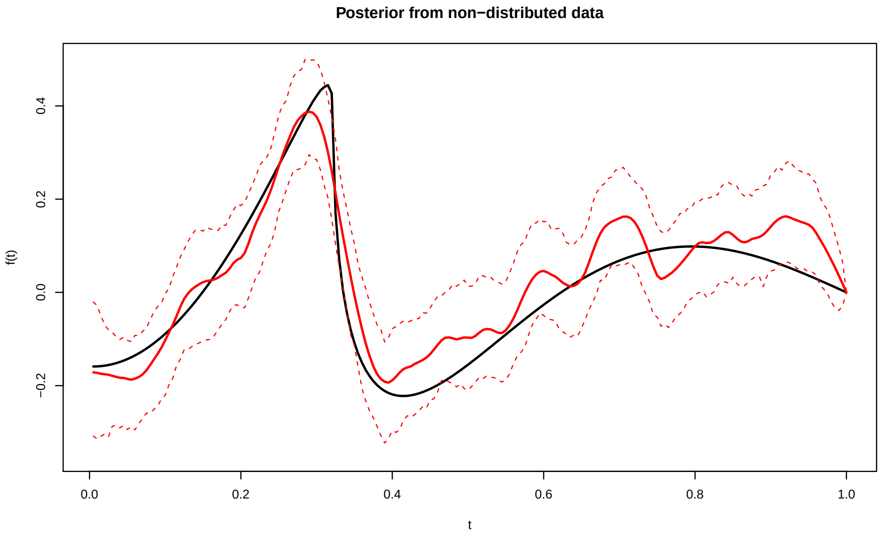
where Z_j are iid $N(0,1)$ and $(\phi_j)_j$ the Fourrier basis.

- Compute locally the posterior (or a modification of it)
- Aggregate the local posteriors into a global one.
- Can we get optimal recovery and reliable uncertainty quantification?

Benchmark: Non-distributed setting I

- One server: $m = 1$.
- Squared bias (of posterior mean): $\|f_0 - E\hat{f}_\alpha\|_2^2 \lesssim n^{-\frac{2\beta}{1+2\alpha}}$
- Variance, posterior spread: $Var(\hat{f}_\alpha) \asymp \sigma^2_{|Y} \asymp n^{-\frac{2\alpha}{1+2\alpha}}$.
- Optimal bias-variance trade-off: at $\alpha = \beta$.

Benchmark: Non-distributed setting II



Distributed naive method

- We have m local machines, with data $(Y^{(1)}, \ldots, Y^{(m)})$.
- Take $\alpha = \beta$.
- Local posteriors:

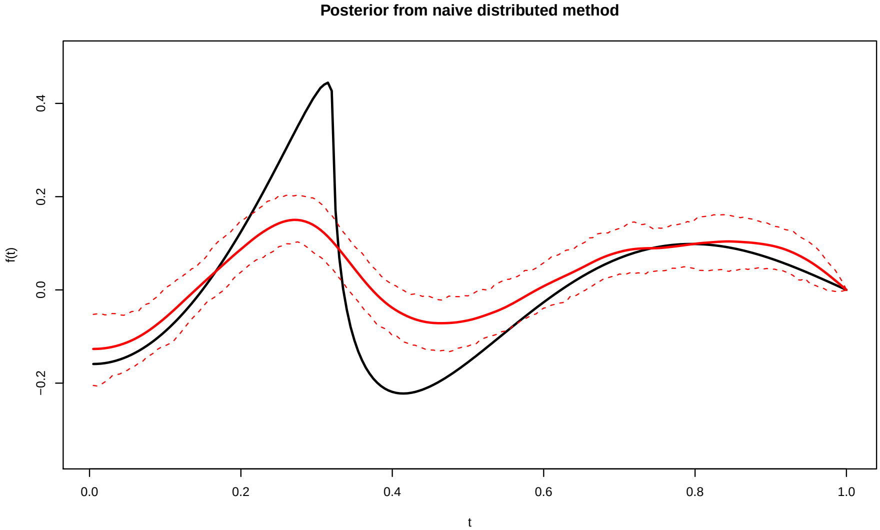
$$\Pi_\beta^{(i)}(f \in B | Y^{(i)}) = \frac{\int_B p_f(Y^{(i)}) d\Pi_\beta(f)}{\int p_f(Y^{(i)}) d\Pi_\beta(f)}.$$

- Aggregate the local posteriors by averaging the draws taken from them.

Result: Sub-optimal contraction, misleading uncertainty quantification.

$$\|f_0 - E\hat{f}\|_2^2 \lesssim (n/m)^{-\frac{2\beta}{1+2\beta}}, \qquad Var(\hat{f}) \asymp \sigma^2_{|Y} \asymp m^{-\frac{1}{1+2\beta}} n^{-\frac{2\beta}{1+2\beta}}.$$

Distributed naive method II



The likelihood approach

- Again m local machines, with data $(Y^{(1)}, ..., Y^{(m)})$ and take $\alpha = \beta$.
- Modify the local likelihoods for each machine:

$$\Pi^{(i)}(f \in B | Y^{(i)}) = \frac{\int_B p_f(Y^{(i)})^m d\Pi(f)}{\int p_f(Y^{(i)})^m d\Pi(f)}.$$

- Aggregate the modified posteriors by averaging the draws taken from them.

The likelihood approach

- Again m local machines, with data $(Y^{(1)}, ..., Y^{(m)})$ and take $\alpha = \beta$.
- Modify the local likelihoods for each machine:

$$\Pi^{(i)}(f \in B | Y^{(i)}) = \frac{\int_B p_f(Y^{(i)})^m d\Pi(f)}{\int p_f(Y^{(i)})^m d\Pi(f)}.$$

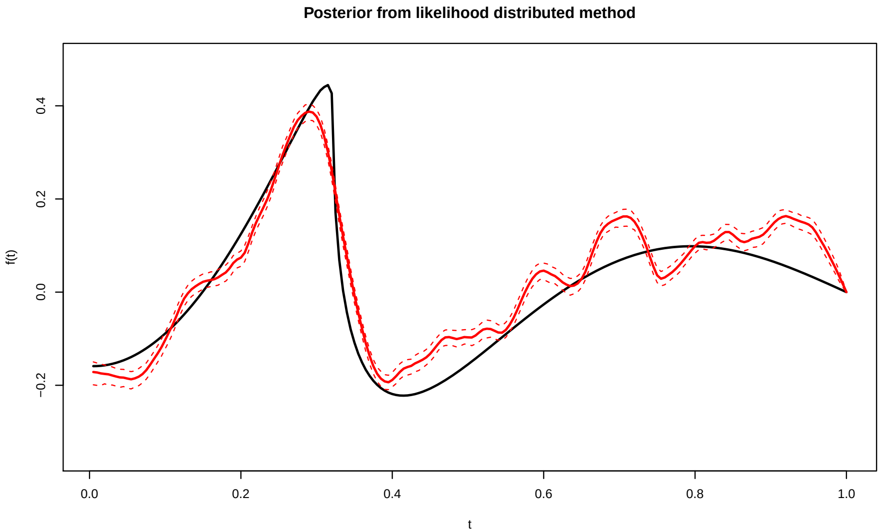
- Aggregate the modified posteriors by averaging the draws taken from them.

Result: Optimal posterior contraction, but bad uncertainty quantification.

$$\|f_0 - E\hat{f}\|_2^2 \lesssim n^{-\frac{2\beta}{1+2\beta}}, \quad Var(\hat{f}) \asymp n^{-\frac{2\beta}{1+2\beta}},, \quad \sigma_{|Y}^2 \asymp m^{-1} n^{-\frac{2\beta}{1+2\beta}}.$$

Universiteit Leiden

The likelihood approach II



The prior rescaling approach

- Again m local machines, with data $(Y^{(1)}, \ldots, Y^{(m)})$.
- Modify the local priors for each machine:

$$\Pi^{(i)}(f \in B | Y^{(i)}) = \frac{\int_B p_f(Y^{(i)}) \pi(f)^{1/m} d\lambda(f)}{\int p_f(Y^{(i)}) \pi(f)^{1/m} d\lambda(f)}.$$

- Aggregate the modified posteriors by averaging the draws taken from them.

The prior rescaling approach

- Again m local machines, with data $(Y^{(1)}, ..., Y^{(m)})$.
- Modify the local priors for each machine:

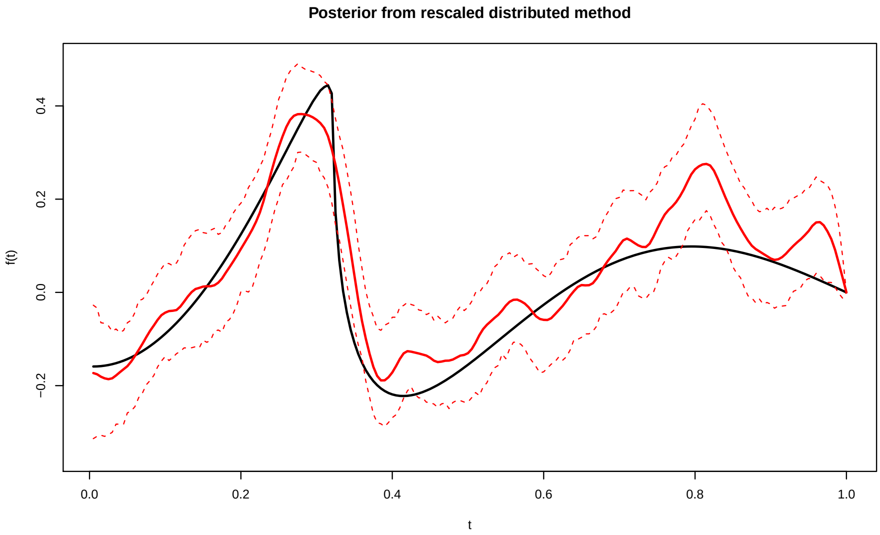
$$\Pi^{(i)}(f \in B | Y^{(i)}) = \frac{\int_B p_f(Y^{(i)})\pi(f)^{1/m} d\lambda(f)}{\int p_f(Y^{(i)})\pi(f)^{1/m} d\lambda(f)}.$$

- Aggregate the modified posteriors by averaging the draws taken from them.

Result: Optimal posterior contraction and uncertainty quantification.

$$\|f_0 - E\hat{f}\|_2^2 \lesssim n^{-\frac{2\beta}{1+2\beta}}, \quad Var(\hat{f}) \asymp \sigma^2_{|Y} \asymp n^{-\frac{2\beta}{1+2\beta}}.$$

The prior rescaling approach II



Other approaches

Methods	posterior contraction rate	coverage
naive, average	sub-optimal	no
naive,Wasserstein	sub-optimal	yes
likelihood, average	minimax	no
likelihood, Wasserstein (WASP)	minimax	yes
scaling, average (consensus MC)	minimax	yes
scaling, Wasserstein	minimax	yes
undersmoothing	minimax (on a range of β, m)	yes (on a range of β, m)
PoE	sub-optimal	no
gPoE	sub-optimal	yes
BCM	minimax	yes
rBCM	sub-optimal	yes

Data-driven methods

Note: All methods above use the knowledge of the true regularity parameter β, which is in practice usually not available.

Solution: Data-driven choice of the regularity-, tunning-hyperparameter.

Data-driven methods

Note: All methods above use the knowledge of the true regularity parameter β, which is in practice usually not available.

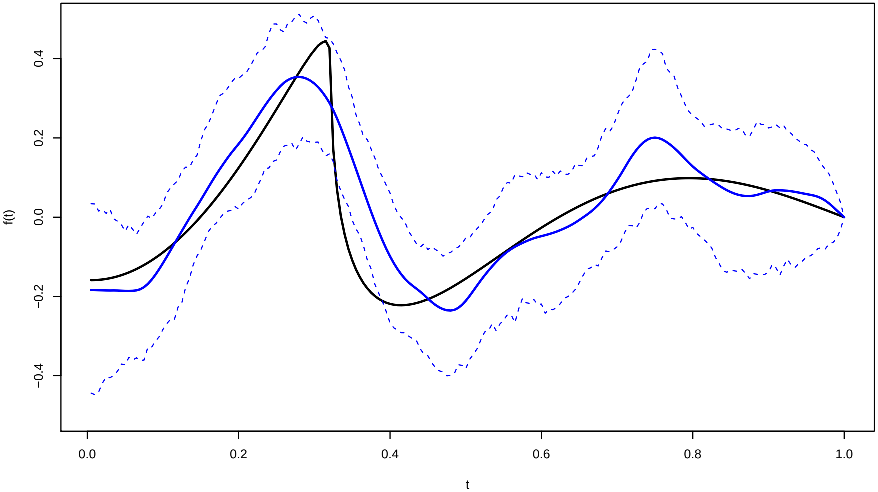
Solution: Data-driven choice of the regularity-, tunning-hyperparameter.

Benchmark: In the non-distributed case $(m = 1)$

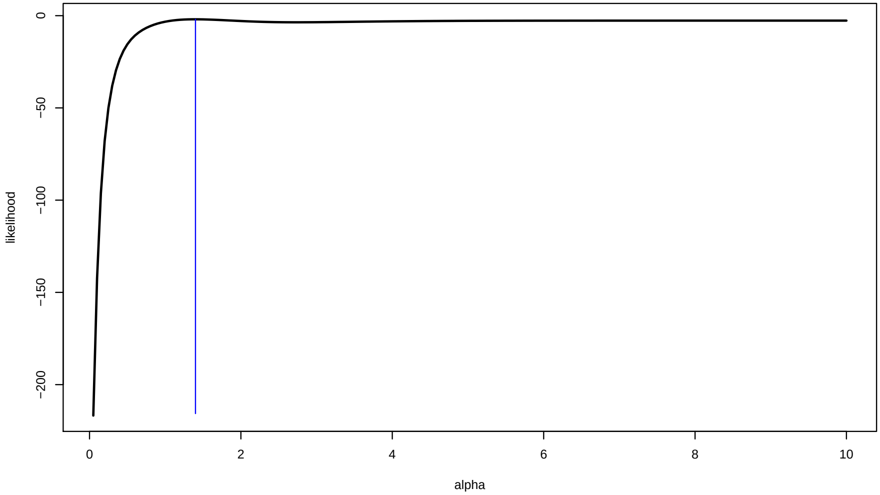
- Hierarchical Bayes: endow α with hyperprior.
- Empirical Bayes: estimate α from the data (marginal maximum likelihood estimator).
- Adaptive minimax posterior contraction rate.
- Coverage of credible sets (under polished tail/self-similarity assumption, using blow-up factors).

Universiteit Leiden

Empirical Bayes posterior



Marginal likelihood



Data driven distributed methods

Proposed methods:

- Naive EB: local MMLE

$$\hat{\alpha}^{(i)} = \arg\max_{\alpha} \int p_f(Y^{(i)}) d\Pi_\alpha(f).$$

- Interactive EB Deisenroth and Ng (2015):

$$\hat{\alpha} = \arg\max_{\alpha} \sum_{i=1}^{m} \log \int p_f(Y^{(i)}) d\Pi_\alpha(f).$$

- Other EB: Lepskii's method $\tilde{\alpha}^{(i)}$ or cross-validation (in the context of ridge regression Zhang, Duchi, Wainwright (2015))

Universiteit Leiden

Counter example

Theorem: Consider $f_0 \in S^\beta(L)$ with Fourrier coefficients

$$f_{0,j}^2 = \begin{cases} j^{-1-2\beta}, & \text{if } j \geq (n/\sqrt{m})^{\frac{1}{1+2\beta}}, \\ 0, & \text{else.} \end{cases}$$

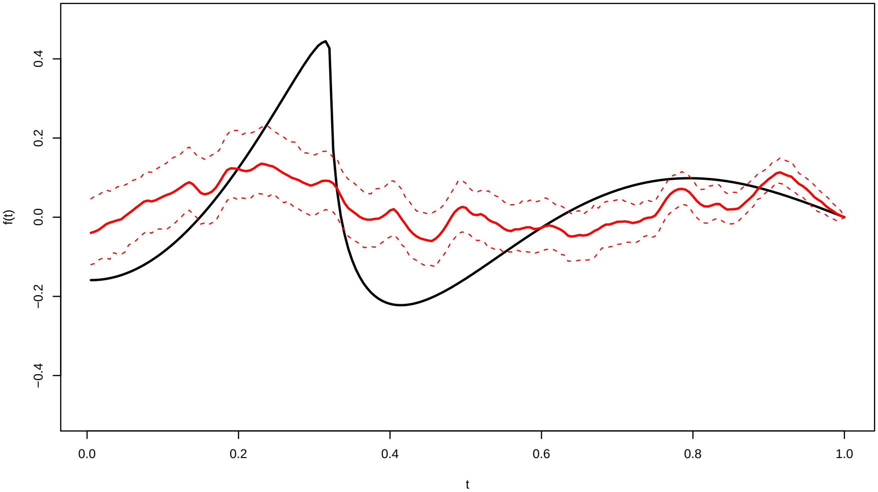
Then for all the above empirical Bayes methods (Naive, Interactive, Lepskii) the regularity hyper-parameter is oversmoothed

$$P(min(\hat{\alpha}^{(i)}, \hat{\alpha}, \tilde{\alpha}^{(i)}) \geq \beta + 1/2) = 1 + o(1).$$

By combining it with any (in non-adaptive case) optimal aggregation methods (above) one gets

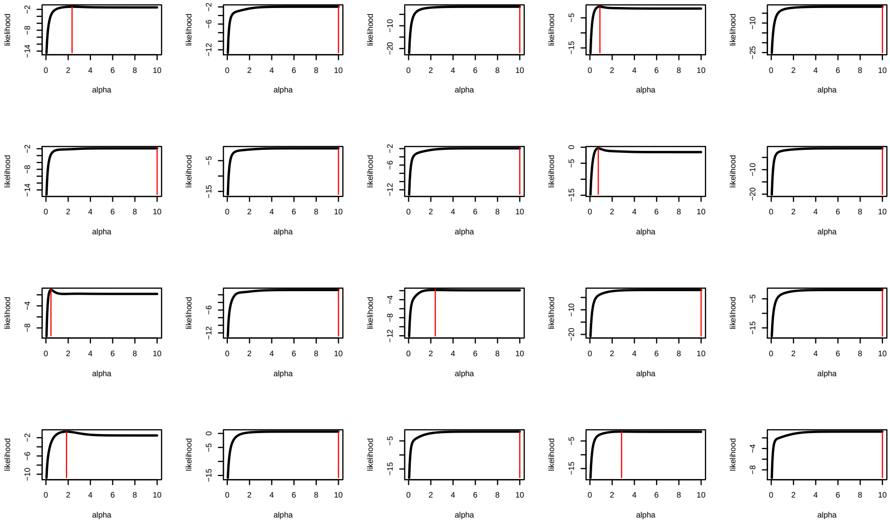
$$\Pi_{aggr,\hat{\alpha}}(f : \|f - f_0\|_2^2 \geq c(n/\sqrt{m})^{-\frac{2\beta}{1+2\beta}} | Y) = 1 + o(1).$$

Aggregated empirical Bayes posterior



Universiteit Leiden

Local marginal likelihoods



Universiteit Leiden

Data-driven methods: constraints

Question: Is it possible to construct data-driven distributed methods with good recovery at all?

Data-driven methods: constraints

Question: Is it possible to construct data-driven distributed methods with good recovery at all?

- Yes: by transferring all data from local machines to central machine and then data-driven method in the centralmachine.

Data-driven methods: constraints

Question: Is it possible to construct data-driven distributed methods with good recovery at all?

- Yes: by transferring all data from local machines to central machine and then data-driven method in the centralmachine.
- BUT this is clearly not what we are looking for...
- In practice there are constraints on the method:
  - **Computational:** in the central machine minimize the amount of computation.
  - **Communication:** as less communication between servers as possible.

Universiteit Leiden

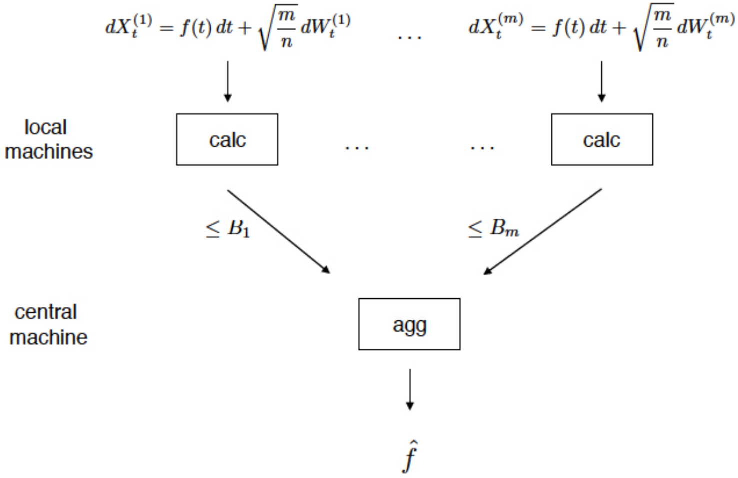
Data-driven methods: constraints

Question: Is it possible to construct data-driven distributed methods with good recovery at all?

- Yes: by transferring all data from local machines to central machine and then data-driven method in the centralmachine.
- BUT this is clearly not what we are looking for...
- In practice there are constraints on the method:
  - **Computational:** in the central machine minimize the amount of computation.
  - **Communication:** as less communication between servers as possible.

New Question: Are there distributed data-driven methods with optimal recovery and **"optimal"** communication/computational costs.

Communication constraints



Communication constraints: minimax rate

- **No restriction** ($B_i = \infty$): back to non-distributed case.
- **No communcation** ($B_i = 0$): no (sensible) inference is possible.
- In parametric models: Zhang et al. (2013). No result in nonparametric models.

Theorem: For $\beta, L > 0$

$$\inf_{\hat{f} \in \mathcal{F}_{dist:B_1,..,B_m}} \sup_{f \in B^{\beta}_{2,\infty}(L)} E_f \|\hat{f} - f\|_2^2 \gtrsim \delta_n^{\frac{2\beta}{1+2\beta}},$$

where δ_n is the solution of

$$\delta_n = \min\left\{ \frac{m}{n \log m}, \frac{m}{n \log m \sum_{i=1}^{m} (\delta_n^{\frac{1}{1+2\beta}} B_i \wedge 1)} \right\}$$

Universiteit Leiden

Communication constraints: minimax rate

- **No restriction** ($B_i = \infty$): back to non-distributed case.
- **No communcation** ($B_i = 0$): no (sensible) inference is possible.
- In parametric models: Zhang et al. (2013). No result in nonparametric models.

Theorem: For $\beta, L > 0$

$$\inf_{f \in \mathcal{F}_{dist:B_1,..,B_m}} \sup_{f \in B^{\beta}_{2,\infty}(L)} E_f \|\hat{f} - f\|_2^2 \gtrsim \delta_n^{\frac{2\beta}{1+2\beta}},$$

where δ_n is the solution of (for $B = B_1 = ... = B_m$)

$$\delta_n = \min\left\{\frac{m}{n\log m}, \frac{1}{n\log m(\delta_n^{\frac{1}{1+2\beta}} B \wedge 1)}\right\}$$

Universiteit Leiden

Remarks

- The proof is via Fano's inequality (using mutual information).
- If $B_i \geq n^{\frac{1}{1+2\beta}}$, then $\delta_n \asymp (\log m)^{\gamma_1}/n$ and the minimax lower bound is $(\log m)^{\gamma_2} n^{-\frac{2\beta}{1+2\beta}}$.
- If $B_i \leq n^{\rho} n^{\frac{1}{1+2\beta}}$ (for some $\rho < 0$), then the lower bound is $n^{\rho_1} n^{-\frac{2\beta}{1+2\beta}}$ (for some $\rho_1 > 0$).
- It is easy to construct estimators, which attain the lower bounds up to logaritmic terms.
- So the optimal communication cost is $B_i = n^{\frac{1}{1+2\beta}}$ (up tp $\log m$ term).
- **Problem:** β is usually not available in practice.

Universiteit Leiden

Adaptive distributed methods - bad news

Question: Is it possible to achieve the minimax (non-distributed) convergence rate and optimal communication at the same time (without knowing β)?

Adaptive distributed methods - bad news

Question: Is it possible to achieve the minimax (non-distributed) convergence rate and optimal communication at the same time (without knowing β)?

Theorem: Let $\beta, L > 0$ be arbitrary. If $m \gg n^{\frac{1}{2+2\beta}}$, then there exist no ideal procedure that can adapt both the transmission rate and the estimation rate uniformly over all $f_0 \in B^{\beta}_{2,\infty}(L)$.

Adaptive distributed methods - bad news

Question: Is it possible to achieve the minimax (non-distributed) convergence rate and optimal communication at the same time (without knowing β)?

Theorem: Let $\beta, L > 0$ be arbitrary. If $m \gg n^{\frac{1}{2+2\beta}}$, then there exist no ideal procedure that can adapt both the transmission rate and the estimation rate uniformly over all $f_0 \in B^{\beta}_{2,\infty}(L)$.

Corollary: Suppose $m = n^p$ for $p \in (0, 1/2)$, let $\beta, L > 0$ be arbitrary. If $\beta > 1/(4p) - 1/2$, then there exist no ideal procedure that can adapt both the transmission rate and the estimation rate uniformly over all $f_0 \in B^{\beta}_{2,\infty}(L)$.

Idea of the proof

One can construct a finite sieve $\mathcal{F} \subset B^{\beta}_{2,\infty}(L)$, such that

- Local machines can not test consistently if $f = 0$ or $f \in \mathcal{F}$ (they are close to 0 and there aren't too many of them).
- The set is large enough, such that the minimax (non-dstributed) rate for estimation is $n^{-\frac{2\beta}{1+2\beta}}$.
- To achieve this rate (up to a logaritmic factor) one has to transmit (in average) $n^{1/(1+2\beta)}$ bits (up to a logaritmic factor).
- Using the number of transmitted bits one could construct tests with higher precision, than possible via the first theoretical limit. Contradiction.

Adaptive distributed methods - good news

Theorem: Assume that $m = n^p$ for $p \in (0, 1/2)$, let $\beta, L > 0$. Then there exists a distributed procedure with transmission rates $\hat{B}_i$ and agrregated estimator $\hat{f}$ such that for all $0 < \underline{\beta} < \overline{\beta} < 1/(4p) - 1/2$

$$\inf_{\underline{\beta} \leq \beta \leq \overline{\beta}} \inf_{f \in B^{\beta}_{2,\infty}(L)} P_f\big(\hat{B}_i \leq C(\log n)^{\delta} n^{\frac{1}{1+2\beta}}\big) \to 1,$$

$$\inf_{\underline{\beta} \leq \beta \leq \overline{\beta}} \inf_{f \in B^{\beta}_{2,\infty}(L)} P_f\big(\|f - \hat{f}\|_2^2 \leq C(\log n)^{\delta} n^{-\frac{2\beta}{1+2\beta}}\big) \to 1.$$

Good news: Idea of the proof I

We show adaptation to two classes indexed with $0 < \beta_1 < \beta_2 < 1/(4p) - 1/2$ (adaptation for continuum classes can be done by introducing a grid).

Good news: Idea of the proof I

We show adaptation to two classes indexed with $0 < \beta_1 < \beta_2 < 1/(4p) - 1/2$ (adaptation for continuum classes can be done by introducing a grid).

Local machines:

- Split data into two iid parts (twice the variance)
- Using the first part construct consistent test ϕ for

$$H_0 : f \in B^{\beta_2}_{2,\infty}(L) \quad \text{vs} \quad H_a : f \in \{f \in B^{\beta_1}_{2,\infty}(L) : \|f - B^{\beta_2}_{2,\infty}(L)\|_2^2 \geq (\frac{n}{m})^{-\frac{\beta_1}{1/2+2\beta_1}}\}.$$

- Turn the test into an estimator for the smoothness

$$\hat{\beta}^{(i)} = \begin{cases} \beta_1, & \text{if } \phi = 0, \\ \beta_2, & \text{if } \phi = 1. \end{cases}$$

- Transmit $\log n$ bits of the first $\hat{N}^{(i)} = n^{\frac{1}{1+2\hat{\beta}^{(i)}}}$ wavelet coefficients of $Y_t^{(i)}$.

Good news: Idea of the proof II

Central machine:

- Compute the median number of transmitted coefficients: $\hat{N}$.
- Define estimator:

$$\hat{f}_{j,k} = \begin{cases} \frac{1}{N_{j,k}} \sum_{i \in N_{j,k}} Y_{j,k}^{(i)}, & \text{if } 2^j \leq \hat{N}, \\ 0, & \text{else,} \end{cases}$$

where $Y_{j,k}^{(i)}$ is the (first $\log n$ bits) of the (j,k)th wavelet coefficient of $Y_t^{(i)}$, $\hat{f}_{j,k}$ the (j,k)th wavelet coefficient of $\hat{f}$, and $N_{j,k} = \{1 \leq i \leq m : \hat{N}^{(i)} \geq 2^j\}$.

Summary

- Several distributed methods proposed in the literature (Bayes and frequentist).
- Compared them on a unified framework (distributed Gaussian white noise).
- Investigated standard data-driven methods: do not work.
- Theoretical limitations: under communication constraints.
- Only on a range of regularity classes exists adaptive estimator with optimal communication costs (in L_2).

Universiteit Leiden

Further results/Ongoing work

- For $f_0 \in B^{\beta}_{\infty,\infty}$ and L_∞ doesn't exist an adaptive procedure (not even on a limited range).
- Under self-similarity assumption there exists an adaptive procedure.
- Similar results can be derived for random design regression (technically more demanding): ongoing.
- Uncertainty quantification in adaptive setting: ongoing.
- Computational constraints: NP vs P, quadratic, linear algorithms: future.
- Combining computational and communication constraints: future.
- General theorem (both Bayesian and non-bayesian): future.