The Pursuit of Truth in a Big (Data) World

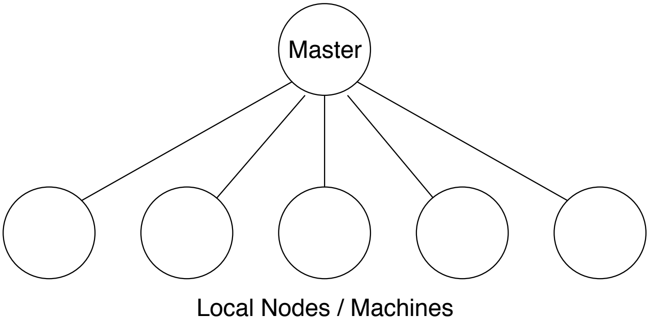
Bobby He, Hector McKimm, Deborah Sulem, James Thornton

February 27, 2019

Motivation and Introduction

Problem: Big data is becoming too large for one machine (in practice).
Solutions: two main methods:

1. Divide and Conquer:
   multi-machine / multi-core approach
2. Sub-sampling [Bardenet et al., 2017]:
   decrease number of individual data point likelihood evaluations



Signal-in-White-Noise Model

Distributed setting:

- n observations
- m machines
- Machine j has data $Y_1^j, Y_2^j, \dots$

$$Y_i^j = \theta_i + \sqrt{\frac{\sigma^2 m}{n}} Z_i \quad ; \quad Z_i \sim \mathcal{N}(0,1).$$

Non-distributed setting:

$$Y_i = \theta_i + \sqrt{\frac{\sigma^2}{n}} \tilde{Z}_i \quad ; \quad \tilde{Z}_i \sim \mathcal{N}(0,1)$$

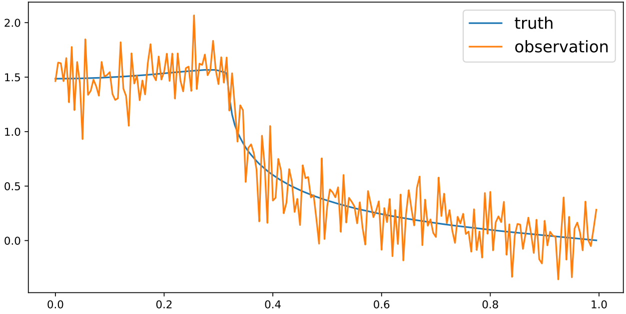
Prior:

$$\theta_i \overset{\text{iid}}{\sim} \mathcal{N}(0, i^{-1-2\alpha}) \quad \forall i$$

Illustrative example

$$f(x) = \sum_{i=1}^{\infty} \theta_i \times \cos(\pi(i - \frac{1}{2})x)$$

$$\text{truth: } \theta_{0,i} = \frac{\sin i}{i^{3/2}}$$



Inference for the Signal-in-White-Noise Model

Definition (Hyper Rectangle)

For $\beta, M > 0$, let $\mathcal{H}_{\beta,M} = \{\theta \in \ell^2 : \sup_i (i^{1+2\beta}\theta_i^2) \leq M^2\}$

In non-distributed case:

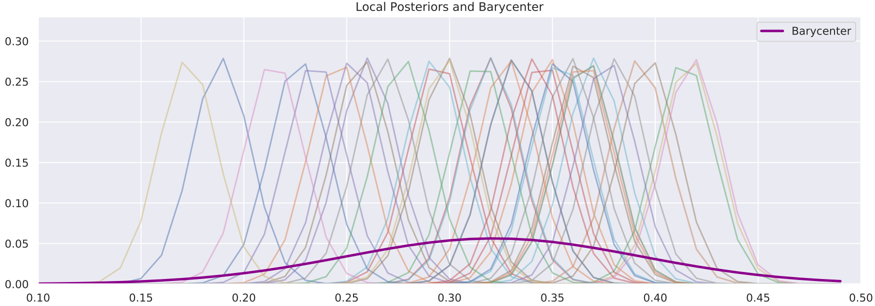
- If $\theta_0 \in \mathcal{H}_{\beta,M}$ for known β, optimal posterior contraction rate: $\mathcal{O}(n^{-\beta/(1+2\beta)})$
- For unknown β there exists adaptive estimators with same optimal rate, [Knapik et al., 2016], - more to come!

In the distributed case:

- When β known, [Szabo and van Zanten, 2017] present 2 successful methods to aggregate local posteriors.
- When β unknown: trickier, also see later!

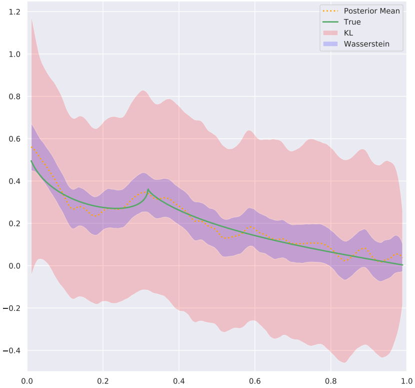
Non-adaptive Method: Adjusted local likelihoods with Wasserstein Barycenters

- Adjust local likelihoods to mimic sample size n on local machines.
- Compute Wasserstein Barycenter $\bar{\mu}_W = \underset{\mu \in I^2}{\operatorname{argmin}} \frac{1}{m} \sum_{j=1}^{m} W_2^2(\mu, \mu_j)$

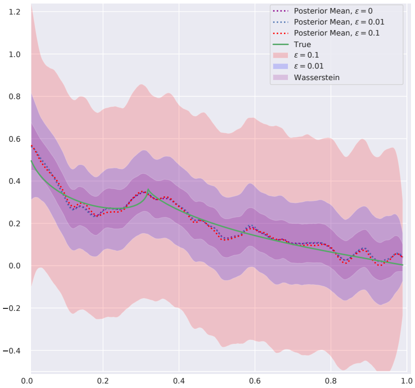


Non-adaptive extensions

Kullback-Leibler Barycenter



Sinkhorn Barycenter



Adaptive method from [Deisenroth and Ng, 2015]

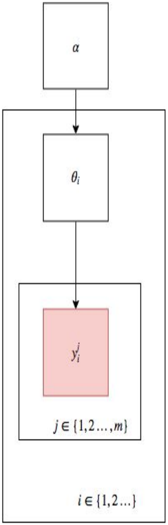
[Deisenroth and Ng, 2015] propose to approximate the map:

$$\hat{\alpha} = \underset{\alpha}{argmax} \log \int \prod_{j=1}^{m} \left(p(Y^j|\theta)\right) \Pi(d\theta|\alpha)$$

by:

$$\tilde{\alpha} = \underset{\alpha}{argmax} \sum_{j=1}^{m} \log \left(\int p(Y^j|\theta)\Pi(d\theta|\alpha) \right),$$

Then optimize the sum of local functions: repeated function and gradient evaluations



Adaptive methods

3 possible solutions:

- Restrict the class of signals
- Use a prior on α
- Find another approximation of the global marginal likelihood

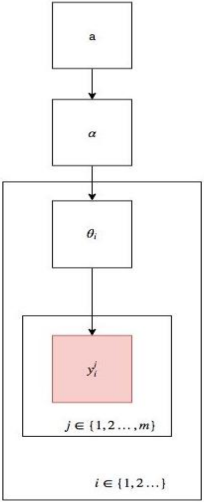
Hierarchical Prior

Consider a Hierarchical prior on θ:

$\forall i \in \mathbb{N}, \forall j \in \{1, \ldots, m\}$:

$$\alpha \sim \lambda(\cdot|a),$$
$$\theta \sim \bigotimes_{i=1}^{\infty} \mathcal{N}(0, \tau i^{-1-2\alpha})$$
$$y_i^j \sim \mathcal{N}\left(\theta, \sigma^2 \frac{n}{m}\right).$$

- Mixture prior given by $\int \Pi(\cdot|\alpha)\lambda(d\alpha|a)$
- Theorem 3 of [Knapik et al., 2016]
- Communication cost and MCMC cost
  - Update α via e.g. rejection sampling
  - Update θ via divide and conquer



Communication-efficient Surrogate Likelihood ([Jordan et al., 2016])

For parametric inference ($\Theta \subset \mathbf{R}^d$) in a distributed setting, with a loss function $\mathcal{L} : \Theta \times \mathcal{Z} \to \mathbf{R}$

$$\text{(global)} \quad \mathcal{L}_n(\theta) = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n/m} \mathcal{L}(\theta, z_i^j)$$

$$\text{(local)} \quad \mathcal{L}_j(\theta) = \frac{m}{n} \sum_{i=1}^{n/m} \mathcal{L}_j(\theta, z_i^j) \quad \text{for } 1 \leq j \leq m$$

Surrogate loss function:

$$\tilde{\mathcal{L}}(\theta) := \mathcal{L}_1(\theta) - < \theta, \nabla \mathcal{L}_1(\bar{\theta}) - \nabla \mathcal{L}_n(\bar{\theta}) >$$

where $\bar{\theta} = \arg\min_\theta \mathcal{L}_1(\theta)$

Surrogate marginal likelihood

Motivations:

- Parametric inference for α
- Good convergence properties of the estimator $\tilde{\theta} = \arg\min_\theta \tilde{\mathcal{L}}(\theta)$
- Low communication cost ($O(nd)$) and easy to implement

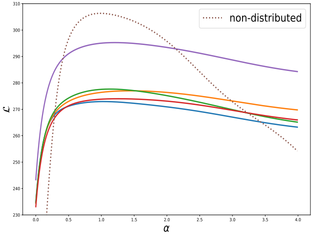
Surrogate marginal log likelihood:

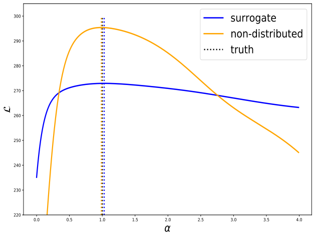
$$\tilde{\ell}(\alpha) := \ell_1(\alpha|Y^1) - \alpha \times (\ell_1'(\bar{\alpha}|Y^1) - \ell'(\bar{\alpha}|Y_{1:n}))$$

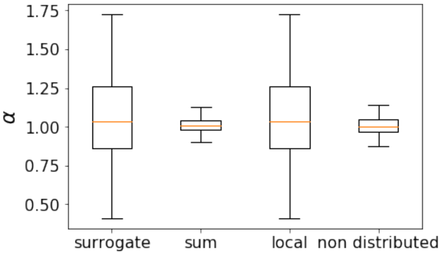
- Affine function of a local marginal likelihood
- Additional approximation of the derivative at $\bar{\alpha} = \arg\max_\alpha \ell_1(\alpha|Y^1)$:

$$\ell'(\bar{\alpha}|Y_{1:n}) \approx \frac{1}{m} \sum_{j=1}^{m} \ell_j'(\bar{\alpha}|Y^j)$$

Surrogate marginal likelihood: Simulation







- Signal with regularity $\beta = 1$
- Dataset of $n = 4000$ observations divided into $m = 40$ machines

Conclusion

- Few theoretical results on adaptive methods in distributed, non-parametric inference
- Failure of some methods used in the non-distributed case

- Add assumptions on the smoothness of the signal
- Enlarge results to other types of prior used in the non-distributed case (uniform)

- Comparison between divide-and-conquer approaches and subsampling

Bardenet, R., Doucet, A., and Holmes, C. (2017).
On markov chain monte carlo methods for tall data.
The Journal of Machine Learning Research, 18(1):1515–1557.

Deisenroth, M. P. and Ng, J. W. (2015).
Distributed gaussian processes.
arXiv preprint arXiv:1502.02843.

Jordan, M., Lee, J., and Yang, Y. (2016).
Communication-efficient distributed statistical inference.
arXiv preprint arXiv:1711.03149.

Knapik, B., Szabó, B., van der Vaart, A., and van Zanten, J. (2016).
Bayes procedures for adaptive inference in inverse problems for the white noise model.
Probability Theory and Related Fields, 164(3-4):771–813.

Szabo, B. and van Zanten, H. (2017).
An asymptotic analysis of distributed nonparametric methods.
arXiv preprint arXiv:1711.03149.

Adaptive method from [Deisenroth and Ng, 2015]

Pathological signal $\theta_0 \in \mathcal{H}_{\beta,M}$:

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

If M is small enough then if $n/m \to \infty$ and $m \to \infty$,

$$\mathbf{P}_{\theta_0}(\tilde{\alpha} \geq \beta + 1/2) \to 1$$

- Overestimates the regularity of the signal
- Sub-optimal rates of convergence and bad coverage probabilities of the aggregated posteriors with the previous "good" methods

Non-adaptive Method 1: Adjusted Prior with Naive Averaging

Assume $\beta \leq 1 + 2\alpha$ and define a new prior:

$$\theta_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \tau i^{-1-2\alpha}) \quad \forall i \quad \text{where} \quad \tau = mn^{\frac{2(\alpha-\beta)}{1+2\beta}}$$

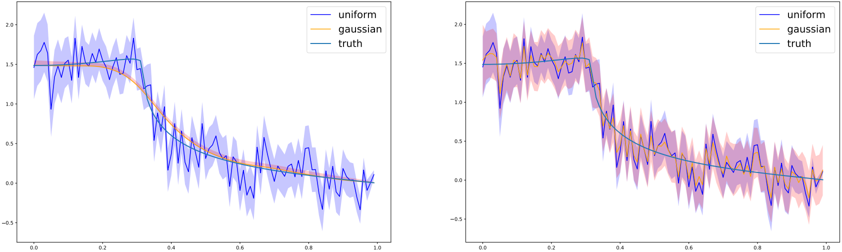


Figure: Estimated signal and pointwise 95% credible intervals with a uniform prior and a Gaussian prior using adjusted likelihoods (left panel) and adjusted priors (right panel)