Scalable Stochastic Gradient-based Inference for Gaussian Processes

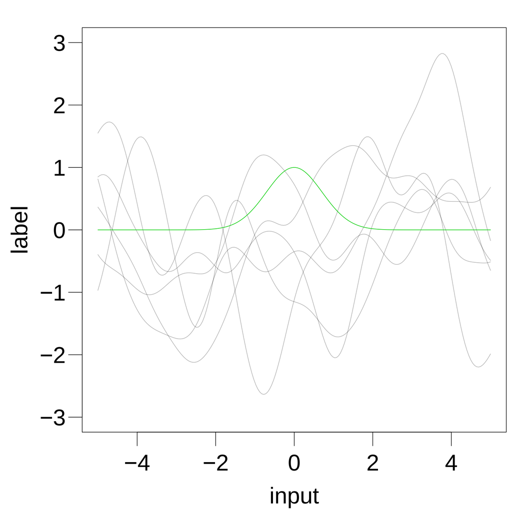
Maurizio Filippone

EURECOM, Sophia Antipolis, France

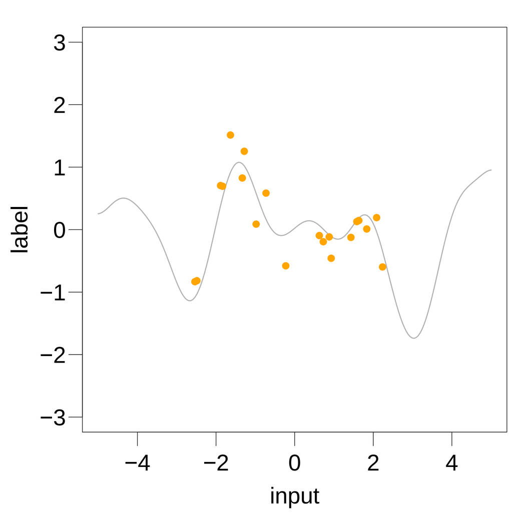
email: maurizio.filippone@eurecom.fr

Gaussian Process (GP) Regression - Illustration

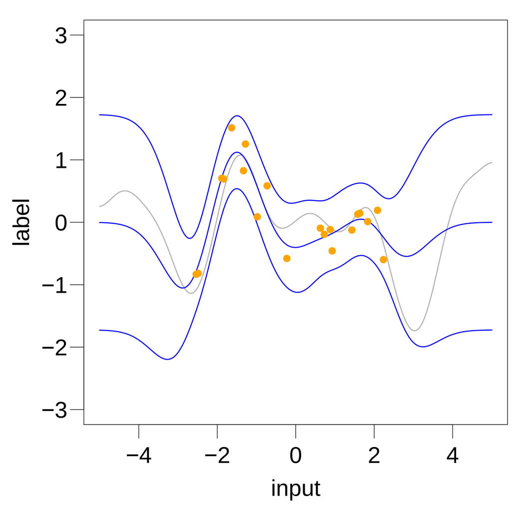
GP prior



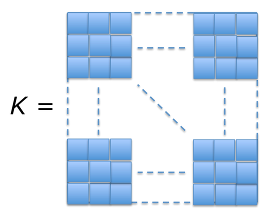
GP regression example



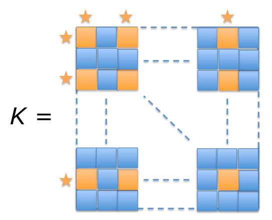
Inference result



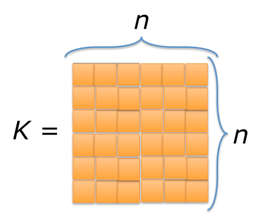
- Grey - Samples from the GP prior
- Green - Radial Basis Function covariance
- Full and "infinite" covariance matrix $K \in M_{\infty,\infty}$



- Data generated from the model
- Marginal distributions of multivariate Gaussian are Gaussian
- It is sufficient to look at the values of the covariance at the input locations



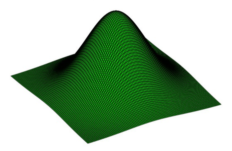
- Blue - Mean prediction $\pm$ 2 Std devs
- Conditional distributions of multivariate Gaussian are Gaussian
- Predictions can be made by calculating conditional distributions



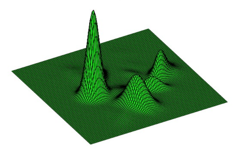
Bayesian Inference for GPs

$$p(\text{par}|\text{data}) = \frac{p(\text{data}|\text{par})p(\text{par})}{\int p(\text{data}|\text{par})p(\text{par})d\text{par}}$$

$p(\text{par})$



$p(\text{par}|\text{data})$



Marginal likelihood

- Marginal likelihood

$$p(\text{data}|\text{par}) = \int p(\text{data}|\text{latent})p(\text{latent}|\text{par})d\text{latent}$$

can only be computed if $p(\text{data}|\text{latent})$ is Gaussian

- ... even then

$$\log[p(\text{data}|\text{par})] = -\frac{1}{2}\log|K| - \frac{1}{2}\mathbf{y}^{\mathrm{T}}K^{-1}\mathbf{y} + \text{const.}$$

where $K = K(\text{par})$ is generally an $n \times n$ dense matrix!

Stochastic Gradient Langevin Dynamics (SGLD) algorithm

- Stochastic gradient ascent optimization with injected noise η_t

$$\text{par}' = \text{par} + \frac{\alpha_t}{2}\widetilde{\nabla_{\text{par}}}\log[p(\text{data}|\text{par})p(\text{par})] + \eta_t \qquad \eta_t \sim \mathcal{N}(0, \alpha_t) \qquad \alpha_t \to 0$$

- First phase – α_t large – Optimization phase
  - Injected noise η_t is smaller than the gradient-based update
  - Behavior similar to stochastic gradient ascent
- Second phase – α_t small – Langevin dynamics phase
  - Injected noise η_t dominates gradient-based update
  - ✓ Acceptance rate reaches one so no need to accept/reject
  - ✓ No need to evaluate $p(\text{data}|\text{par})$
  - ✓ We only need stochastic gradients to obtain samples from $p(\text{par}|\text{data})$

Stochastic gradients for GPs

- Marginal likelihood

$$\log[p(\text{data}|\text{par})] = -\frac{1}{2}\log|K| - \frac{1}{2}\mathbf{y}^{\mathrm{T}}K^{-1}\mathbf{y} + \text{const.}$$

- Derivatives wrt par

$$\frac{\partial \log[p(\text{data}|\text{par})]}{\partial \text{par}_i} = -\frac{1}{2}\mathrm{Tr}\left(K^{-1}\frac{\partial K}{\partial \text{par}_i}\right) + \frac{1}{2}\mathbf{y}^{\mathrm{T}}K^{-1}\frac{\partial K}{\partial \text{par}_i}K^{-1}\mathbf{y}$$

- Stochastic estimate of the trace

$$\mathrm{Tr}\left(K^{-1}\frac{\partial K}{\partial \text{par}_i}\right) = \mathrm{Tr}\left(K^{-1}\frac{\partial K}{\partial \text{par}_i}\mathrm{E}[\mathbf{r}\mathbf{r}^{\mathrm{T}}]\right) = \mathrm{E}\left[\mathbf{r}^{\mathrm{T}}K^{-1}\frac{\partial K}{\partial \text{par}_i}\mathbf{r}\right]$$

with $\mathrm{E}[\mathbf{r}\mathbf{r}^{\mathrm{T}}] = I$ - e.g., r_j drawn from $\{-1, 1\}$ with $p = 1/2$

- Stochastic gradient

$$-\frac{1}{2N_{\mathbf{r}}}\sum_{i=1}^{N_{\mathbf{r}}}\mathbf{r}^{(i)\mathrm{T}}K^{-1}\frac{\partial K}{\partial \text{par}_i}\mathbf{r}^{(i)} + \frac{1}{2}\mathbf{y}^{\mathrm{T}}K^{-1}\frac{\partial K}{\partial \text{par}_i}K^{-1}\mathbf{y}$$

- Linear systems only!

Solving linear systems

- Linear systems:

$$K\mathbf{s} = \mathbf{b}$$

- Can be solved using the Conjugate Gradient algorithm:

$$\mathbf{s} = \arg\min_{\mathbf{x}}\left(\frac{1}{2}\mathbf{x}^{\mathrm{T}}K\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{b}\right)$$

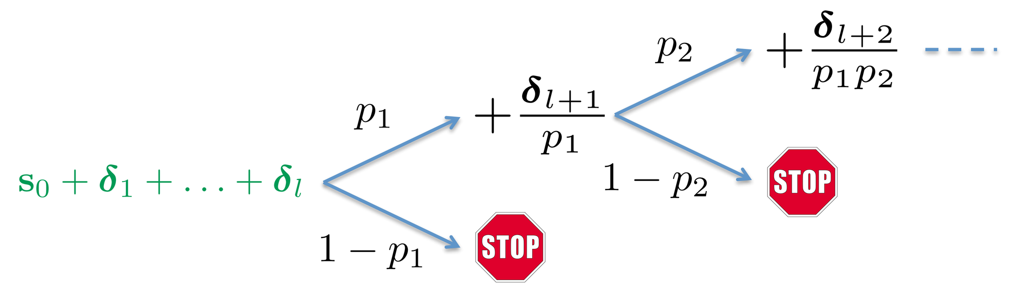
- Iterative update $\mathbf{s} = \mathbf{s}_0 + \boldsymbol{\delta}_1 + \ldots + \boldsymbol{\delta}_T$
- Requires only Covariance Matrix Vector Products (CMVPs)! $O(n^2)$ time
- No need to store K! $O(n)$ space

ULISSE - the Unbiased LInear System SolvEr

- Accelerate the solution of dense linear systems
- ... returning an unbiased estimate of the solution
- Full CG solution:

$$\mathbf{s} = \mathbf{s}_0 + \boldsymbol{\delta}_1 + \ldots + \boldsymbol{\delta}_l + \boldsymbol{\delta}_{l+1}\ldots + \boldsymbol{\delta}_T$$

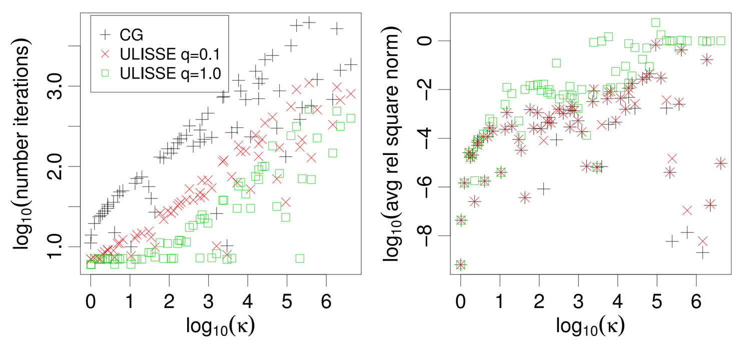
- ULISSE:



In this work:

$p_i = \exp(-\beta i)$

- Final solution is an unbiased estimate of s!



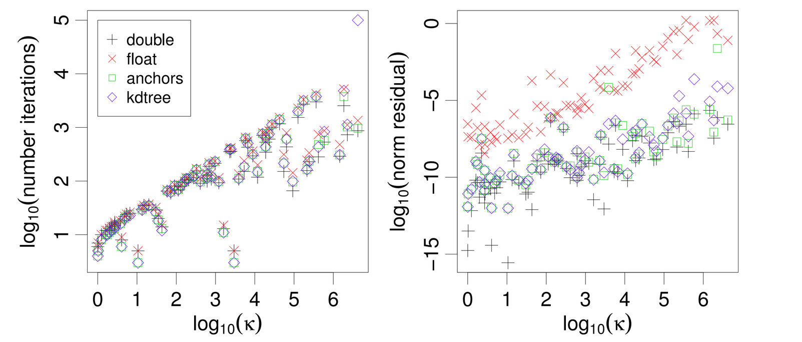
✓ Fast computation of stochastic gradients

✓ Small relative error wrt exact gradients

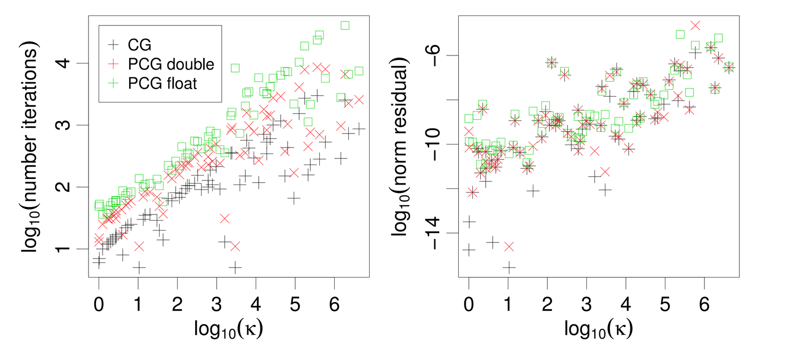
$$\text{rel square norm} = \frac{\|\mathbf{g}(\boldsymbol{\theta}) - \tilde{\mathbf{g}}(\boldsymbol{\theta})\|^2}{\|\mathbf{g}(\boldsymbol{\theta})\|^2}$$

Traditional solvers vs ULISSE

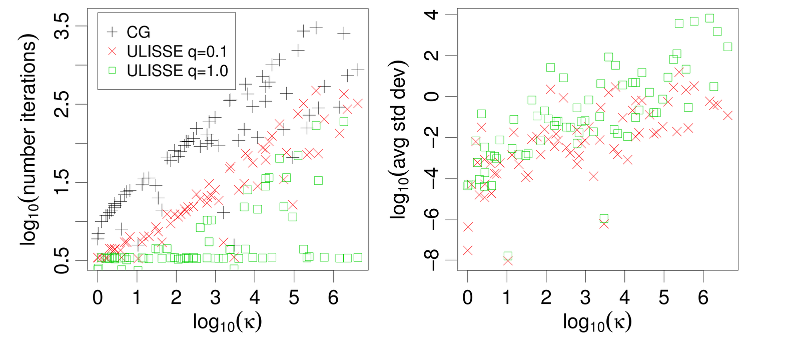
Fast CMVPs



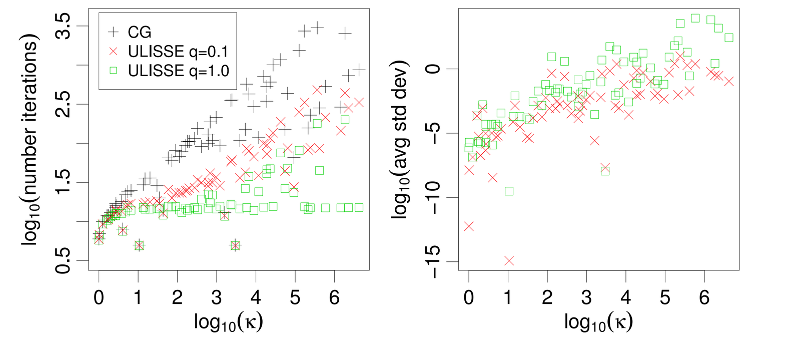
Preconditioning



CG vs ULISSE with $\beta = 1$

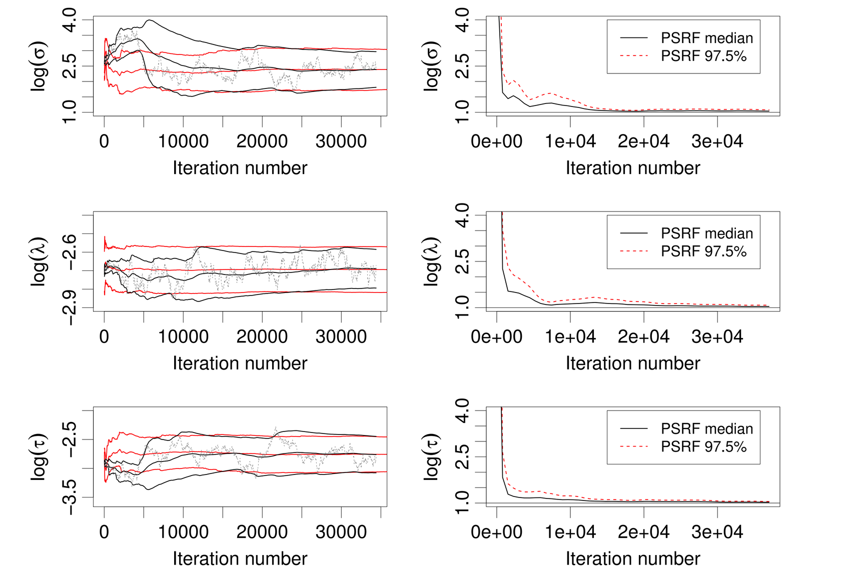


CG vs ULISSE with $\beta = 100$

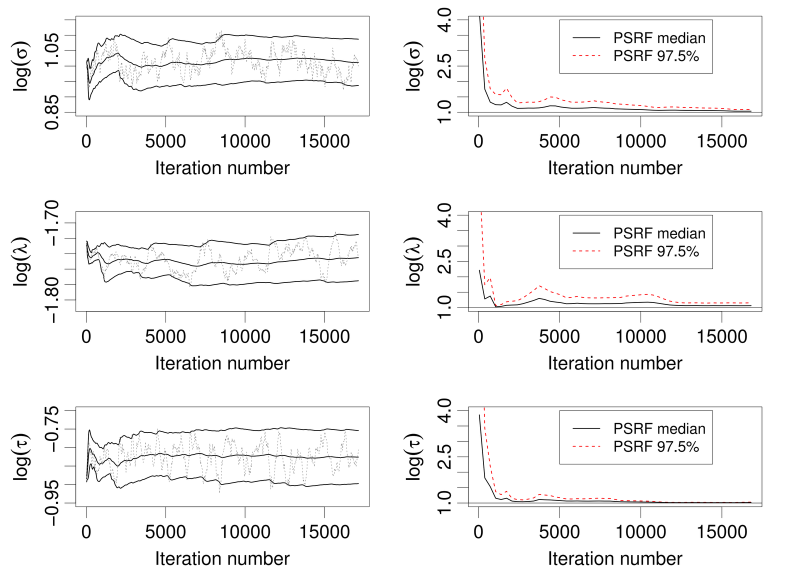


Inference Results

Concrete data $n \approx 1K$



Census data $n \approx 23K$



Conclusions

- Novel adaptation of SGLD to infer covariance parameters in Gaussian processes
  - ✓ Accurate in characterizing the posterior distribution over covariance parameters
  - ✓ Scales with $O(n)$ in space and with $O(n^2)$ in time
  - ✓ Massively parallelizable
  - ✓ Without assuming factorization of the likelihood (mini-batches)
  - ✓ Without considering subsets of the data or inducing points
  - ✓ Without considering subsets of the spectrum of the covariance
  - ✓ Without imposing sparsity on the covariance or its inverse
- Novel linear solver - ULISSE
  - ✓ Early stop of iterative linear solver that yields an unbiased solution
  - ✓ Can be adopted to accelerate **any** iterative solver
- Ongoing work
  - Extension to Gaussian Markov Random Fields and other likelihoods
  - Tuning of a preconditioner in SGLD
  - Mixed precision calculations within the Conjugate Gradient algorithm

References

[1] M. Filippone and R. Engler, Enabling scalable stochastic gradient-based inference for Gaussian processes by employing the Unbiased LInear System SolvEr (ULISSE). *ICML 2015*, to appear.

[2] M. Filippone and M. Girolami, Pseudo-marginal Bayesian inference for Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2214–2226, 2014.

[3] M. Filippone et al. Probabilistic Prediction of Neurological Disorders with a Statistical Assessment of Neuroimaging Data Modalities. *Annals of Applied Statistics*, 6(4):1883–1905, 2012.

[4] M. N. Gibbs, *Bayesian Gaussian processes for regression and classification*. PhD thesis, University of Cambridge, 1997.

[5] M. Welling and Y. W. Teh, Bayesian Learning via Stochastic Gradient Langevin Dynamics. *ICML 2011*, pp. 681–688. Omnipress, 2011.