Enabling scalable stochastic gradient-based inference for Gaussian processes by employing the Unbiased LLinear System SolvEr (ULISSE)

Maurizio Filippone

EURECOM, Sophia Antipolis, France & University of Glasgow, Glasgow, UK

Maurizio.Filippone@eurecom.fr

ICML 2015, Lille
July 9th, 2015
Gaussian Processes

\[ K = \]

\[
\begin{bmatrix}
K_{11} & K_{12} & \ldots & K_{1n} \\
K_{21} & K_{22} & \ldots & K_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
K_{n1} & K_{n2} & \ldots & K_{nn}
\end{bmatrix}
\]
Bayesian Inference

- Inputs = \( X \)  
- Labels = \( y \)
- \( K = K(X, \text{par}) \)

\[
p(par|y) = \frac{p(y|par)p(par)}{\int p(y|par)p(par)dpar}
\]
Acceptance probability: \[
\min\left(1, \frac{p(y|\text{par}')p(\text{par}')}{p(y|\text{par})p(\text{par})}\right)
\]
Gaussian likelihood case

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

where \( K = K(X, \text{par}) \) is an \( n \times n \) dense matrix!
1 INTRODUCTION

Non-parametric or kernel based models are a successful class of statistical modelling methods. To focus ideas throughout the paper we used an example of a multivariate Gaussian process (GP) in particular. Implications of kernel-based classifier models are simple GP classification and function classification. The Gaussian process classifier [1] is a model that is based on different model selection and paradigms of statistical inference, the importance of a kernel function or covariance is assessed by a kernel function or covariance or analogous to build nonlinear classifiers based on kernel methods. Pseudo-marginal approaches to Markov chain Monte Carlo methods are employed to perform posterior inference over the model. This paper reports a statistical assessment of multiple neuroimaging modalities applied to the discrimination of three Parkinsonian neurological disorders from one another and healthy controls, showing promising predictive performance of disease states when compared to nonprobabilistic classifiers based on multiple modalities. The statistical analysis also quantifies the relative importance of different neuroimaging measures and brain regions in discriminating between these diseases and suggests that for prediction there is little benefit in acquiring multiple neuroimaging sequences. Finally, the predictive capability of different
Gradient ascent

\[
\text{par}' = \text{par} + \frac{\alpha}{2} \nabla_{\text{par}} \log[p(y|\text{par})p(\text{par})]
\]
Stochastic Gradient ascent

\[
E\left\{ \widehat{\nabla}_{\text{par}} \log[p(y|\text{par})] \right\} = \nabla_{\text{par}} \log[p(y|\text{par})]
\]

Robbins and Monro, AoMS, 1951
\[ \text{par}' = \text{par} + \frac{\alpha_t}{2} \nabla_{\text{par}} \log[p(y|\text{par})p(\text{par})] \quad \alpha_t \to 0 \]

Robbins and Monro, *AoMS*, 1951
Stochastic Gradient Langevin Dynamics (SGLD) algorithm

\[ \text{par}' = \text{par} + \frac{\alpha_t}{2} \nabla_{\text{par}} \log[p(y|\text{par})p(\text{par})] + \eta_t \quad \eta_t \sim \mathcal{N}(0, \alpha_t) \]

Welling and Teh, *ICML*, 2011
Traditionally, in SGLD stochastic gradients

\[ \nabla_{\text{par}} \log [p(y|\text{par})p(\text{par})] \]

are computed based on mini-batches of data

In GPs the likelihood DOES NOT factorize

What can we do?
Marginal likelihood

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

Derivatives wrt $\text{par}$

$$\frac{\partial \log[p(y|\text{par})]}{\partial \text{par}_i} = -\frac{1}{2} \text{Tr} \left( K^{-1} \frac{\partial K}{\partial \text{par}_i} \right) + \frac{1}{2} y^T K^{-1} \frac{\partial K}{\partial \text{par}_i} K^{-1} y$$
Stochastic Gradients in GP regression

- Stochastic estimate of the trace

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

with \( \mathbb{E}[rr^T] = I \)

- For example \( r_j \) drawn from \( \{-1, 1\} \) with \( p = 1/2 \)
Stochastic Gradients in GP regression

- Stochastic estimate of the trace

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

with $\mathbb{E}[rr^T] = I$

- For example $r_j$ drawn from $\{-1, 1\}$ with $p = 1/2$

- Stochastic gradient

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

- Linear systems only!

Filippone and Engler, *ICML*, 2015
Solving linear systems

- Linear systems:
  \[ Ks = b \]

- Can be solved using conjugate gradient:
  \[ s = \arg \min_x \left( \frac{1}{2} x^T K x - x^T b \right) \]

- Iterative update \( s = s_0 + \delta_1 + \ldots + \delta_T \)

- Requires only \( Kv \) multiplications! \( O(n^2) \) time

- No need to store \( K \)! \( O(n) \) space

Filippone and Engler, *ICML*, 2015
Accelerate the solution of dense linear systems

... returning an unbiased estimate of the solution
Accelerate the solution of dense linear systems
... returning an unbiased estimate of the solution
Basic idea - unbiased estimator for generic sums $a + b$:
Full CG solution:

\[ s = s_0 + \delta_1 + \ldots + \delta_l + \delta_{l+1} \ldots + \delta_T \]

ULISSE:

\[ s_0 + \delta_1 + \ldots + \delta_l \]

\[ 1 - p_1 \]

Final solution is an unbiased estimate of \( s \)!

Filippone and Engler, ICML, 2015
Comparison with MCMC - Concrete dataset - \( n \approx 1K \)
M. Filippone  Bayesian inference for Gaussian processes
Conclusions and ongoing work

• “Noisy” MCMC offers a practical and scalable way to carry out “exact” Bayesian computations for GPs
Conclusions and ongoing work

• “Noisy” MCMC offers a practical and scalable way to carry out “exact” Bayesian computations for GPs
• Novel adaptation of SGLD when the likelihood does not factorize
• Novel linear solver ULISSE to speed up computations of stochastic gradients
Conclusions and ongoing work

- “Noisy” MCMC offers a practical and scalable way to carry out “exact” Bayesian computations for GPs
- Novel adaptation of SGLD when the likelihood does not factorize
- Novel linear solver ULISSE to speed up computations of stochastic gradients
- General likelihoods?
- Preconditioners?
**Joint work with Raphael Engler**


