Sparse Gaussian process inference: how many inducing points do you really need?

David Burt, Carl E. Rasmussen and Mark van der Wilk
Cambridge University

December 2^{nd} 2018

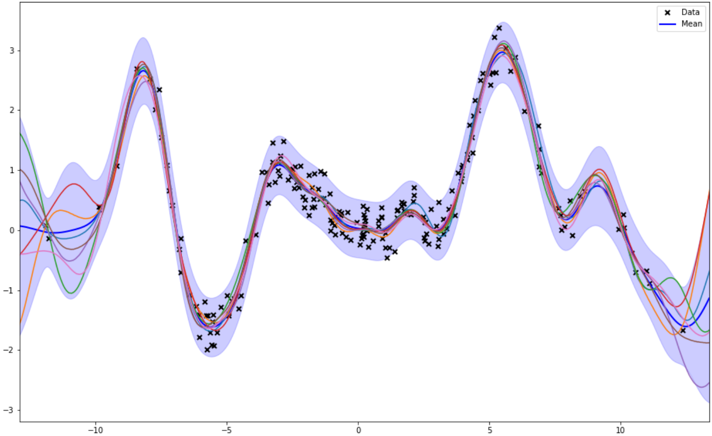
UNIVERSITY OF CAMBRIDGE

Outline

Introduction

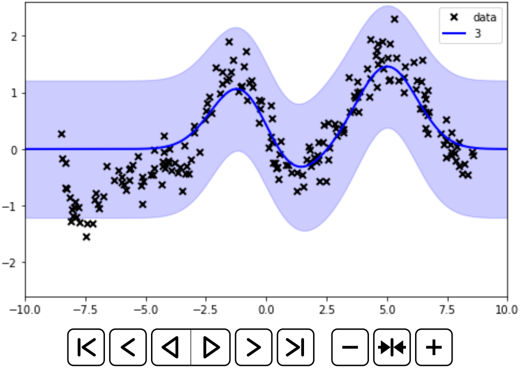
Gaussian Process Regression

Gaussian process priors allow us to perform exact Bayesian inference in regression by directly placing a prior on functions instead of parameters.



Sparse Gaussian Inference

Idea: Make a variational approximation to the posterior process that is Gaussian and only uses $M \ll N$ inducing features.



Main Question

Question: How many inducing points are needed to be sure that a sparse approximation accurately approximates inference in the full model?

Main Question

Question: How many inducing points are needed to be sure that a sparse approximation accurately approximates inference in the full model?

- We approach this problem by proving bounds on the KL-divergence that hold with high probability for large N, with training inputs i.i.d draws from some distribution.

Main Question

Question: How many inducing points are needed to be sure that a sparse approximation accurately approximates inference in the full model?

- We approach this problem by proving bounds on the KL-divergence that hold with high probability for large N, with training inputs i.i.d draws from some distribution.
- In separate work, we introduce *eigenfunction inducing features* an interdomain feature with diagonal covariance matrix. We derive bounds using these features.

Bounds on the marginal likelihood and understanding the KL-divergence

The Variational Lower Bound

Recall,

$$\log(p(\mathbf{y})) = \mathcal{L} = \log\left(\mathcal{N}\left(\mathbf{y}; 0, \mathbf{K}_{f,f} + \sigma^2_{noise}\mathbf{I}\right)\right).$$

The Variational Lower Bound

Recall,

$$\log(p(\mathbf{y})) = \mathcal{L} = \log\left(\mathcal{N}\left(\mathbf{y}; 0, \mathbf{K}_{f,f} + \sigma^2_{noise}\mathbf{I}\right)\right).$$

The variational lower bound for sparse GP regression is [Titsias, 2009]:

$$\begin{aligned}\mathcal{L}_{lower} = &\log\left(\mathcal{N}\left(\mathbf{y}; 0, \mathbf{Q}_{f,f} + \sigma^2_{noise}\mathbf{I}\right)\right)\\ &- \frac{1}{2\sigma^2_{noise}}\underbrace{\mathrm{tr}\left(\mathbf{K}_{f,f} - \mathbf{Q}_{f,f}\right)}_{\mathrm{t}}\end{aligned}$$

$$\text{with} \quad \mathbf{Q}_{f,f} = \mathbf{K}_{u,f}^T \mathbf{K}_{u,u}^{-1} \mathbf{K}_{u,f}.$$

The Variational Lower Bound

Recall,

$$\log(p(\mathbf{y})) = \mathcal{L} = \log\left(\mathcal{N}\left(\mathbf{y}; 0, \mathbf{K}_{f,f} + \sigma^2_{noise}\mathbf{I}\right)\right).$$

The variational lower bound for sparse GP regression is [Titsias, 2009]:

$$\begin{aligned}\mathcal{L}_{lower} &= \log\left(\mathcal{N}\left(\mathbf{y}; 0, \mathbf{Q}_{f,f} + \sigma^2_{noise}\mathbf{I}\right)\right) \\ &\quad - \frac{1}{2\sigma^2_{noise}}\underbrace{\mathrm{tr}\left(\mathbf{K}_{f,f} - \mathbf{Q}_{f,f}\right)}_{\mathrm{t}}\end{aligned}$$

$$\text{with} \qquad \mathbf{Q}_{f,f} = \mathbf{K}^T_{u,f}\mathbf{K}^{-1}_{u,u}\mathbf{K}_{u,f}.$$

$\mathcal{L} - \mathcal{L}_{lower}$ is a KL-divergence from the variational approximation to the full posterior [Matthews et al., 2016].

An Upper Bound on the Marginal Likelihood

We want to bound $\mathcal{L} - \mathcal{L}_{lower}$.

We have the upper bound [Titsias, 2014]:

$$\begin{aligned}\mathcal{L} \leq \mathcal{L}_{upper} :=& \log\left(\mathcal{N}(\mathbf{y}; 0, \mathbf{Q}_{f,f} + t\mathbf{I} + \sigma^2_{noise}\mathbf{I})\right) \\ &+ \frac{1}{2}\log\left(|\mathbf{Q}_{f,f} + t\mathbf{I} + \sigma^2_{noise}\mathbf{I}|\right) \\ &- \frac{1}{2}\log\left(|\mathbf{Q}_{f,f} + \sigma^2_{noise}\mathbf{I}|\right),\end{aligned}$$

- This upper bound depends on $\mathbf{Q}_{f,f}$ and t, making it easier to compare to the lower bound.

A Bound on the Gap (KL-divergence)

Using the upper and lower bound and manipulating matrices,

$$KL(Q \| \hat{P}) \leq \frac{t}{2\sigma^2_{noise}} \underbrace{\left(1 + \frac{\|\mathbf{y}\|_2^2}{\sigma^2_{noise} + t}\right)}_{O(N)}.$$

- Under weak assumptions $\|\mathbf{y}\|^2 = O(N)$.
- In order to show the KL divergence tends to 0 it suffices to choose $M = M(N)$ so that $t = o(1/N)$.

Eigenfunction inducing features

Bounds with Standard Inducing Points

Recall,

$$t := \operatorname{tr}\left(\mathbf{K}_{f,f} - \mathbf{K}_{u,f}^{T}\mathbf{K}_{u,u}^{-1}\mathbf{K}_{u,f}\right).$$

- Subsampled inducing points $\Rightarrow$ Nyström approximation.

Bounds with Standard Inducing Points

Recall,

$$t := \operatorname{tr}\left(\mathbf{K}_{f,f} - \mathbf{K}_{u,f}^{T}\mathbf{K}_{u,u}^{-1}\mathbf{K}_{u,f}\right).$$

- Subsampled inducing points $\Rightarrow$ Nyström approximation.
- Bounds tend to depend heavily on $\mathbf{K}_{f,f}$ being 'well-behaved' (coherence).

Bounds with Standard Inducing Points

Recall,

$$t := \operatorname{tr}\left(\mathbf{K}_{f,f} - \mathbf{K}_{u,f}^{T}\mathbf{K}_{u,u}^{-1}\mathbf{K}_{u,f}\right).$$

- Subsampled inducing points $\Rightarrow$ Nyström approximation.
- Bounds tend to depend heavily on $\mathbf{K}_{f,f}$ being 'well-behaved' (coherence).
- A major obstacle in a more direct approach to understanding this bound when M is allowed to grow as a function of N is $\mathbf{K}_{u,u}^{-1}$.

The Covariance Operator

Given a kernel k, assume the training inputs are are i.i.d. draws according to measure ρ. The covariance operator associated to k, ρ is:

$$\mathcal{K} : [\mathcal{K}f](\mathbf{x}') = \int_{\mathbf{x} \in \mathcal{X}} k(\mathbf{x}, \mathbf{x}') f(\mathbf{x}) d\rho(\mathbf{x}).$$

The Covariance Operator

Given a kernel k, assume the training inputs are are i.i.d. draws according to measure ρ. The covariance operator associated to k, ρ is:

$$\mathcal{K} : [\mathcal{K}f](\mathbf{x}') = \int_{\mathbf{x} \in \mathcal{X}} k(\mathbf{x}, \mathbf{x}') f(\mathbf{x}) d\rho(\mathbf{x}).$$

- As $N \to \infty$ the eigenvalues of $\mathbf{K}_{f,f}$ approach the eigenvalues of $\mathcal{K}$.

The Covariance Operator

Given a kernel k, assume the training inputs are are i.i.d. draws according to measure ρ. The covariance operator associated to k, ρ is:

$$\mathcal{K} : [\mathcal{K}f](\mathbf{x}') = \int_{\mathbf{x} \in \mathcal{X}} k(\mathbf{x}, \mathbf{x}') f(\mathbf{x}) d\rho(\mathbf{x}).$$

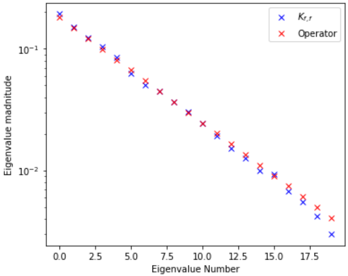
- As $N \to \infty$ the eigenvalues of $\mathbf{K}_{f,f}$ approach the eigenvalues of $\mathcal{K}$.
- For the SE-kernel and a normal input density, the eigenvalues of $\mathcal{K}$ have a closed form and decay exponentially.

The Covariance Operator

Given a kernel k, assume the training inputs are are i.i.d. draws according to measure ρ. The covariance operator associated to k, ρ is:

$$\mathcal{K} : [\mathcal{K}f](\mathbf{x}') = \int_{\mathbf{x}\in\mathcal{X}} k(\mathbf{x}, \mathbf{x}')f(\mathbf{x})d\rho(\mathbf{x}).$$

- As $N \to \infty$ the eigenvalues of $\mathbf{K}_{f,f}$ approach the eigenvalues of $\mathcal{K}$.
- For the SE-kernel and a normal input density, the eigenvalues of $\mathcal{K}$ have a closed form and decay exponentially.



Eigenfunction Inducing Features

We define *eigenfunction inducing features* by,

$$\mathbf{u}_m = \int_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) \frac{1}{\sqrt{\lambda_m}} \phi_m(\mathbf{x}) d\mu(\mathbf{x}).$$

where μ is a probability measure and the (λ_m, ϕ_m) come from the eigendecompostion of $\mathcal{K}_\mu$.

Eigenfunction Inducing Features

We define *eigenfunction inducing features* by,

$$\mathbf{u}_m = \int_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) \frac{1}{\sqrt{\lambda_m}} \phi_m(\mathbf{x}) d\mu(\mathbf{x}).$$

where μ is a probability measure and the (λ_m, ϕ_m) come from the eigendecompostion of $\mathcal{K}_\mu$.

- These are an example of *interdomain inducing features*:
  - Same argument as for inducing points can be used to establish the variational lower bound for these features.

Eigenfunction Inducing Features

We define *eigenfunction inducing features* by,

$$\mathbf{u}_m = \int_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \frac{1}{\sqrt{\lambda_m}} \phi_m(\mathbf{x}) d\mu(\mathbf{x}).$$

where μ is a probability measure and the (λ_m, ϕ_m) come from the eigendecompostion of $\mathcal{K}_\mu$.

- These are an example of *interdomain inducing features*:
  - Same argument as for inducing points can be used to establish the variational lower bound for these features.
  - Need to compute covariance matrices: $\mathbf{K}_{u,f}$ and $\mathbf{K}_{u,u}$.

Covariances

$$\begin{aligned}\mathrm{cov}(\mathbf{u}_m, \mathbf{u}_n) &= \int_{\mathbf{x}\in\mathcal{X}} \int_{\mathbf{x}'\in\mathcal{X}} \mathbb{E}[f(\mathbf{x})f(\mathbf{x}')] \frac{1}{\sqrt{\lambda_m}\sqrt{\lambda_n}} \phi_m(\mathbf{x})\phi_n(\mathbf{x}') d\mu(\mathbf{x}') d\mu(\mathbf{x}) \\ &= \delta_{m,n}.\end{aligned}$$

So $\mathbf{K}_{u,u} = \mathbf{I}$!

Covariances

$$\begin{aligned}\mathrm{cov}(\mathbf{u}_m, \mathbf{u}_n) &= \int_{\mathbf{x}\in\mathcal{X}} \int_{\mathbf{x}'\in\mathcal{X}} \mathbb{E}[f(\mathbf{x})f(\mathbf{x}')] \frac{1}{\sqrt{\lambda_m}\sqrt{\lambda_n}} \phi_m(\mathbf{x})\phi_n(\mathbf{x}') d\mu(\mathbf{x}') d\mu(\mathbf{x}) \\ &= \delta_{m,n}.\end{aligned}$$

So $\mathbf{K}_{u,u} = \mathbf{I}$!

$$\begin{aligned}\mathrm{cov}(\mathbf{u}_m, f(\mathbf{x})) &= \int_{\mathbf{x}\in\mathcal{X}} \mathbb{E}[f(\mathbf{x})f(\mathbf{x}')] \frac{1}{\sqrt{\lambda_m}} \phi_m(\mathbf{x}') d\mu(\mathbf{x}') \\ &= \sqrt{\lambda_m}\phi_m(\mathbf{x}).\end{aligned}$$

These are the entries in $\mathbf{K}_{u,f}$.

Mercer's Theorem

Recall Mercer's Theorem,

$$(\mathbf{K}_{f,f})_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{\infty} \lambda_m \phi_m(\mathbf{x}_i)\phi_m(\mathbf{x}_j),$$

where the λ_m, ϕ_m are eigenvalue, eigenfunction pairs of the operator $\mathcal{K}_\mu$.

Mercer's Theorem

Recall Mercer's Theorem,

$$(\mathbf{K}_{f,f})_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{\infty} \lambda_m \phi_m(\mathbf{x}_i)\phi_m(\mathbf{x}_j),$$

where the λ_m, ϕ_m are eigenvalue, eigenfunction pairs of the operator $\mathcal{K}_\mu$. From the covariance calculation

$$(\mathbf{Q}_{f,f})_{i,j} = \left(\mathbf{K}_{u,f}^T \mathbf{K}_{u,f}\right)_{i,j} = \sum_{m=1}^{M} \lambda_m \phi_m(\mathbf{x}_i)\phi_m(\mathbf{x}_j),$$

for the eigenfunction inducing features defined with respect to the measure μ.

Number of features needed for sparse inference

A Bound on the Trace

Using the two formulas from the previous slide,

$$t_i = (\mathbf{K}_{f,f})_{i,i} - \mathbf{K}_{u,f}^T \mathbf{K}_{u,f} = \sum_{m=M+1}^{\infty} \lambda_m \phi_m(\mathbf{x}_i)^2.$$

A Bound on the Trace

Using the two formulas from the previous slide,

$$t_i = (\mathbf{K}_{f,f})_{i,i} - \mathbf{K}_{u,f}^T \mathbf{K}_{u,f} = \sum_{m=M+1}^{\infty} \lambda_m \phi_m(\mathbf{x}_i)^2.$$

Consider the expected value of t_i, (now assuming the training examples are drawn i.i.d. according to μ).

$$\mathbb{E}[t_i] = \sum_{m=M+1}^{\infty} \int \lambda_m \phi_m(\mathbf{x})^2 d\mu(\mathbf{x}) = \sum_{m=M+1}^{\infty} \lambda_m.$$

A Bound on the Trace

Using the two formulas from the previous slide,

$$t_i = (\mathbf{K}_{f,f})_{i,i} - \mathbf{K}_{u,f}^T \mathbf{K}_{u,f} = \sum_{m=M+1}^{\infty} \lambda_m \phi_m(\mathbf{x}_i)^2.$$

Consider the expected value of t_i, (now assuming the training examples are drawn i.i.d. according to μ).

$$\mathbb{E}[t_i] = \sum_{m=M+1}^{\infty} \int \lambda_m \phi_m(\mathbf{x})^2 d\mu(\mathbf{x}) = \sum_{m=M+1}^{\infty} \lambda_m.$$

The strong law of large numbers tells us that $\frac{1}{N} t$ tends to its expected value as $N \to \infty$.

A Bound on the Trace

Using the two formulas from the previous slide,

$$t_i = (\mathbf{K}_{f,f})_{i,i} - \mathbf{K}_{u,f}^T \mathbf{K}_{u,f} = \sum_{m=M+1}^{\infty} \lambda_m \phi_m(\mathbf{x}_i)^2.$$

Consider the expected value of t_i, (now assuming the training examples are drawn i.i.d. according to μ).

$$\mathbb{E}[t_i] = \sum_{m=M+1}^{\infty} \int \lambda_m \phi_m(\mathbf{x})^2 d\mu(\mathbf{x}) = \sum_{m=M+1}^{\infty} \lambda_m.$$

The strong law of large numbers tells us that $\frac{1}{N} t$ tends to its expected value as $N \to \infty$.

As long as the eigenfunctions are "well-behaved" we can use this to bound t.

Example: SE Kernel, Normal Inputs

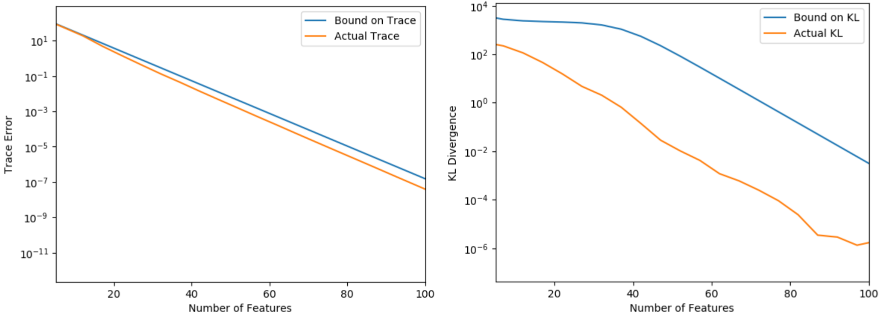


Figure: Bounds for normally distributed data with 200 training inputs.

For small M the bound is not useful, but in this case with $M = 50$ to 60 features the bound gives strong guarantees.

Number of Inducing Features Needed

Theorem

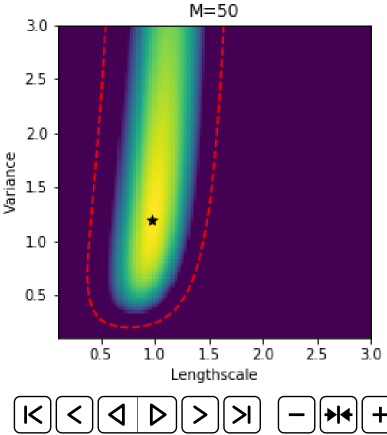
For inference using a squared exponential kernel, if the $x_i \sim \mathcal{N}(0, s'^2)$ i.i.d. For sparse inference with eigenfunction inducing features defined with respect to $q(x) \sim \mathcal{N}(0, s^2)$ with $2s'^2 < s^2$ there exists an N_0 such that for all $N > N_0$ inducing point inference with a set of $M = c\log(N)$ features results in:

$$Pr(KL(Q\|\hat{P}) > \epsilon) < \delta.$$

- Straightforward to address multidimensional data with additive or multiplicative kernel in a similar framework.

What does this tell us about hyperparameter selection?

With an ideal optimizer and assuming the error surface is continuous, if $\mathcal{L}_{lower} \approx \mathcal{L}$ we must be near the optimal choice of hyperparameters!



Conclusion

Convergence guarantees can be obtained for sparse variational GP regression. For a certain set of inducing features with normally distributed training inputs we proved:

- A probabilistic upper bound on the KL-divergence.
- The loss from sparsity can be made arbitrarily small with high probability using $M = O(\log(N))$ features for the SE-kernel.

Extensions

- Can anything be said about sparse non-conjugate variational inference (how much more do we lose from sparsity)?
- Bounds on Nyström approximation tailored to optimized inducing points?
- How tight are these bounds?

References

A. G. d. G. Matthews, J. Hensman, R. Turner, and Z. Ghahramani. On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. *Journal of Machine Learning Research*, 51:231–239, 2016.

M. K. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009.

M. K. Titsias. *Variational Inference for Gaussian and Determinantal Point Processes*. Dec. 2014. Published: Workshop on Advances in Variational Inference (NIPS 2014).