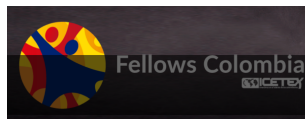


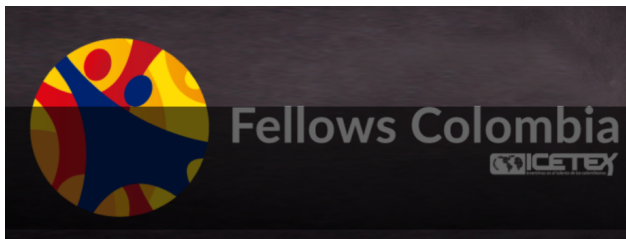
Variational inference for Gaussian processes

Mauricio A. Álvarez

Topics on Deep Probabilistic Models



Acknowledgements



Contents

Motivation

Variational inference for sparse GPs

Stochastic variational inference for sparse GPs

“Collapsed” variational inference for sparse GPs

Contents

Motivation

Variational inference for sparse GPs

Stochastic variational inference for sparse GPs

“Collapsed” variational inference for sparse GPs

Posterior inference

- When using Bayesian inference, we need to compute the posterior distribution of \mathbf{f} given the data.
- We then use that posterior distribution to compute the predictive distribution.
- Reasons as why computing the posterior distribution is an issue for GPs.
 - Computational complexity.
 - Non-Gaussian likelihood.
 - Both of the above.

Computational complexity

- To compute the predictive mean and the predictive covariance we need to compute $[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}]^{-1}$
- The usual way to do this is using the Cholesky decomposition which costs $\mathcal{O}(n^3)$.
- If $n = 1000$, then we need to perform 10^9 operations.

Non-Gaussian likelihoods

- In Bayesian inference we want to compute

$$p(\mathbf{f}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{p(\mathbf{y})},$$

where $p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})d\mathbf{f}$.

- When $p(\mathbf{y}|\mathbf{f})$ is a Gaussian likelihood, then we can compute $p(\mathbf{y})$ and $p(\mathbf{f}|\mathbf{y})$ analytically.
- When $p(\mathbf{y}|\mathbf{f})$ is non-Gaussian (e.g. Bernoulli with a sigmoid link function) both $p(\mathbf{y})$ and $p(\mathbf{f}|\mathbf{y})$ are intractable.

How to address these issues?

- ❑ One successful approach is by using the idea of *inducing variables* or *pseudo-variables*.
- ❑ The idea in itself was quite well known in the GP literature. See for example Chapter 8 of the GPML book and in the paper Quiñero-Candela and Rasmussen (2005).
- ❑ However, if we couple this idea with a variational inference approach, we have a powerful tool to build complex GP models.

Contents

Motivation

Variational inference for sparse GPs

Stochastic variational inference for sparse GPs

“Collapsed” variational inference for sparse GPs

Auxiliary variables

- We introduce a new set of M variables $\mathbf{u} = \{u(\mathbf{z}_m)\}_{m=1}^M$ that we refer to as inducing variables or pseudo-variables.
- The set of points $\mathbf{Z} = \{\mathbf{z}_m\}_{m=1}^M$ is usually known as inducing inputs.
- We augment the original prior $p(\mathbf{f})$ to $p(\mathbf{f}, \mathbf{u})$ such that

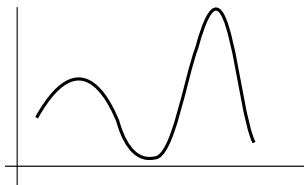
$$p(\mathbf{f}) = \int p(\mathbf{f}, \mathbf{u}) d\mathbf{u} = \int p(\mathbf{f}|\mathbf{u})p(\mathbf{u})d\mathbf{u},$$

where $p(\mathbf{u})$ and $p(\mathbf{f}, \mathbf{u})$ are both Gaussians.

- The auxiliary variables \mathbf{u} can be part of the GP $f(\mathbf{x})$ or they can be linearly related to $f(\mathbf{x})$ (sometimes known as interdomain inducing variables).
- In the former case, $p(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}(\mathbf{Z}, \mathbf{Z}))$.

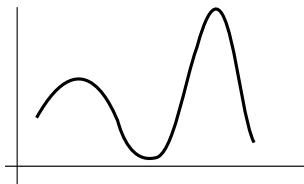
Auxiliary variables

A sample from $p(f)$

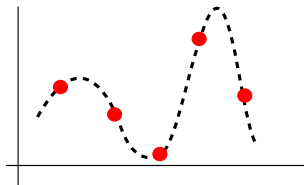


Auxiliary variables

A sample from $p(f)$

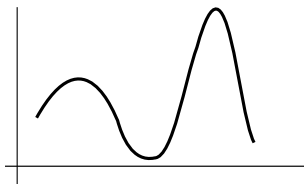


Inducing variables \mathbf{u}

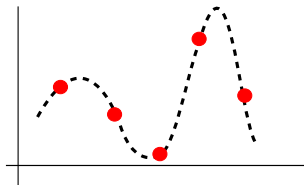


Auxiliary variables

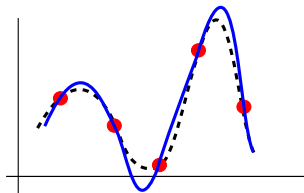
A sample from $p(f)$



Inducing variables \mathbf{u}



A sample from $p(f|\mathbf{u})$



Variational lower-bound for the marginal likelihood (I)

- We can write the log marginal probability for \mathbf{y} using

$$\log p(\mathbf{y}) = \mathcal{L}(q(\mathbf{f})) + \text{KL}(q(\mathbf{f})\|p(\mathbf{f}|\mathbf{y})),$$

where

$$\begin{aligned}\mathcal{L}(q(\mathbf{f})) &= \int q(\mathbf{f}) \log \left\{ \frac{p(\mathbf{y}, \mathbf{f})}{q(\mathbf{f})} \right\} d\mathbf{f}, \\ \text{KL}(q(\mathbf{f})\|p(\mathbf{f}|\mathbf{y})) &= - \int q(\mathbf{f}) \log \left\{ \frac{p(\mathbf{f}|\mathbf{y})}{q(\mathbf{f})} \right\} d\mathbf{f},\end{aligned}$$

with $q(\mathbf{f})$ the approximated posterior, $\text{KL}(q\|p)$ is the Kullback-Leibler divergence between q and p and $p(\mathbf{f}|\mathbf{y})$ is the true posterior.

- The KL divergence is zero when $q = p$. In that case, $\log p(\mathbf{y}) = \mathcal{L}(q(\mathbf{f}))$.
- If this is not the case $\text{KL}(q(\mathbf{f})\|p(\mathbf{f}|\mathbf{y}))$ and $\log p(\mathbf{y}) > \mathcal{L}(q(\mathbf{f}))$.

Variational lower-bound for the marginal likelihood (II)

- We have two ways to find the optimal $q(\mathbf{f})$
 1. We find $q(\mathbf{f})$ by minimising $\text{KL}(q(\mathbf{f})\|p(\mathbf{f}|\mathbf{y}))$.
 2. We find $q(\mathbf{f})$ by maximising $\mathcal{L}(q(\mathbf{f}))$.
- Option 1 is not possible since $p(\mathbf{f}|\mathbf{y})$ is unknown.
- So, in general, we appeal to option 2

$$\log p(\mathbf{y}) \geq \mathcal{L}(q(\mathbf{f})).$$

Lower-bound with inducing variables

- For our augmented model we want to find an approximated posterior $q(\mathbf{f}, \mathbf{u})$ by maximising

$$\mathcal{L}(q(\mathbf{f}, \mathbf{u})) = \int \int q(\mathbf{f}, \mathbf{u}) \log \left\{ \frac{p(\mathbf{y}, \mathbf{f}, \mathbf{u})}{q(\mathbf{f}, \mathbf{u})} \right\} d\mathbf{u} d\mathbf{f},$$

where $p(\mathbf{y}, \mathbf{f}, \mathbf{u}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})$ and we will approximate the posterior $q(\mathbf{f}, \mathbf{u})$ as $q(\mathbf{f}, \mathbf{u}) \approx p(\mathbf{f}|\mathbf{u})q(\mathbf{u})$.

- Since we know that $q(\mathbf{f}) = \int_{\mathbf{u}} p(\mathbf{f}|\mathbf{u})q(\mathbf{u})d\mathbf{u}$, the bound above really only depends on $q(\mathbf{u})$

$$\begin{aligned} \mathcal{L}(q(\mathbf{u})) &= \int \int p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) \log \left\{ \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})}{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} \right\} d\mathbf{u} d\mathbf{f}, \\ &= \int \int p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) \log \left\{ \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{u})}{q(\mathbf{u})} \right\} d\mathbf{u} d\mathbf{f}. \end{aligned}$$

Two approaches for optimising $\mathcal{L}(q(\mathbf{u}))$

- There are two approaches for optimising $q(\mathbf{u})$ in $\mathcal{L}(q(\mathbf{u}))$.
- First approach (Hensman et al., 2013):
 - We assume a multi-variate Gaussian form for $q(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\boldsymbol{\mu}, \mathbf{S})$ with $\boldsymbol{\mu} \in \mathbb{R}^{M \times 1}$ and $\mathbf{S} \in \mathbb{R}^{M \times M}$.
 - We then find $\boldsymbol{\mu}$ and \mathbf{S} by numerically optimising $\mathcal{L}(q(\mathbf{u}))$.
- Second approach (Titsias, 2009):
 - We marginalise $q(\mathbf{u})$ from the bound and then compute it by using Jensen's inequality.
 - We then find $\boldsymbol{\mu}$ and \mathbf{S} by using the rules of probability.

Contents

Motivation

Variational inference for sparse GPs

Stochastic variational inference for sparse GPs

“Collapsed” variational inference for sparse GPs

Stochastic variational inference

- Stochastic variational inference (SVI) allows (Hoffman et al., 2013) the use of stochastic gradients over variational lower bounds.
- Hensman et al. (2013) proposed the use of SVI for sparse GPs.
- The idea is to use stochastic gradients for optimising $\mathcal{L}(q(\mathbf{u}))$ with respect to $q(\mathbf{u})$, this is, μ and \mathbf{S} .

Lower bound $\mathcal{L}(q(\mathbf{u}))$ (I)

- From a previous slide,

$$\mathcal{L}(q(\mathbf{u})) = \int \int p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) \log \left\{ \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{u})}{q(\mathbf{u})} \right\} d\mathbf{u}d\mathbf{f}.$$

- We can re-arrange the expression above using

$$\begin{aligned}\mathcal{L}(q(\mathbf{u})) &= \int \int p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) \log \left\{ \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{u})}{q(\mathbf{u})} \right\} d\mathbf{u}d\mathbf{f}, \\ &= \int \int p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) \left[\log p(\mathbf{y}|\mathbf{f}) + \log \frac{p(\mathbf{u})}{q(\mathbf{u})} \right] d\mathbf{u}d\mathbf{f}, \\ &= \int \int p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) \log p(\mathbf{y}|\mathbf{f}) d\mathbf{u}d\mathbf{f} + \int q(\mathbf{u}) \log \frac{p(\mathbf{u})}{q(\mathbf{u})} d\mathbf{u}, \\ &= \int \log p(\mathbf{y}|\mathbf{f}) \int p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) d\mathbf{u}d\mathbf{f} + \int q(\mathbf{u}) \log \frac{p(\mathbf{u})}{q(\mathbf{u})} d\mathbf{u}, \\ &= \int \log p(\mathbf{y}|\mathbf{f})q(\mathbf{f})d\mathbf{f} - \text{KL}(q(\mathbf{u})\|p(\mathbf{u}))d\mathbf{u}, \\ &= \mathbb{E}_{q(\mathbf{f})}[\log p(\mathbf{y}|\mathbf{f})] - \text{KL}(q(\mathbf{u})\|p(\mathbf{u})).\end{aligned}$$

Lower bound $\mathcal{L}(q(\mathbf{u}))$ (II)

- The lower bound is

$$\mathcal{L}(q(\mathbf{u})) = \mathbb{E}_{q(\mathbf{f})}[\log p(\mathbf{y}|\mathbf{f})] - \text{KL}(q(\mathbf{u})\|p(\mathbf{u})).$$

- We can find an estimate for μ and \mathbf{S} by maximising $\mathcal{L}(q(\mathbf{u}))$ using numerical optimisation.
- We need to compute the gradients

$$\frac{\partial \mathcal{L}}{\partial \mu}, \frac{\partial \mathcal{L}}{\partial \mathbf{S}}$$

Exercises

- Recall that $q(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\boldsymbol{\mu}, \mathbf{S})$ and $p(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}(\mathbf{Z}, \mathbf{Z}))$. Now, for the regression case,

$$p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma_n^2 \mathbf{I}).$$

1. Write the expression for the bound in terms of $\boldsymbol{\mu}$, \mathbf{S} and σ_n^2 .
 2. Compute $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}}$ and $\frac{\partial \mathcal{L}}{\partial \mathbf{S}}$.
- For the case above, what is the computational complexity of this method?

Stochastic gradient descend

- It can be shown that the lower bound can be written as

$$\mathcal{L}(q(\mathbf{u})) = \sum_{i=1}^n \ell(y_i, \mathbf{x}_i, \theta) - \text{KL}(q(\mathbf{u}) \| p(\mathbf{u})),$$

where $\ell(y_i, \mathbf{x}_i, \theta)$ is a function that depends on the data, the variational parameters μ , \mathbf{S} , and any other (hyper) parameters in the model (e.g. the hyperparameters of the kernel).

- For n large, we could only use a subset of the data to compute the gradients to be used in numerical optimisation.
- This is usually known as *stochastic gradient descend*.
- The computational complexity of this model is $\mathcal{O}(nM^2)$, where M is the number of inducing points.

Example SVI for Sparse GPs

Lab SVI for Sparse GPs (GPFlow)

Contents

Motivation

Variational inference for sparse GPs

Stochastic variational inference for sparse GPs

“Collapsed” variational inference for sparse GPs

Marginalising $q(\mathbf{u})$ from the bound (I)

- Instead of finding particular parameters μ and \mathbf{S} as before, we can marginalise $q(\mathbf{u})$ from the lower bound and then use probability rules to compute $q(\mathbf{u})$.

- Let us go back to the general expression for the bound

$$\mathcal{L}(q(\mathbf{u})) = \int \int p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) \log \left\{ \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{u})}{q(\mathbf{u})} \right\} d\mathbf{u}d\mathbf{f}.$$

- Let us assume that $q(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mu, \mathbf{S})$ and find expressions for μ and \mathbf{S} .
- We first integrate over \mathbf{f} .

Marginalising $q(\mathbf{u})$ from the bound (II)

- The bound can be expressed as

$$\begin{aligned}\mathcal{L}(q(\mathbf{u})) &= \int \int p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) \log \left\{ \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{u})}{q(\mathbf{u})} \right\} d\mathbf{u}d\mathbf{f} \\ &= \int q(\mathbf{u}) \int p(\mathbf{f}|\mathbf{u}) \left\{ \log p(\mathbf{y}|\mathbf{f}) + \log \left[\frac{p(\mathbf{u})}{q(\mathbf{u})} \right] \right\} d\mathbf{f}d\mathbf{u}.\end{aligned}$$

- Let us focus on the integral over \mathbf{f}

$$\log T(\mathbf{y}, \mathbf{u}) = \int \log p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})d\mathbf{f}.$$

Exercise

- $p(\mathbf{f}|\mathbf{u})$ is a conditional Gaussian distribution given as,

$$\mathcal{N}(\mathbf{f}|\mathbf{K}(\mathbf{X}, \mathbf{Z})\mathbf{K}^{-1}(\mathbf{Z}, \mathbf{Z})\mathbf{u}, \mathbf{K}(\mathbf{X}, \mathbf{X}) - \mathbf{K}(\mathbf{X}, \mathbf{Z})\mathbf{K}^{-1}(\mathbf{Z}, \mathbf{Z})\mathbf{K}(\mathbf{X}, \mathbf{Z})^{\top}),$$

and $p(\mathbf{y}|\mathbf{f})$ is again a Gaussian given as $\mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma_n^2\mathbf{I})$.

- Compute the expression $\log T(\mathbf{y}, \mathbf{u})$

Marginalising $q(\mathbf{u})$ from the bound (III)

- The bound can now be expressed as

$$\mathcal{L}(q(\mathbf{u})) = \int q(\mathbf{u}) \left\{ \log \mathcal{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma_n^2 \mathbf{I}) - \frac{1}{2} \text{trace}(\sigma_n^{-2} \tilde{\mathbf{K}}) + \log \left[\frac{p(\mathbf{u})}{q(\mathbf{u})} \right] \right\} d\mathbf{u},$$

where

$$\boldsymbol{\alpha} = \mathbf{K}(\mathbf{X}, \mathbf{Z}) \mathbf{K}^{-1}(\mathbf{Z}, \mathbf{Z}) \mathbf{u}$$

$$\tilde{\mathbf{K}} = \mathbf{K}(\mathbf{X}, \mathbf{X}) - \mathbf{K}(\mathbf{X}, \mathbf{Z}) \mathbf{K}^{-1}(\mathbf{Z}, \mathbf{Z}) \mathbf{K}(\mathbf{X}, \mathbf{Z})^\top.$$

- It follows that

$$\mathcal{L}(q(\mathbf{u})) = \int q(\mathbf{u}) \left\{ \log \left[\frac{\mathcal{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma_n^2 \mathbf{I}) p(\mathbf{u})}{q(\mathbf{u})} \right] \right\} d\mathbf{u} - \frac{1}{2} \text{trace}(\sigma_n^{-2} \tilde{\mathbf{K}})$$

Jensen's inequality

- A function φ is *convex* if

$$\varphi(\lambda a + (1 - \lambda)b) \leq \lambda\varphi(a) + (1 - \lambda)\varphi(b).$$

- A function φ is *concave* if

$$\varphi(\lambda a + (1 - \lambda)b) \geq \lambda\varphi(a) + (1 - \lambda)\varphi(b).$$

- Let φ be a convex function. It can be shown that

$$\begin{aligned}\varphi(\mathbb{E}(\mathbf{x})) &\leq \mathbb{E}(\varphi(\mathbf{x})) \\ \varphi\left(\int \mathbf{x}p(\mathbf{x})d\mathbf{x}\right) &\leq \int \varphi(\mathbf{x})p(\mathbf{x})d\mathbf{x}.\end{aligned}$$

This inequality is known as the *Jensen's inequality*.

- If φ is a concave function then

$$\begin{aligned}\varphi(\mathbb{E}(\mathbf{x})) &\geq \mathbb{E}(\varphi(\mathbf{x})) \\ \varphi\left(\int \mathbf{x}p(\mathbf{x})d\mathbf{x}\right) &\geq \int \varphi(\mathbf{x})p(\mathbf{x})d\mathbf{x}.\end{aligned}$$

Jensen's inequality applied to $\mathcal{L}(\mathbf{q}(\mathbf{u}))$

- Reversing Jensen's inequality, we can write

$$\log \left[\int q(\mathbf{u}) \left[\frac{\mathcal{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma_n^2 \mathbf{I}) p(\mathbf{u})}{q(\mathbf{u})} \right] d\mathbf{u} \right] \geq \int q(\mathbf{u}) \left\{ \log \left[\frac{\mathcal{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma_n^2 \mathbf{I}) p(\mathbf{u})}{q(\mathbf{u})} \right] \right\} d\mathbf{u}$$

- The expression above can be simplified as

$$\log \left[\int \mathcal{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma_n^2 \mathbf{I}) p(\mathbf{u}) d\mathbf{u} \right] \geq \int q(\mathbf{u}) \left\{ \log \left[\frac{\mathcal{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma_n^2 \mathbf{I}) p(\mathbf{u})}{q(\mathbf{u})} \right] \right\} d\mathbf{u}$$

A tighter bound $\mathcal{L}(q(\mathbf{u}))$

- Reversing Jensen's inequality, we can write

$$\log \left[\int q(\mathbf{u}) \left[\frac{\mathcal{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma_n^2 \mathbf{I}) p(\mathbf{u})}{q(\mathbf{u})} \right] d\mathbf{u} \right] \geq \int q(\mathbf{u}) \left\{ \log \left[\frac{\mathcal{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma_n^2 \mathbf{I}) p(\mathbf{u})}{q(\mathbf{u})} \right] \right\} d\mathbf{u}$$

- The expression above can be simplified as

$$\log \left[\int \mathcal{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma_n^2 \mathbf{I}) p(\mathbf{u}) d\mathbf{u} \right] \geq \int q(\mathbf{u}) \left\{ \log \left[\frac{\mathcal{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma_n^2 \mathbf{I}) p(\mathbf{u})}{q(\mathbf{u})} \right] \right\} d\mathbf{u}$$

- If we define

$$\mathcal{L}_2 = \log \left[\int \mathcal{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma_n^2 \mathbf{I}) p(\mathbf{u}) d\mathbf{u} \right] - \frac{1}{2} \text{trace}(\sigma_n^{-2} \tilde{\mathbf{K}}),$$

then

$$\mathcal{L}_2 \geq \mathcal{L}(q(\mathbf{u})).$$

And \mathcal{L}_2 is closer to $\log p(\mathbf{y})$ than $\mathcal{L}(q(\mathbf{u}))$.

Exercise

- Show that \mathcal{L}_2 can be written as

$$\mathcal{L}_2 = \log \mathcal{N}(\mathbf{y} | \mathbf{0}, \mathbf{K}(\mathbf{X}, \mathbf{Z})\mathbf{K}^{-1}(\mathbf{Z}, \mathbf{Z})\mathbf{K}^\top(\mathbf{X}, \mathbf{Z}) + \sigma_n^2 \mathbf{I}) - \frac{1}{2} \text{trace}(\sigma_n^{-2} \tilde{\mathbf{K}}).$$

References I

- James Hensman, Nicolò Fusi, and Neil D. Lawrence. Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, pages 282–290, 2013.
- Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *J. Mach. Learn. Res.*, 14(1):1303–1347, 2013.
- Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.
- Michalis K. Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial Intelligence and Statistics 12*, pages 567–574, 2009.