Modern Gaussian Processes: Scalable Inference and Novel Applications

(Part II-b) Approximate Inference

Edwin V. Bonilla and Maurizio Filippone

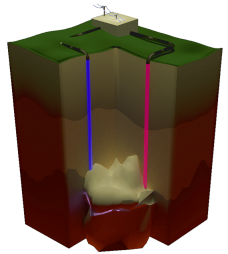
CSIRO's Data61, Sydney, Australia and EURECOM, Sophia Antipolis, France

July 14th, 2019

Challenges in Bayesian Reasoning with Gaussian Process Priors

$p(\mathbf{f})$: prior over geology and rock properties

$p(\mathbf{y} \mid \mathbf{f})$: observation model's likelihood



$20 Million geothermal well

Geol. surveys and explorations

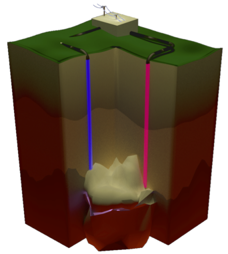
Challenges in Bayesian Reasoning with Gaussian Process Priors

$p(\mathbf{f})$: prior over geology and rock properties

$p(\mathbf{y} \mid \mathbf{f})$: observation model's likelihood

$p(\mathbf{f}|\mathbf{y})$: posterior geological model:

$$p(\mathbf{f} \mid \mathbf{y}, \theta) = \frac{p(\mathbf{f} \mid \theta) p(\mathbf{y} \mid \mathbf{f})}{\underbrace{\int p(\mathbf{f} \mid \theta) p(\mathbf{y} \mid \mathbf{f}) d\mathbf{f}}_{\text{hard bit}}}$$



$20 Million geothermal well

Geol. surveys and explorations

Challenges in Bayesian Reasoning with Gaussian Process Priors

$p(\mathbf{f})$: prior over geology and rock properties

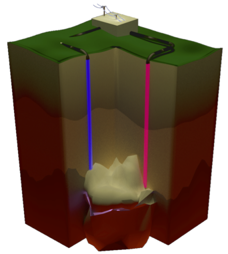
$p(\mathbf{y} \mid \mathbf{f})$: observation model's likelihood

$p(\mathbf{f}|\mathbf{y})$: posterior geological model:

$$p(\mathbf{f} \mid \mathbf{y}, \theta) = \frac{p(\mathbf{f} \mid \theta) p(\mathbf{y} \mid \mathbf{f})}{\underbrace{\int p(\mathbf{f} \mid \theta) p(\mathbf{y} \mid \mathbf{f}) d\mathbf{f}}_{\text{hard bit}}}$$

Challenges:

- Non-linear likelihood models
- Large datasets

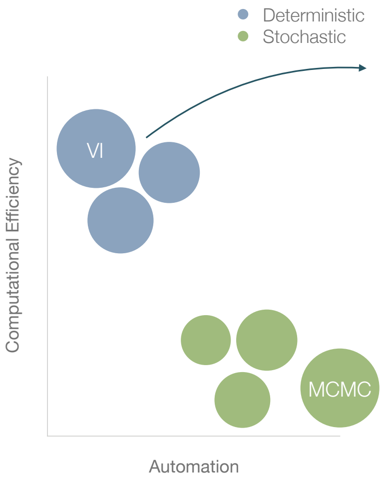


$20 Million geothermal well

Geol. surveys and explorations

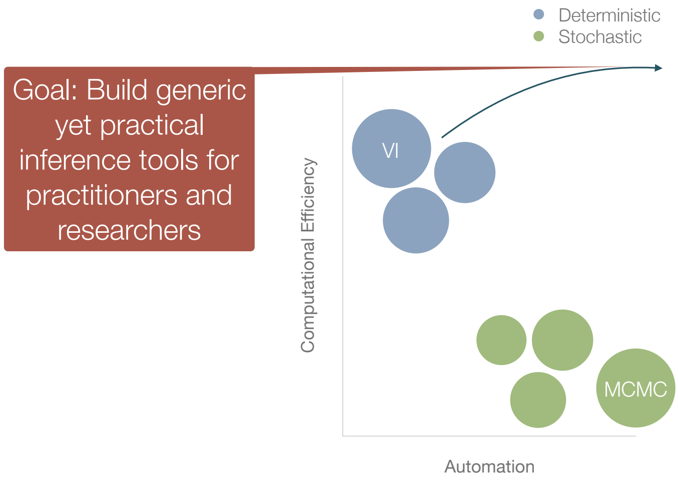
Automated Probabilistic Reasoning

- Approximate inference



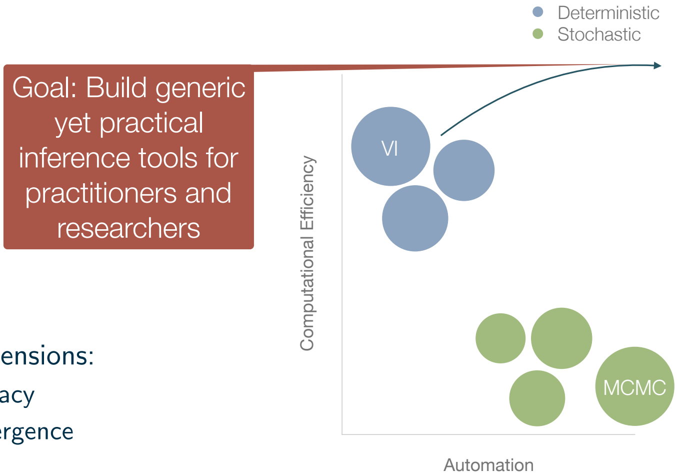
Automated Probabilistic Reasoning

- Approximate inference



Automated Probabilistic Reasoning

- Approximate inference



- Other dimensions:
  - Accuracy
  - Convergence

Outline

1. Latent Gaussian Process Models (LGPMs)
2. Variational Inference
3. Scalability through Inducing Variables and Stochastic Variational Inference (SVI)

Latent Gaussian Process Models (LGPMs)

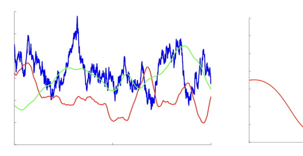
Latent Gaussian Process Models (LGPMs)

Supervised learning $\mathcal{D} = \{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^N$

- Factorised GP priors over Q latent functions:

$$f_j(\mathbf{x}) \sim \mathcal{GP}(0, \kappa_j(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}))$$

$$p(\mathbf{F} \mid \mathbf{X}, \boldsymbol{\theta}) = \prod_{j=1}^{Q} \mathcal{N}(\mathbf{F}_{\cdot j}; \mathbf{0}, \mathbf{K}_j)$$



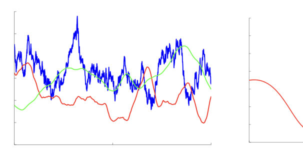
Latent Gaussian Process Models (LGPMs)

Supervised learning $\mathcal{D} = \{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^{N}$

- Factorised GP priors over Q latent functions:

$$f_j(\mathbf{x}) \sim \mathcal{GP}(0, \kappa_j(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}))$$

$$p(\mathbf{F} \mid \mathbf{X}, \boldsymbol{\theta}) = \prod_{j=1}^{Q} \mathcal{N}(\mathbf{F}_{\cdot j}; \mathbf{0}, \mathbf{K}_j)$$



- Factorised likelihood over observations

$$p(\mathbf{Y} \mid \mathbf{X}, \mathbf{F}, \phi) = \prod_{n=1}^{N} p(\mathbf{Y}_{n\cdot} \mid \mathbf{F}_{n\cdot}, \phi)$$

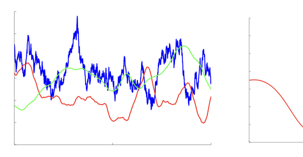
Latent Gaussian Process Models (LGPMs)

Supervised learning $\mathcal{D} = \{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^{N}$

- Factorised GP priors over Q latent functions:

$$f_j(\mathbf{x}) \sim \mathcal{GP}(0, \kappa_j(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}))$$

$$p(\mathbf{F} \mid \mathbf{X}, \boldsymbol{\theta}) = \prod_{j=1}^{Q} \mathcal{N}(\mathbf{F}_{\cdot j}; \mathbf{0}, \mathbf{K}_j)$$



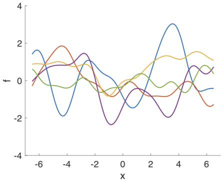
- Factorised likelihood over observations

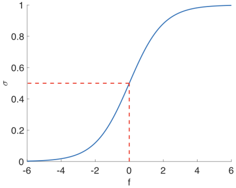
$$p(\mathbf{Y} \mid \mathbf{X}, \mathbf{F}, \phi) = \prod_{n=1}^{N} p(\mathbf{Y}_{n\cdot} \mid \mathbf{F}_{n\cdot}, \phi)$$

What can we model within this framework?

Examples of LGPMs (1)

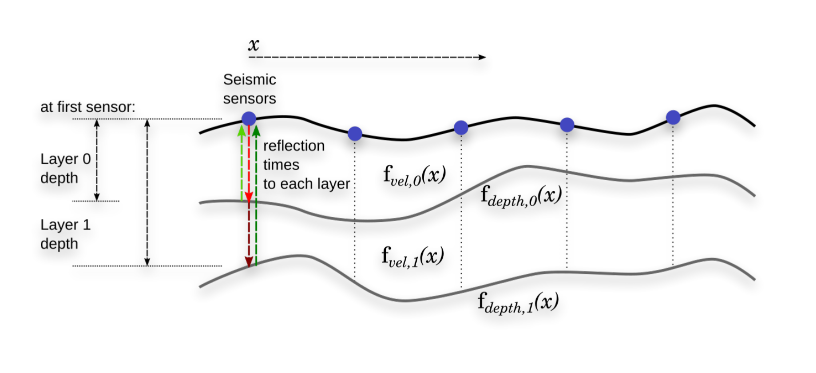
- Multi-output regression
- Multi-class classification
  - $P = Q$ classes
  - softmax likelihood





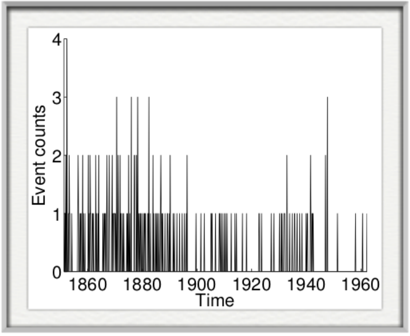
Examples of LGPMs (2)

- Inversion problems



Examples of LGPMs (3)

- Log Gaussian Cox processes (LGCPs)



Inference in LGPMs

We only require access to ‘black-box’ likelihoods. *How can we carry out inference in these general models?*

Variational Inference

Variational Inference (VI): Optimise Rather than Integrate

Recall our posterior estimation problem:

$$\underbrace{p(\mathbf{F} \mid \mathbf{Y})}_{\text{posterior}} = \frac{1}{\underbrace{p(\mathbf{Y})}_{\substack{\text{marginal}\\\text{likelihood}}}} \underbrace{p(\mathbf{F})}_{\text{prior}} \underbrace{p(\mathbf{Y} \mid \mathbf{F})}_{\substack{\text{conditional}\\\text{likelihood}}}$$

Variational Inference (VI): Optimise Rather than Integrate

Recall our posterior estimation problem:

$$\underbrace{p(\mathbf{F} \mid \mathbf{Y})}_{\text{posterior}} = \frac{1}{\underbrace{p(\mathbf{Y})}_{\substack{\text{marginal}\\ \text{likelihood}}}} \underbrace{p(\mathbf{F})}_{\text{prior}} \underbrace{p(\mathbf{Y} \mid \mathbf{F})}_{\substack{\text{conditional}\\ \text{likelihood}}}$$

- Estimating $p(\mathbf{Y}) = \int p(\mathbf{F}) p(\mathbf{Y} \mid \mathbf{F}) d\mathbf{F}$ is hard

Variational Inference (VI): Optimise Rather than Integrate

Recall our posterior estimation problem:

$$\underbrace{p(\mathbf{F} \mid \mathbf{Y})}_{\text{posterior}} = \frac{1}{\underbrace{p(\mathbf{Y})}_{\substack{\text{marginal} \\ \text{likelihood}}}} \underbrace{p(\mathbf{F})}_{\text{prior}} \underbrace{p(\mathbf{Y} \mid \mathbf{F})}_{\substack{\text{conditional} \\ \text{likelihood}}}$$

- Estimating $p(\mathbf{Y}) = \int p(\mathbf{F}) p(\mathbf{Y} \mid \mathbf{F}) d\mathbf{F}$ is hard
- Instead, approximate $q(\mathbf{F} \mid \boldsymbol{\lambda}) \approx p(\mathbf{F} \mid \mathbf{Y})$ to minimize:

$$\mathrm{KL}\left[q(\mathbf{F} \mid \boldsymbol{\lambda}) \parallel p(\mathbf{F} \mid \mathbf{Y})\right] \stackrel{\text{def}}{=} \mathbb{E}_{q(\mathbf{F} \mid \boldsymbol{\lambda})} \log \frac{q(\mathbf{F} \mid \boldsymbol{\lambda})}{p(\mathbf{F} \mid \mathbf{Y})}$$

Variational Inference (VI): Optimise Rather than Integrate

Recall our posterior estimation problem:

$$\underbrace{p(\mathbf{F} \mid \mathbf{Y})}_{\text{posterior}} = \frac{1}{\underbrace{p(\mathbf{Y})}_{\substack{\text{marginal}\\\text{likelihood}}}} \underbrace{p(\mathbf{F})}_{\text{prior}} \underbrace{p(\mathbf{Y} \mid \mathbf{F})}_{\substack{\text{conditional}\\\text{likelihood}}}$$

- Estimating $p(\mathbf{Y}) = \int p(\mathbf{F}) p(\mathbf{Y} \mid \mathbf{F}) d\mathbf{F}$ is hard

- Instead, approximate $q(\mathbf{F} \mid \boldsymbol{\lambda}) \approx p(\mathbf{F} \mid \mathbf{Y})$ to minimize:

$$\mathrm{KL}\left[q(\mathbf{F} \mid \boldsymbol{\lambda}) \parallel p(\mathbf{F} \mid \mathbf{Y})\right] \stackrel{\text{def}}{=} \mathbb{E}_{q(\mathbf{F} \mid \boldsymbol{\lambda})} \log \frac{q(\mathbf{F} \mid \boldsymbol{\lambda})}{p(\mathbf{F} \mid \mathbf{Y})}$$

Properties:

$$\mathrm{KL}\left[q \parallel p\right] \geq 0,$$

$$\mathrm{KL}\left[q \parallel p\right] = 0 \text{ iff } q = p.$$

Decomposition of the Marginal Likelihood

$$\log p(\mathbf{Y}) = \mathrm{KL}\left[q(\mathbf{F} \mid \boldsymbol{\lambda}) \parallel p(\mathbf{F} \mid \mathbf{Y})\right] + \mathcal{L}_{\mathrm{ELBO}}(\boldsymbol{\lambda})$$

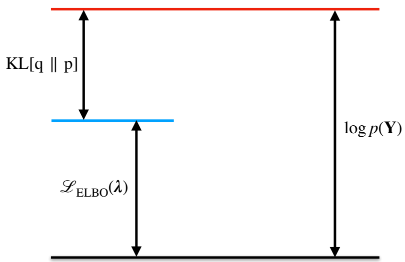


Fig reproduced from Bishop (2006)

- $\mathcal{L}_{\mathrm{ELBO}}(\boldsymbol{\lambda})$ is a lower bound on the log marginal likelihood
- The optimum is achieved when $q = p$
- Maximizing $\mathcal{L}_{\mathrm{ELBO}}(\boldsymbol{\lambda}) \equiv$ minimizing $\mathrm{KL}\left[q(\mathbf{F} \mid \boldsymbol{\lambda}) \parallel p(\mathbf{F} \mid \mathbf{Y})\right]$

Variational Inference Strategy

- The evidence lower bound $\mathcal{L}_{\mathrm{ELBO}}(\boldsymbol{\lambda})$ can be written as:

$$\mathcal{L}_{\mathrm{ELBO}}(\boldsymbol{\lambda}) \stackrel{\text{def}}{=} \underbrace{\mathbb{E}_{q(\mathbf{F} \mid \boldsymbol{\lambda})} \log p(\mathbf{Y} \mid \mathbf{F})}_{\text{expected log likelihood (ELL)}} - \underbrace{\mathrm{KL}\left[q(\mathbf{F} \mid \boldsymbol{\lambda}) \parallel p(\mathbf{F})\right]}_{\text{KL(approx. posterior } \parallel \text{ prior)}}$$

- ELL is a model-fit term and KL is a penalty term

Variational Inference Strategy

- The evidence lower bound $\mathcal{L}_{\text{ELBO}}(\boldsymbol{\lambda})$ can be written as:

$$\mathcal{L}_{\text{ELBO}}(\boldsymbol{\lambda}) \stackrel{\text{def}}{=} \underbrace{\mathbb{E}_{q(\mathbf{F} \mid \boldsymbol{\lambda})} \log p(\mathbf{Y} \mid \mathbf{F})}_{\text{expected log likelihood (ELL)}} - \underbrace{\text{KL}\left[q(\mathbf{F} \mid \boldsymbol{\lambda}) \parallel p(\mathbf{F})\right]}_{\text{KL(approx. posterior} \parallel \text{prior)}}$$

- ELL is a model-fit term and KL is a penalty term

- What family of distributions?
  - As flexible as possible
  - Tractability is the main constraint
  - No risk of over-fitting

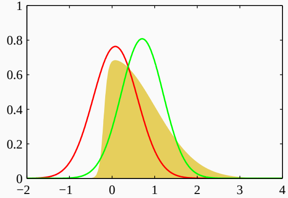


Fig from Bishop (2006)

Variational Inference Strategy

- The evidence lower bound $\mathcal{L}_{\mathrm{ELBO}}(\boldsymbol{\lambda})$ can be written as:

$$\mathcal{L}_{\mathrm{ELBO}}(\boldsymbol{\lambda}) \stackrel{\text{def}}{=} \underbrace{\mathbb{E}_{q(\mathbf{F} \mid \boldsymbol{\lambda})} \log p(\mathbf{Y} \mid \mathbf{F})}_{\text{expected log likelihood (ELL)}} - \underbrace{\mathrm{KL}\left[q(\mathbf{F} \mid \boldsymbol{\lambda}) \,\|\, p(\mathbf{F})\right]}_{\text{KL(approx. posterior } \| \text{ prior)}}$$

- ELL is a model-fit term and KL is a penalty term

- What family of distributions?
  - As flexible as possible
  - Tractability is the main constraint
  - No risk of over-fitting

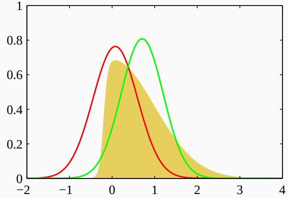


Fig from Bishop (2006)

We want to maximise $\mathcal{L}_{\mathrm{ELBO}}(\boldsymbol{\lambda})$ wrt variational parameters $\boldsymbol{\lambda}$

Automated VI for LGPMs

(Nguyen and Bonilla, NeurIPS, 2014)

Goal: Approximate intractable posterior $p(\mathbf{F} \mid \mathbf{Y})$ with variational distribution

$$q(\mathbf{F} \mid \boldsymbol{\lambda}) = \sum_{k=1}^{K} \pi_k q_k(\mathbf{F} \mid \boldsymbol{\lambda}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{Q} \mathcal{N}(\mathbf{F}_k; \mathbf{m}_{kj}, \mathbf{S}_{kj})$$

with variational parameters $\boldsymbol{\lambda} = \{\mathbf{m}_{kj}, \mathbf{S}_{kj}\}$,

Automated VI for LGPMs

(Nguyen and Bonilla, NeurIPS, 2014)

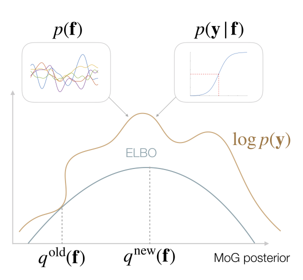
Goal: Approximate intractable posterior $p(\mathbf{F} \mid \mathbf{Y})$ with variational distribution

$$q(\mathbf{F} \mid \boldsymbol{\lambda}) = \sum_{k=1}^{K} \pi_k q_k(\mathbf{F} \mid \boldsymbol{\lambda}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{Q} \mathcal{N}(\mathbf{F}_k; \mathbf{m}_{kj}, \mathbf{S}_{kj})$$

with variational parameters $\boldsymbol{\lambda} = \{\mathbf{m}_{kj}, \mathbf{S}_{kj}\}$,

Recall $\mathcal{L}_{\text{ELBO}}(\boldsymbol{\lambda}) =$ - KL + ELL:

- KL term can be bounded using Jensen's inequality
  - Exact gradients of parameters



Automated VI for LGPMs

(Nguyen and Bonilla, NeurIPS, 2014)

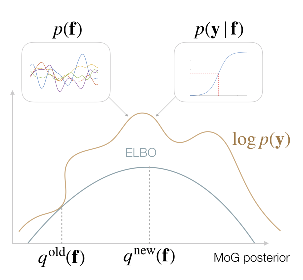
Goal: Approximate intractable posterior $p(\mathbf{F} \mid \mathbf{Y})$ with variational distribution

$$q(\mathbf{F} \mid \boldsymbol{\lambda}) = \sum_{k=1}^{K} \pi_k q_k(\mathbf{F} \mid \boldsymbol{\lambda}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{Q} \mathcal{N}(\mathbf{F}_{k}; \mathbf{m}_{kj}, \mathbf{S}_{kj})$$

with variational parameters $\boldsymbol{\lambda} = \{\mathbf{m}_{kj}, \mathbf{S}_{kj}\}$,

Recall $\mathcal{L}_{\text{ELBO}}(\boldsymbol{\lambda})$ = - KL + ELL:

- KL term can be bounded using Jensen's inequality
  - Exact gradients of parameters



ELL and its gradients can be estimated *efficiently*

Expected Log Likelihood Term

Th.1: Efficient estimation

The ELL and its gradients can be estimated using expectations over univariate Gaussian distributions.

$$q_{k(n)} \stackrel{\text{def}}{=} q_{k(n)}(\mathbf{F}_{\cdot n} \mid \boldsymbol{\lambda}_{k(n)})$$

$$\mathbb{E}_{q_k} \log p(\mathbf{Y} \mid \mathbf{F}) = \sum_{n=1}^{N} \mathbb{E}_{q_{k(n)}} \log p(\mathbf{Y}_{n\cdot} \mid \mathbf{F}_{n\cdot})$$

$$\nabla_{\boldsymbol{\lambda}_{k(n)}} \mathbb{E}_{q_{k(n)}} \log p(\mathbf{Y}_{n\cdot} \mid \mathbf{F}_{n\cdot}) = \mathbb{E}_{q_{k(n)}} \nabla_{\boldsymbol{\lambda}_{k(n)}} \log q_{k(n)}(\mathbf{F}_{\cdot n} \mid \boldsymbol{\lambda}_{k(n)}) \log p(\mathbf{Y}_{n\cdot} \mid \mathbf{F}$$

Expected Log Likelihood Term

Th.1: Efficient estimation

The ELL and its gradients can be estimated using expectations over univariate Gaussian distributions.

$$q_{k(n)} \stackrel{\text{def}}{=} q_{k(n)}(\mathbf{F}_{\cdot n} \,|\, \boldsymbol{\lambda}_{k(n)})$$

$$\mathbb{E}_{q_k} \log p(\mathbf{Y} \,|\, \mathbf{F}) = \sum_{n=1}^{N} \mathbb{E}_{q_{k(n)}} \log p(\mathbf{Y}_{n\cdot} \,|\, \mathbf{F}_{n\cdot})$$

$$\nabla_{\boldsymbol{\lambda}_{k(n)}} \mathbb{E}_{q_{k(n)}} \log p(\mathbf{Y}_{n\cdot} \,|\, \mathbf{F}_{n\cdot}) = \mathbb{E}_{q_{k(n)}} \nabla_{\boldsymbol{\lambda}_{k(n)}} \log q_{k(n)}(\mathbf{F}_{\cdot n} \,|\, \boldsymbol{\lambda}_{k(n)}) \log p(\mathbf{Y}_{n\cdot} \,|\, \mathbf{F}$$

Practical consequences

- Can use unbiased Monte Carlo estimates
- Gradients of the likelihood are not required (only likelihood evaluations)
- Holds $\forall Q \geq 1$

Scalability through Inducing Variables and Stochastic Variational Inference (SVI)

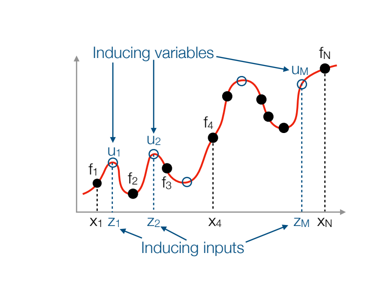
Inducing Variables in GP Models

Inducing variables $\mathbf{u}$

- Latent values of the GP, as $\mathbf{f}$ and $\mathbf{f}_*$
- Usually marginalized (integrated out)

Inducing inputs $\mathbf{Z}$

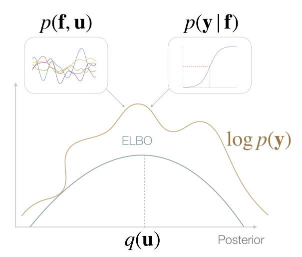
- Corresponding input location, as $\mathbf{x}$
- Imprint on final solution



Generalization of "support points", "active set", "pseudo-inputs"

Variational Learning of Inducing Variables (Titsias, AISTATS, 2009)

- Augmented prior $p(\mathbf{f}, \mathbf{u}) = p(\mathbf{f} \mid \mathbf{u})p(\mathbf{u})$, exact marginal $p(\mathbf{f})$
- Approximate posterior $q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f} \mid \mathbf{u})q(\mathbf{u})$

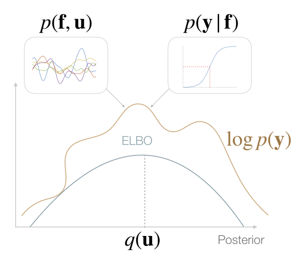


Variational Learning of Inducing Variables

(Titisias, AISTATS, 2009)

- Augmented prior $p(\mathbf{f}, \mathbf{u}) = p(\mathbf{f} \mid \mathbf{u})p(\mathbf{u})$, exact marginal $p(\mathbf{f})$
- Approximate posterior $q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f} \mid \mathbf{u})q(\mathbf{u})$

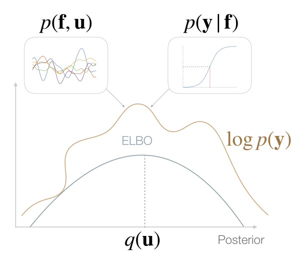
- Cubic operations on N 'vanish'



Variational Learning of Inducing Variables (Titisias, AISTATS, 2009)

- Augmented prior $p(\mathbf{f}, \mathbf{u}) = p(\mathbf{f} \mid \mathbf{u})p(\mathbf{u})$, exact marginal $p(\mathbf{f})$
- Approximate posterior $q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f} \mid \mathbf{u})q(\mathbf{u})$

- Cubic operations on N 'vanish'
- Exact optimal solution for Gaussian likelihood

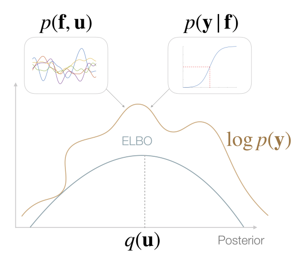


Variational Learning of Inducing Variables

(Titisias, AISTATS, 2009)

- Augmented prior $p(\mathbf{f}, \mathbf{u}) = p(\mathbf{f} \mid \mathbf{u})p(\mathbf{u})$, exact marginal $p(\mathbf{f})$
- Approximate posterior $q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f} \mid \mathbf{u})q(\mathbf{u})$

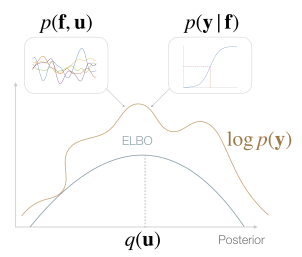
- Cubic operations on N ‘vanish’
- Exact optimal solution for Gaussian likelihood
- Hyper-parameters and inducing inputs optimized *jointly*



Variational Learning of Inducing Variables (Titisias, AISTATS, 2009)

- Augmented prior $p(\mathbf{f}, \mathbf{u}) = p(\mathbf{f} \mid \mathbf{u}) p(\mathbf{u})$, exact marginal $p(\mathbf{f})$
- Approximate posterior $q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f} \mid \mathbf{u}) q(\mathbf{u})$

- Cubic operations on N 'vanish'
- Exact optimal solution for Gaussian likelihood
- Hyper-parameters and inducing inputs optimized *jointly*



Computation dominated by:

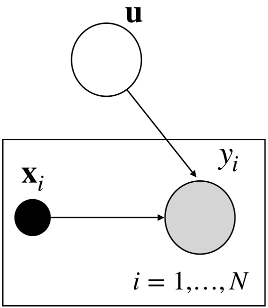
$$\mathbf{K}_{\mathbf{XZ}} \mathbf{K}_{\mathbf{ZZ}}^{-1} \mathbf{K}_{\mathbf{ZX}}$$

Time cost $\mathcal{O}(NM^2)$, *can we do better?*

Stochastic Variational Inference for GP Models

Maintain an explicit representation of $q(\mathbf{u}) = \mathcal{N}(\mathbf{m}, \mathbf{S})$

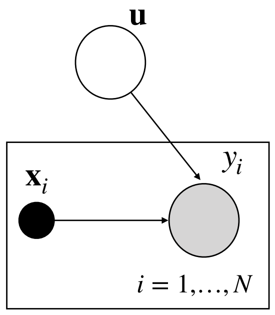
- Inducing variables act as global variables
- ELBO decomposes across observations
- Use stochastic optimization
- $\mathbf{K}_{\mathbf{x}_i \mathbf{Z}} \mathbf{K}_{\mathbf{ZZ}}^{-1} \mathbf{K}_{\mathbf{Z} \mathbf{x}_i}$: Time cost $\mathcal{O}(M^3) \rightarrow$ big data!



Stochastic Variational Inference for GP Models

Maintain an explicit representation of $q(\mathbf{u}) = \mathcal{N}(\mathbf{m}, \mathbf{S})$

- Inducing variables act as global variables
- ELBO decomposes across observations
- Use stochastic optimization
- $\mathbf{K}_{\mathbf{x}_i \mathbf{Z}} \mathbf{K}_{\mathbf{Z}\mathbf{Z}}^{-1} \mathbf{K}_{\mathbf{Z}\mathbf{x}_i}$: Time cost $\mathcal{O}(M^3) \rightarrow$ big data!

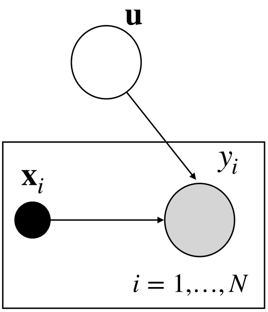


- Converge to optimal solution for Gaussian likelihoods (Hensman et al, UAI, 2013)

Stochastic Variational Inference for GP Models

Maintain an explicit representation of $q(\mathbf{u}) = \mathcal{N}(\mathbf{m}, \mathbf{S})$

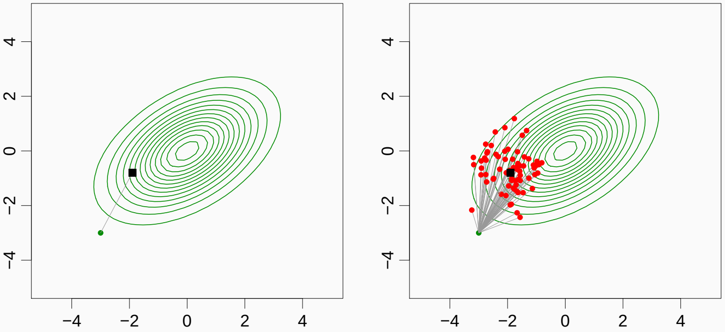
- Inducing variables act as global variables
- ELBO decomposes across observations
- Use stochastic optimization
- $\mathbf{K}_{\mathbf{x}_i \mathbf{Z}} \mathbf{K}_{\mathbf{ZZ}}^{-1} \mathbf{K}_{\mathbf{Z}\mathbf{x}_i}$: Time cost $\mathcal{O}(M^3) \rightarrow$ big data!



- Converge to optimal solution for Gaussian likelihoods (Hensman et al, UAI, 2013)
- Generalization to LGPMs (Dezfouli & Bonilla, NeurIPS, 2015)

Stochastic Gradient Optimization

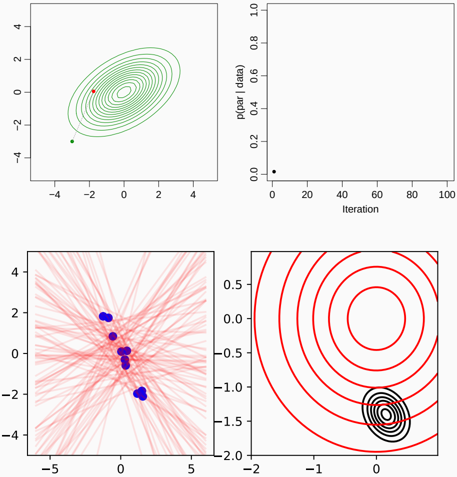
$$\mathbb{E}\left\{\widetilde{\nabla_{\text{vpar}}\text{LowerBound}}\right\} = \nabla_{\text{vpar}}\text{LowerBound}$$



Robbins and Monro, *AoMS*, 1951

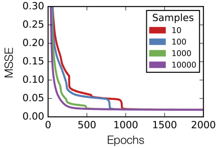
Stochastic Variational Inference

$$\text{vpar}' = \text{vpar} + \frac{\alpha_t}{2}\widetilde{\nabla_{\text{vpar}}}(\text{LowerBound}) \qquad \alpha_t \to 0$$



Further Developments: AutoGP

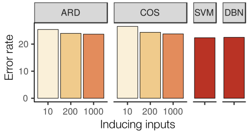
(Krauth et al UAI, 2017)



Scalability & efficient computation

Low-variance gradient estimates

- ⋆ Breaks error-barrier on MNIST for GP models
- ⋆ Unprecedented scale



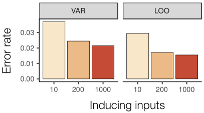
Well-targeted objective functions

Leave-one-out hyper-parameter learning

The holy trinity of machine learning

Representational power

Flexible kernels



Conclusion

- LGPMs: General framework for GP priors and non-linear likelihoods
- Applications in multi-class classification, multi-output regression, modelling count data and more
- Generic inference via optimisation of the variational objective (ELBO)
- Scalability via inducing-variable approach
- AutoGP