Variational Inference and Natural Gradients for GPs

Juan José Giraldo Gutierrez

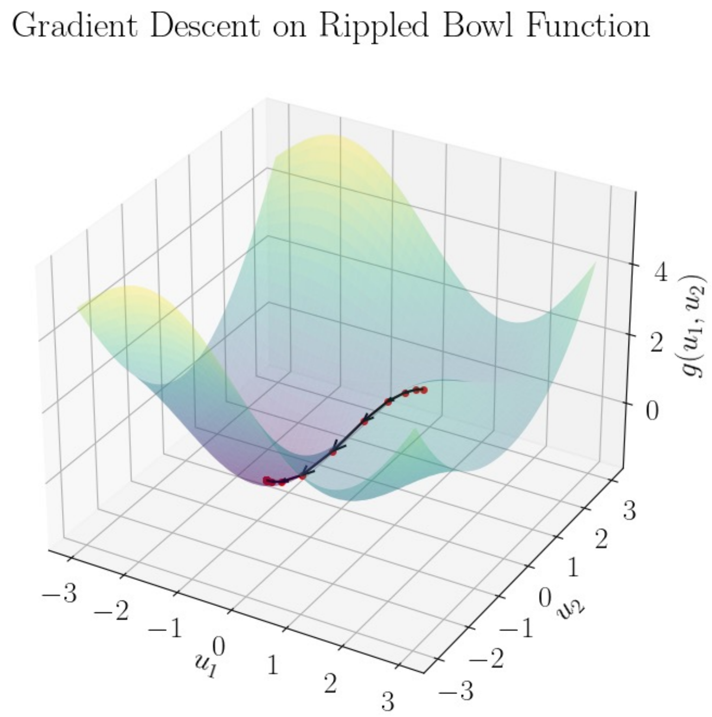
Gaussian Processes Summer School
Sept. 9, 2025

Imperial College London

Gradient Descent

$$\min_{\mathbf{u}\in\mathbb{R}^M} g(\mathbf{u})$$

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \alpha \nabla_{\mathbf{u}} g(\mathbf{u}_k)$$



α is a step-size or learning rate

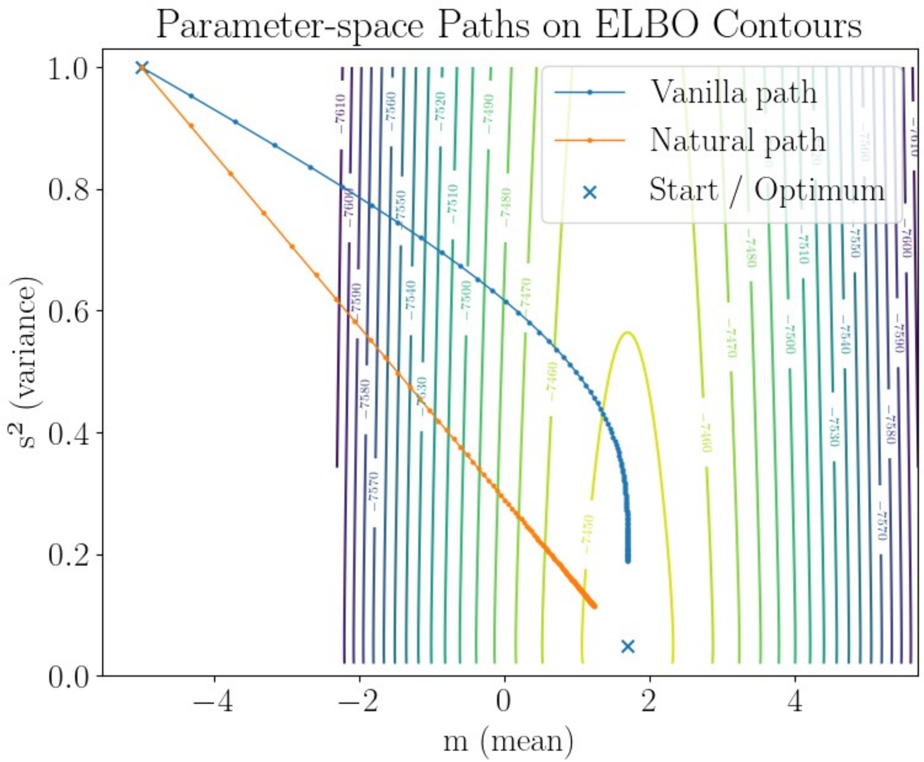
Natural Gradient

Parameters update

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha \mathbf{F}^{-1} \nabla_{\boldsymbol{\theta}} \tilde{\mathcal{L}}$$

Fisher Information Matrix (FIM)

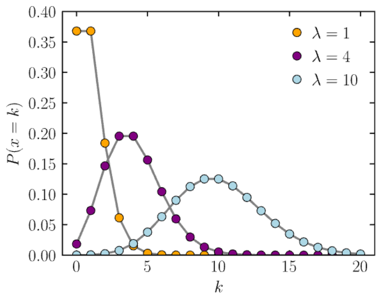
$$\mathbf{F} = -\mathbb{E}_{q(\mathbf{u}|\boldsymbol{\theta})}[\nabla^2_{\boldsymbol{\theta}} \log q(\mathbf{u}|\boldsymbol{\theta})]$$



Amari, S. (1998).

Exponential Family:

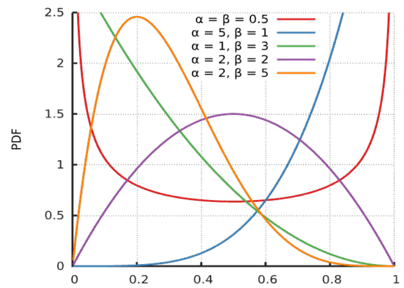
Poisson



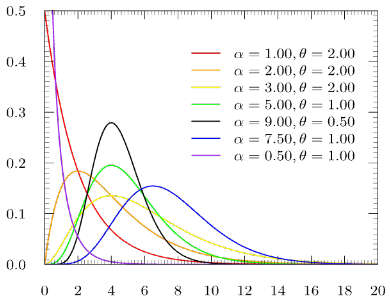
$$q(\mathbf{u}|\boldsymbol{\theta}) = h(\mathbf{u}) \exp(\boldsymbol{\theta}^\top \mathbf{t}(\mathbf{u}) - A(\boldsymbol{\theta}))$$

$\boldsymbol{\theta}$	$\mathbf{t}(\mathbf{u})$	$A(\boldsymbol{\theta})$	$h(\mathbf{u})$
Natural parameters	Sufficient statistics	Log partition function	Base measure

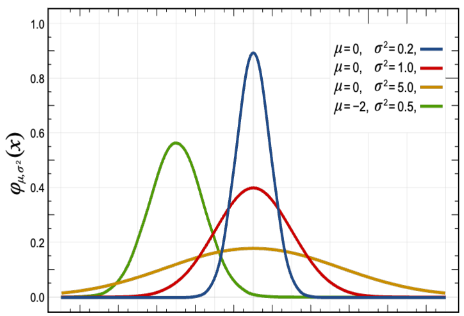
Beta



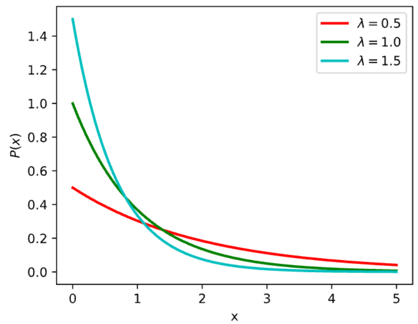
Gamma



Gaussian



Exponential



Kevin P. Murphy, (2022)

Exponential Family:

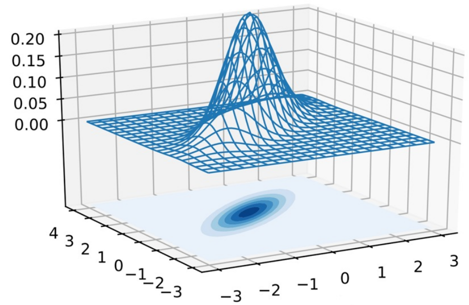
$$q(\mathbf{u}|\boldsymbol{\theta}) = h(\mathbf{u}) \exp(\boldsymbol{\theta}^\top \mathbf{t}(\mathbf{u}) - A(\boldsymbol{\theta}))$$

$$\log q(\mathbf{u}|\boldsymbol{\theta}) = \log h(\mathbf{u}) + \boldsymbol{\theta}^\top \mathbf{t}(\mathbf{u}) - A(\boldsymbol{\theta})$$

Fisher Information

$$\boxed{\mathbf{F} = \nabla^2_{\boldsymbol{\theta}} A(\boldsymbol{\theta})}$$

$$q(\mathbf{u}|\mathbf{m}, \mathbf{S}) = \mathcal{N}(\mathbf{u}|\mathbf{m}, \mathbf{S})$$



Natural Parameters

$$\boxed{\boldsymbol{\theta}_1 = \mathbf{S}^{-1}\mathbf{m}}$$

$$\boxed{\boldsymbol{\theta}_2 = -\tfrac{1}{2}\mathbf{S}^{-1}}$$

Sufficient statistics

$$\boxed{\mathbf{t_1}(\mathbf{u}) = \mathbf{u}}$$

$$\boxed{\mathbf{t_2}(\mathbf{u}) = \mathbf{u}\mathbf{u}^\top}$$

Deriving the FIM using the Exponential Family:

Goal

$$\boxed{\mathbf{F} = \nabla^2_{\boldsymbol{\theta}} A(\boldsymbol{\theta})}$$

$$\log q(\mathbf{u}|\boldsymbol{\theta}) = \log h(\mathbf{u}) + \boldsymbol{\theta}^\top \mathbf{t}(\mathbf{u}) - A(\boldsymbol{\theta})$$

Computing Derivative:

$$\nabla_{\boldsymbol{\theta}} \log q(\mathbf{u}|\boldsymbol{\theta}) = \mathbf{t}(\mathbf{u}) - \nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta})$$

Deriving the FIM using the Exponential Family:

Goal

$$\boxed{\mathbf{F} = \nabla^2_{\boldsymbol{\theta}} A(\boldsymbol{\theta})}$$

$$\log q(\mathbf{u}|\boldsymbol{\theta}) = \log h(\mathbf{u}) + \boldsymbol{\theta}^\top \mathbf{t}(\mathbf{u}) - A(\boldsymbol{\theta})$$

Computing Derivative:

Mean Parameters

$$\boldsymbol{\eta}_1 = \mathbb{E}[\mathbf{u}] = \mathbf{m}$$

$$\boldsymbol{\eta}_2 = \mathbb{E}[\mathbf{u}\mathbf{u}^\top] = \mathbf{S} + \mathbf{m}\mathbf{m}^\top$$

$$\nabla_{\boldsymbol{\theta}} \log q(\mathbf{u}|\boldsymbol{\theta}) = \mathbf{t}(\mathbf{u}) - \nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta})$$

Cool Property:

$$\mathbb{E}_{q(\mathbf{u}|\boldsymbol{\theta})}[\nabla_{\boldsymbol{\theta}} \log q(\mathbf{u}|\boldsymbol{\theta})] = \mathbb{E}_{q(\mathbf{u}|\boldsymbol{\theta})}[\mathbf{t}(\mathbf{u})] - \nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta})$$

$$\nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta}) = \mathbb{E}_{q(\mathbf{u}|\boldsymbol{\theta})}[\mathbf{t}(\mathbf{u})]$$

$$\boxed{\nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta}) = \boldsymbol{\eta}}$$

Deriving the FIM using the Exponential Family:

$$\log q(\mathbf{u}|\boldsymbol{\theta}) = \log h(\mathbf{u}) + \boldsymbol{\theta}^\top \mathbf{t}(\mathbf{u}) - A(\boldsymbol{\theta})$$

Computing Derivative:

$$\nabla_{\boldsymbol{\theta}} \log q(\mathbf{u}|\boldsymbol{\theta}) = \mathbf{t}(\mathbf{u}) - \nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta})$$

Goal

$$\mathbf{F} = \nabla^2_{\boldsymbol{\theta}} A(\boldsymbol{\theta})$$

Computing Second Derivative:

Deriving the FIM using the Exponential Family:

$$\log q(\mathbf{u}|\boldsymbol{\theta}) = \log h(\mathbf{u}) + \boldsymbol{\theta}^\top \mathbf{t}(\mathbf{u}) - A(\boldsymbol{\theta})$$

Computing Derivative:

$$\nabla_{\boldsymbol{\theta}} \log q(\mathbf{u}|\boldsymbol{\theta}) = \mathbf{t}(\mathbf{u}) - \nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta})$$

Goal

$$\boxed{\mathbf{F} = \nabla^2_{\boldsymbol{\theta}} A(\boldsymbol{\theta})}$$

Computing Second Derivative:

$$\nabla^2_{\boldsymbol{\theta}} \log q(\mathbf{u}|\boldsymbol{\theta}) = -\nabla^2_{\boldsymbol{\theta}} A(\boldsymbol{\theta})$$

$$-\mathbb{E}_{q(\mathbf{u}|\boldsymbol{\theta})}[\nabla^2_{\boldsymbol{\theta}} \log q(\mathbf{u}|\boldsymbol{\theta})] = \nabla^2_{\boldsymbol{\theta}} A(\boldsymbol{\theta})$$

$$\boxed{\mathbf{F} = -\mathbb{E}_{q(\mathbf{u}|\boldsymbol{\theta})}[\nabla^2_{\boldsymbol{\theta}} \log q(\mathbf{u}|\boldsymbol{\theta})]}$$

Amari, S. (1998).

Now what? Do we have to compute the inverse of the FIM?

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha \mathbf{F}^{-1} \nabla_{\boldsymbol{\theta}} \tilde{\mathcal{L}}$$

Elegant Gradient update using the "Inverse of the FIM"

Cool Property: $\boxed{\nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta}) = \boldsymbol{\eta}}$

$$\mathbf{F} = \nabla_{\boldsymbol{\theta}}[\nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta})] = \nabla_{\boldsymbol{\theta}} \boldsymbol{\eta} = \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}} \qquad \mathbf{F}^{-1} = \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}}\right)^{-1} = \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}}$$

Elegant Gradient update using the "Inverse of the FIM"

Cool Property: $\boxed{\nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta}) = \boldsymbol{\eta}}$

$$\mathbf{F} = \nabla_{\boldsymbol{\theta}}[\nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta})] = \nabla_{\boldsymbol{\theta}} \boldsymbol{\eta} = \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}} \qquad \mathbf{F}^{-1} = \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}}\right)^{-1} = \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}}$$

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha \mathbf{F}^{-1} \nabla_{\boldsymbol{\theta}} \tilde{\mathcal{L}}$$

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}} \frac{\partial \tilde{\mathcal{L}}}{\partial \boldsymbol{\theta}} \qquad \boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha \frac{\partial \tilde{\mathcal{L}}}{\partial \boldsymbol{\eta}}$$

$$\boxed{\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha \nabla_{\boldsymbol{\eta}} \tilde{\mathcal{L}}}$$

Salimbeni, H. et al (2018)

Computing the updates for: $q(\mathbf{u}|\mathbf{m}, \mathbf{S}) = \mathcal{N}(\mathbf{u}|\mathbf{m}, \mathbf{S})$

Natural Parameters

$$\boxed{\boldsymbol{\theta}_1 = \mathbf{S}^{-1}\mathbf{m}}$$

$$\boxed{\boldsymbol{\theta}_2 = -\tfrac{1}{2}\mathbf{S}^{-1}}$$

Mean Parameters

$$\boldsymbol{\eta}_1 = \mathbb{E}[\mathbf{u}] = \mathbf{m}$$

$$\boldsymbol{\eta}_2 = \mathbb{E}[\mathbf{u}\mathbf{u}^\top] = \mathbf{S} + \mathbf{m}\mathbf{m}^\top$$

Relations and Grads

$$\mathbf{S} = \boldsymbol{\eta}_2 - \boldsymbol{\eta}_1\boldsymbol{\eta}_1{}^\top$$

$$\mathbf{m} = \boldsymbol{\eta}_1$$

$$\frac{\partial \tilde{\mathcal{L}}}{\partial \boldsymbol{\eta}_1} = \nabla_{\mathbf{m}}\tilde{\mathcal{L}} - 2\,\nabla_{\mathbf{S}}\tilde{\mathcal{L}}\,\mathbf{m}$$

$$\frac{\partial \tilde{\mathcal{L}}}{\partial \boldsymbol{\eta}_2} = \nabla_{\mathbf{S}}\tilde{\mathcal{L}}$$

Natural Gradient Updates

$$\boxed{\boldsymbol{\theta}_2^+ = \boldsymbol{\theta}_2 - \alpha\nabla_{\boldsymbol{\eta}_2}\tilde{\mathcal{L}}}$$

$$-\tfrac{1}{2}\,\mathbf{S}_{k+1}^{-1} = -\tfrac{1}{2}\,\mathbf{S}_k^{-1} - \alpha\nabla_{\mathbf{S}}\tilde{\mathcal{L}}_k$$

$$\boxed{\mathbf{S}_{k+1}^{-1} = \mathbf{S}_k^{-1} + 2\alpha\nabla_{\mathbf{S}}\tilde{\mathcal{L}}_k}$$

$$\boxed{\boldsymbol{\theta}_1^+ = \boldsymbol{\theta}_1 - \alpha\nabla_{\boldsymbol{\eta}_1}\tilde{\mathcal{L}}}$$

$$\mathbf{S}_{k+1}^{-1}\mathbf{m}_{k+1} = \mathbf{S}_k^{-1}\mathbf{m}_k - \alpha\nabla_{\mathbf{m}}\tilde{\mathcal{L}}_k + 2\alpha\nabla_{\mathbf{S}}\tilde{\mathcal{L}}_k\,\mathbf{m}_k$$

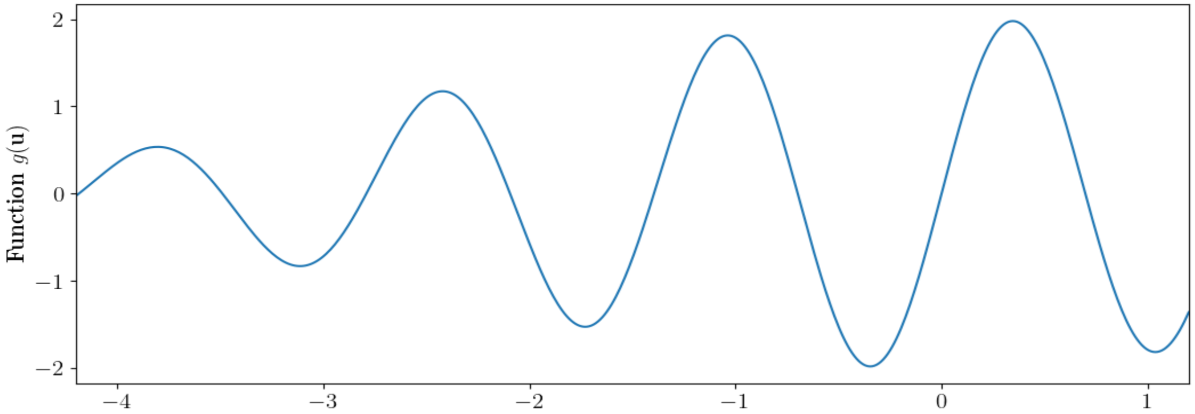
$$\mathbf{S}_{k+1}^{-1}\mathbf{m}_{k+1} = (\mathbf{S}_k^{-1} + 2\alpha\nabla_{\mathbf{S}}\tilde{\mathcal{L}}_k)\mathbf{m}_k - \alpha\nabla_{\mathbf{m}}\tilde{\mathcal{L}}_k$$

$$\mathbf{m}_{k+1} = \mathbf{S}_{k+1}[(\mathbf{S}_k^{-1} + 2\alpha\nabla_{\mathbf{S}}\tilde{\mathcal{L}}_k)\mathbf{m}_k - \alpha\nabla_{\mathbf{m}}\tilde{\mathcal{L}}_k]$$

$$\boxed{\mathbf{m}_{k+1} = \mathbf{m}_k - \alpha\mathbf{S}_{k+1}\nabla_{\mathbf{m}}\tilde{\mathcal{L}}_k}$$

Khan and Lin (2017, 2018, 2019); J. -J. Giraldo and M. A. Álvarez (2022)

Variational Optimization



Upper bound:

$$g^\star = \min_{\mathbf{u}\in\mathbb{R}^M} g(\mathbf{u}) \leq \mathbb{E}_{q(\mathbf{u}|\boldsymbol{\theta})}[g(\mathbf{u})]$$

$$\tilde{\mathcal{L}} := \mathbb{E}_{q(\mathbf{u}|\boldsymbol{\theta})}[g(\mathbf{u})]$$

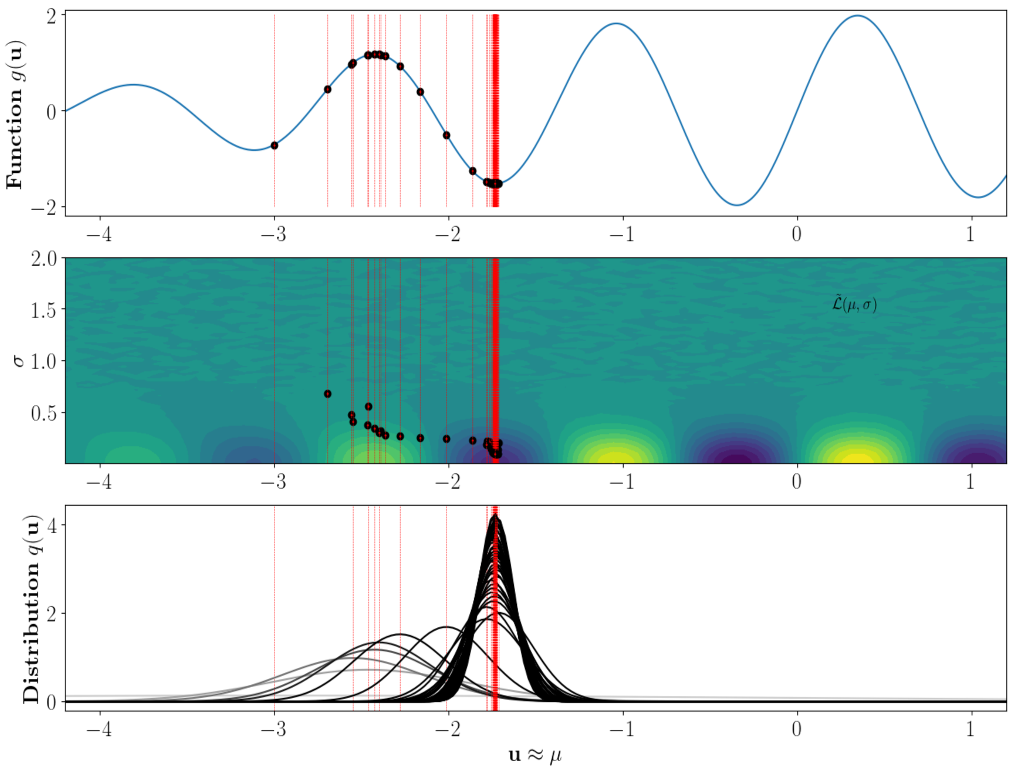
Staines, J. and Barber, D. (2013)

Variational Optimization

$$g^\star = \min_{\mathbf{u}\in\mathbb{R}^M} g(\mathbf{u}) \leq \mathbb{E}_{q(\mathbf{u}|\boldsymbol{\theta})}[g(\mathbf{u})]$$

$$\tilde{\mathcal{L}} := \mathbb{E}_{q(\mathbf{u}|\boldsymbol{\theta})}[g(\mathbf{u})]$$

$$q(\mathbf{u}|\boldsymbol{\theta}) \rightarrow q(u|\mu, \sigma^2)$$



Variational Optimization

Using **penalization** to avoid the (co) variance to collapse (i.e.,):

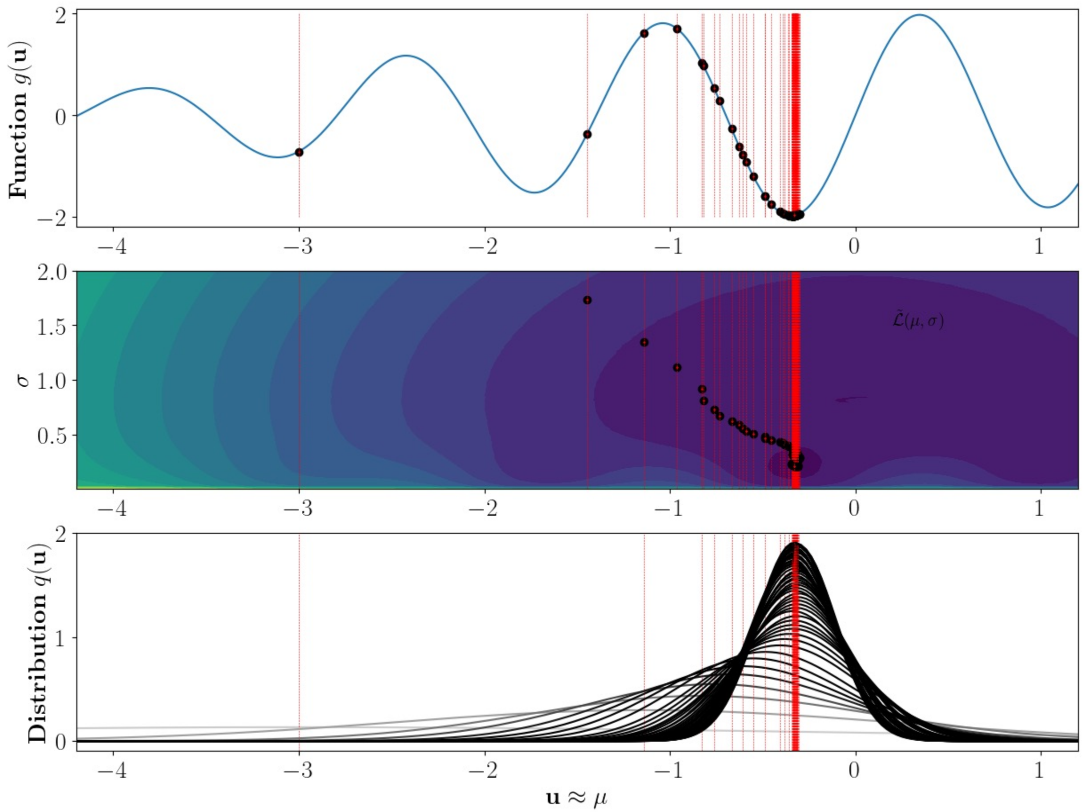
$$\tilde{\mathcal{L}} = \mathbb{E}_{q(\mathbf{u}|\boldsymbol{\theta})}[g(\mathbf{u})] + \mathrm{KL}[q(\mathbf{u}|\boldsymbol{\theta})||p(\mathbf{u})]$$

Variational Optimization

$$g^{\star} = \min_{\mathbf{u}\in\mathbb{R}^M} g(\mathbf{u}) \leq \mathbb{E}_{q(\mathbf{u}|\boldsymbol{\theta})}[g(\mathbf{u})] + \mathrm{KL}[q(\mathbf{u}|\boldsymbol{\theta})||p(\mathbf{u})]$$

$$\tilde{\mathcal{L}} = \mathbb{E}_{q(\mathbf{u}|\boldsymbol{\theta})}[g(\mathbf{u})] + \mathrm{KL}[q(\mathbf{u}|\boldsymbol{\theta})||p(\mathbf{u})]$$

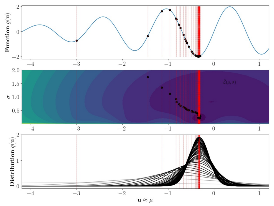
$$q(\mathbf{u}|\boldsymbol{\theta}) \rightarrow q(u|\mu, \sigma^2)$$



J. -J. Giraldo and M. A. Álvarez, (2022)

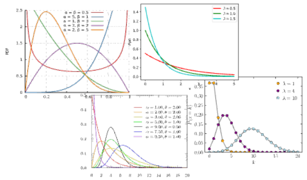
Summary of Key Points:

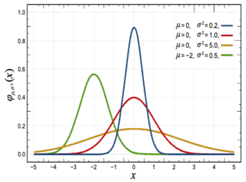
Variational Optimization:



$$\tilde{\mathcal{L}} = \mathbb{E}_{q(\mathbf{u}|\boldsymbol{\theta})}[g(\mathbf{u})] + \mathrm{KL}[q(\mathbf{u}|\boldsymbol{\theta})||p(\mathbf{u})]$$

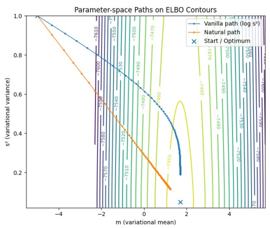
Exponential Family:





$$q(\mathbf{u}|\boldsymbol{\theta}) = h(\mathbf{u})\exp(\boldsymbol{\theta}^\top \mathbf{t}(\mathbf{u}) - A(\boldsymbol{\theta}))$$

Natural Gradient Optimization Tool:



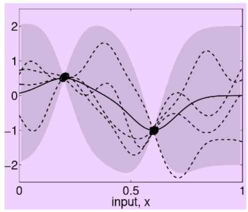
$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha \mathbf{F}^{-1}\nabla_{\boldsymbol{\theta}}\tilde{\mathcal{L}} \longleftrightarrow \boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha\nabla_{\boldsymbol{\eta}}\tilde{\mathcal{L}}$$

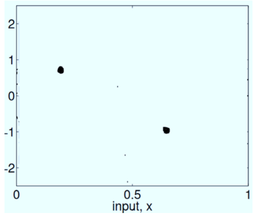
Grads w.r.t natural parameters $\boldsymbol{\theta}$

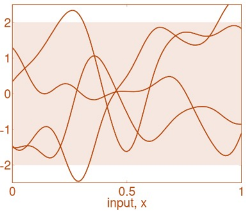
Grads w.r.t mean parameters $\boldsymbol{\eta} = \mathbb{E}_{q(\mathbf{u}|\boldsymbol{\theta})}[\mathbf{t}(\mathbf{u})]$

Bayes' Theorem applied to a Gaussian Process Model

$y \in \mathbb{R}$







$$\underbrace{p(\mathbf{f}|\mathbf{y}, X)}_{\text{Posterior}} \propto \underbrace{p(\mathbf{y}|\mathbf{f}, X)}_{\text{Likelihood}} \underbrace{p(\mathbf{f}|X)}_{\text{Prior}}$$

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

Gaussian Process (GP)

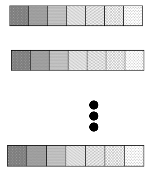
$f(\cdot) \sim \mathcal{GP}(0, k(\cdot, \cdot'))$

$\mathbf{x}_1$

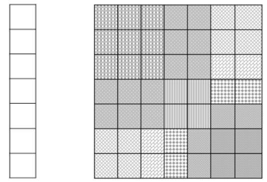
$\mathbf{x}_2$

$\vdots$

$\mathbf{x}_N$



$\mathbf{f} = f(\mathbf{X})$



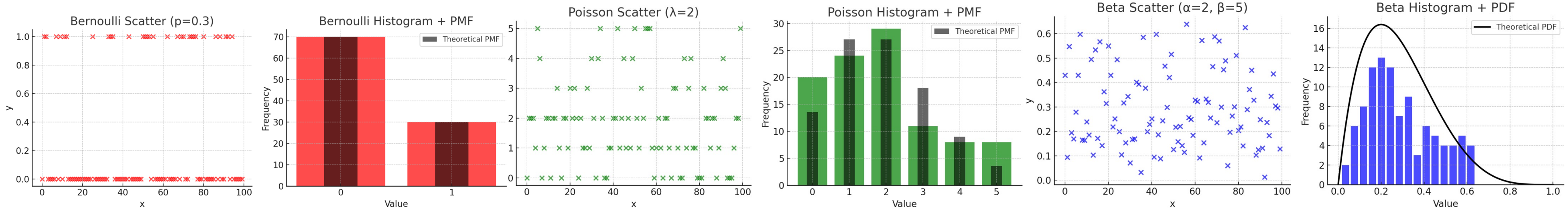
$$p(\mathbf{f}|X) = \mathcal{N}(\mathbf{0}, K)$$

Rasmussen C. E. and Williams C.K.I. (2006)

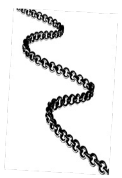
Bayes' Theorem applied to a Gaussian Process Model

Likelihood:

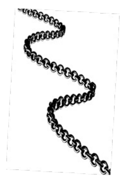
$y \in \{0, 1\}$ $\qquad$ $y \in \mathbb{N}_0$ $\qquad$ $y \in [0, 1]$



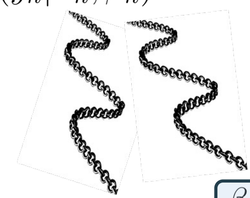
$$\prod_{n=1}^{N} \mathrm{Bernoulli}(y_n|\pi_n) \qquad \prod_{n=1}^{N} \mathrm{Poisson}(y_n|\lambda_n) \qquad \prod_{n=1}^{N} \mathrm{Beta}(y_n|\alpha_n, \beta_n)$$



$$\pi_n = \sigma(f_1(\mathbf{x}_n))$$



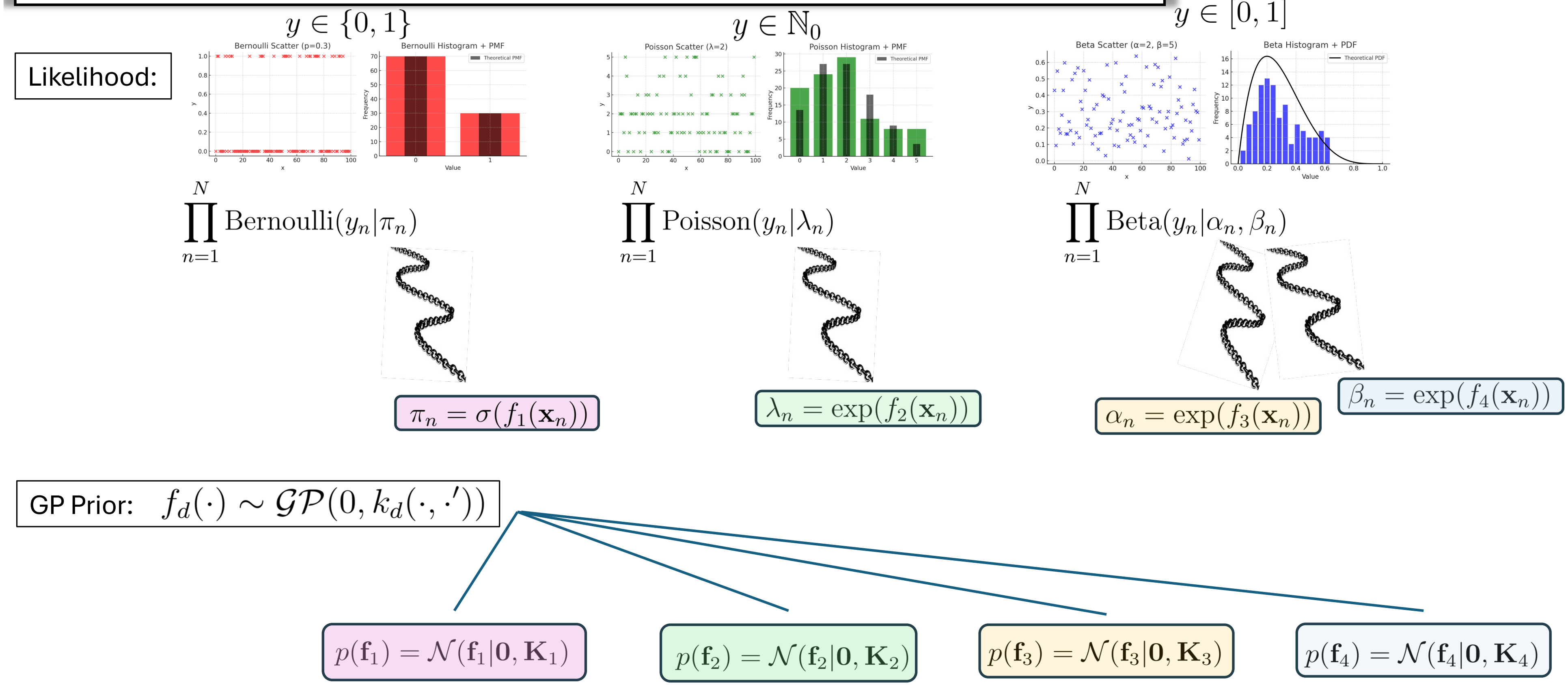
$$\lambda_n = \exp(f_2(\mathbf{x}_n))$$



$$\alpha_n = \exp(f_3(\mathbf{x}_n)) \qquad \beta_n = \exp(f_4(\mathbf{x}_n))$$

Saul, A. D. et al (2016)

The Bayes' Theorem applied to a Gaussian Process Model



The Bayes' Theorem applied to a Gaussian Process Model

$f_d(\cdot) \sim \mathcal{GP}(0, k_d(\cdot, \cdot'))$

Posterior Likelihood Prior

$$\mathcal{N}(\mathbf{f}_0|\mathbf{y}) \propto \prod_{n=1}^{N} \mathcal{N}(y_n|\mu_n, \sigma^2)\, \mathcal{N}(\mathbf{f}_0|\mathbf{0}, \mathbf{K}_0)$$

$$p(\mathbf{f}_1|\mathbf{y}) \propto \prod_{n=1}^{N} \mathrm{Bernoulli}(y_n|\pi_n)\, \mathcal{N}(\mathbf{f}_1|\mathbf{0}, \mathbf{K}_1)$$

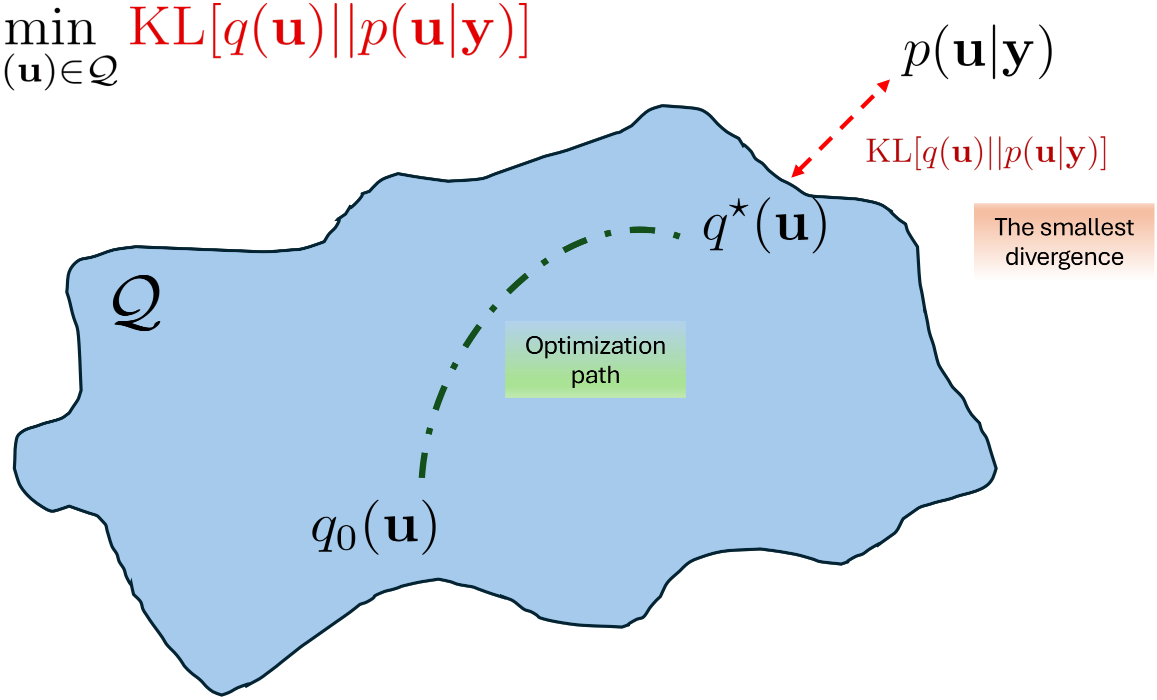
$$p(\mathbf{f}_2|\mathbf{y}) \propto \prod_{n=1}^{N} \mathrm{Poisson}(y_n|\lambda_n)\, \mathcal{N}(\mathbf{f}_2|\mathbf{0}, \mathbf{K}_2)$$

$$p(\mathbf{f}_3, \mathbf{f}_4|\mathbf{y}) \propto \prod_{n=1}^{N} \mathrm{Beta}(y_n|\alpha_n, \beta_n)\, \mathcal{N}(\mathbf{f}_3|\mathbf{0}, \mathbf{K}_3)\, \mathcal{N}(\mathbf{f}_4|\mathbf{0}, \mathbf{K}_4)$$

Moreno-Muñoz, P. et al (2018)

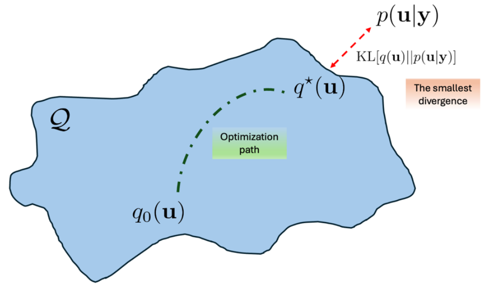
Variational Inference

$$q^{\star}(\mathbf{u}) = \arg \min_{q(\mathbf{u}) \in \mathcal{Q}} \mathrm{KL}[q(\mathbf{u}) || p(\mathbf{u}|\mathbf{y})]$$



Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017)

Variational Inference



$$
\begin{aligned}
\mathrm{KL}[q(\mathbf{u})||p(\mathbf{u}|\mathbf{y})] &= \mathbb{E}_q[\log q(\mathbf{u})] - \mathbb{E}_q[\log p(\mathbf{u}|\mathbf{y})] \\
&= \mathbb{E}_q[\log q(\mathbf{u})] - \mathbb{E}_q[\log p(\mathbf{y},\mathbf{u})] + \mathbb{E}_q[\log p(\mathbf{y})] \\
&= \mathbb{E}_q[\log q(\mathbf{u})] - \mathbb{E}_q[\log p(\mathbf{y},\mathbf{u})] + \log p(\mathbf{y})
\end{aligned}
$$

Given that,

$$\mathrm{KL}[q(\mathbf{u})||p(\mathbf{u}|\mathbf{y})] > 0$$

Thus,

$$\mathbb{E}_q[\log q(\mathbf{u})] - \mathbb{E}_q[\log p(\mathbf{y},\mathbf{u})] + \log p(\mathbf{y}) \geq 0$$

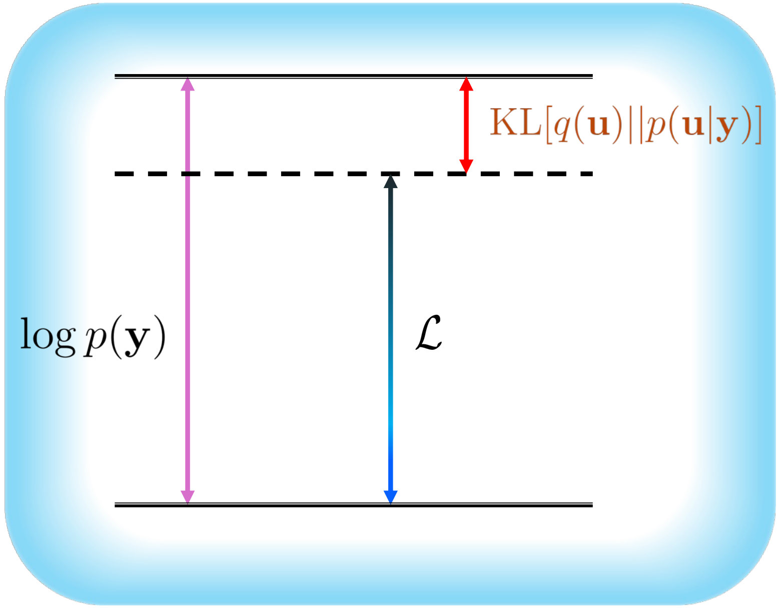
Variational Inference

Bound to the Log marginal likelihood

$$\log p(\mathbf{y}) \geq \mathbb{E}_q[\log p(\mathbf{y}|\mathbf{u})p(\mathbf{u})] - \mathbb{E}_q[\log q(\mathbf{u})]$$

$$\geq \mathbb{E}_q[\log p(\mathbf{y}|\mathbf{u})] - \mathbb{E}_q[\log q(\mathbf{u})] + \mathbb{E}_q[\log p(\mathbf{u})]$$

$$\geq \mathbb{E}_q[\log p(\mathbf{y}|\mathbf{u})] - \mathrm{KL}[q(\mathbf{u})||p(\mathbf{u})]$$



Also known as the Evidence Lower Bound (ELBO)

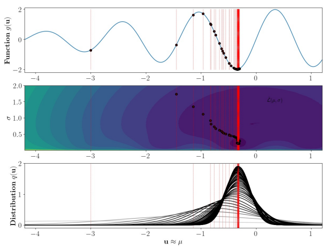
$$\mathcal{L} = \mathbb{E}_{q(\mathbf{u}|\boldsymbol{\theta})}[\log p(\mathbf{y}|\mathbf{u})] - \mathrm{KL}[q(\mathbf{u}|\boldsymbol{\theta})||p(\mathbf{u})]$$

or

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{u}|\boldsymbol{\theta})}\left[\log \frac{p(\mathbf{y}|\mathbf{u})p(\mathbf{u})}{q(\mathbf{u}|\boldsymbol{\theta})}\right]$$

Key concept to take home:

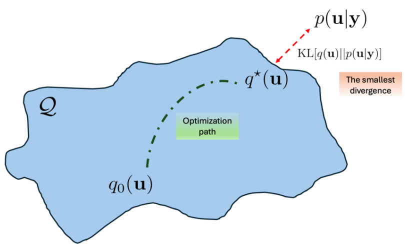
VO



VI is a VO (with penalization) where the objective function is the Negative Log Likelihood

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{u}|\boldsymbol{\theta})}[\log p(\mathbf{y}|\mathbf{u})] - \mathrm{KL}[q(\mathbf{u}|\boldsymbol{\theta})||p(\mathbf{u})]$$

VI



$$-\mathcal{L} = \mathbb{E}_{q(\mathbf{u}|\boldsymbol{\theta})}[\underbrace{-\log p(\mathbf{y}|\mathbf{u})}_{g(\mathbf{u})}] + \mathrm{KL}[q(\mathbf{u}|\boldsymbol{\theta})||p(\mathbf{u})]$$

Computing our Variational Posterior: $q(\mathbf{u}|\mathbf{m}, \mathbf{S}) = \mathcal{N}(\mathbf{u}|\mathbf{m}, \mathbf{S})$

Natural Parameters

$$\boxed{\boldsymbol{\theta}_1 = \mathbf{S}^{-1}\mathbf{m}}$$

$$\boxed{\boldsymbol{\theta}_2 = -\tfrac{1}{2}\mathbf{S}^{-1}}$$

Mean Parameters

$$\boldsymbol{\eta}_1 = \mathbb{E}[\mathbf{u}] = \mathbf{m}$$

$$\boldsymbol{\eta}_2 = \mathbb{E}[\mathbf{u}\mathbf{u}^\top] = \mathbf{S} + \mathbf{m}\mathbf{m}^\top$$

Relations and Grads

$$\mathbf{S} = \boldsymbol{\eta}_2 - \boldsymbol{\eta}_1\boldsymbol{\eta}_1^\top$$

$$\mathbf{m} = \boldsymbol{\eta}_1$$

$$\frac{\partial \tilde{\mathcal{L}}}{\partial \boldsymbol{\eta}_1} = \nabla_{\mathbf{m}}\tilde{\mathcal{L}} - 2\,\nabla_{\mathbf{S}}\tilde{\mathcal{L}}\,\mathbf{m}$$

$$\frac{\partial \tilde{\mathcal{L}}}{\partial \boldsymbol{\eta}_2} = \nabla_{\mathbf{S}}\tilde{\mathcal{L}}$$

Natural Gradient Updates

$$\boxed{\boldsymbol{\theta}_2^{+} = \boldsymbol{\theta}_2 - \alpha\nabla_{\boldsymbol{\eta}_2}\tilde{\mathcal{L}}}$$

$$-\tfrac{1}{2}\mathbf{S}_{k+1}^{-1} = -\tfrac{1}{2}\mathbf{S}_k^{-1} - \alpha\nabla_{\mathbf{S}}\tilde{\mathcal{L}}_k$$

$$\boxed{\mathbf{S}_{k+1}^{-1} = \mathbf{S}_k^{-1} + 2\alpha\nabla_{\mathbf{S}}\tilde{\mathcal{L}}_k}$$

$$\boxed{\boldsymbol{\theta}_1^{+} = \boldsymbol{\theta}_1 - \alpha\nabla_{\boldsymbol{\eta}_1}\tilde{\mathcal{L}}}$$

$$\mathbf{S}_{k+1}^{-1}\mathbf{m}_{k+1} = \mathbf{S}_k^{-1}\mathbf{m}_k - \alpha\nabla_{\mathbf{m}}\tilde{\mathcal{L}}_k + 2\alpha\nabla_{\mathbf{S}}\tilde{\mathcal{L}}_k\,\mathbf{m}_k$$

$$\mathbf{S}_{k+1}^{-1}\mathbf{m}_{k+1} = (\mathbf{S}_k^{-1} + 2\alpha\nabla_{\mathbf{S}}\tilde{\mathcal{L}}_k)\mathbf{m}_k - \alpha\nabla_{\mathbf{m}}\tilde{\mathcal{L}}_k$$

$$\mathbf{m}_{k+1} = \mathbf{S}_{k+1}[(\mathbf{S}_k^{-1} + 2\alpha\nabla_{\mathbf{S}}\tilde{\mathcal{L}}_k)\mathbf{m}_k - \alpha\nabla_{\mathbf{m}}\tilde{\mathcal{L}}_k]$$

$$\boxed{\mathbf{m}_{k+1} = \mathbf{m}_k - \alpha\mathbf{S}_{k+1}\nabla_{\mathbf{m}}\tilde{\mathcal{L}}_k}$$

Khan and Lin (2017, 2018, 2019); J. -J. Giraldo and M. A. Álvarez (2022)

Variational Inference for a GP model

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{u}|\boldsymbol{\theta})}[\log p(\mathbf{y}|\mathbf{u})] - \mathrm{KL}[q(\mathbf{u}|\boldsymbol{\theta})||p(\mathbf{u})]$$

$$\mathcal{L} = \sum_{n=1}^{N} \mathbb{E}_{q(u_n)}\Big[\log p(y_n|u_n)\Big] - \mathrm{KL}\Big[q(\mathbf{u})||p(\mathbf{u})\Big]$$

Approximate 1d-Gaussian Expectation*

$$\mathcal{L} = \frac{N}{B} \sum_{i=1}^{B} \mathbb{E}_{q(u_i)}\Big[\log p(y_i|u_i)\Big] - \underbrace{\mathrm{KL}\Big[q(\mathbf{u})||p(\mathbf{u})\Big]}_{\mathcal{O}(N^3)}$$

Mini-batching to get stochastic gradients

* Use Gauss-Hermite Quadrature or MCMC; Solve 2d-Gaussian Expectation when using two Chained GPs.

Natural Gradients in Practice: Non-Conjugate Variational Inference in Gaussian Process Models

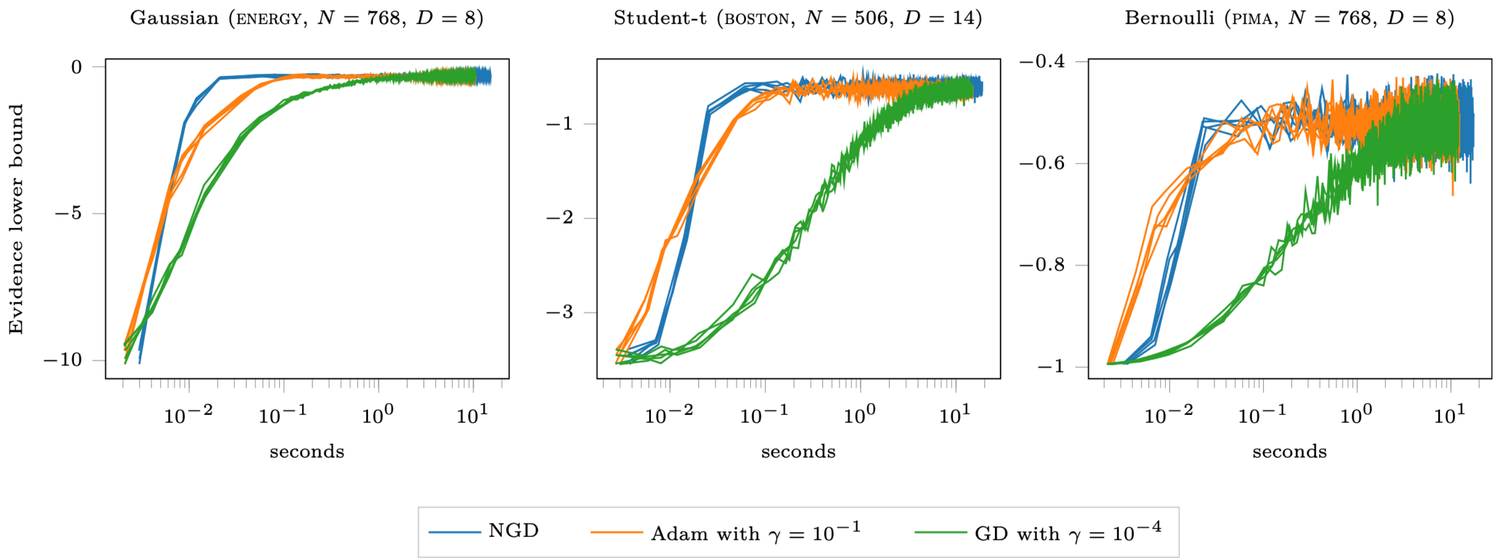


Figure 3: Stochastic optimization of the lower bound for fixed hyperparameters. The batch size is 256 and 5000 iterations are shown for five splits.

Figure taken from Salimbeni, H. et al (2018)

Sparse Variational GPs

$$p(\mathbf{f}, \mathbf{u}) = \mathcal{N}\left(\begin{bmatrix} \mathbf{f} \\ \mathbf{u} \end{bmatrix} \middle| \, \mathbf{0}, \begin{bmatrix} \mathbf{K}_{NN} & \boxed{\mathbf{K}_{NM}} \\ \boxed{\mathbf{K}_{NM}^{\top}} & \mathbf{K}_{MM} \end{bmatrix} \right)$$

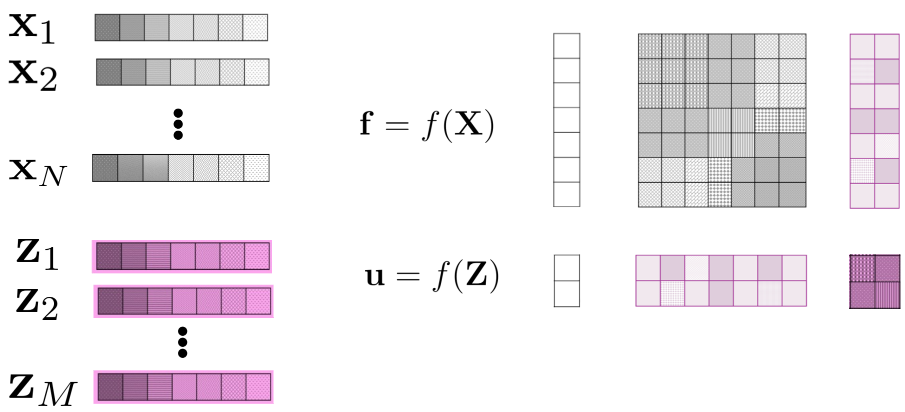
Joint distribution

$$\boxed{p(\mathbf{y}|\mathbf{f}, \mathbf{u})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})}$$

$$f(\cdot) \sim \mathcal{GP}(0, k(\cdot, \cdot'))$$

ELBO

$$\mathcal{L} = \mathbb{E}_q \left[\log \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})}{q} \right]$$



Approx. Posterior

$$\boxed{p(\mathbf{f}, \mathbf{u}|\mathbf{y}) \approx q(\mathbf{f}, \mathbf{u})}$$

$\mathbf{Z}$ are unknown ”inducing points”

Titsias, M. K. (2009); Hensman, J. et al (2015a)

Sparse Variational GPs

Approx. Posterior

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

ELBO

$$\mathcal{L} = \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})}\left[\log \frac{p(\mathbf{y}|\mathbf{f})\cancel{p(\mathbf{f}|\mathbf{u})}p(\mathbf{u})}{\cancel{p(\mathbf{f}|\mathbf{u})}q(\mathbf{u})}\right]$$

Key marginalization

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

Apply **Stochastic Variational Inference** (SVI)

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

$$\mathcal{L} = \sum_{n=1}^{N} \mathbb{E}_{q(f_n)}\big[\log p(y_n|f_n)\big] - \mathrm{KL}\big[q(\mathbf{u})||p(\mathbf{u})\big]$$

Approximate 1d-Gaussian Expectation*

$$\mathcal{L} = \frac{N}{B}\sum_{i=1}^{B} \mathbb{E}_{q(f_i)}\big[\log p(y_i|f_i)\big] - \underbrace{\mathrm{KL}\big[q(\mathbf{u})||p(\mathbf{u})\big]}_{\mathcal{O}(M^3)}$$

* Use Gauss-Hermite Quadrature or MCMC; Solve 2d-Gaussian Expectation when using two Chained GPs.

Hensman J. et al (2013)

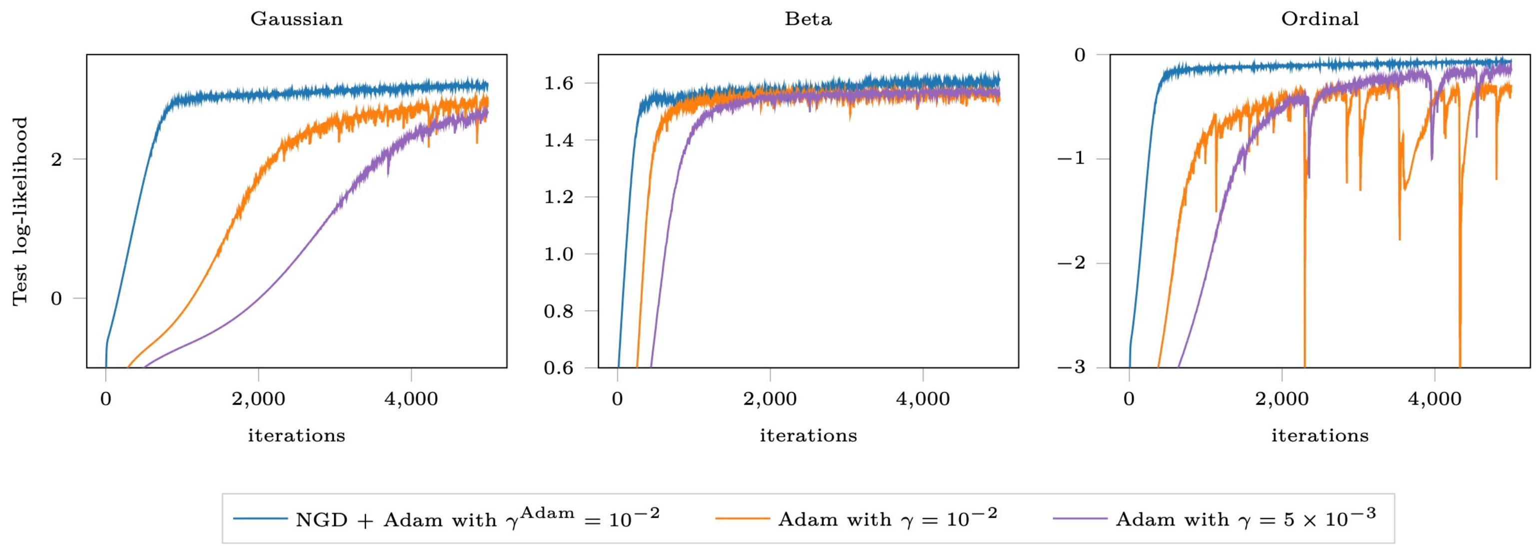


Figure 5: Optimization of the NAVAL dataset ($N = 11\text{K}$, $D = 16$), with three different likelihoods. The ill-conditioning of the variational distributions renders the optimization using ordinary gradients extremely difficult, even given a large number of iterations and different values for the Adam learning rate. The batch size is 256 and 5000 iterations are shown for a single split.

Figure taken from Salimbeni, H. et al (2018)

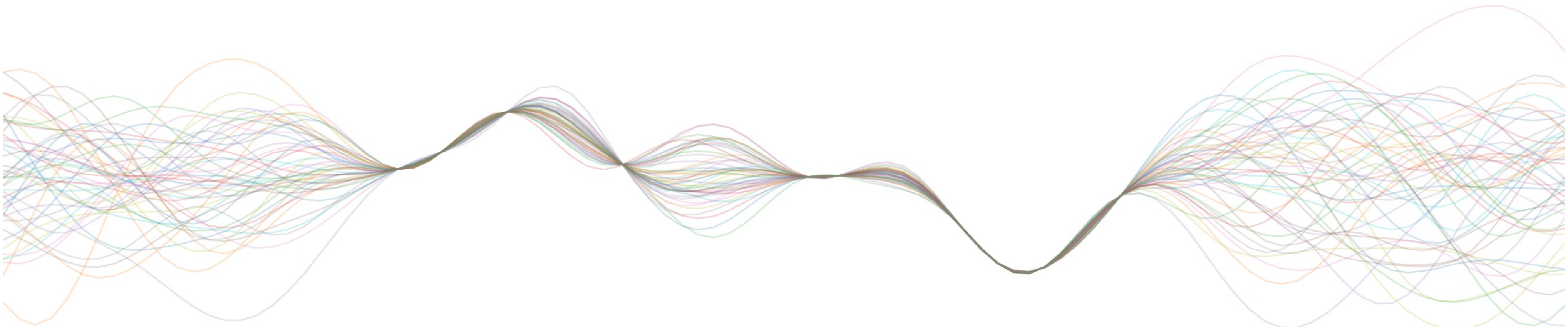
Key Points to takeaway:

Variational Inference:

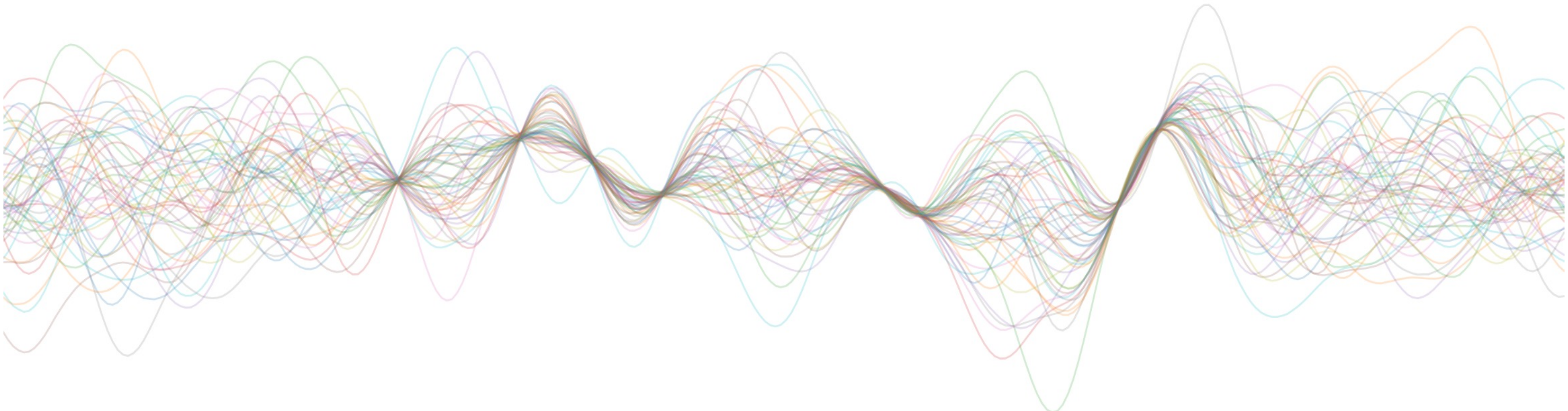
- Allows to approximate an intractable posterior via an optimization process
- Easier to scale to large data scenarios providing faster convergence than sampling methods like MCMC
- The quality of the approximate posterior depends on the choice of the variational family
- Tends to underestimate variances of the true posterior leading to ”overconfident” posteriors

Natural Gradients:

- Better convergence rates than using ordinary gradient methods
- In Likelihood-based probabilistic objectives like SVI and Gaussian processes, the Fisher matrix matches the shape of the model’s distribution space (not just parameter space)
- Thus, NGs are more robust to ill-conditioning settings than common gradient approaches
- Tolerate bigger step sizes than standard gradient updates



Thanks!!!



References

- Staines, J. and Barber, D. (2013). Optimization by variational bounding. In ESANN, pages 473–478.

- Titsias, M. K. (2009). Variational learning of inducing variables in sparse Gaussian processes. In International Conference on Artificial Intelligence and Statistics, pages 567–574.

- Amari, S. (1998). Natural gradient works efficiently in learning. Neural Computation, 10(2):251–276.

- Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. JASA, 112(518):859–877.

- Hensman, J., de G. Matthews, A. G., and Ghahramani, Z. (2015a). Scalable variational Gaussian process classification. In International Conference on Artificial Intelligence and Statistics.

- Khan, M. E. and Lin, W. (2017). Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. In International Conference on Artificial Intelligence and Statistics, pages 878–887.

- Hensman, James and Fusi, Nicolo and Lawrence, Neil D., Gaussian processes for Big data. AUAI Press, 2013

- Moreno-Muñoz, P., Artés-Rodríguez, A., and Alvarez, M. A. (2018). Heterogeneous multi-output Gaussian process prediction. In Conference on Neural Information Processing Systems, pages 6712–6721.

- Salimbeni, H., Eleftheriadis, S., and Hensman, J. (2018). Natural gradients in practice: Non-conjugate variational inference in Gaussian process models. In International Conference on Artificial Intelligence and Statistics, pages 689–697.

- Saul, A. D., Hensman, J., Vehtari, A., and Lawrence, N. D. (2016). Chained Gaussian processes. In Proceedings of the Nineteenth International Workshop on Artificial Intelligence and Statistics, volume 51, pages 1431–1440. Proceedings of Machine Learning Research.

- J. -J. Giraldo and M. A. Álvarez, "A Fully Natural Gradient Scheme for Improving Inference of the Heterogeneous Multioutput Gaussian Process Model," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 11, pp. 6429-6442, Nov. 2022

- J. Gil-González , J-J. Giraldo, A. M. Álvarez-Meza, A. Orozco-Gutiérrez, and M. A. Álvarez , Correlated Chained Gaussian Processes for Datasets With Multiple Annotators, IEEE Transactions on Neural Networks and Learning Systems, 20223.