Special course on Gaussian processes: Session #4

Vincent Adam

Aalto University

vincent.adam@aalto.fi

21/01/2021

Roadmap for today

1. Computational challenges
   - Computational complexity of GP regression
   - Non-Gaussian likelihoods: GP classification

2. Approximate inference
   - Variational inference: scratching the surface
   - Inducing points approximations

Computational complexity of Gaussian process regression

- The key equations for predictions at new input x^*, given $\boldsymbol{x}, \boldsymbol{y}$ (Gaussian noise)

$$
\begin{aligned}
p(f_*|\boldsymbol{y}) &= \mathcal{N}\left(f_* \middle| \mu_*, \sigma_*^2\right) \\
\mu_* &= \boldsymbol{k}_{f_* f}\left(\boldsymbol{K}_{ff} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{y} \\
\sigma_*^2 &= K_{f_* f_*} - \boldsymbol{k}_{f_* f}\left(\boldsymbol{K}_{ff} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{k}_{f_* f}^T
\end{aligned}
$$

Computational complexity of Gaussian process regression

- The key equations for predictions at new input x^*, given $\boldsymbol{x}, \boldsymbol{y}$ (Gaussian noise)

$$
\begin{aligned}
p(f_*|\boldsymbol{y}) &= \mathcal{N}\left(f_* \middle| \mu_*, \sigma_*^2\right) \\
\mu_* &= \boldsymbol{k}_{f_* f}\left(\boldsymbol{K}_{ff} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{y} \\
\sigma_*^2 &= K_{f_* f_*} - \boldsymbol{k}_{f_* f}\left(\boldsymbol{K}_{ff} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{k}_{f_* f}^T
\end{aligned}
$$

- Recall: If $\boldsymbol{A} \in \mathbb{R}^{N \times M}$ and $\boldsymbol{b} \in \mathbb{R}^M$, then the cost of computing $\boldsymbol{Ab}$ is $\mathcal{O}(NM)$

- Recall: If $\boldsymbol{C} \in \mathbb{R}^{N \times N}$, then the cost of computing $\boldsymbol{C}^{-1}$ is $\mathcal{O}\left(N^3\right)$

Computational complexity of Gaussian process regression

- The key equations for predictions at new input x^*, given $\boldsymbol{x}, \boldsymbol{y}$ (Gaussian noise)

$$
\begin{aligned}
p(f_*|\boldsymbol{y}) &= \mathcal{N}\left(f_* \middle| \mu_*, \sigma_*^2\right) \\
\mu_* &= \boldsymbol{k}_{f_* f}\left(\boldsymbol{K}_{ff} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{y} \\
\sigma_*^2 &= K_{f_* f_*} - \boldsymbol{k}_{f_* f}\left(\boldsymbol{K}_{ff} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{k}_{f_* f}^T
\end{aligned}
$$

- Recall: If $\boldsymbol{A} \in \mathbb{R}^{N \times M}$ and $\boldsymbol{b} \in \mathbb{R}^M$, then the cost of computing $\boldsymbol{Ab}$ is $\mathcal{O}(NM)$

- Recall: If $\boldsymbol{C} \in \mathbb{R}^{N \times N}$, then the cost of computing $\boldsymbol{C}^{-1}$ is $\mathcal{O}\left(N^3\right)$

- **Questions:** What is computational complexity for computing the posterior distribution for 1 test point based on a data set with N observations? What is the dominating operation?

Computational complexity of Gaussian process regression

- The key equations for predictions at new input x^*, given $\boldsymbol{x}, \boldsymbol{y}$ (Gaussian noise)

$$
\begin{aligned}
p(f_* | \boldsymbol{y}) &= \mathcal{N}\left(f_* \middle| \mu_*, \sigma_*^2\right) \\
\mu_* &= \boldsymbol{k}_{f_* f} \left(\boldsymbol{K}_{ff} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{y} \\
\sigma_*^2 &= K_{f_* f_*} - \boldsymbol{k}_{f_* f} \left(\boldsymbol{K}_{ff} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{k}_{f_* f}^T
\end{aligned}
$$

- Recall: If $\boldsymbol{A} \in \mathbb{R}^{N \times M}$ and $\boldsymbol{b} \in \mathbb{R}^M$, then the cost of computing $\boldsymbol{Ab}$ is $\mathcal{O}(NM)$
- Recall: If $\boldsymbol{C} \in \mathbb{R}^{N \times N}$, then the cost of computing $\boldsymbol{C}^{-1}$ is $\mathcal{O}\left(N^3\right)$
- **Questions:** What is computational complexity for computing the posterior distribution for 1 test point based on a data set with N observations? What is the dominating operation?
- $\boldsymbol{h} = \left(\boldsymbol{K}_{ff} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{y}$ scales as $\mathcal{O}\left(N^3\right)$

Computational complexity of Gaussian process regression

- The key equations for predictions at new input x^*, given $\boldsymbol{x}, \boldsymbol{y}$ (Gaussian noise)

$$p(f_*|\boldsymbol{y}) = \mathcal{N}\left(f_* \middle| \mu_*, \sigma_*^2\right)$$
$$\mu_* = \boldsymbol{k}_{f_* f}\left(\boldsymbol{K}_{ff} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{y}$$
$$\sigma_*^2 = K_{f_* f_*} - \boldsymbol{k}_{f_* f}\left(\boldsymbol{K}_{ff} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{k}_{f_* f}^T$$

- Recall: If $\boldsymbol{A} \in \mathbb{R}^{N\times M}$ and $\boldsymbol{b} \in \mathbb{R}^M$, then the cost of computing $\boldsymbol{Ab}$ is $\mathcal{O}(NM)$

- Recall: If $\boldsymbol{C} \in \mathbb{R}^{N\times N}$, then the cost of computing $\boldsymbol{C}^{-1}$ is $\mathcal{O}\left(N^3\right)$

- **Questions:** What is computational complexity for computing the posterior distribution for 1 test point based on a data set with N observations? What is the dominating operation?

- $\boldsymbol{h} = \left(\boldsymbol{K}_{ff} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{y}$ scales as $\mathcal{O}\left(N^3\right)$, $\mu_* = \boldsymbol{k}_{f_* f}\boldsymbol{h}$ scales as $\mathcal{O}(N)$

Computational complexity of Gaussian process regression

- The key equations for predictions at new input x^*, given $\boldsymbol{x}, \boldsymbol{y}$ (Gaussian noise)

$$
\begin{aligned}
p(f_*|\boldsymbol{y}) &= \mathcal{N}\left(f_* \middle| \mu_*, \sigma_*^2\right) \\
\mu_* &= \boldsymbol{k}_{f_* f}\left(\boldsymbol{K}_{ff} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{y} \\
\sigma_*^2 &= K_{f_* f_*} - \boldsymbol{k}_{f_* f}\left(\boldsymbol{K}_{ff} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{k}_{f_* f}^T
\end{aligned}
$$

- Recall: If $\boldsymbol{A} \in \mathbb{R}^{N \times M}$ and $\boldsymbol{b} \in \mathbb{R}^M$, then the cost of computing $\boldsymbol{Ab}$ is $\mathcal{O}(NM)$
- Recall: If $\boldsymbol{C} \in \mathbb{R}^{N \times N}$, then the cost of computing $\boldsymbol{C}^{-1}$ is $\mathcal{O}\left(N^3\right)$
- **Questions:** What is computational complexity for computing the posterior distribution for 1 test point based on a data set with N observations? What is the dominating operation?
- $\boldsymbol{h} = \left(\boldsymbol{K}_{ff} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{y}$ scales as $\mathcal{O}\left(N^3\right)$, $\mu_* = \boldsymbol{k}_{f_* f} \boldsymbol{h}$ scales as $\mathcal{O}(N)$
- $N \leq 1000$: Fine, $N \leq 10000$: Slow, but possible, $N > 10000$: Prohibitively slow

Regression vs classification

- Response variable $\boldsymbol{y}$ is continuous in regression problems

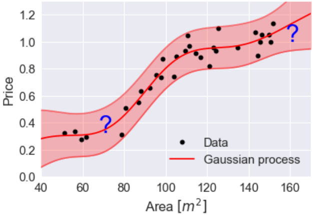
$$y_n \in \mathbb{R}$$

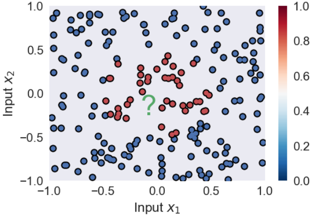
- Response variable $\boldsymbol{y}$ is discrete in classification problems

$$y_n \in \{c_1, c_2, \ldots, c_K\}$$

- Classification problems

$$\begin{aligned}
\boldsymbol{X} &= \text{images}, & y_n &\in \{\text{cat}, \text{dog}\} \\
\boldsymbol{X} &= \text{X-ray scan}, & y_n &\in \{\text{tumor}, \text{no tumor}\} \\
\boldsymbol{X} &= \text{images of digits}, & y_n &\in \{0, 1, 2, \ldots, 9\} \\
\boldsymbol{X} &= \text{emails}, & y_n &\in \{\text{spam}, \text{not spam}\}
\end{aligned}$$





Regression vs classification

- Response variable $\boldsymbol{y}$ is continuous in regression problems

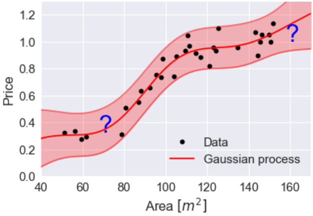
$$y_n \in \mathbb{R}$$

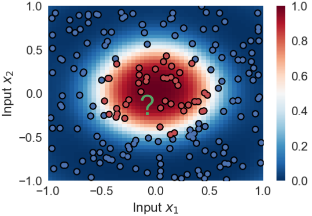
- Response variable $\boldsymbol{y}$ is discrete in classification problems

$$y_n \in \{c_1, c_2, \ldots, c_K\}$$

- Classification problems

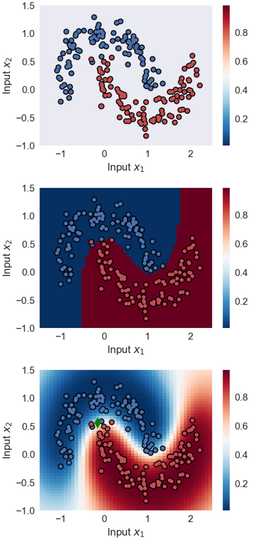
$$\begin{aligned}
\boldsymbol{X} &= \text{images}, & y_n &\in \{\text{cat}, \text{dog}\} \\
\boldsymbol{X} &= \text{X-ray scan}, & y_n &\in \{\text{tumor}, \text{no tumor}\} \\
\boldsymbol{X} &= \text{images of digits}, & y_n &\in \{0, 1, 2, \ldots, 9\} \\
\boldsymbol{X} &= \text{emails}, & y_n &\in \{\text{spam}, \text{not spam}\}
\end{aligned}$$





Why Gaussian processes for classification?

- Complex decision boundaries
  1. Non-linear boundary
  2. Can learn complexity of decision boundary from data
- Probabilistic classification
  1. How would you classify the green point?
  2. We want to model the uncertainty



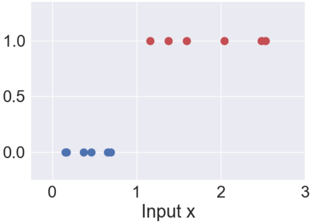
Why don't we use regression models for classification?

- We focus on binary classification: $y_n \in \{0, 1\}$ or $y_n \in \{-1, 1\}$
- We are given a data set $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$ and we want to model

$$p(y_n = +1 | \boldsymbol{x}_n)$$

- What's wrong with simply using the GP regression model with labels: $y_n \in \{0, 1\}$:

$$p(y_n = +1 | \boldsymbol{x}_n) = f(\boldsymbol{x_n})$$



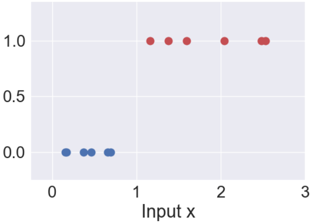
Why don't we use regression models for classification?

- We focus on binary classification: $y_n \in \{0, 1\}$ or $y_n \in \{-1, 1\}$
- We are given a data set $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$ and we want to model

$$p(y_n = +1|\boldsymbol{x}_n)$$

- What's wrong with simply using the GP regression model with labels: $y_n \in \{0, 1\}$:

$$p(y_n = +1|\boldsymbol{x}_n) = f(\boldsymbol{x_n})$$



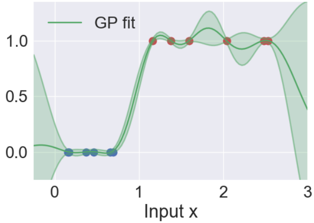
Why don't we use regression models for classification?

- We focus on binary classification: $y_n \in \{0,1\}$ or $y_n \in \{-1,1\}$
- We are given a data set $\{\boldsymbol{x}_n, y_n\}_{n=1}^{N}$ and we want to model

$$p(y_n = +1|\boldsymbol{x}_n)$$

- What's wrong with simply using the GP regression model with labels: $y_n \in \{0,1\}$:

$$p(y_n = +1|\boldsymbol{x}_n) = f(\boldsymbol{x_n})$$



Gaussian process classification setup (I)

- We'll use a 'squashing function' $\phi : \mathbb{R} \to (0,1)$ with $y_n \in \{-1, 1\}$

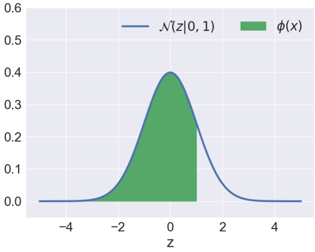
$$p(y_n | \boldsymbol{x}_n) = \phi\left(y_n \cdot f\left(\boldsymbol{x}_n\right)\right) \in (0,1)$$

- Multiple possible choices for $\phi(\cdot)$, we'll use the standard normal CDF

$$\phi\left(x\right) = \int_{-\infty}^{x} \mathcal{N}\left(z|0,1\right) \mathrm{d}z$$

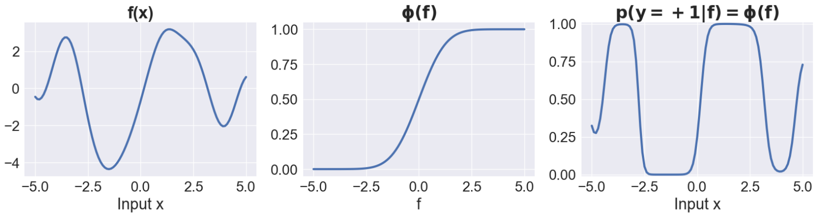
Can you figure it out?

1. What is $\phi(0)$?
2. What is $\phi(-\infty)$?
3. What is $\phi(\infty)$?
4. What is $\phi(x) + \phi(-x)$?
5. Is $\phi\left(y_n f\left(\boldsymbol{x}_n\right)\right)$ normalized wrt. y_n?

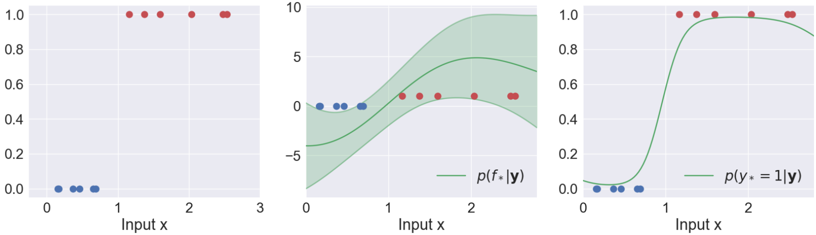


Gaussian process classification setup (II)

- We map the unknown function $f(\boldsymbol{x})$ through the squashing function



- Example re-visited



Gaussian process classification: Inference

Three steps to compute the predictive distribution for a new test point $\boldsymbol{x}_*$

$$p(\mathbf{y}, \boldsymbol{f}) = \prod_{n=1}^{N} p(y_n|f_n)p(\boldsymbol{f}) = \prod_{n=1}^{N} \phi(y_n \cdot f_n)\, \mathcal{N}(\boldsymbol{f}|\mathbf{0}, \boldsymbol{K})$$

- Step 1: Compute posterior distribution of $p(\boldsymbol{f}|\mathbf{y})$:

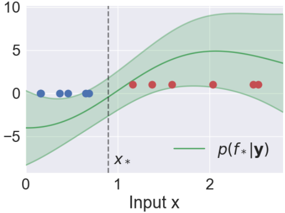
$$p(\boldsymbol{f}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{f})p(\boldsymbol{f})}{p(\mathbf{y})}$$

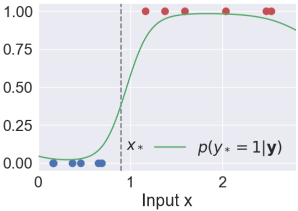
- Step 2: Compute posterior of f_* for new test point $\boldsymbol{x}_*$:

$$p(f_*|\mathbf{y}) = \int p(f_*|\boldsymbol{f})\, p(\boldsymbol{f}|\mathbf{y})\, \mathrm{d}\boldsymbol{f}$$

- Step 3: Compute predictive distribution

$$p(y_*|\mathbf{y}) = \int \phi(y_* \cdot f_*)\, p(f_*|\mathbf{y})\mathrm{d}f_*$$





Gaussian process classification: Inference

Three steps to compute the predictive distribution for a new test point $\boldsymbol{x}_*$

$$p(\boldsymbol{y}, \boldsymbol{f}) = \prod_{n=1}^{N} p(y_n|f_n)p(\boldsymbol{f}) = \prod_{n=1}^{N} \phi(y_n \cdot f_n)\,\mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{K})$$

- Step 1: Compute posterior distribution of $p(\boldsymbol{f}|\boldsymbol{y})$:

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

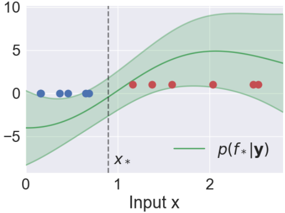
- Step 2: Compute posterior of f_* for new test point $\boldsymbol{x}_*$:

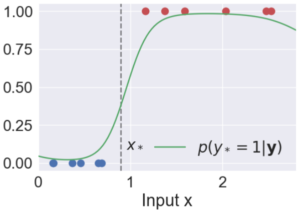
$$p(f_*|\boldsymbol{y}) = \int p(f_*|\boldsymbol{f})\, p(\boldsymbol{f}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{f}$$

- Step 3: Compute predictive distribution

$$p(y_*|\boldsymbol{y}) = \int \phi(y_* \cdot f_*)\, p(f_*|\boldsymbol{y})\mathrm{d}f_*$$

- Unfortunately, these distributions are analytically intractable.





Gaussian process classification: Inference

Three steps to compute the predictive distribution for a new test point $\boldsymbol{x}_*$

$$p(\boldsymbol{y}, \boldsymbol{f}) = \prod_{n=1}^{N} p(y_n|f_n)p(\boldsymbol{f}) = \prod_{n=1}^{N} \phi\left(y_n \cdot f_n\right) \mathcal{N}\left(\boldsymbol{f}\middle|\boldsymbol{0}, \boldsymbol{K}\right)$$

- Step 1: Compute posterior distribution of $p(\boldsymbol{f}|\boldsymbol{y})$:

$$p(\boldsymbol{f}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})}{p(\boldsymbol{y})} \approx q(\boldsymbol{f})$$

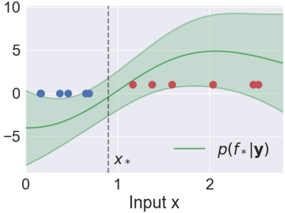
- Step 2: Compute posterior of f_* for new test point $\boldsymbol{x}_*$:

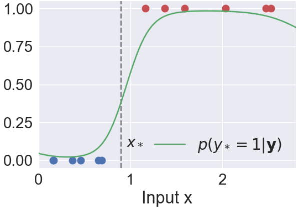
$$p(f_*|\boldsymbol{y}) = \int p\left(f_*|\boldsymbol{f}\right) p\left(\boldsymbol{f}\middle|\boldsymbol{y}\right) \mathrm{d}\boldsymbol{f} \approx \int p\left(f_*|\boldsymbol{f}\right) q\left(\boldsymbol{f}\right) \mathrm{d}\boldsymbol{f}$$

- Step 3: Compute predictive distribution

$$p(y_*|\boldsymbol{y}) = \int \phi\left(y_* \cdot f_*\right) p(f_*|\boldsymbol{y})\mathrm{d}f_*$$

- Unfortunately, these distributions are analytically intractable.





Computational problems

We need to figure out what to do when

- ... likelihood is non-Gaussian?
- ... inference becomes slow due to large N?

Computational problems

We need to figure out what to do when

- ... likelihood is non-Gaussian?
- ... inference becomes slow due to large N?

Variational inference

Computational problems

We need to figure out what to do when

- ... likelihood is non-Gaussian?
- ... inference becomes slow due to large N?

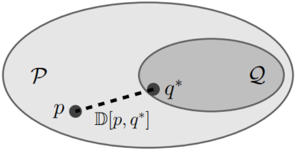
Variational inference

- General framework for approximate Bayesian inference
- Many recent application in the machine learning literature:
  1. GPs for big data
  2. GPs with non-Gaussian likelihoods
  3. Deep Gaussian processes
  4. Convolutional Gaussian processes
  5. Variational autoencoders (VAEs)
  6. ...

Variational inference: the big picture

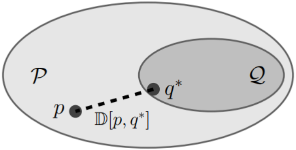
Recipe for approximating intractable distribution $p \in \mathcal{P}$

1. Define some "simple" family of distribution $\mathcal{Q}$.



Variational inference: the big picture

Recipe for approximating intractable distribution $p \in \mathcal{P}$

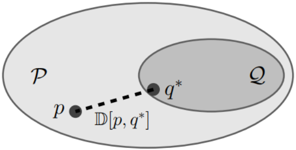


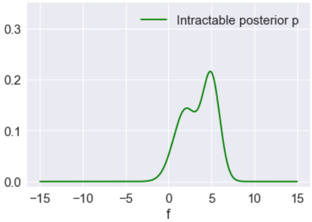
1. Define some "simple" family of distribution $\mathcal{Q}$.
2. Define some way to compute a "distance" $\mathbb{D}[q, p]$ between each of the distribution $q \in \mathcal{Q}$ and the intractable distribution p

Variational inference: the big picture

Recipe for approximating intractable distribution $p \in \mathcal{P}$

1. Define some "simple" family of distribution $\mathcal{Q}$.
2. Define some way to compute a "distance" $\mathbb{D}[q, p]$ between each of the distribution $q \in \mathcal{Q}$ and the intractable distribution p

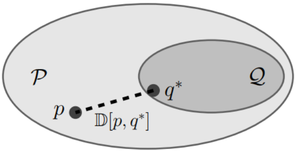


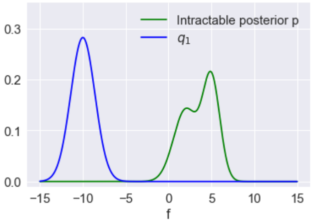


Variational inference: the big picture

Recipe for approximating intractable distribution $p \in \mathcal{P}$

1. Define some "simple" family of distribution $\mathcal{Q}$.
2. Define some way to compute a "distance" $\mathbb{D}[q, p]$ between each of the distribution $q \in \mathcal{Q}$ and the intractable distribution p

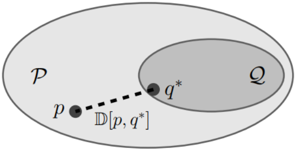


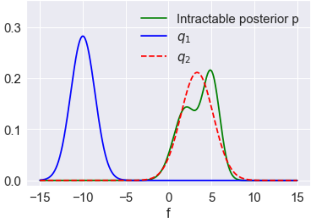


Variational inference: the big picture

Recipe for approximating intractable distribution $p \in \mathcal{P}$

1. Define some "simple" family of distribution $\mathcal{Q}$.
2. Define some way to compute a "distance" $\mathbb{D}[q, p]$ between each of the distribution $q \in \mathcal{Q}$ and the intractable distribution p





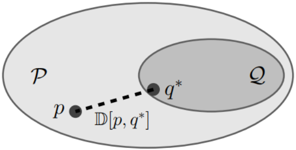
Variational inference: the big picture

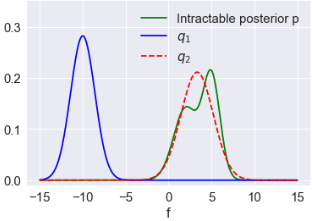
Recipe for approximating intractable distribution $p \in \mathcal{P}$

1. Define some "simple" family of distribution $\mathcal{Q}$.

2. Define some way to compute a "distance" $\mathbb{D}[q, p]$ between each of the distribution $q \in \mathcal{Q}$ and the intractable distribution p

$$\mathbb{D}[q_1, p] > \mathbb{D}[q_2, p]$$

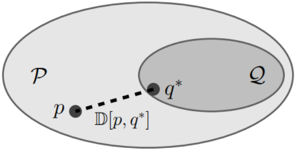




Variational inference: the big picture

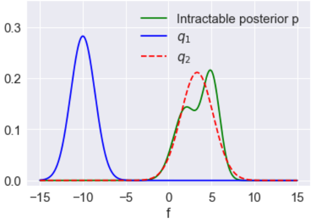
Recipe for approximating intractable distribution $p \in \mathcal{P}$

1. Define some "simple" family of distribution $\mathcal{Q}$.



2. Define some way to compute a "distance" $\mathbb{D}[q, p]$ between each of the distribution $q \in \mathcal{Q}$ and the intractable distribution p

$$\mathbb{D}[q_1, p] > \mathbb{D}[q_2, p]$$

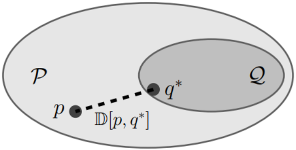


3. Search for the distribution in $q \in \mathcal{Q}$ such that $\mathbb{D}[q, p]$ is minimized

$$q^* = \arg\min_{q \in \mathcal{Q}} \mathbb{D}[q, p]$$

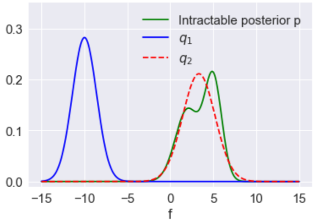
Variational inference: the big picture

Recipe for approximating intractable distribution $p \in \mathcal{P}$



1. Define some "simple" family of distribution $\mathcal{Q}$.
2. Define some way to compute a "distance" $\mathbb{D}[q, p]$ between each of the distribution $q \in \mathcal{Q}$ and the intractable distribution p

$$\mathbb{D}[q_1, p] > \mathbb{D}[q_2, p]$$



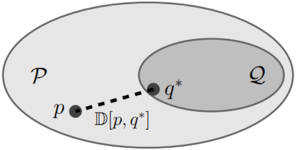
3. Search for the distribution in $q \in \mathcal{Q}$ such that $\mathbb{D}[q, p]$ is minimized

$$q^* = \arg\min_{q \in \mathcal{Q}} \mathbb{D}[q, p]$$

4. Use q^* as an approximation of p

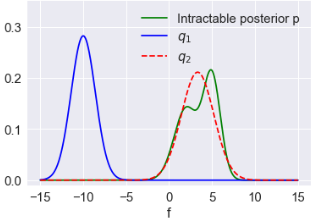
Variational inference: the big picture

Recipe for approximating intractable distribution $p \in \mathcal{P}$



1. Define some "simple" family of distribution $\mathcal{Q}$.
2. Define some way to compute a "distance" $\mathbb{D}[q, p]$ between each of the distribution $q \in \mathcal{Q}$ and the intractable distribution p

$$\mathbb{D}[q_1, p] > \mathbb{D}[q_2, p]$$



3. Search for the distribution in $q \in \mathcal{Q}$ such that $\mathbb{D}[q, p]$ is minimized

$$q^* = \arg\min_{q \in \mathcal{Q}} \mathbb{D}[q, p]$$

4. Use q^* as an approximation of p

Here we will always choose $\mathcal{Q}$ to be the set of multivariate Gaussian distributions.

Variational inference I

- We will use to the *Kullback-Leibler divergence* to "measure distances" between distributions

$$\mathbb{D}\left[q||p\right] = \int q(\boldsymbol{f}) \ln \frac{q(\boldsymbol{f})}{p(\boldsymbol{f})} \mathrm{d}\boldsymbol{f} = \mathbb{E}_q \left[\ln \frac{q(\boldsymbol{f})}{p(\boldsymbol{f})}\right]$$

Variational inference I

- We will use to the *Kullback-Leibler divergence* to "measure distances" between distributions

$$\mathbb{D}\left[q||p\right] = \int q(\boldsymbol{f}) \ln \frac{q(\boldsymbol{f})}{p(\boldsymbol{f})} \mathrm{d}\boldsymbol{f} = \mathbb{E}_q \left[\ln \frac{q(\boldsymbol{f})}{p(\boldsymbol{f})} \right]$$

- Most important properties for our purpose:
  1. Always positive: $\mathbb{D}\left[q||p\right] \geq 0$
  2. Identity of indiscernibles: $\mathbb{D}\left[q||p\right] = 0 \iff p = q \quad \text{(a.e.)}$
  3. Not-symmetric: $\mathbb{D}\left[q||p\right] \neq \mathbb{D}\left[p||q\right]$

Variational inference II

Our goal is to minimize the KL divergence between some approximation $q \in \mathcal{Q}$ and some posterior distribution $p(\boldsymbol{f}|\boldsymbol{y})$

Variational inference II

Our goal is to minimize the KL divergence between some approximation $q \in \mathcal{Q}$ and some posterior distribution $p(\boldsymbol{f}|\boldsymbol{y})$

$$\mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right] = \mathbb{E}_q\left[\ln \frac{q(\boldsymbol{f})}{p(\boldsymbol{f}|\boldsymbol{y})}\right]$$

Variational inference II

Our goal is to minimize the KL divergence between some approximation $q \in \mathcal{Q}$ and some posterior distribution $p(\boldsymbol{f}|\boldsymbol{y})$

$$\begin{aligned}\mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right] &= \mathbb{E}_q\left[\ln \frac{q(\boldsymbol{f})}{p(\boldsymbol{f}|\boldsymbol{y})}\right] \\ &= \mathbb{E}_q\left[\ln q(\boldsymbol{f}) - \ln p(\boldsymbol{f}|\boldsymbol{y})\right]\end{aligned}$$

Variational inference II

Our goal is to minimize the KL divergence between some approximation $q \in \mathcal{Q}$ and some posterior distribution $p(\boldsymbol{f}|\boldsymbol{y})$

$$\begin{aligned}
\mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right] &= \mathbb{E}_q\left[\ln \frac{q(\boldsymbol{f})}{p(\boldsymbol{f}|\boldsymbol{y})}\right] \\
&= \mathbb{E}_q\left[\ln q(\boldsymbol{f}) - \ln p(\boldsymbol{f}|\boldsymbol{y})\right] \\
&= \mathbb{E}_q\left[\ln q(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln p(\boldsymbol{f}|\boldsymbol{y})\right]
\end{aligned}$$

Variational inference II

Our goal is to minimize the KL divergence between some approximation $q \in \mathcal{Q}$ and some posterior distribution $p(\boldsymbol{f}|\boldsymbol{y})$

$$\begin{aligned}
\mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right] &= \mathbb{E}_q\left[\ln \frac{q(\boldsymbol{f})}{p(\boldsymbol{f}|\boldsymbol{y})}\right] \\
&= \mathbb{E}_q\left[\ln q(\boldsymbol{f}) - \ln p(\boldsymbol{f}|\boldsymbol{y})\right] \\
&= \mathbb{E}_q\left[\ln q(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln p(\boldsymbol{f}|\boldsymbol{y})\right]
\end{aligned}$$

Variational inference II

Our goal is to minimize the KL divergence between some approximation $q \in \mathcal{Q}$ and some posterior distribution $p(\boldsymbol{f}|\boldsymbol{y})$

$$\begin{aligned}
\mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right] &= \mathbb{E}_q\left[\ln \frac{q(\boldsymbol{f})}{p(\boldsymbol{f}|\boldsymbol{y})}\right] \\
&= \mathbb{E}_q\left[\ln q(\boldsymbol{f}) - \ln p(\boldsymbol{f}|\boldsymbol{y})\right] \\
&= \mathbb{E}_q\left[\ln q(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln p(\boldsymbol{f}|\boldsymbol{y})\right]
\end{aligned}$$

Variational inference II

Our goal is to minimize the KL divergence between some approximation $q \in \mathcal{Q}$ and some posterior distribution $p(\boldsymbol{f}|\boldsymbol{y})$

$$
\begin{aligned}
\mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right] &= \mathbb{E}_q\left[\ln \frac{q(\boldsymbol{f})}{p(\boldsymbol{f}|\boldsymbol{y})}\right] \\
&= \mathbb{E}_q\left[\ln q(\boldsymbol{f}) - \ln p(\boldsymbol{f}|\boldsymbol{y})\right] \\
&= \mathbb{E}_q\left[\ln q(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln p(\boldsymbol{f}|\boldsymbol{y})\right]
\end{aligned}
$$

Last term depends on the exact posterior $p(\boldsymbol{f}|\boldsymbol{y})$, which is intractable.

Variational inference III

We can rewrite the posterior: $p(\boldsymbol{f}|\boldsymbol{y}) = \frac{p(\boldsymbol{y},\boldsymbol{f})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})}{p(\boldsymbol{y})}$

$$\mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right] = \mathbb{E}_q\left[\ln q(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln p(\boldsymbol{f}|\boldsymbol{y})\right]$$

Variational inference III

We can rewrite the posterior: $p(\boldsymbol{f}|\boldsymbol{y}) = \frac{p(\boldsymbol{y},\boldsymbol{f})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})}{p(\boldsymbol{y})}$

$$\begin{aligned}\mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right] &= \mathbb{E}_q\left[\ln q(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln p(\boldsymbol{f}|\boldsymbol{y})\right] \\ &= \mathbb{E}_q\left[\ln q(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln \frac{p(\boldsymbol{y},\boldsymbol{f})}{p(\boldsymbol{y})}\right]\end{aligned}$$

Variational inference III

We can rewrite the posterior: $p(\boldsymbol{f}|\boldsymbol{y}) = \frac{p(\boldsymbol{y},\boldsymbol{f})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})}{p(\boldsymbol{y})}$

$$\begin{aligned}
\mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right] &= \mathbb{E}_q\left[\ln q(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln p(\boldsymbol{f}|\boldsymbol{y})\right] \\
&= \mathbb{E}_q\left[\ln q(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln \frac{p(\boldsymbol{y},\boldsymbol{f})}{p(\boldsymbol{y})}\right] \\
&= \mathbb{E}_q\left[\ln q(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln p(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] + \ln p(\boldsymbol{y})
\end{aligned}$$

Variational inference III

We can rewrite the posterior: $p(\boldsymbol{f}|\boldsymbol{y}) = \frac{p(\boldsymbol{y},\boldsymbol{f})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})}{p(\boldsymbol{y})}$

$$
\begin{aligned}
\mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right] &= \mathbb{E}_q\left[\ln q(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln p(\boldsymbol{f}|\boldsymbol{y})\right] \\
&= \mathbb{E}_q\left[\ln q(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln \frac{p(\boldsymbol{y},\boldsymbol{f})}{p(\boldsymbol{y})}\right] \\
&= \mathbb{E}_q\left[\ln q(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln p(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] + \ln p(\boldsymbol{y}) \\
&= \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] + \ln p(\boldsymbol{y})
\end{aligned}
$$

Variational inference III

We can rewrite the posterior: $p(\boldsymbol{f}|\boldsymbol{y}) = \frac{p(\boldsymbol{y},\boldsymbol{f})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})}{p(\boldsymbol{y})}$

$$
\begin{aligned}
\mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right] &= \mathbb{E}_q\left[\ln q(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln p(\boldsymbol{f}|\boldsymbol{y})\right] \\
&= \mathbb{E}_q\left[\ln q(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln \frac{p(\boldsymbol{y},\boldsymbol{f})}{p(\boldsymbol{y})}\right] \\
&= \mathbb{E}_q\left[\ln q(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln p(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] + \ln p(\boldsymbol{y}) \\
&= \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] + \ln p(\boldsymbol{y})
\end{aligned}
$$

Let's re-arrange the terms

$$
\ln p(\boldsymbol{y}) = \mathbb{E}_q\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f})\right] + \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right]
$$

Variational inference III

We can rewrite the posterior: $p(\boldsymbol{f}|\boldsymbol{y}) = \frac{p(\boldsymbol{y},\boldsymbol{f})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})}{p(\boldsymbol{y})}$

$$\begin{aligned}
\mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right] &= \mathbb{E}_q\left[\ln q(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln p(\boldsymbol{f}|\boldsymbol{y})\right] \\
&= \mathbb{E}_q\left[\ln q(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln \frac{p(\boldsymbol{y},\boldsymbol{f})}{p(\boldsymbol{y})}\right] \\
&= \mathbb{E}_q\left[\ln q(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln p(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] + \ln p(\boldsymbol{y}) \\
&= \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f})\right] - \mathbb{E}_q\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] + \ln p(\boldsymbol{y})
\end{aligned}$$

Let's re-arrange the terms

$$\ln p(\boldsymbol{y}) = \underbrace{\mathbb{E}_q\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f})\right]}_{\mathcal{L}[q]} + \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right]$$

$\mathcal{L}[q]$ does not depend on the posterior $p(\boldsymbol{f}|\boldsymbol{y})$, but only separately on the conditional density $p(\boldsymbol{y}|\boldsymbol{f})$ and the prior $p(\boldsymbol{f})$.

Variational inference IV

$$\ln p(\boldsymbol{y}) = \underbrace{\mathbb{E}_q \left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f})\right]}_{\mathcal{L}[q]} + \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right]$$

Variational inference IV

$$\ln p(\boldsymbol{y}) = \underbrace{\mathbb{E}_q \left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f})\right]}_{\mathcal{L}[q]} + \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right]$$

Let's make a few observations

Variational inference IV

$$\ln p(\boldsymbol{y}) = \underbrace{\mathbb{E}_q\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f})\right]}_{\mathcal{L}[q]} + \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right]$$

Let's make a few observations

1. $\ln p(\boldsymbol{y})$ is a constant

Variational inference IV

$$\ln p(\boldsymbol{y}) = \underbrace{\mathbb{E}_q\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f})\right]}_{\mathcal{L}[q]} + \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right]$$

Let's make a few observations

1. $\ln p(\boldsymbol{y})$ is a constant
2. $\mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right] \geq 0$ is non-negative

Variational inference IV

$$\ln p(\boldsymbol{y}) = \underbrace{\mathbb{E}_q\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f})\right]}_{\mathcal{L}[q]} + \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right]$$

Let's make a few observations

1. $\ln p(\boldsymbol{y})$ is a constant
2. $\mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right] \geq 0$ is non-negative
3. $\mathcal{L}\left[q\right]$ only depends on q and the joint density $p(\boldsymbol{y}, \boldsymbol{f})$

Variational inference IV

$$\ln p(\boldsymbol{y}) = \underbrace{\mathbb{E}_q\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f})\right]}_{\mathcal{L}[q]} + \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right]$$

Let's make a few observations

1. $\ln p(\boldsymbol{y})$ is a constant
2. $\mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right] \geq 0$ is non-negative
3. $\mathcal{L}\left[q\right]$ only depends on q and the joint density $p(\boldsymbol{y}, \boldsymbol{f})$

Some consequences

Variational inference IV

$$\ln p(\boldsymbol{y}) = \underbrace{\mathbb{E}_q\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f})\right]}_{\mathcal{L}[q]} + \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right]$$

Let's make a few observations

1. $\ln p(\boldsymbol{y})$ is a constant
2. $\mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right] \geq 0$ is non-negative
3. $\mathcal{L}\left[q\right]$ only depends on q and the joint density $p(\boldsymbol{y}, \boldsymbol{f})$

Some consequences

1. $\mathcal{L}\left[q\right]$ is a *lower bound* of $\ln p(\boldsymbol{y})$. That is: $\ln p(\boldsymbol{y}) \geq \mathcal{L}\left[q\right]$

Variational inference IV

$$\ln p(\boldsymbol{y}) = \underbrace{\mathbb{E}_q\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f})\right]}_{\mathcal{L}[q]} + \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right]$$

Let's make a few observations

1. $\ln p(\boldsymbol{y})$ is a constant
2. $\mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right] \geq 0$ is non-negative
3. $\mathcal{L}[q]$ only depends on q and the joint density $p(\boldsymbol{y}, \boldsymbol{f})$

Some consequences

1. $\mathcal{L}[q]$ is a *lower bound* of $\ln p(\boldsymbol{y})$. That is: $\ln p(\boldsymbol{y}) \geq \mathcal{L}[q]$
2. Maximizing $\mathcal{L}[q]$ is equivalent to minizing $\mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right]$

Variational inference IV

$$\ln p(\boldsymbol{y}) = \underbrace{\mathbb{E}_q\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f})\right]}_{\mathcal{L}[q]} + \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right]$$

Let's make a few observations

1. $\ln p(\boldsymbol{y})$ is a constant
2. $\mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right] \geq 0$ is non-negative
3. $\mathcal{L}[q]$ only depends on q and the joint density $p(\boldsymbol{y}, \boldsymbol{f})$

Some consequences

1. $\mathcal{L}[q]$ is a *lower bound* of $\ln p(\boldsymbol{y})$. That is: $\ln p(\boldsymbol{y}) \geq \mathcal{L}[q]$
2. Maximizing $\mathcal{L}[q]$ is equivalent to minizing $\mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right]$

Key take-away: we can fit the variational approx. q by optimizing $\mathcal{L}$

Variational inference III-bis

We can derive the ELBO via Jensen's inequality:
if ϕ concave, f a function, then $\phi[\mathbb{E}_{p(x)} f(x)] > \mathbb{E}_{p(x)} \phi[f(x)]$

The ln function is concave so,

$$\ln p(\boldsymbol{y}) = \ln \int p(\boldsymbol{f}, \boldsymbol{y}) d\boldsymbol{f}$$

Variational inference III-bis

We can derive the ELBO via Jensen's inequality:
if ϕ concave, f a function, then $\phi[\mathbb{E}_{p(x)} f(x)] > \mathbb{E}_{p(x)} \phi[f(x)]$

The ln function is concave so,

$$\begin{aligned} \ln p(\boldsymbol{y}) &= \ln \int p(\boldsymbol{f}, \boldsymbol{y}) d\boldsymbol{f} \\ &= \ln \int q(\boldsymbol{f}) \frac{p(\boldsymbol{f}, \boldsymbol{y})}{q(\boldsymbol{f})} d\boldsymbol{f} \end{aligned}$$

Variational inference III-bis

We can derive the ELBO via Jensen's inequality:
if ϕ concave, f a function, then $\phi[\mathbb{E}_{p(x)} f(x)] > \mathbb{E}_{p(x)} \phi[f(x)]$

The ln function is concave so,

$$\begin{aligned}
\ln p(\boldsymbol{y}) &= \ln \int p(\boldsymbol{f}, \boldsymbol{y}) d\boldsymbol{f} \\
&= \ln \int q(\boldsymbol{f}) \frac{p(\boldsymbol{f}, \boldsymbol{y})}{q(\boldsymbol{f})} d\boldsymbol{f} \\
&= \ln \mathbb{E}_q \frac{p(\boldsymbol{f}, \boldsymbol{y})}{q(\boldsymbol{f})}
\end{aligned}$$

Variational inference III-bis

We can derive the ELBO via Jensen's inequality:
if ϕ concave, f a function, then $\phi[\mathbb{E}_{p(x)} f(x)] > \mathbb{E}_{p(x)} \phi[f(x)]$

The ln function is concave so,

$$
\begin{aligned}
\ln p(\boldsymbol{y}) &= \ln \int p(\boldsymbol{f}, \boldsymbol{y}) d\boldsymbol{f} \\
&= \ln \int q(\boldsymbol{f}) \frac{p(\boldsymbol{f}, \boldsymbol{y})}{q(\boldsymbol{f})} d\boldsymbol{f} \\
&= \ln \mathbb{E}_q \frac{p(\boldsymbol{f}, \boldsymbol{y})}{q(\boldsymbol{f})} \\
(Jensen) &\geq \mathbb{E}_q \ln \left[\frac{p(\boldsymbol{f}, \boldsymbol{y})}{q(\boldsymbol{f})} \right]
\end{aligned}
$$

Variational inference III-bis

We can derive the ELBO via Jensen's inequality:
if ϕ concave, f a function, then $\phi[\mathbb{E}_{p(x)} f(x)] > \mathbb{E}_{p(x)} \phi[f(x)]$

The ln function is concave so,

$$\begin{aligned}
\ln p(\boldsymbol{y}) &= \ln \int p(\boldsymbol{f}, \boldsymbol{y}) d\boldsymbol{f} \\
&= \ln \int q(\boldsymbol{f}) \frac{p(\boldsymbol{f}, \boldsymbol{y})}{q(\boldsymbol{f})} d\boldsymbol{f} \\
&= \ln \mathbb{E}_q \frac{p(\boldsymbol{f}, \boldsymbol{y})}{q(\boldsymbol{f})} \\
(Jensen) &\geq \mathbb{E}_q \ln \left[\frac{p(\boldsymbol{f}, \boldsymbol{y})}{q(\boldsymbol{f})} \right] \\
&= \mathbb{E}_q \ln p(\boldsymbol{y}|\boldsymbol{f}) + \mathbb{E}_q \ln \left[\frac{p(\boldsymbol{f})}{q(\boldsymbol{f})} \right]
\end{aligned}$$

Variational inference III-bis

We can derive the ELBO via Jensen's inequality:
if ϕ concave, f a function, then $\phi[\mathbb{E}_{p(x)} f(x)] > \mathbb{E}_{p(x)} \phi[f(x)]$

The ln function is concave so,

$$
\begin{aligned}
\ln p(\boldsymbol{y}) &= \ln \int p(\boldsymbol{f}, \boldsymbol{y}) d\boldsymbol{f} \\
&= \ln \int q(\boldsymbol{f}) \frac{p(\boldsymbol{f}, \boldsymbol{y})}{q(\boldsymbol{f})} d\boldsymbol{f} \\
&= \ln \mathbb{E}_q \frac{p(\boldsymbol{f}, \boldsymbol{y})}{q(\boldsymbol{f})} \\
(Jensen) &\geq \mathbb{E}_q \ln \left[\frac{p(\boldsymbol{f}, \boldsymbol{y})}{q(\boldsymbol{f})} \right] \\
&= \mathbb{E}_q \ln p(\boldsymbol{y}|\boldsymbol{f}) + \mathbb{E}_q \ln \left[\frac{p(\boldsymbol{f})}{q(\boldsymbol{f})} \right] \\
&= \mathcal{L}(q)
\end{aligned}
$$

Variational inference V

$$\ln p(\boldsymbol{y}) = \underbrace{\mathbb{E}_q\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f})\right]}_{\mathcal{L}[q]} + \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right]$$

- $\mathcal{L}[q]$ is often called the *Evidence Lower Bound* (ELBO)

Variational inference V

$$\ln p(\boldsymbol{y}) = \underbrace{\mathbb{E}_q\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f})\right]}_{\mathcal{L}[q]} + \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right]$$

- $\mathcal{L}[q]$ is often called the *Evidence Lower Bound* (ELBO)
- The first term in $\mathcal{L}[q]$ can be interpreted as a data fit term and the second term can be interpreted as a regularization term (staying close to the prior)

Variational inference V

$$\ln p(\boldsymbol{y}) = \underbrace{\mathbb{E}_q \left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f})\right]}_{\mathcal{L}[q]} + \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right]$$

- $\mathcal{L}[q]$ is often called the *Evidence Lower Bound* (ELBO)
- The first term in $\mathcal{L}[q]$ can be interpreted as a data fit term and the second term can be interpreted as a regularization term (staying close to the prior)
- If we want to approximate $p(\boldsymbol{f}|\boldsymbol{y})$, then $q(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{m}, \boldsymbol{V})$

Variational inference V

$$\ln p(\boldsymbol{y}) = \underbrace{\mathbb{E}_q\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f})\right]}_{\mathcal{L}[q]} + \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right]$$

- $\mathcal{L}[q]$ is often called the *Evidence Lower Bound* (ELBO)
- The first term in $\mathcal{L}[q]$ can be interpreted as a data fit term and the second term can be interpreted as a regularization term (staying close to the prior)
- If we want to approximate $p(\boldsymbol{f}|\boldsymbol{y})$, then $q(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{m}, \boldsymbol{V})$
- Define $\boldsymbol{\lambda} = \{\boldsymbol{m}, \boldsymbol{V}\}$, then we can write $\mathcal{L}[q] = \mathcal{L}[\boldsymbol{\lambda}]$

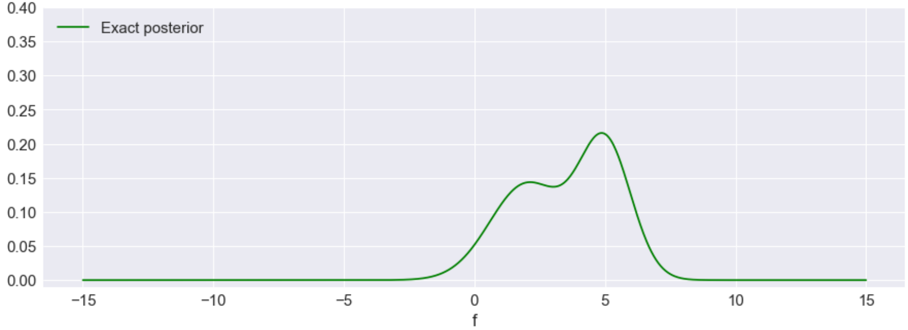
Variational inference V

$$\ln p(\boldsymbol{y}) = \underbrace{\mathbb{E}_q\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f})\right]}_{\mathcal{L}[q]} + \mathbb{D}\left[q(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})\right]$$

- $\mathcal{L}[q]$ is often called the *Evidence Lower Bound* (ELBO)
- The first term in $\mathcal{L}[q]$ can be interpreted as a data fit term and the second term can be interpreted as a regularization term (staying close to the prior)
- If we want to approximate $p(\boldsymbol{f}|\boldsymbol{y})$, then $q(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{m}, \boldsymbol{V})$
- Define $\boldsymbol{\lambda} = \{\boldsymbol{m}, \boldsymbol{V}\}$, then we can write $\mathcal{L}[q] = \mathcal{L}[\boldsymbol{\lambda}]$
- In practice, we optimize $\mathcal{L}[\boldsymbol{\lambda}]$ using gradient-based methods

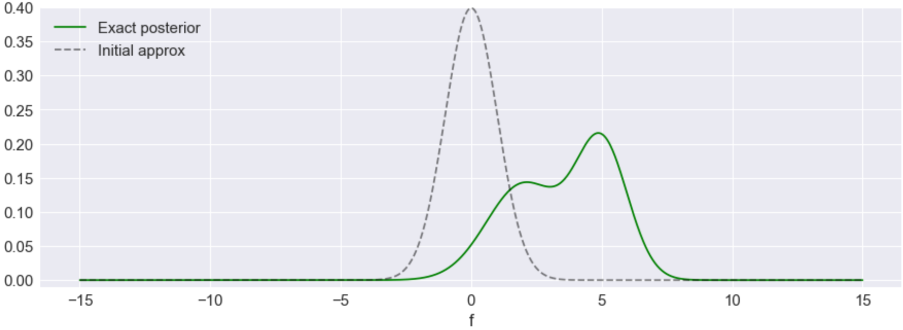
1D Toy example I

- Assume we have some model $p(y, f)$ that gives rise to some intractable posterior $p(f|y)$
- We want to approximate $p(f|y)$ using a variational approximation
- In 1D: $\mathcal{Q}$ is the the set of univariate Gaussian, i.e. $q_\lambda(x) = \mathcal{N}(x|m, v)$, where we denote $\boldsymbol{\lambda} = \{m, v\}$
- We initialize our approximation as $q(f) = \mathcal{N}(f|0, 1)$



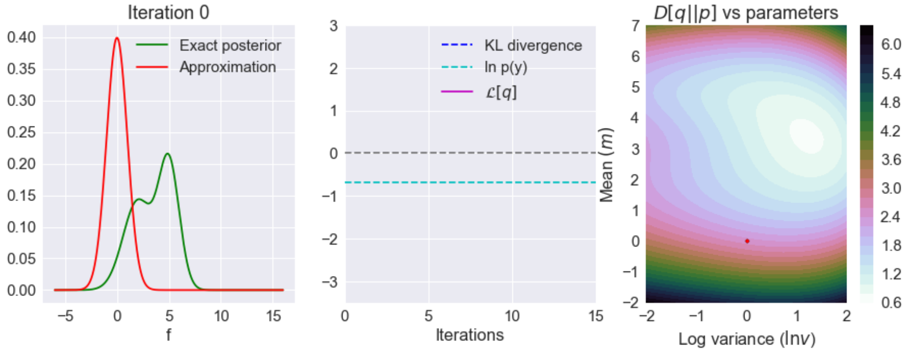
1D Toy example I

- Assume we have some model $p(y, f)$ that gives rise to some intractable posterior $p(f|y)$
- We want to approximate $p(f|y)$ using a variational approximation
- In 1D: $\mathcal{Q}$ is the the set of univariate Gaussian, i.e. $q_\lambda(x) = \mathcal{N}(x|m, v)$, where we denote $\boldsymbol{\lambda} = \{m, v\}$
- We initialize our approximation as $q(f) = \mathcal{N}(f|0, 1)$



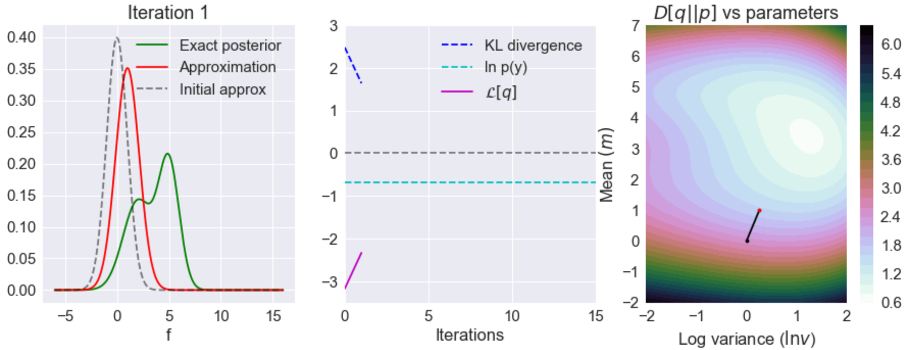
1D Toy example II

- Gradient ascent: $\boldsymbol{\lambda}_{i+1} = \boldsymbol{\lambda}_i + \eta \nabla_{\boldsymbol{\lambda}} \mathcal{L}[\boldsymbol{\lambda}]$
- $\ln p(\boldsymbol{y}) = \mathcal{L}[\boldsymbol{\lambda}] + \mathbb{D}[q_\lambda(\boldsymbol{f}) || p(\boldsymbol{f}|\boldsymbol{y})] \geq \mathcal{L}[\boldsymbol{\lambda}]$



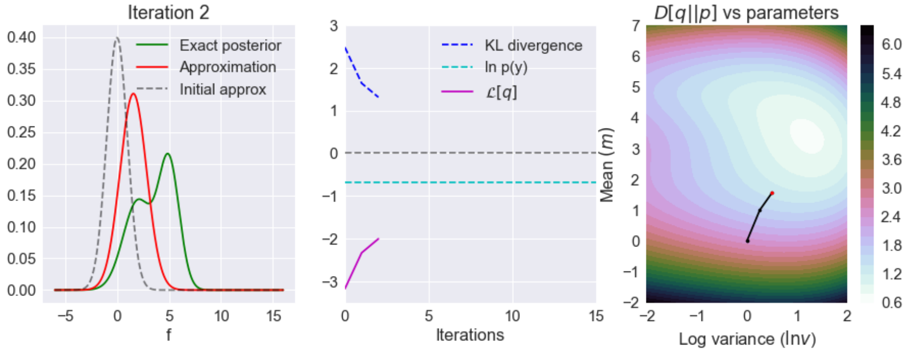
1D Toy example II

- Gradient ascent: $\boldsymbol{\lambda}_{i+1} = \boldsymbol{\lambda}_i + \eta \nabla_{\boldsymbol{\lambda}} \mathcal{L}[\boldsymbol{\lambda}]$
- $\ln p(\boldsymbol{y}) = \mathcal{L}[\boldsymbol{\lambda}] + \mathbb{D}[q_\lambda(\boldsymbol{f})||p(\boldsymbol{f}|\boldsymbol{y})] \geq \mathcal{L}[\boldsymbol{\lambda}]$



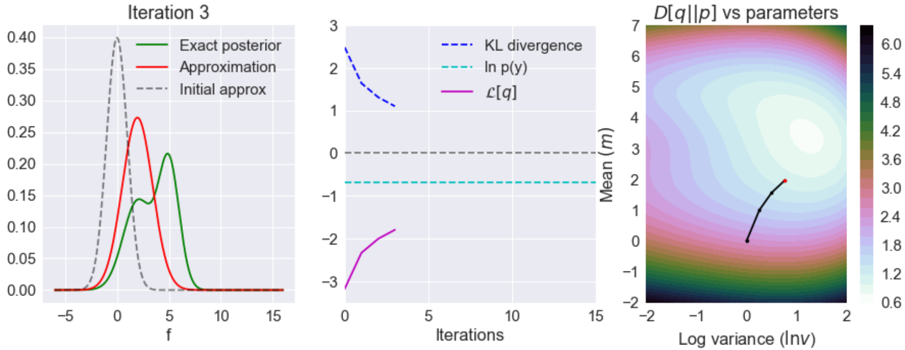
1D Toy example II

- Gradient ascent: $\boldsymbol{\lambda}_{i+1} = \boldsymbol{\lambda}_i + \eta \nabla_{\boldsymbol{\lambda}} \mathcal{L}[\boldsymbol{\lambda}]$
- $\ln p(\boldsymbol{y}) = \mathcal{L}[\boldsymbol{\lambda}] + \mathbb{D}[q_\lambda(\boldsymbol{f}) || p(\boldsymbol{f}|\boldsymbol{y})] \geq \mathcal{L}[\boldsymbol{\lambda}]$

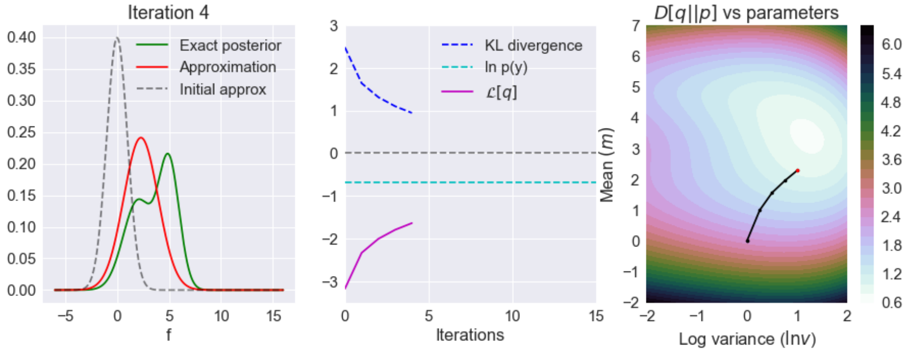


1D Toy example II

- Gradient ascent: $\boldsymbol{\lambda}_{i+1} = \boldsymbol{\lambda}_i + \eta \nabla_{\boldsymbol{\lambda}} \mathcal{L}[\boldsymbol{\lambda}]$
- $\ln p(\boldsymbol{y}) = \mathcal{L}[\boldsymbol{\lambda}] + \mathbb{D}[q_\lambda(\boldsymbol{f}) || p(\boldsymbol{f}|\boldsymbol{y})] \geq \mathcal{L}[\boldsymbol{\lambda}]$

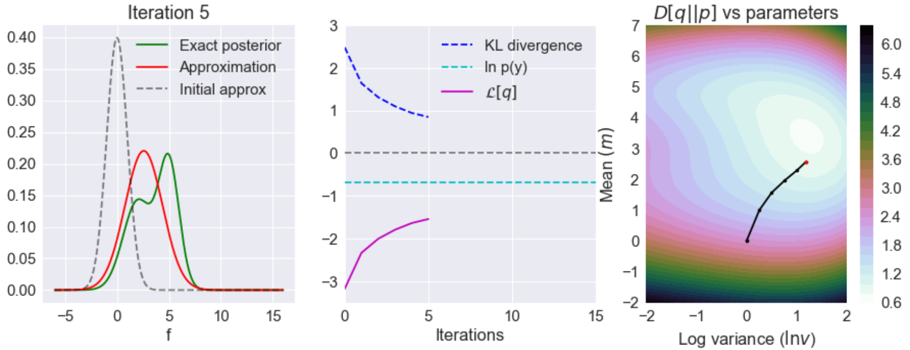


- Gradient ascent: $\boldsymbol{\lambda}_{i+1} = \boldsymbol{\lambda}_i + \eta \nabla_{\boldsymbol{\lambda}} \mathcal{L}[\boldsymbol{\lambda}]$
- $\ln p(\boldsymbol{y}) = \mathcal{L}[\boldsymbol{\lambda}] + \mathbb{D}[q_\lambda(\boldsymbol{f}) || p(\boldsymbol{f}|\boldsymbol{y})] \geq \mathcal{L}[\boldsymbol{\lambda}]$



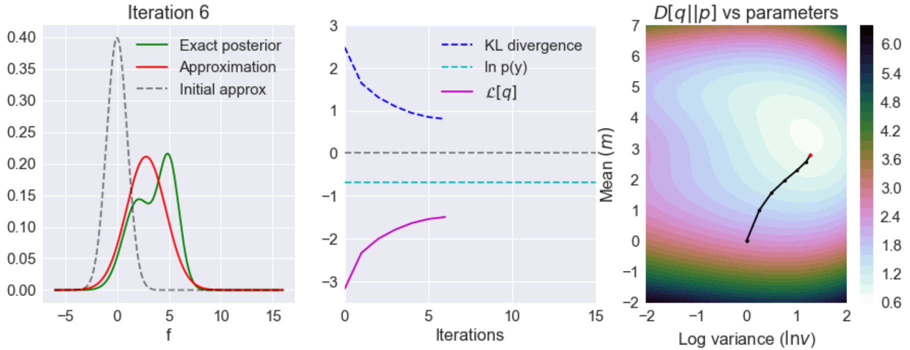
1D Toy example II

- Gradient ascent: $\boldsymbol{\lambda}_{i+1} = \boldsymbol{\lambda}_i + \eta \nabla_{\boldsymbol{\lambda}} \mathcal{L}[\boldsymbol{\lambda}]$
- $\ln p(\boldsymbol{y}) = \mathcal{L}[\boldsymbol{\lambda}] + \mathbb{D}[q_\lambda(\boldsymbol{f}) || p(\boldsymbol{f}|\boldsymbol{y})] \geq \mathcal{L}[\boldsymbol{\lambda}]$



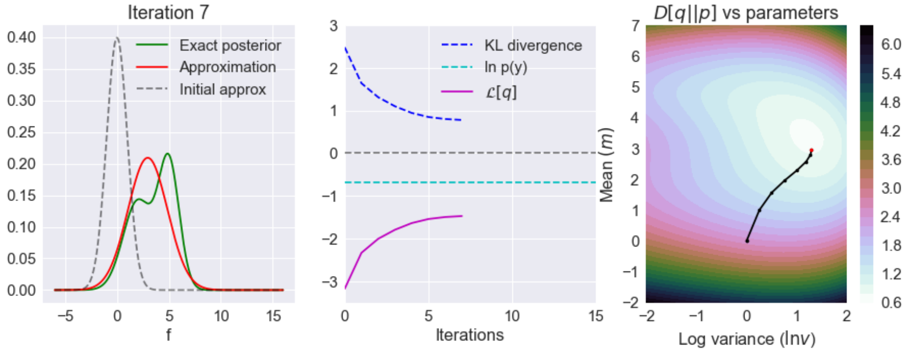
1D Toy example II

- Gradient ascent: $\boldsymbol{\lambda}_{i+1} = \boldsymbol{\lambda}_i + \eta \nabla_{\boldsymbol{\lambda}} \mathcal{L}[\boldsymbol{\lambda}]$
- $\ln p(\boldsymbol{y}) = \mathcal{L}[\boldsymbol{\lambda}] + \mathbb{D}[q_\lambda(\boldsymbol{f}) || p(\boldsymbol{f}|\boldsymbol{y})] \geq \mathcal{L}[\boldsymbol{\lambda}]$



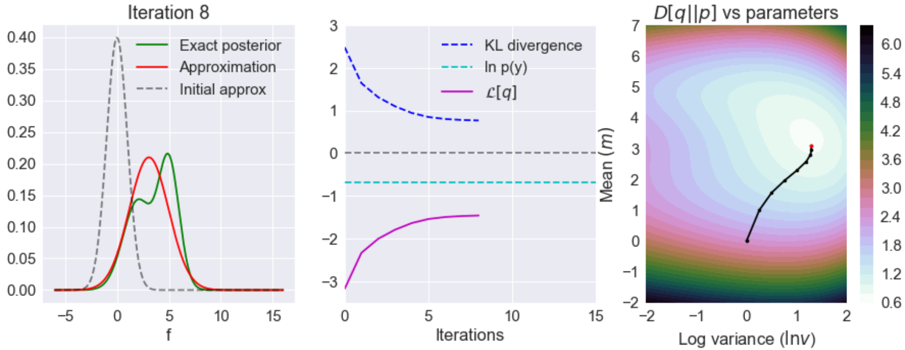
1D Toy example II

- Gradient ascent: $\boldsymbol{\lambda}_{i+1} = \boldsymbol{\lambda}_i + \eta \nabla_{\boldsymbol{\lambda}} \mathcal{L}[\boldsymbol{\lambda}]$
- $\ln p(\boldsymbol{y}) = \mathcal{L}[\boldsymbol{\lambda}] + \mathbb{D}[q_\lambda(\boldsymbol{f}) || p(\boldsymbol{f}|\boldsymbol{y})] \geq \mathcal{L}[\boldsymbol{\lambda}]$



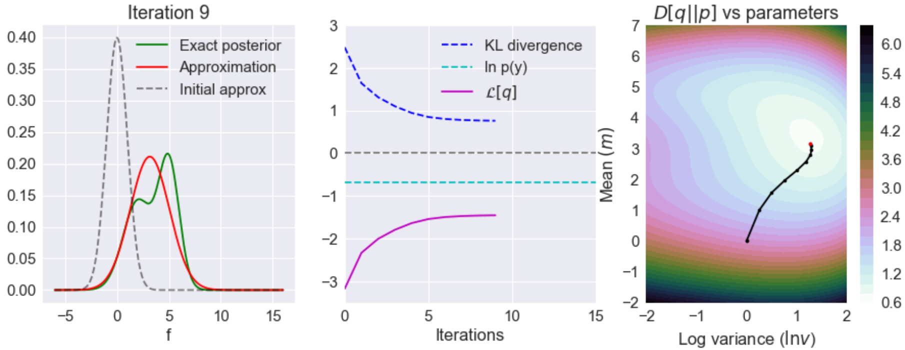
1D Toy example II

- Gradient ascent: $\boldsymbol{\lambda}_{i+1} = \boldsymbol{\lambda}_i + \eta \nabla_{\boldsymbol{\lambda}} \mathcal{L}[\boldsymbol{\lambda}]$
- $\ln p(\boldsymbol{y}) = \mathcal{L}[\boldsymbol{\lambda}] + \mathbb{D}[q_\lambda(\boldsymbol{f}) || p(\boldsymbol{f}|\boldsymbol{y})] \geq \mathcal{L}[\boldsymbol{\lambda}]$



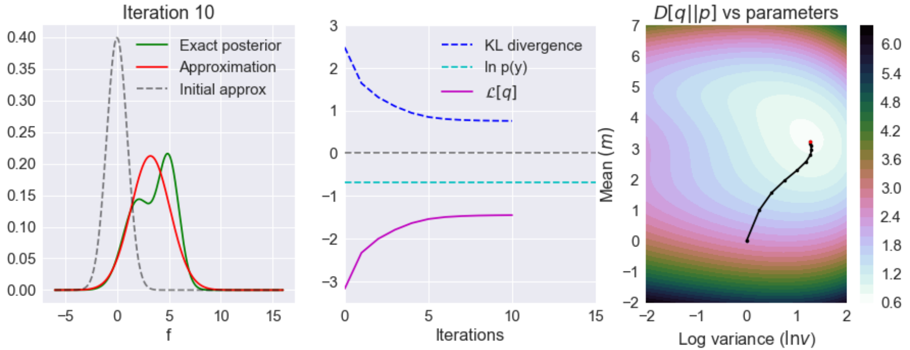
1D Toy example II

- Gradient ascent: $\boldsymbol{\lambda}_{i+1} = \boldsymbol{\lambda}_i + \eta \nabla_{\boldsymbol{\lambda}} \mathcal{L}[\boldsymbol{\lambda}]$
- $\ln p(\boldsymbol{y}) = \mathcal{L}[\boldsymbol{\lambda}] + \mathbb{D}[q_\lambda(\boldsymbol{f}) || p(\boldsymbol{f}|\boldsymbol{y})] \geq \mathcal{L}[\boldsymbol{\lambda}]$



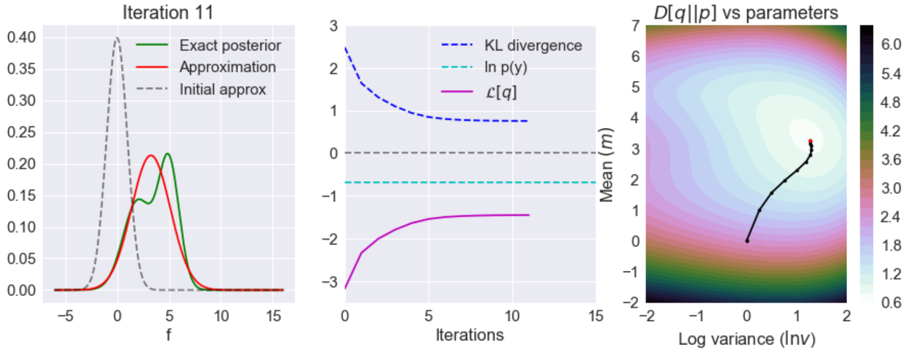
1D Toy example II

- Gradient ascent: $\boldsymbol{\lambda}_{i+1} = \boldsymbol{\lambda}_i + \eta \nabla_{\boldsymbol{\lambda}} \mathcal{L}[\boldsymbol{\lambda}]$
- $\ln p(\boldsymbol{y}) = \mathcal{L}[\boldsymbol{\lambda}] + \mathbb{D}[q_{\lambda}(\boldsymbol{f}) || p(\boldsymbol{f}|\boldsymbol{y})] \geq \mathcal{L}[\boldsymbol{\lambda}]$



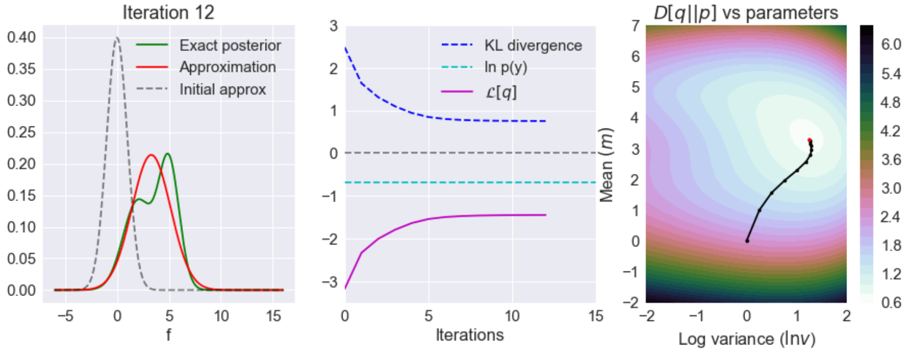
1D Toy example II

- Gradient ascent: $\boldsymbol{\lambda}_{i+1} = \boldsymbol{\lambda}_i + \eta \nabla_{\boldsymbol{\lambda}} \mathcal{L}[\boldsymbol{\lambda}]$
- $\ln p(\boldsymbol{y}) = \mathcal{L}[\boldsymbol{\lambda}] + \mathbb{D}[q_\lambda(\boldsymbol{f}) || p(\boldsymbol{f}|\boldsymbol{y})] \geq \mathcal{L}[\boldsymbol{\lambda}]$



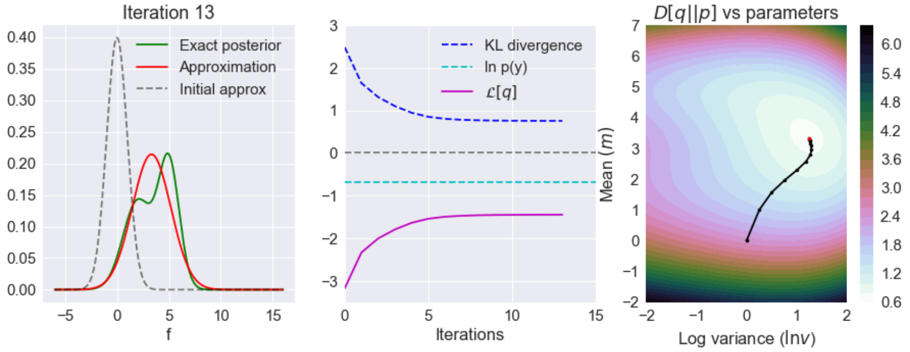
1D Toy example II

- Gradient ascent: $\boldsymbol{\lambda}_{i+1} = \boldsymbol{\lambda}_i + \eta \nabla_{\boldsymbol{\lambda}} \mathcal{L}[\boldsymbol{\lambda}]$
- $\ln p(\boldsymbol{y}) = \mathcal{L}[\boldsymbol{\lambda}] + \mathbb{D}[q_\lambda(\boldsymbol{f}) || p(\boldsymbol{f}|\boldsymbol{y})] \geq \mathcal{L}[\boldsymbol{\lambda}]$



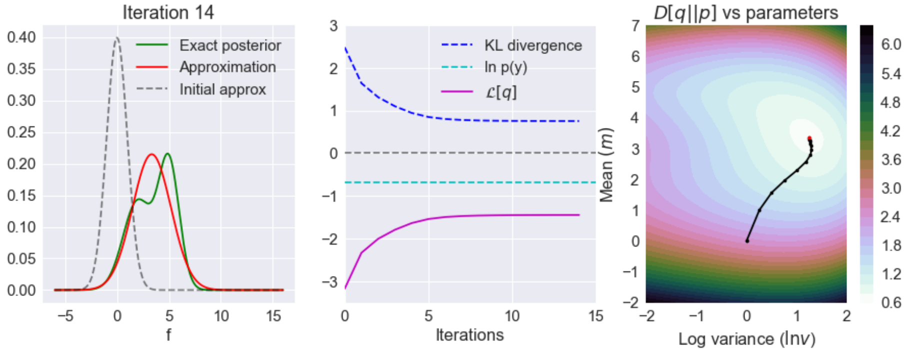
1D Toy example II

- Gradient ascent: $\boldsymbol{\lambda}_{i+1} = \boldsymbol{\lambda}_i + \eta \nabla_{\boldsymbol{\lambda}} \mathcal{L}[\boldsymbol{\lambda}]$
- $\ln p(\boldsymbol{y}) = \mathcal{L}[\boldsymbol{\lambda}] + \mathbb{D}[q_\lambda(\boldsymbol{f}) || p(\boldsymbol{f}|\boldsymbol{y})] \geq \mathcal{L}[\boldsymbol{\lambda}]$



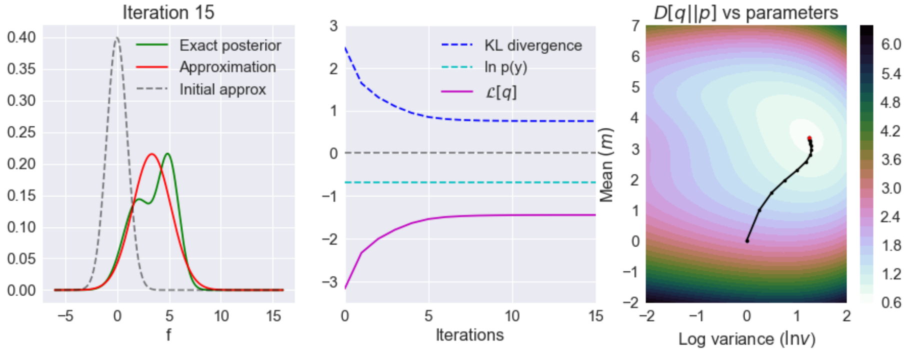
1D Toy example II

- Gradient ascent: $\boldsymbol{\lambda}_{i+1} = \boldsymbol{\lambda}_i + \eta \nabla_{\boldsymbol{\lambda}} \mathcal{L}[\boldsymbol{\lambda}]$
- $\ln p(\boldsymbol{y}) = \mathcal{L}[\boldsymbol{\lambda}] + \mathbb{D}[q_\lambda(\boldsymbol{f}) || p(\boldsymbol{f}|\boldsymbol{y})] \geq \mathcal{L}[\boldsymbol{\lambda}]$



1D Toy example II

- Gradient ascent: $\boldsymbol{\lambda}_{i+1} = \boldsymbol{\lambda}_i + \eta \nabla_{\boldsymbol{\lambda}} \mathcal{L}[\boldsymbol{\lambda}]$
- $\ln p(\boldsymbol{y}) = \mathcal{L}[\boldsymbol{\lambda}] + \mathbb{D}[q_\lambda(\boldsymbol{f}) || p(\boldsymbol{f}|\boldsymbol{y})] \geq \mathcal{L}[\boldsymbol{\lambda}]$



Computational challenges

- Let's see how we can use combine the ideas from variational inference with inducing points methods to solve the two computational problems:

  1. The computational complexity of GPs is $\mathcal{O}(N^3)$
  2. How to handle non-Gaussian likelihoods

Solution: Inducing point methods

- The main idea is to "represent" the information from the full dataset using a smaller "virtual" dataset

Solution: Inducing point methods

- The main idea is to "represent" the information from the full dataset using a smaller "virtual" dataset
- Recall our GP model:

$$p(\boldsymbol{y}, \boldsymbol{f}) = p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}), \quad \text{where} \quad \boldsymbol{f} = [f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \dots, f(\boldsymbol{x}_N)]$$

Solution: Inducing point methods

- The main idea is to "represent" the information from the full dataset using a smaller "virtual" dataset

- Recall our GP model:

$$p(\boldsymbol{y}, \boldsymbol{f}) = p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}), \quad \text{where} \quad \boldsymbol{f} = [f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \ldots, f(\boldsymbol{x}_N)]$$

- We will now introduce a set of *inducing points* $\{\boldsymbol{z}_m\}_{m=1}^{M}$

- They live in the same space as the input points, i.e. $\boldsymbol{x}_i, \boldsymbol{z}_j \in \mathbb{R}^D$

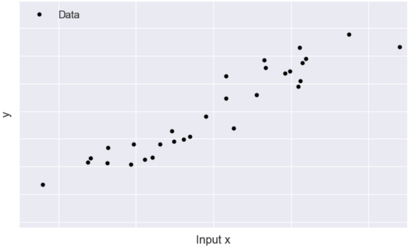
Solution: Inducing point methods

- The main idea is to "represent" the information from the full dataset using a smaller "virtual" dataset
- Recall our GP model:

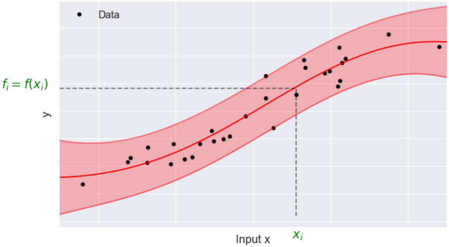
$$p(\boldsymbol{y}, \boldsymbol{f}) = p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}), \quad \text{where} \quad \boldsymbol{f} = [f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \ldots, f(\boldsymbol{x}_N)]$$

- We will now introduce a set of *inducing points* $\{\boldsymbol{z}_m\}_{m=1}^{M}$
- They live in the same space as the input points, i.e. $\boldsymbol{x}_i, \boldsymbol{z}_j \in \mathbb{R}^D$
- Let u_m denote the value of the function f evaluated at each $\boldsymbol{z}_m$, i.e. $u_m = f(\boldsymbol{z}_m)$
- ... and $\boldsymbol{u} = [f(\boldsymbol{z}_1), f(\boldsymbol{z}_2), \ldots, f(\boldsymbol{z}_M)]$

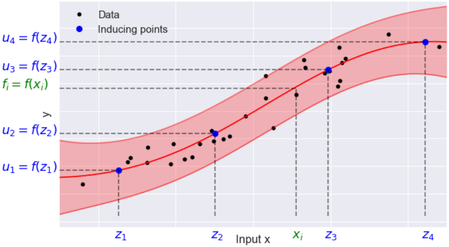
Inducing point methods



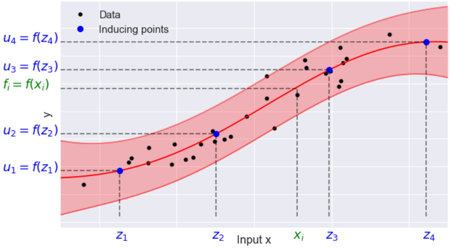
Inducing point methods



Inducing point methods

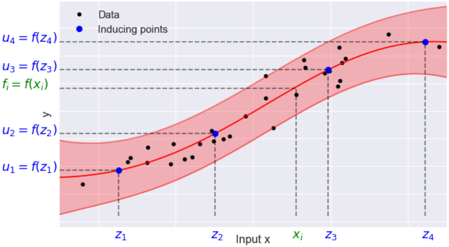


Inducing point methods



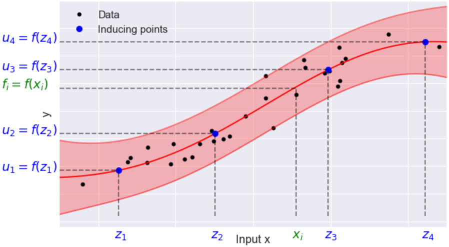
- Goal: choose the set of inducing points such that it contains the same information as the full dataset

Inducing point methods



- Goal: choose the set of inducing points such that it contains the same information as the full dataset
- Remember: Both $u_j = f(\boldsymbol{z}_j)$ and $f_i = f(\boldsymbol{x}_i)$ are random variables

Inducing point methods



- Goal: choose the set of inducing points such that it contains the same information as the full dataset
- Remember: Both $u_j = f(\boldsymbol{z}_j)$ and $f_i = f(\boldsymbol{x}_i)$ are random variables
- Next step: Formulate joint model $p(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{u})$

Inducing point methods: the joint model

- The augmented model

$$p(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{u}) = p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}, \boldsymbol{u})$$

- Let's decompose the "augmented" model as follows

$$p(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{u}) = p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u})$$

- We can get back to the original model by marginalizing over $\boldsymbol{u}$

$$p(\boldsymbol{y}, \boldsymbol{f}) = \int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}, \boldsymbol{u})\mathrm{d}\boldsymbol{u} = p(\boldsymbol{y}|\boldsymbol{f}) \int p(\boldsymbol{f}, \boldsymbol{u})\mathrm{d}\boldsymbol{u} = p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})$$

Setting up the approximation

- The idea is now to derive a variational approximation for the posterior $p(\boldsymbol{f}, \boldsymbol{u}|\boldsymbol{y})$

Setting up the approximation

- The idea is now to derive a variational approximation for the posterior $p(\boldsymbol{f}, \boldsymbol{u}|\boldsymbol{y})$
- We choose $\mathcal{Q}$ be the set of all distributions of the form $q(\boldsymbol{f}, \boldsymbol{u}) = p(\boldsymbol{f}|\boldsymbol{u})q(\boldsymbol{u})$, where $q(\boldsymbol{u}) = \mathcal{N}\left(\boldsymbol{u}|\boldsymbol{m}, \boldsymbol{S}\right)$

Setting up the approximation

- The idea is now to derive a variational approximation for the posterior $p(\boldsymbol{f}, \boldsymbol{u}|\boldsymbol{y})$
- We choose $\mathcal{Q}$ be the set of all distributions of the form $q(\boldsymbol{f}, \boldsymbol{u}) = p(\boldsymbol{f}|\boldsymbol{u})q(\boldsymbol{u})$, where $q(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{u}|\boldsymbol{m}, \boldsymbol{S})$
- Let's derive the ELBO, introducing $q(\boldsymbol{f}, \boldsymbol{u})$

$$
\begin{aligned}
\ln p(\boldsymbol{y}) &\geq \mathbb{E}_{q(\boldsymbol{u},\boldsymbol{f})} \ln p(\boldsymbol{y}|\boldsymbol{f}) - \mathbb{E}_{q(\boldsymbol{u},\boldsymbol{f})} \frac{q(\boldsymbol{f}, \boldsymbol{u})}{p(\boldsymbol{f}, \boldsymbol{u})} \\
&= \mathbb{E}_{q(\boldsymbol{f})} \ln p(\boldsymbol{y}|\boldsymbol{f}) - \mathbb{E}_{q(\boldsymbol{u},\boldsymbol{f})} \frac{p(\boldsymbol{f}|\boldsymbol{u})q(\boldsymbol{u})}{p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u})} \\
&= \mathbb{E}_{q(\boldsymbol{f})} \ln p(\boldsymbol{y}|\boldsymbol{f}) - \mathbb{E}_{q(\boldsymbol{u})} \frac{q(\boldsymbol{u})}{p(\boldsymbol{u})} \\
&= \mathbb{E}_{q(\boldsymbol{f})} \ln p(\boldsymbol{y}|\boldsymbol{f}) - \mathbb{D}[q(\boldsymbol{u})||p(\boldsymbol{u})] = \mathcal{L}
\end{aligned}
$$

The inducing points approximation

- **Take-away #1**: We can now tractably optimize the lower bound wrt. $\boldsymbol{m}$, $\boldsymbol{S}$, and even $\boldsymbol{z}$

$$\ln p(\boldsymbol{y}) \geq \mathbb{E}_{q(\boldsymbol{f})}\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}[q(\boldsymbol{u})||p(\boldsymbol{u})] \equiv \mathcal{L}$$

The inducing points approximation

- **Take-away #1**: We can now tractably optimize the lower bound wrt. $\boldsymbol{m}$, $\boldsymbol{S}$, and even $\boldsymbol{z}$

$$\ln p(\boldsymbol{y}) \geq \mathbb{E}_{q(\boldsymbol{f})}\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}[q(\boldsymbol{u})||p(\boldsymbol{u})] \equiv \mathcal{L}$$

- We will now show that the first decomposes in a very convenient way

The inducing points approximation

- **Take-away #1**: We can now tractably optimize the lower bound wrt. $\boldsymbol{m}$, $\boldsymbol{S}$, and even $\boldsymbol{z}$
$$\ln p(\boldsymbol{y}) \geq \mathbb{E}_{q(\boldsymbol{f})}\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}[q(\boldsymbol{u})||p(\boldsymbol{u})] \equiv \mathcal{L}$$
- We will now show that the first decomposes in a very convenient way
- Remember: $p(\boldsymbol{y}|\boldsymbol{f}) = \prod_{i=1}^{N} p(y_i|f_i)$

The inducing points approximation

- **Take-away #1**: We can now tractably optimize the lower bound wrt. $\boldsymbol{m}$, $\boldsymbol{S}$, and even $\boldsymbol{z}$

$$\ln p(\boldsymbol{y}) \geq \mathbb{E}_{q(\boldsymbol{f})}\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}[q(\boldsymbol{u})||p(\boldsymbol{u})] \equiv \mathcal{L}$$

- We will now show that the first decomposes in a very convenient way
- Remember: $p(\boldsymbol{y}|\boldsymbol{f}) = \prod_{i=1}^{N} p(y_i|f_i)$
- Let's have a closer look at the first term

$$\mathbb{E}_{q(\boldsymbol{f})}\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] = \mathbb{E}_{q(\boldsymbol{f})}\left[\ln \prod_{i=1}^{N} p(y_i|f_i)\right] = \sum_{i=1}^{N} \mathbb{E}_{q(f_i)}\left[\ln p(y_i|f_i)\right]$$

The inducing points approximation

- **Take-away #1**: We can now tractably optimize the lower bound wrt. $\boldsymbol{m}$, $\boldsymbol{S}$, and even $\boldsymbol{z}$
$$\ln p(\boldsymbol{y}) \geq \mathbb{E}_{q(\boldsymbol{f})}\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}[q(\boldsymbol{u})||p(\boldsymbol{u})] \equiv \mathcal{L}$$
- We will now show that the first decomposes in a very convenient way
- Remember: $p(\boldsymbol{y}|\boldsymbol{f}) = \prod_{i=1}^{N} p(y_i|f_i)$
- Let's have a closer look at the first term
$$\mathbb{E}_{q(\boldsymbol{f})}\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] = \mathbb{E}_{q(\boldsymbol{f})}\left[\ln \prod_{i=1}^{N} p(y_i|f_i)\right] = \sum_{i=1}^{N} \mathbb{E}_{q(f_i)}\left[\ln p(y_i|f_i)\right]$$
  where
$$q(f_i) = \int p(f_i|\boldsymbol{u})\mathcal{N}\left(\boldsymbol{u}|\boldsymbol{m}, \boldsymbol{S}\right) \mathrm{d}\boldsymbol{u} = \mathcal{N}\left(f_i|\boldsymbol{k}_{im}\boldsymbol{K}_{mm}^{-1}\boldsymbol{m}, \tilde{K}_{ii} + \boldsymbol{k}_{im}\boldsymbol{K}_{mm}^{-1}\boldsymbol{S}\boldsymbol{K}_{mm}^{-1}\boldsymbol{k}_{mi}\right)$$

The inducing points approximation

- **Take-away #1**: We can now tractably optimize the lower bound wrt. $\boldsymbol{m}$, $\boldsymbol{S}$, and even $\boldsymbol{z}$

$$\ln p(\boldsymbol{y}) \geq \mathbb{E}_{q(\boldsymbol{f})}\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}[q(\boldsymbol{u})||p(\boldsymbol{u})] \equiv \mathcal{L}$$

- We will now show that the first decomposes in a very convenient way
- Remember: $p(\boldsymbol{y}|\boldsymbol{f}) = \prod_{i=1}^{N} p(y_i|f_i)$
- Let's have a closer look at the first term

$$\mathbb{E}_{q(\boldsymbol{f})}\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] = \mathbb{E}_{q(\boldsymbol{f})}\left[\ln \prod_{i=1}^{N} p(y_i|f_i)\right] = \sum_{i=1}^{N} \mathbb{E}_{q(f_i)}\left[\ln p(y_i|f_i)\right]$$

where

$$q(f_i) = \int p(f_i|\boldsymbol{u})\mathcal{N}\left(\boldsymbol{u}|\boldsymbol{m},\boldsymbol{S}\right)\mathrm{d}\boldsymbol{u} = \mathcal{N}\left(f_i|\boldsymbol{k}_{im}\boldsymbol{K}_{mm}^{-1}\boldsymbol{m}, \tilde{K}_{ii} + \boldsymbol{k}_{im}\boldsymbol{K}_{mm}^{-1}\boldsymbol{S}\boldsymbol{K}_{mm}^{-1}\boldsymbol{k}_{mi}\right)$$

Thus, the "likelihood term"

The inducing points approximation

- **Take-away #1**: We can now tractably optimize the lower bound wrt. $\boldsymbol{m}$, $\boldsymbol{S}$, and even $\boldsymbol{z}$

$$\ln p(\boldsymbol{y}) \geq \mathbb{E}_{q(\boldsymbol{f})}\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}[q(\boldsymbol{u})||p(\boldsymbol{u})] \equiv \mathcal{L}$$

- We will now show that the first decomposes in a very convenient way
- Remember: $p(\boldsymbol{y}|\boldsymbol{f}) = \prod_{i=1}^{N} p(y_i|f_i)$
- Let's have a closer look at the first term

$$\mathbb{E}_{q(\boldsymbol{f})}\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] = \mathbb{E}_{q(\boldsymbol{f})}\left[\ln \prod_{i=1}^{N} p(y_i|f_i)\right] = \sum_{i=1}^{N} \mathbb{E}_{q(f_i)}\left[\ln p(y_i|f_i)\right]$$

where

$$q(f_i) = \int p(f_i|\boldsymbol{u})\mathcal{N}\left(\boldsymbol{u}|\boldsymbol{m},\boldsymbol{S}\right)\mathrm{d}\boldsymbol{u} = \mathcal{N}\left(f_i|\boldsymbol{k}_{im}\boldsymbol{K}_{mm}^{-1}\boldsymbol{m}, \tilde{K}_{ii} + \boldsymbol{k}_{im}\boldsymbol{K}_{mm}^{-1}\boldsymbol{S}\boldsymbol{K}_{mm}^{-1}\boldsymbol{k}_{mi}\right)$$

Thus, the "likelihood term"

- decomposes into a sum over 1D integrals

The inducing points approximation

- **Take-away #1**: We can now tractably optimize the lower bound wrt. $\boldsymbol{m}$, $\boldsymbol{S}$, and even $\boldsymbol{z}$

$$\ln p(\boldsymbol{y}) \geq \mathbb{E}_{q(\boldsymbol{f})}\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}[q(\boldsymbol{u})||p(\boldsymbol{u})] \equiv \mathcal{L}$$

- We will now show that the first decomposes in a very convenient way
- Remember: $p(\boldsymbol{y}|\boldsymbol{f}) = \prod_{i=1}^{N} p(y_i|f_i)$
- Let's have a closer look at the first term

$$\mathbb{E}_{q(\boldsymbol{f})}\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] = \mathbb{E}_{q(\boldsymbol{f})}\left[\ln \prod_{i=1}^{N} p(y_i|f_i)\right] = \sum_{i=1}^{N} \mathbb{E}_{q(f_i)}\left[\ln p(y_i|f_i)\right]$$

where

$$q(f_i) = \int p(f_i|\boldsymbol{u})\mathcal{N}\left(\boldsymbol{u}|\boldsymbol{m}, \boldsymbol{S}\right) \mathrm{d}\boldsymbol{u} = \mathcal{N}\left(f_i|\boldsymbol{k}_{im}\boldsymbol{K}_{mm}^{-1}\boldsymbol{m}, \tilde{K}_{ii} + \boldsymbol{k}_{im}\boldsymbol{K}_{mm}^{-1}\boldsymbol{S}\boldsymbol{K}_{mm}^{-1}\boldsymbol{k}_{mi}\right)$$

Thus, the "likelihood term"

- decomposes into a sum over 1D integrals
- Can be solved analytically for Gaussian likelihoods and some classification likelihoods

The inducing points approximation

- **Take-away #1**: We can now tractably optimize the lower bound wrt. $\boldsymbol{m}$, $\boldsymbol{S}$, and even $\boldsymbol{z}$

$$\ln p(\boldsymbol{y}) \geq \mathbb{E}_{q(\boldsymbol{f})}\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] - \mathbb{D}[q(\boldsymbol{u})||p(\boldsymbol{u})] \equiv \mathcal{L}$$

- We will now show that the first decomposes in a very convenient way
- Remember: $p(\boldsymbol{y}|\boldsymbol{f}) = \prod_{i=1}^{N} p(y_i|f_i)$
- Let's have a closer look at the first term

$$\mathbb{E}_{q(\boldsymbol{f})}\left[\ln p(\boldsymbol{y}|\boldsymbol{f})\right] = \mathbb{E}_{q(\boldsymbol{f})}\left[\ln \prod_{i=1}^{N} p(y_i|f_i)\right] = \sum_{i=1}^{N} \mathbb{E}_{q(f_i)}\left[\ln p(y_i|f_i)\right]$$

where

$$q(f_i) = \int p(f_i|\boldsymbol{u})\mathcal{N}\left(\boldsymbol{u}|\boldsymbol{m}, \boldsymbol{S}\right) \mathrm{d}\boldsymbol{u} = \mathcal{N}\left(f_i|\boldsymbol{k}_{im}\boldsymbol{K}_{mm}^{-1}\boldsymbol{m}, \tilde{K}_{ii} + \boldsymbol{k}_{im}\boldsymbol{K}_{mm}^{-1}\boldsymbol{S}\boldsymbol{K}_{mm}^{-1}\boldsymbol{k}_{mi}\right)$$

Thus, the "likelihood term"

- decomposes into a sum over 1D integrals
- Can be solved analytically for Gaussian likelihoods and some classification likelihoods
- But it is fast to approximate 1D integrals using numerical integration for other likelihoods

The inducing points approximation

- **Take-away #1**: We can now tractably optimize the lower bound wrt. $\boldsymbol{m}$, $\boldsymbol{S}$, and even $\boldsymbol{z}$

$$\ln p(\boldsymbol{y}) \geq \mathbb{E}_{q(\boldsymbol{f})}[\ln p(\boldsymbol{y}|\boldsymbol{f})] - \mathbb{D}[q(\boldsymbol{u})||p(\boldsymbol{u})] \equiv \mathcal{L}$$

- We will now show that the first decomposes in a very convenient way
- Remember: $p(\boldsymbol{y}|\boldsymbol{f}) = \prod_{i=1}^{N} p(y_i|f_i)$
- Let's have a closer look at the first term

$$\mathbb{E}_{q(\boldsymbol{f})}[\ln p(\boldsymbol{y}|\boldsymbol{f})] = \mathbb{E}_{q(\boldsymbol{f})}\left[\ln \prod_{i=1}^{N} p(y_i|f_i)\right] = \sum_{i=1}^{N} \mathbb{E}_{q(f_i)}[\ln p(y_i|f_i)]$$

where

$$q(f_i) = \int p(f_i|\boldsymbol{u})\mathcal{N}(\boldsymbol{u}|\boldsymbol{m}, \boldsymbol{S})\,\mathrm{d}\boldsymbol{u} = \mathcal{N}\left(f_i|\boldsymbol{k}_{im}\boldsymbol{K}_{mm}^{-1}\boldsymbol{m}, \tilde{K}_{ii} + \boldsymbol{k}_{im}\boldsymbol{K}_{mm}^{-1}\boldsymbol{S}\boldsymbol{K}_{mm}^{-1}\boldsymbol{k}_{mi}\right)$$

Thus, the "likelihood term"

- decomposes into a sum over 1D integrals
- Can be solved analytically for Gaussian likelihoods and some classification likelihoods
- But it is fast to approximate 1D integrals using numerical integration for other likelihoods
- **Take away #2**: We can tractably optimize the bound even with non-Gaussian likelihoods

The resulting bound

- Substituting back into $\mathcal{L}$

$$\ln p(\boldsymbol{y}) \geq \mathcal{L} = \sum_{i=1}^{N} \int q(f_i) \ln p(y_i|f_i) \mathrm{d}f_i - \mathbb{D}[q(\boldsymbol{u})||p(\boldsymbol{u})]$$

- We want to optimize $\mathcal{L}$ wrt. $\boldsymbol{\lambda} = \{\boldsymbol{m}, \boldsymbol{S}, \boldsymbol{z}\}$ using gradient-based methods

$$\nabla_{\boldsymbol{\lambda}} \mathcal{L} = \nabla_{\boldsymbol{\lambda}} \sum_{i=1}^{N} \int q(f_i) \ln p(y_i|f_i) \mathrm{d}f_i - \nabla_{\boldsymbol{\lambda}} \mathbb{D}[q(\boldsymbol{u})||p(\boldsymbol{u})]$$

The resulting bound

- Substituting back into $\mathcal{L}$

$$\ln p(\boldsymbol{y}) \geq \mathcal{L} = \sum_{i=1}^{N} \int q(f_i) \ln p(y_i|f_i) \mathrm{d}f_i - \mathbb{D}[q(\boldsymbol{u})||p(\boldsymbol{u})]$$

- We want to optimize $\mathcal{L}$ wrt. $\boldsymbol{\lambda} = \{\boldsymbol{m}, \boldsymbol{S}, \boldsymbol{z}\}$ using gradient-based methods

$$\nabla_{\boldsymbol{\lambda}} \mathcal{L} = \nabla_{\boldsymbol{\lambda}} \sum_{i=1}^{N} \int q(f_i) \ln p(y_i|f_i) \mathrm{d}f_i - \nabla_{\boldsymbol{\lambda}} \mathbb{D}[q(\boldsymbol{u})||p(\boldsymbol{u})]$$

- We can approximate the gradient as follows (mini-batching)

$$\nabla_{\boldsymbol{\lambda}} \sum_{i=1}^{N} \int q(f_i) \ln p(y_i|f_i) \mathrm{d}f_i \approx \frac{N}{|S|} \sum_{i \in S} \nabla_{\boldsymbol{\lambda}} \int q(f_i) \ln p(y_i|f_i) \mathrm{d}f_i$$

The resulting bound

- Substituting back into $\mathcal{L}$

$$\ln p(\boldsymbol{y}) \geq \mathcal{L} = \sum_{i=1}^{N} \int q(f_i) \ln p(y_i|f_i) \mathrm{d}f_i - \mathbb{D}[q(\boldsymbol{u})||p(\boldsymbol{u})]$$

- We want to optimize $\mathcal{L}$ wrt. $\boldsymbol{\lambda} = \{\boldsymbol{m}, \boldsymbol{S}, \boldsymbol{z}\}$ using gradient-based methods

$$\nabla_{\boldsymbol{\lambda}} \mathcal{L} = \nabla_{\boldsymbol{\lambda}} \sum_{i=1}^{N} \int q(f_i) \ln p(y_i|f_i) \mathrm{d}f_i - \nabla_{\boldsymbol{\lambda}} \mathbb{D}[q(\boldsymbol{u})||p(\boldsymbol{u})]$$

- We can approximate the gradient as follows (mini-batching)

$$\nabla_{\boldsymbol{\lambda}} \sum_{i=1}^{N} \int q(f_i) \ln p(y_i|f_i) \mathrm{d}f_i \approx \frac{N}{|S|} \sum_{i \in S} \nabla_{\boldsymbol{\lambda}} \int q(f_i) \ln p(y_i|f_i) \mathrm{d}f_i$$

- **Take away #3**: Because it decomposes as a sum over the data points, the bound becomes amendable to stochastic gradient descent (mini-batching) and hence, we can scale the method to really really large datasets!

Example from the paper

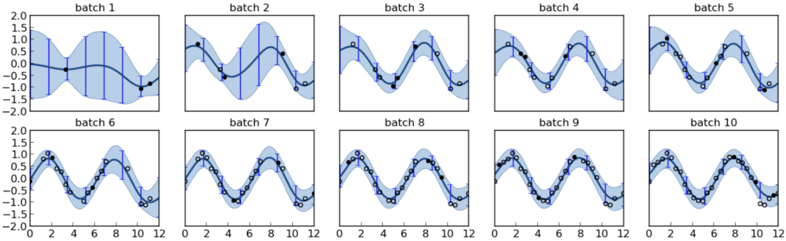


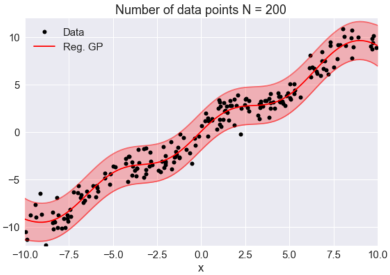
Figure 2: Stochastic variational inference on a trivial GP regression problem. Each pane shows the posterior of the GP after a batch of data, marked as solid points. Previoulsy seen (and discarded) data are marked as empty points, the distribution $q(\mathbf{u})$ is represented by vertical errorbars.

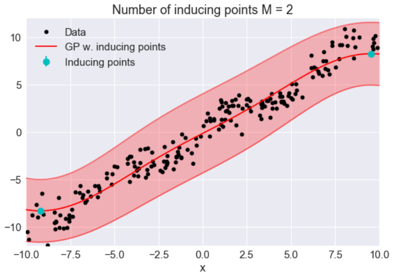
(from Hensman et al: Gaussian processes for big data)

Inducing points method summary

- The inducing point approximation allows us to
  - ... scale Gaussian processes to big data
  - ... use non-Gaussian likelihoods
- It reduces the computational complexity from $\mathcal{O}(N^3)$ to $\mathcal{O}(M^3)$, where $M \ll N$
- It's implemented in most GP toolboxes, e.g. GPy (numpy) and gpflow (tensorflow)

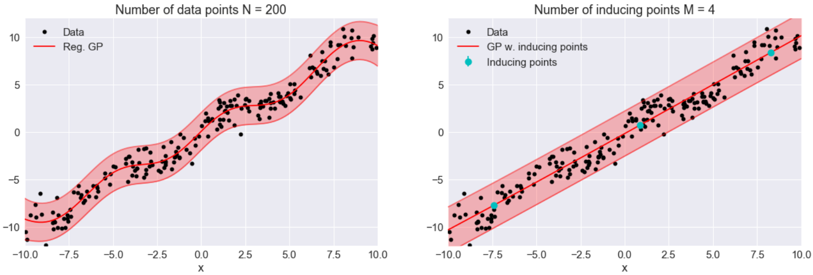
Example: Number of inducing points





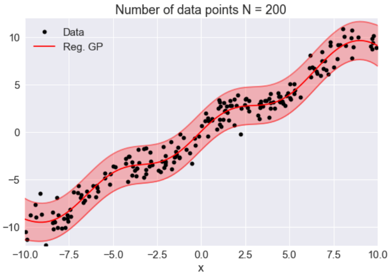
- We can think of the number of inducing points as a parameter that trades off speed for accuracy

Example: Number of inducing points



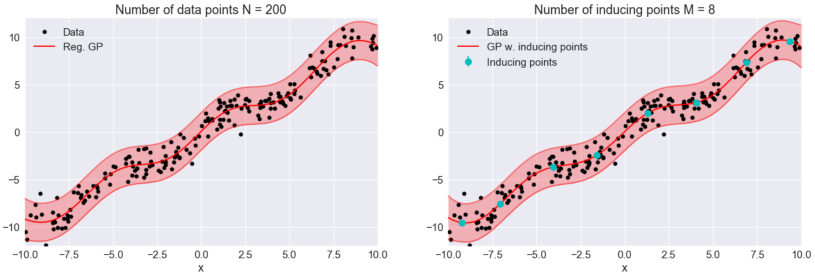
- We can think of the number of inducing points as a parameter that trades off speed for accuracy

Example: Number of inducing points



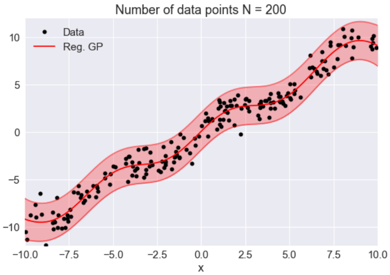
- We can think of the number of inducing points as a parameter that trades off speed for accuracy

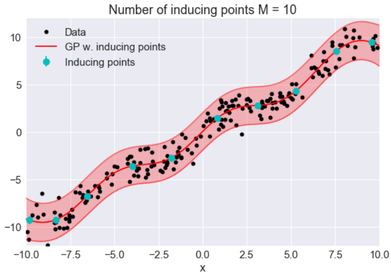
Example: Number of inducing points



- We can think of the number of inducing points as a parameter that trades off speed for accuracy

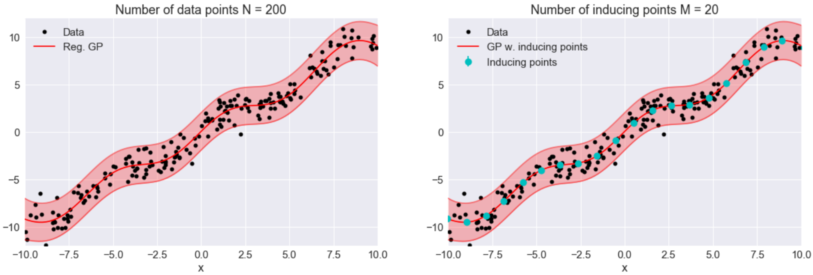
Example: Number of inducing points





- We can think of the number of inducing points as a parameter that trades off speed for accuracy

Example: Number of inducing points



- We can think of the number of inducing points as a parameter that trades off speed for accuracy

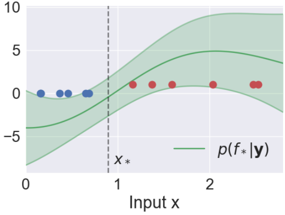
Gaussian process classification: Inference

Three steps to compute the predictive distribution for a new test point $\boldsymbol{x}_*$

$$p(\boldsymbol{y}, \boldsymbol{f}) = \prod_{n=1}^{N} p(y_n|f_n)p(\boldsymbol{f}) = \prod_{n=1}^{N} \phi(y_n \cdot f_n) \mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{K})$$

- Step 1: Compute posterior distribution of $p(\boldsymbol{f}|\boldsymbol{y})$:

$$p(\boldsymbol{f}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})}{p(\boldsymbol{y})} \approx q(\boldsymbol{f})$$

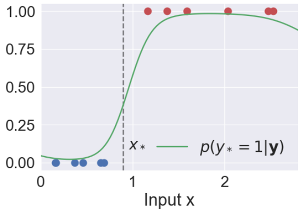


- Step 2: Compute posterior of f_* for new test point $\boldsymbol{x}_*$:

$$p(f_*|\boldsymbol{y}) = \int p(f_*|\boldsymbol{f})\, p(\boldsymbol{f}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{f} \approx \int p(f_*|\boldsymbol{f})\, q(\boldsymbol{f})\, \mathrm{d}\boldsymbol{f}$$

- Step 3: Compute predictive distribution

$$p(y_*|\boldsymbol{y}) = \int \phi(y_* \cdot f_*)\, p(f_*|\boldsymbol{y}) \mathrm{d}f_*$$



Predictive distribution

- Using the (approximate) posterior $q(f_*)$, we can compute $p(y_*|\mathbf{y})$

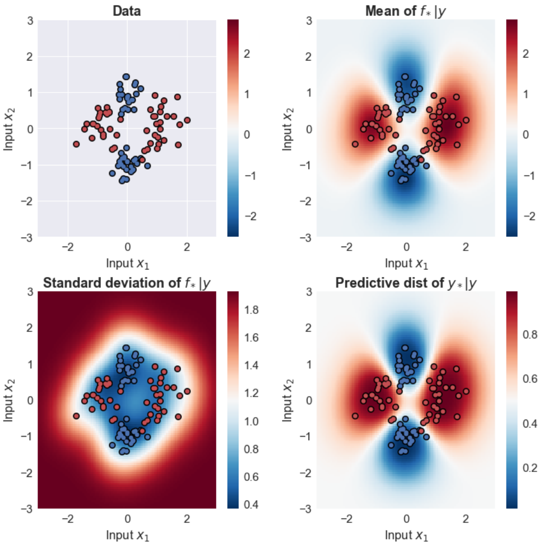
$$
\begin{aligned}
p(y_* = 1|\mathbf{y}) &= \int p(y_*|f_*)p(f_*\big|\mathbf{y})\mathrm{d}f_* \\
&= \int \phi\left(y_* \cdot f_*\right) p(f_*\big|\mathbf{y})\mathrm{d}f_* \\
&\approx \int \phi\left(y_* \cdot f_*\right) q\left(f_*\right) \mathrm{d}f_* \\
&= \int \phi\left(y_* \cdot f_*\right) \mathcal{N}\left(f_*\big|\mu_*, \sigma_*^2\right) \mathrm{d}f_* \\
&= \phi\left(\frac{\mu_*}{\sqrt{1+\sigma_*^2}}\right)
\end{aligned}
$$

Can you figure it out?

- What can we say about the predictive distributions for y_* when μ_* is positive? or negative?
- How does the uncertainty of the posterior distribution of f_* influence the predictions for y_*? What happens as σ_*^2 approaches ∞?

Gaussian process classification example

- Non-linear classification problem
- $N = 100$ data points
- Squared exponential kernel
- Hyperparameters are chosen by optimizing $\mathcal{L}$



Next time

Next Monday Charles Gadd will talk about

- latent variable modelling (GPs for unsupervised learning),
- Multi-Output GPs

Read:

- Michalis Titsias, Neil D. Lawrence (2010), *Bayesian Gaussian Process Latent Variable Model*, ICML
- Andrew Gordon Wilson, David A. Knowles, Zoubin Ghahramani (2012), *Gaussian Process Regression Networks*, ICML

Assignments

- Assignment #1: done
- Assignment #2: deadline 27th of January
- Assignment #3:
  - handed: 25th of January
  - due: 3rd of February