Probabilistic Programming

Rajbir-Singh Nirwan

October 22, 2019

Outline

- Probabilistic Modelling
- Sampling
- Variational Inference
- Probabilistic Programming

Why do we need it?

- Uncertainty estimation
- Intrinsic Regularization
- Explicit assumptions
- More interpretable models

Recap

Observed Data

$$\mathscr{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$$

$$\boldsymbol{x}_n \in \mathbb{R}^D, \; y_n \in \mathbb{R}$$

Model

$$y_n = f_{\boldsymbol{w}}(\boldsymbol{x}_n) + \epsilon_n = \boldsymbol{w}^T\boldsymbol{x}_n + \epsilon_n$$

$$E_{\mathscr{D}}(\boldsymbol{w}) = \sum_{n=1}^{N} \left(y_n - \boldsymbol{w}^T\boldsymbol{x}_n\right)^2$$

$$+\lambda \|\boldsymbol{w}\|_2^2$$

Fit

$$\boldsymbol{w}^* = \min_{\boldsymbol{w}} E_{\mathscr{D}}(\boldsymbol{w})$$

Prediction

$$y_{new} = f_{\boldsymbol{w}^*}(\boldsymbol{x}_{new})$$

Recap

Observed Data $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$

$\boldsymbol{x}_n \in \mathbb{R}^D,\ y_n \in \mathbb{R}$

Model $y_n = f_{\boldsymbol{w}}(\boldsymbol{x}_n) + \epsilon_n = \boldsymbol{w}^T\boldsymbol{x}_n + \epsilon_n$

$$E_{\mathcal{D}}(\boldsymbol{w}) = \sum_{n=1}^{N}\left(y_n - \boldsymbol{w}^T\boldsymbol{x}_n\right)^2 + \lambda\|\boldsymbol{w}\|_2^2$$

Fit $\boldsymbol{w}^* = \min_{\boldsymbol{w}} E_{\mathcal{D}}(\boldsymbol{w})$

Prediction $y_{new} = f_{\boldsymbol{w}^*}(\boldsymbol{x}_{new})$

Observed Data $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$

$\boldsymbol{x}_n \in \mathbb{R}^D,\ y_n \in \mathbb{R}$

Model $y_n = f_{\boldsymbol{w}}(\boldsymbol{x}_n) + \epsilon_n = \boldsymbol{w}^T\boldsymbol{x}_n + \epsilon_n$

$$p(\boldsymbol{y}\,|\,\boldsymbol{w}, \sigma^2) = \prod_{n=1}^{N}\mathcal{N}(y_n\,|\,\boldsymbol{w}^T\boldsymbol{x}_n, \sigma^2)$$

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}\,|\,\boldsymbol{0}, \boldsymbol{I})$$

Inference $p(\boldsymbol{w}\,|\,\boldsymbol{y}) = \dfrac{p(\boldsymbol{y}\,|\,\boldsymbol{w})\ p(\boldsymbol{w})}{p(\boldsymbol{y})}$

Prediction $p(y_{new}\,|\,\boldsymbol{y}) = \int p(y_{new}\,|\,\boldsymbol{w})\ p(\boldsymbol{w}\,|\,\boldsymbol{y})d\boldsymbol{w}$

Probabilistic Modelling

$$p(x) = \int p(x, y)\ dy \qquad \& \qquad p(x, y) = p(y \mid x)\ p(x)$$

- Inference

$$p(\theta \mid \mathscr{D}) = \frac{p(\mathscr{D} \mid \theta)\ p(\theta)}{p(\mathscr{D})}$$

- Prediction

$$p(y \mid \mathscr{D}) = \int p(y \mid \theta)\ p(\theta \mid \mathscr{D})\ d\theta$$

Probabilistic Modelling

$$p(x) = \int p(x, y)\ dy \qquad \& \qquad p(x, y) = p(y \mid x)\ p(x)$$

- Inference

$$p(\theta \mid \mathscr{D}) = \frac{p(\mathscr{D} \mid \theta)\ p(\theta)}{p(\mathscr{D})}$$

- Prediction

$$p(y \mid \mathscr{D}) = \int p(y \mid \theta)\ p(\theta \mid \mathscr{D})\ d\theta$$

- Not tractable most of the time
  ➡ Approximation Methods
    - Sampling
    - Variational Inference

Sampling (HMC)

- Approximate $p(\boldsymbol{\theta} \,|\, \mathscr{D})$ by N samples

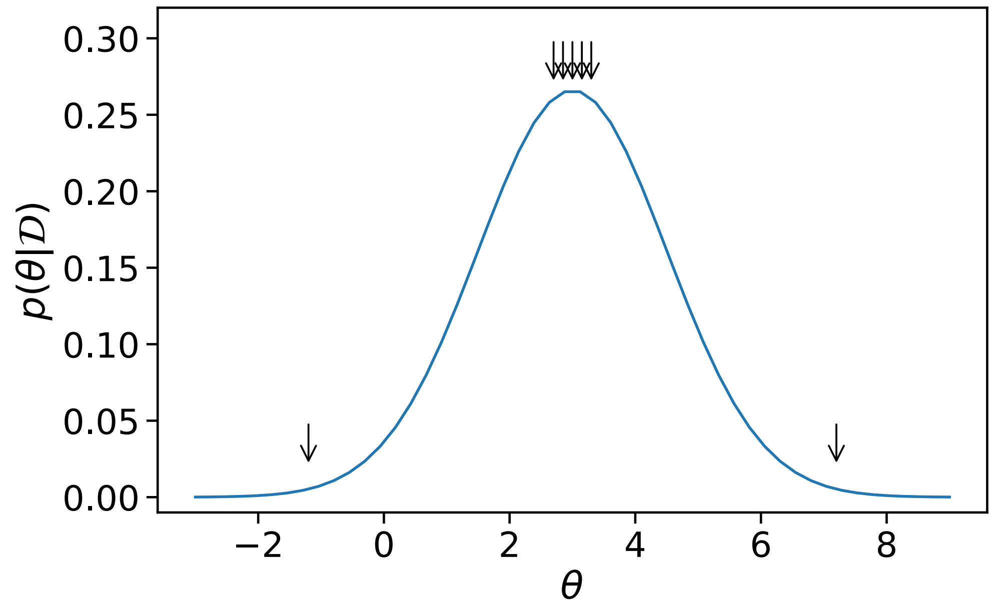
$$\mathbb{E}_p[f] = \int p(\boldsymbol{\theta} \,|\, \mathscr{D}) \; f(\boldsymbol{\theta}) \; d\boldsymbol{\theta} \approx \frac{1}{N} \sum_{n=1}^{N} f(\boldsymbol{\theta}_n)$$

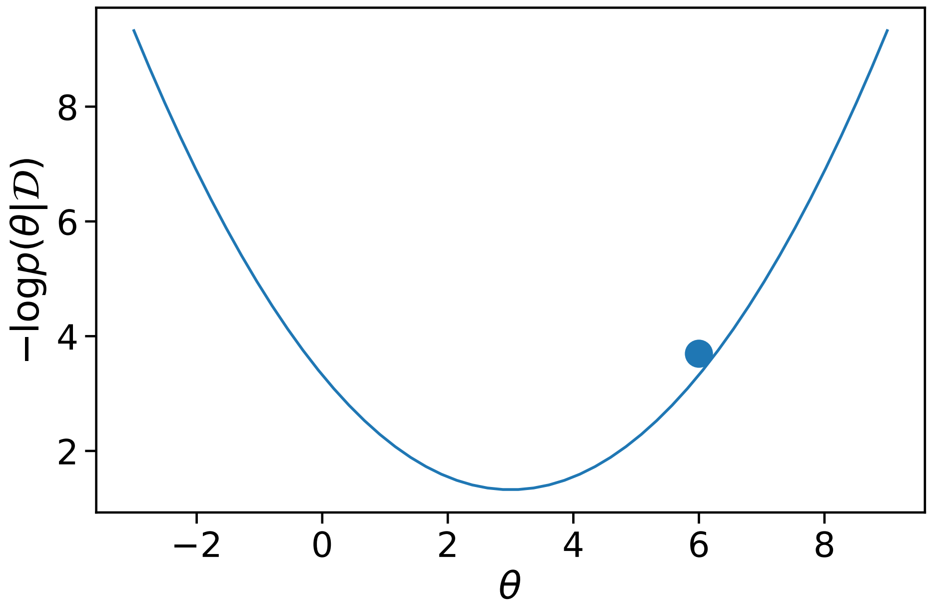
Sampling (HMC)

- Approximate $p(\boldsymbol{\theta} \mid \mathscr{D})$ by N samples

$$\mathbb{E}_p[f] = \int p(\boldsymbol{\theta} \mid \mathscr{D}) \, f(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \approx \frac{1}{N} \sum_{n=1}^{N} f(\boldsymbol{\theta}_n)$$

- HMC: Particle moving in probability landscape according to Hamiltons equations with random Gaussian "kicks"





Sampling (HMC)

$$\mathbf{H}(\boldsymbol{\theta}, \boldsymbol{p}) = \mathbf{U}(\boldsymbol{\theta}) + \mathbf{K}(\boldsymbol{p}) = \mathbf{const}$$

$$\mathbf{U}(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta} \,|\, \mathscr{D}) \qquad \mathbf{K}(\boldsymbol{p}) = \frac{\boldsymbol{p}^2}{2m}$$

Sampling (HMC)

$$\mathsf{H}(\boldsymbol{\theta}, \boldsymbol{p}) = \mathsf{U}(\boldsymbol{\theta}) + \mathsf{K}(\boldsymbol{p}) = \textbf{const}$$

$$\mathsf{U}(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta}\,|\,\mathscr{D}) \qquad \mathsf{K}(\boldsymbol{p}) = \frac{\boldsymbol{p}^2}{2m}$$

$$p(\boldsymbol{\theta}, \boldsymbol{p}) \propto \exp(-\mathsf{H}) = p(\boldsymbol{\theta}\,|\,\mathscr{D})\ e^{-\boldsymbol{p}^2/2m}$$

$\rightarrow$ $\boldsymbol{\theta}$ and $\boldsymbol{p}$ are uncorrelated

$\rightarrow$ $\boldsymbol{p}$ is Gaussian

Sampling (HMC)

$$\mathbf{H}(\boldsymbol{\theta}, \boldsymbol{p}) = \mathbf{U}(\boldsymbol{\theta}) + \mathbf{K}(\boldsymbol{p}) = \mathbf{const}$$

$$\mathbf{U}(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta} \,|\, \mathscr{D}) \qquad \mathbf{K}(\boldsymbol{p}) = \frac{\boldsymbol{p}^2}{2m}$$

$$p(\boldsymbol{\theta}, \boldsymbol{p}) \propto \exp(-\mathbf{H}) = p(\boldsymbol{\theta} \,|\, \mathscr{D})\ e^{-\boldsymbol{p}^2/2m}$$

$\rightarrow \boldsymbol{\theta}$ and $\boldsymbol{p}$ are uncorrelated

$\rightarrow \boldsymbol{p}$ is Gaussian

Hamilton's Equations:

$$\dot{\boldsymbol{\theta}} = \frac{\partial \mathbf{H}}{\partial \boldsymbol{p}} \qquad \dot{\boldsymbol{p}} = -\frac{\partial \mathbf{H}}{\partial \boldsymbol{\theta}}$$

Sampling (HMC)

$$\mathbf{H}(\boldsymbol{\theta}, \boldsymbol{p}) = \mathbf{U}(\boldsymbol{\theta}) + \mathbf{K}(\boldsymbol{p}) = \mathbf{const}$$

$$\mathbf{U}(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta}\,|\,\mathscr{D}) \qquad \mathbf{K}(\boldsymbol{p}) = \frac{\boldsymbol{p}^2}{2m}$$

$$p(\boldsymbol{\theta}, \boldsymbol{p}) \propto \exp(-\mathbf{H}) = p(\boldsymbol{\theta}\,|\,\mathscr{D})\; e^{-\boldsymbol{p}^2/2m}$$

→ $\boldsymbol{\theta}$ and $\boldsymbol{p}$ are uncorrelated

→ $\boldsymbol{p}$ is Gaussian

Hamilton's Equations:

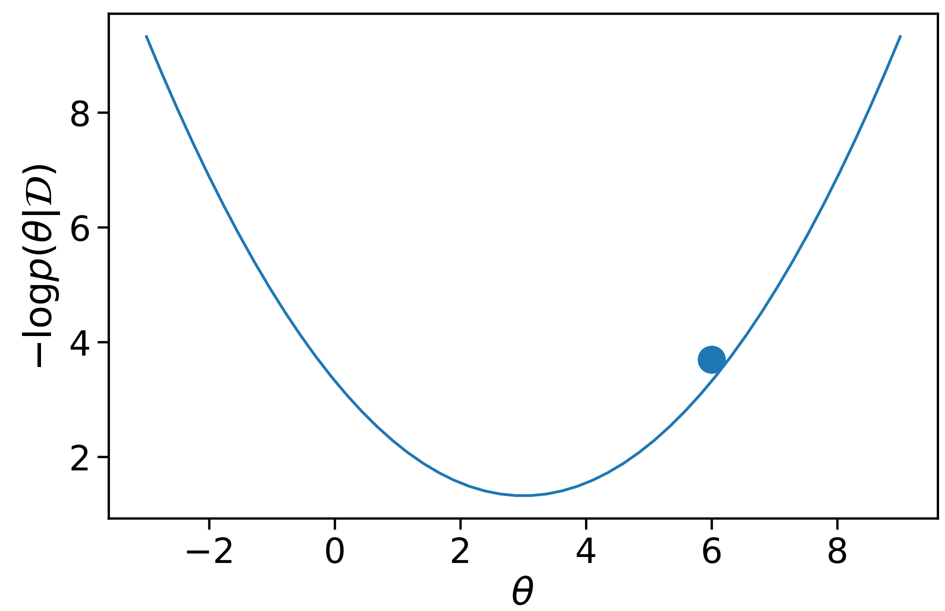
$$\dot{\boldsymbol{\theta}} = \frac{\partial \mathbf{H}}{\partial \boldsymbol{p}} \qquad \dot{\boldsymbol{p}} = -\frac{\partial \mathbf{H}}{\partial \boldsymbol{\theta}}$$

https://github.com/PyMLVizard/PyMLViz

Sampling (in Theory):

1. Choose some $\boldsymbol{\theta}$ at random
2. N times:
    1. $\boldsymbol{p} \sim \mathcal{N}(0,1)$
    2. Solve for path with $\mathbf{H}(\boldsymbol{\theta}, \boldsymbol{p})$ for fixed amount of time
    3. Save $\boldsymbol{\theta}$ and $\boldsymbol{p}$ as a sample

Chain of $\boldsymbol{\theta}$s converge to $p(\boldsymbol{\theta}\,|\,\mathscr{D})$



Variational Bayes

- Approximate $p(\boldsymbol{\theta} \,|\, \mathscr{D})$ by $q_{\boldsymbol{\nu}}(\boldsymbol{\theta})$

Variational Bayes

- Approximate $p(\boldsymbol{\theta} \mid \mathscr{D})$ by $q_{\boldsymbol{\nu}}(\boldsymbol{\theta})$

- Goodness is measured in e.g. Kullback-Leibler divergence

$$\textbf{KL}[q_{\boldsymbol{\nu}}(\boldsymbol{\theta}) \| p(\boldsymbol{\theta} \mid \mathscr{D})] = \int q_{\boldsymbol{\nu}}(\boldsymbol{\theta}) \ln \frac{q_{\boldsymbol{\nu}}(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \mathscr{D})} d\boldsymbol{\theta}$$

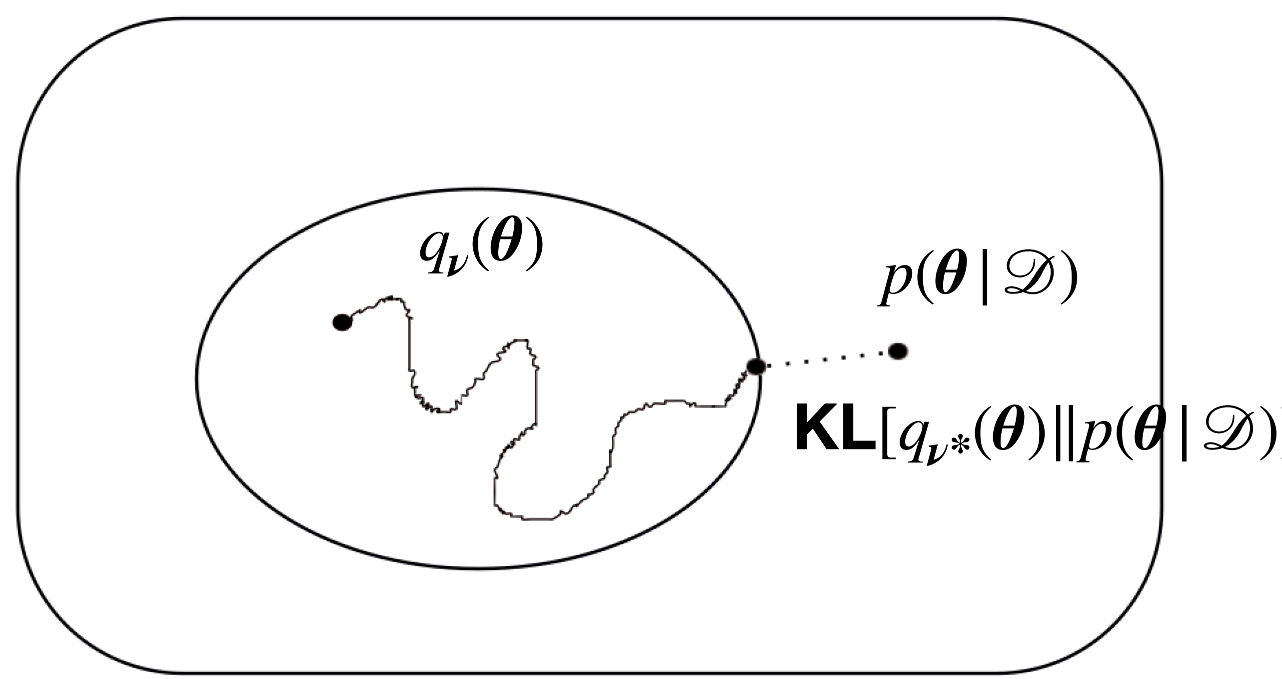
Variational Bayes

- Approximate $p(\boldsymbol{\theta}\,|\,\mathscr{D})$ by $q_{\boldsymbol{\nu}}(\boldsymbol{\theta})$

- Goodness is measured in e.g. Kullback-Leibler divergence

$$\textbf{KL}[q_{\boldsymbol{\nu}}(\boldsymbol{\theta}) \| p(\boldsymbol{\theta}\,|\,\mathscr{D})] = \int q_{\boldsymbol{\nu}}(\boldsymbol{\theta}) \ln \frac{q_{\boldsymbol{\nu}}(\boldsymbol{\theta})}{p(\boldsymbol{\theta}\,|\,\mathscr{D})} d\boldsymbol{\theta}$$

Turns Inference Into Optimization



Variational Bayes

$$\ln p(\mathscr{D}) = \textbf{ELBO} + \textbf{KL}[q_{\boldsymbol{\nu}}(\boldsymbol{\theta}) \| p(\boldsymbol{\theta} \mid \mathscr{D})]$$

Variational Bayes

$$\ln p(\mathscr{D}) = \textbf{ELBO} + \textbf{KL}[q_{\boldsymbol{\nu}}(\boldsymbol{\theta}) \| p(\boldsymbol{\theta} \,|\, \mathscr{D})]$$

$$\textbf{ELBO} = \int q_{\boldsymbol{\nu}}(\boldsymbol{\theta}) \ln \frac{p(\mathscr{D}) \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{q_{\boldsymbol{\nu}}(\boldsymbol{\theta})} d\boldsymbol{\theta} = -\int q_{\boldsymbol{\nu}}(\boldsymbol{\theta}) \ln \frac{q_{\boldsymbol{\nu}}(\boldsymbol{\theta})}{\tilde{p}(\boldsymbol{\theta} \,|\, \mathscr{D})} d\boldsymbol{\theta}$$

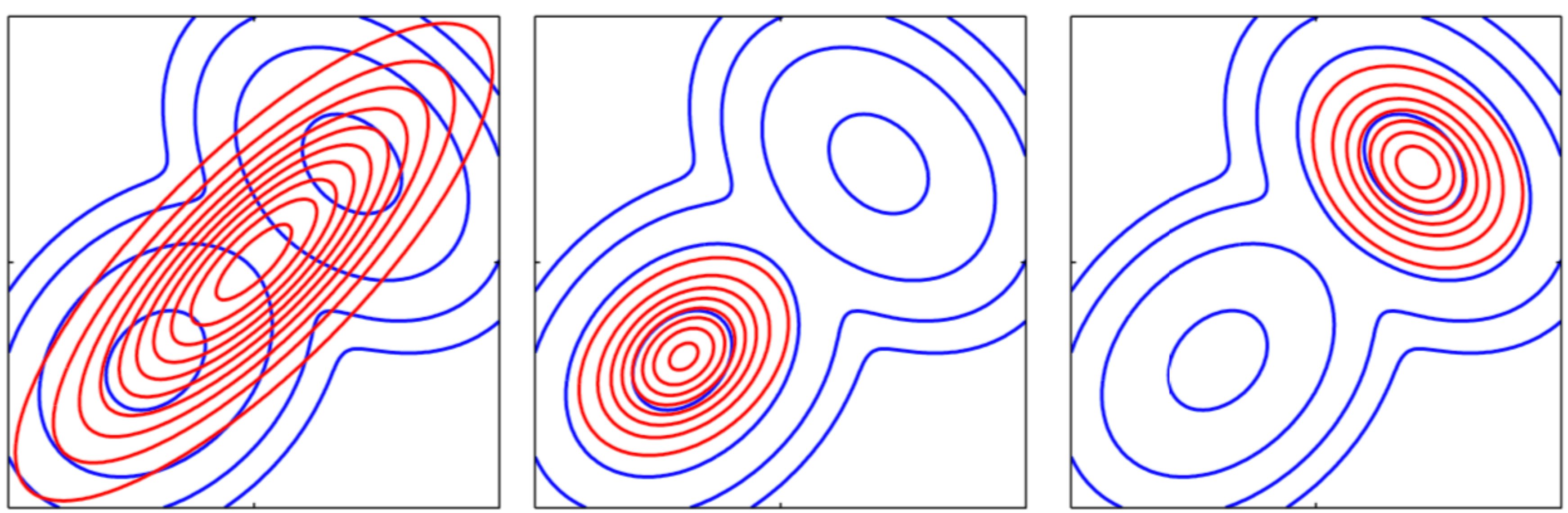
$$\textbf{KL}[q_{\boldsymbol{\nu}}(\boldsymbol{\theta}) \| p(\boldsymbol{\theta} \,|\, \mathscr{D})] = \int q_{\boldsymbol{\nu}}(\boldsymbol{\theta}) \ln \frac{q_{\boldsymbol{\nu}}(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \,|\, \mathscr{D})} d\boldsymbol{\theta}$$

Variational Bayes

$$\ln p(\mathscr{D}) = \textbf{ELBO} + \textbf{KL}[q_{\boldsymbol{\nu}}(\boldsymbol{\theta}) \| p(\boldsymbol{\theta} \,|\, \mathscr{D})]$$

$$\textbf{ELBO} = \int q_{\boldsymbol{\nu}}(\boldsymbol{\theta}) \ln \frac{p(\mathscr{D}) \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{q_{\boldsymbol{\nu}}(\boldsymbol{\theta})} d\boldsymbol{\theta} = - \int q_{\boldsymbol{\nu}}(\boldsymbol{\theta}) \ln \frac{q_{\boldsymbol{\nu}}(\boldsymbol{\theta})}{\tilde{p}(\boldsymbol{\theta} \,|\, \mathscr{D})} d\boldsymbol{\theta}$$

$$\textbf{KL}[q_{\boldsymbol{\nu}}(\boldsymbol{\theta}) \| p(\boldsymbol{\theta} \,|\, \mathscr{D})] = \int q_{\boldsymbol{\nu}}(\boldsymbol{\theta}) \ln \frac{q_{\boldsymbol{\nu}}(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \,|\, \mathscr{D})} d\boldsymbol{\theta}$$



Pattern Recognition and Machine Learning, Christopher M. Bishop

Thanks for your attention!