Advanced Probabilistic Machine Learning

Exercise session 10 – Variational inference

UPPSALA
UNIVERSITET

Aleix Nieto Juscafresa
Division of Systems and Control
Department of Information Technology Uppsala University

aleix.nieto-juscafresa@it.uu.se
`aleixnieto.github.io/`

Bayesian framework reminder

In this course, we solve problems using Bayes' theorem

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$$

Objective: *Update the belief about the parameters θ of our model based on observed data $\mathcal{D}$ using Bayes' theorem.*

- $\mathcal{D}$: observed data
- θ : parameters of our model
- $p(\theta)$: **prior** belief of parameters before we collected any data
- $p(\theta|\mathcal{D})$: **posterior** belief of parameters after inferring data
- $p(\mathcal{D}|\theta)$: **likelihood** of the data in view of the parameters

Inference methods

Broad set of techniques for making inferences about model parameters or latent variables based on observed data.

- **Exact inference**: **Conjugate priors**
  - \+ Exact
  - \+ Fast
  - − Limited applicability (there is not always a closed form sol.)
- **Stochastic approximate inference**: **Monte Carlo**
  - \+ Asymptotically exact (converges to the true posterior)
  - − Computationally costly

↓ ***Today:***

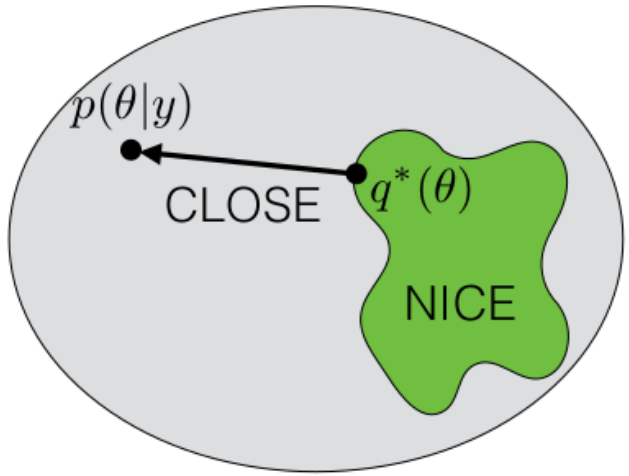
- **Deterministic approximate inference**: **Expectation propagation (EI)** and **Variational inference (VI)**
  - \+ Fast (compared to MC)
  - − Not asymptotically exact.

Deterministic approx. inference: The idea

Instead of sampling (as in Monte Carlo methods), we turn the inference problem into an optimization problem.

Approximate the posterior $p(\boldsymbol{\theta}|\mathcal{D})$ with tractable distribution $q(\boldsymbol{\theta}) \in Q$

$$\hat{q}(\boldsymbol{\theta}) = \underset{q(\boldsymbol{\theta}) \in Q}{\arg\min}\, D\left(p(\boldsymbol{\theta}|\mathcal{D}), q(\boldsymbol{\theta})\right)$$



Short summary of Tamara's Variational Bayes tutorial in ICML2018 `medium.com/@aminamollaysa/`

Kullback-Leibler divergence

A common choice is the **Kullback-Leibler divergence**

$$\mathrm{KL}\left[p \parallel q\right] = -\int p(\mathbf{x}) \ln \frac{q(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x}.$$

Some properties

- Non-negative $\mathrm{KL}\left[p(\mathbf{x}) \parallel q(\mathbf{x})\right] \geq 0$
- $\mathrm{KL}\left[p(\mathbf{x}) \parallel q(\mathbf{x})\right] = 0$ if and only if $p(\mathbf{x}) = q(\mathbf{x})$
- Non-symmetric $\mathrm{KL}\left[p(\mathbf{x}) \parallel q(\mathbf{x})\right] \neq \mathrm{KL}\left[q(\mathbf{x}) \parallel p(\mathbf{x})\right]$

EP vs. VI

The non-symmetry leads to two different classes of variational approximations:

Expectation Propagation (EP)

$$D\left(p(\boldsymbol{\theta}|\mathcal{D}), q(\boldsymbol{\theta})\right) = \mathrm{KL}\left[p(\boldsymbol{\theta}|\mathcal{D}) \parallel q(\boldsymbol{\theta})\right]$$

Variational Inference (VI)

$$D\left(p(\boldsymbol{\theta}|\mathcal{D}), q(\boldsymbol{\theta})\right) = \mathrm{KL}\left[q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta}|\mathcal{D})\right]$$

- Tends to underestimate variance $\rightarrow$ narrower approximation leads to overconfidence
- May not explore lower probability but important regions, limiting the posterior's uncertainty representation.

EP vs. VI

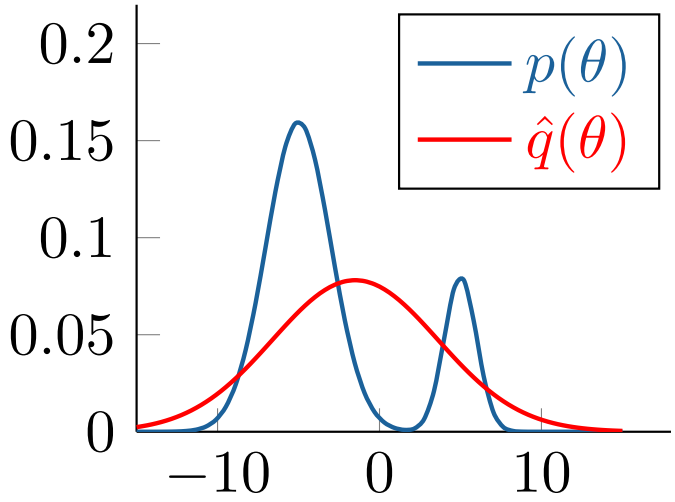
Let

$$p(\theta) = 0.2\mathcal{N}\left(\theta; 5, 1^2\right) + 0.8\mathcal{N}\left(\theta; -5, 2^2\right)$$

$$q(\theta) = \mathcal{N}\left(\theta; \mu, \sigma^2\right)$$

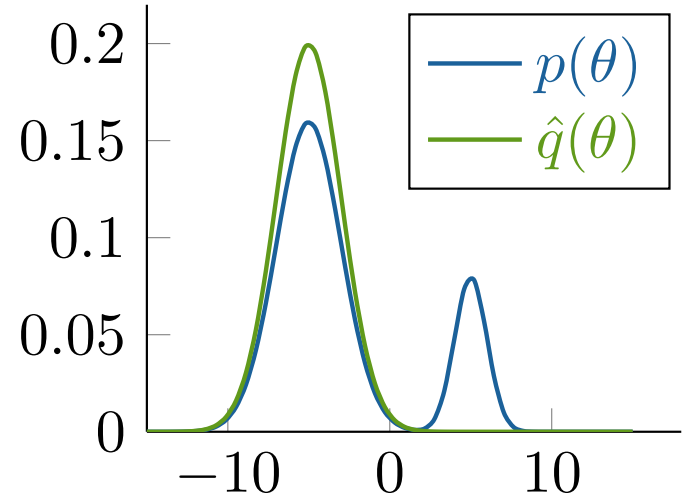
Expectation propagation

$\hat{q} = \min_{\mu,\sigma} \mathrm{KL}\left[p \parallel q_{\mu,\sigma}\right]$



Variational inference

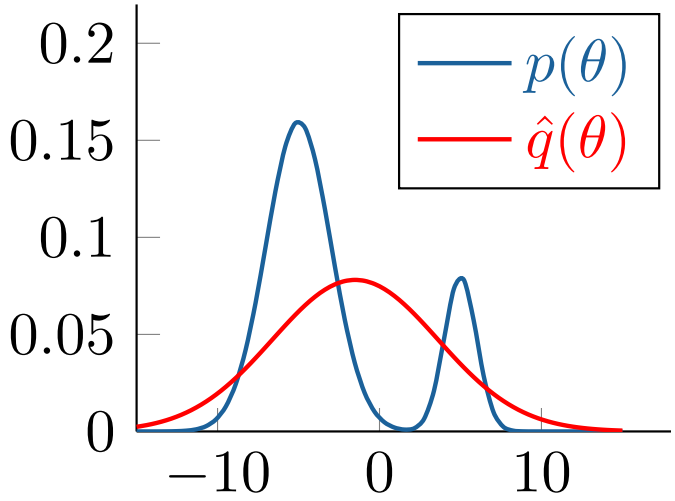
$\hat{q} = \min_{\mu,\sigma} \mathrm{KL}\left[q_{\mu,\sigma} \parallel p\right]$



EP vs. VI

Expectation propagation

$\hat{q} = \min_{\mu,\sigma} \mathrm{KL}\left[p \parallel q_{\mu,\sigma}\right]$



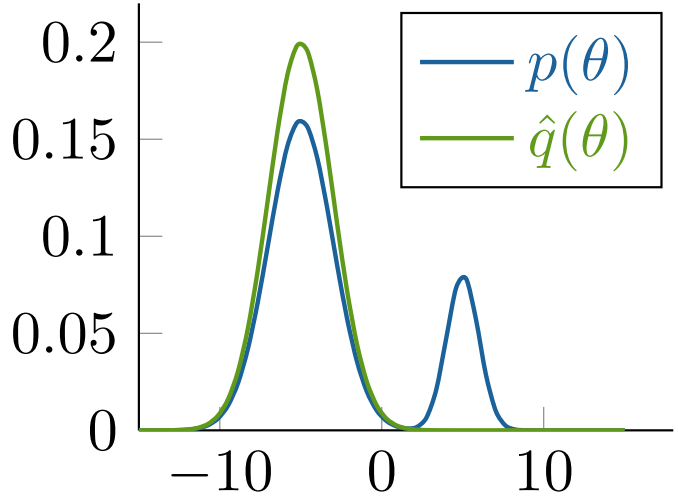
$$\mathrm{KL}\left[p \parallel q_{\mu,\sigma}\right] = -\int p(\theta) \ln \frac{q_{\mu,\sigma}(\theta)}{p(\theta)} d\theta$$

non-zero-forcing

Where $p \gg 0$, q needs to be $\gg 0$.

Variational inference

$\hat{q} = \min_{\mu,\sigma} \mathrm{KL}\left[q_{\mu,\sigma} \parallel p\right]$



$$\mathrm{KL}\left[q_{\mu,\sigma} \parallel p\right] = -\int q_{\mu,\sigma}(\theta) \ln \frac{p(\theta)}{q_{\mu,\sigma}(\theta)} d\theta$$

zero-forcing

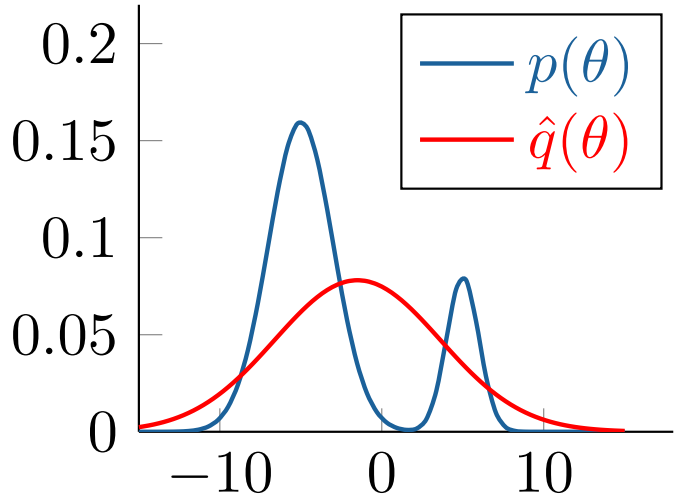
Where $p \approx 0$, q needs to be ≈ 0.

Moment matching

For the first form

$$\mathrm{KL}\left[p \parallel q_{\mu,\sigma}\right] = -\int p(\theta) \ln \frac{q_{\mu,\sigma}(\theta)}{p(\theta)} d\theta$$

$\hat{q} = \min_{\mu,\sigma} \mathrm{KL}\left[p \parallel q_{\mu,\sigma}\right]$



we have that $\hat{\mu}, \hat{\sigma} = \arg\min_{\mu,\sigma} \mathrm{KL}\left[p \parallel q_{\mu,\sigma}\right]$ gives

$$\hat{\mu} = \int \theta p(\theta) d\theta = \mathbb{E}_p[\theta] = \mu_p$$

$$\hat{\sigma}^2 = \int (\theta - \hat{\mu})^2 p(\theta) d\theta = \mathbb{E}_p[(\theta - \hat{\mu})^2] = \sigma_p^2$$

We call this **moment matching** (*Exercise 2!*)

Variational inference

The KL divergence is tricky to compute since it contains an expression of the posterior which we do not have access to.

However, it can be reformulated as:

$$\ln p(\mathcal{D}) = \underbrace{\int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{\theta}, \mathcal{D})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}}_{\mathcal{L}(q) = \text{ELBO}} - \underbrace{\mathrm{KL}\left[q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta}|\mathcal{D})\right]}_{\bigstar \geq 0}.$$

- $\ln p(\mathcal{D}) \geq \mathcal{L}(q)$, i.e., it is a lower bound on the worst-case for the log-likelihood of the data $\ln p(\mathcal{D})$.
- By maximizing $\mathcal{L}(q)$ we get a minimum of $\bigstar$.
- ELBO is usually tractable to estimate via Monte-Carlo.

Why do we choose KL divergence?

Why do we choose KL divergence?

1 $\rightarrow$ Captures the essential "information"

Goal: Make $q(\theta)$ reflect the key characteristics of $p(\theta|\mathcal{D})$.

- KL divergence measures how one probability distribution diverges from a second reference distribution.
- In the context of variational inference:
  - **True posterior** $p(\theta|\mathcal{D})$ contains the full information about parameters given data.
  - **Approximate posterior** $q(\theta)$ tries to capture the essential information while constrained by its form (e.g., Gaussian).

Why do we choose KL divergence?

2 → KL minimization and low-probability regions

- Minimizing $\mathrm{KL}(q(\theta) \parallel p(\theta|\mathcal{D}))$ prevents $q(\theta)$ from assigning mass to low-probability regions of $p(\theta|\mathcal{D})$.
- KL divergence is:

$$\mathrm{KL}(q(\theta) \parallel p(\theta|\mathcal{D})) = \int q(\theta) \log \left(\frac{p(\theta|\mathcal{D})}{q(\theta)} \right) d\theta$$

- A small ratio $\frac{p(\theta|\mathcal{D})}{q(\theta)}$ (when $p(\theta|\mathcal{D})$ is small) leads to a high KL cost.
- This encourages $q(\theta)$ to focus on regions where $p(\theta|\mathcal{D})$ has significant mass.

Why do we choose KL divergence?

3 $\rightarrow$ **Intractability of** $\log p(\mathcal{D})$

- The marginal likelihood $p(\mathcal{D})$ is:

$$p(\mathcal{D}) = \int p(\mathcal{D}|\theta)p(\theta)d\theta$$

- This involves integrating over the entire parameter space θ.
- Reasons for intractability:
  - $p(\theta)$ and $p(\mathcal{D}|\theta)$ can be complex, high-dimensional distributions.
  - Does not have a closed-form solution in most cases.
  - Numerical methods become inefficient in high-dimensional parameter spaces.
- Variational inference avoids direct computation by approximating the evidence.

UPPSALA UNIVERSITET

Why do we choose KL divergence?

4 → VI as regularized maximum likelihood

- Maximum likelihood estimation maximizes the likelihood:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} p(\mathcal{D}|\theta)$$

- Variational inference instead maximizes the ELBO:

$$\text{ELBO} = \mathbb{E}_q[\log p(\mathcal{D}|\theta)] - \text{KL}(q(\theta) \parallel p(\theta))$$

- KL divergence acts as a regularization term, penalizing $q(\theta)$ if it deviates too much from the prior $p(\theta)$.
- This balances fitting the data well (likelihood) with staying close to prior beliefs (regularization), preventing overfitting.