Variational inference

Dmitry Vetrov

Research professor at HSE

Head of ML lab in SAIC-Moscow

Special thanks to Ekaterina Lobacheva for assistance with slides

Deep|Bayes

Outline: Variational Inference

- Variational lower bound derivation
- Variational mean field approximation

Full Bayesian inference

Training stage:

$$p\left(\theta \mid X_{tr}, Y_{tr}\right) = \frac{p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta)}{\int p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta) d\theta}$$

Testing stage:

$$p\left(y \mid x, X_{tr}, Y_{tr}\right) = \int p(y \mid x, \theta) p\left(\theta \mid X_{tr}, Y_{tr}\right) d\theta$$

Full Bayesian inference

Training stage:

$$p\left(\theta \mid X_{tr}, Y_{tr}\right) = \frac{p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta)}{\boxed{\int p\left(Y_{tr} \mid X_{tr}, \theta\right) p(\theta) d\theta}}$$

Testing stage:

May be intractable

$$p\left(y \mid x, X_{tr}, Y_{tr}\right) = \boxed{\int p(y \mid x, \theta) p\left(\theta \mid X_{tr}, Y_{tr}\right) d\theta}$$

Posterior distributions can be calculated analytically only for simple conjugate models!

Approximate inference

Probabilistic model: $p(x, \theta) = p(x \mid \theta)p(\theta)$

Variational Inference

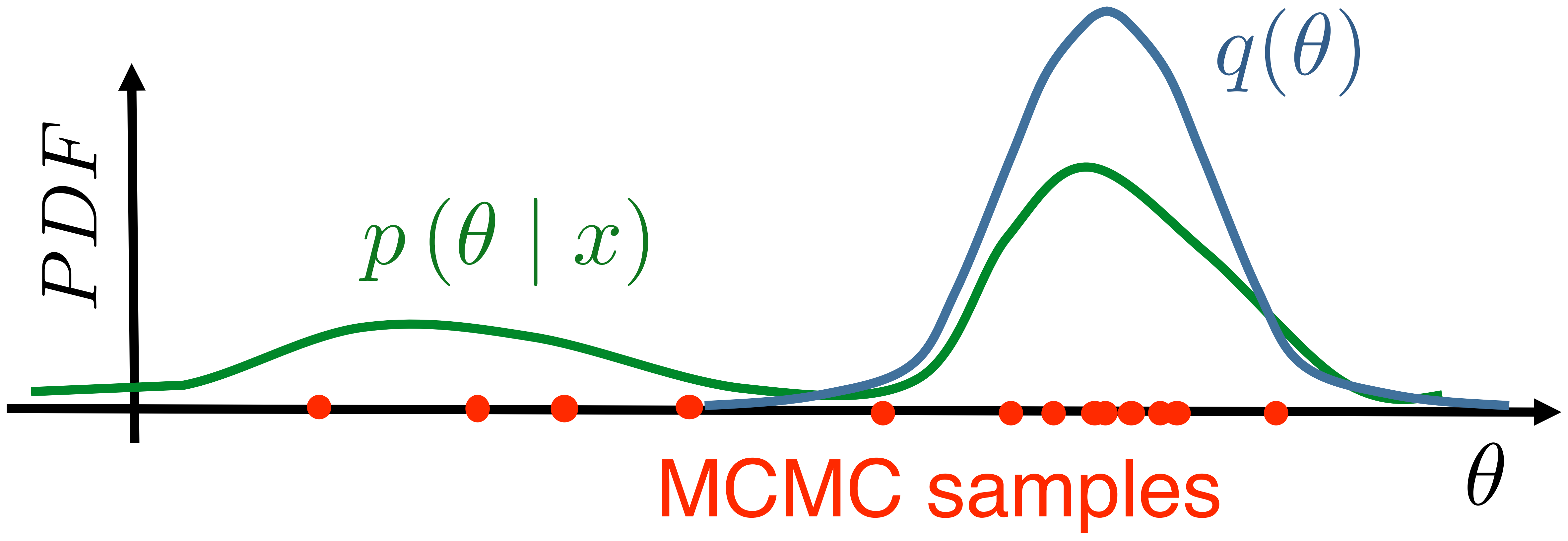
Approximate $p(\theta \mid x) \approx q(\theta) \in \mathcal{Q}$

- Biased
- Faster and more scalable

MCMC

Samples from unnormalized $p(\theta \mid x)$

- Unbiased
- Need a lot of samples



Variational inference

Probabilistic model: $p(x, \theta) = p(x \mid \theta) p(\theta)$

Main idea: find posterior approximation $p(\theta \mid x) \approx q(\theta) \in \mathcal{Q}$, using the following criterion function:

$$F(q) := KL(q(\theta) \| p(\theta \mid x)) \to \min_{q(\theta) \in \mathcal{Q}}$$

Kullback-Leibler divergence
a good mismatch measure between
two distributions over the **same domain**

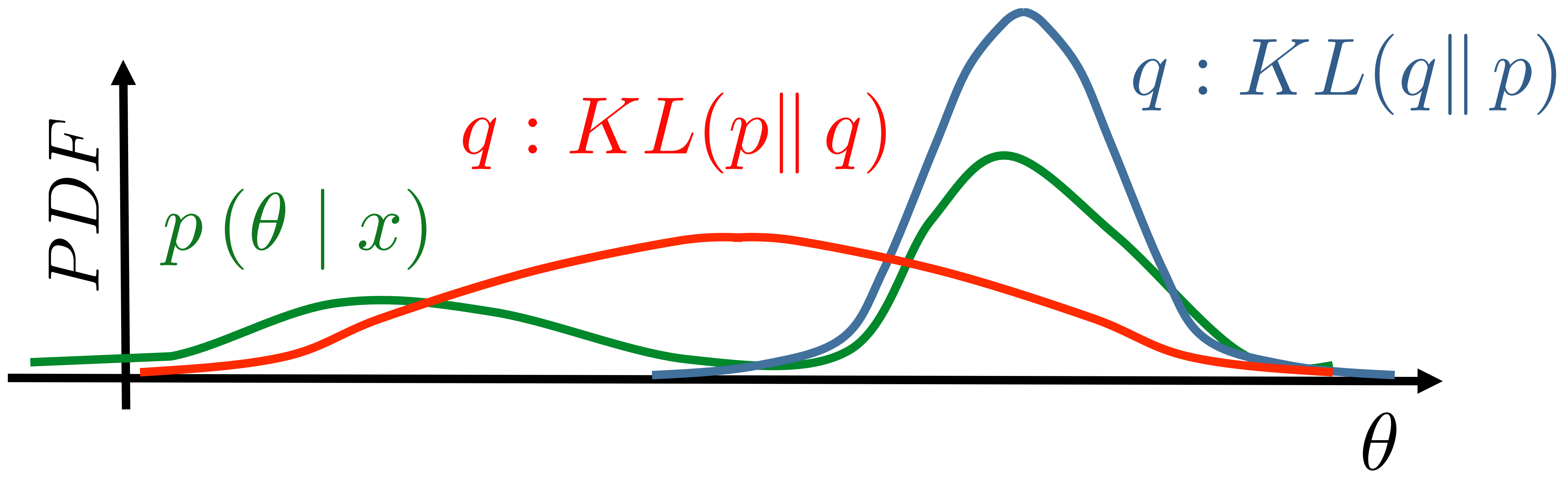
Kullback-Leibler divergence

A good mismatch measure between two distributions over the **same domain**

$$KL(q(\theta) \| p(\theta \mid x)) = \int q(\theta) \log \frac{q(\theta)}{p(\theta \mid x)} d\theta$$

Properties:

- $KL(q \| p) \geq 0$
- $KL(q \| p) = 0 \Leftrightarrow q = p$
- $KL(q \| p) \neq KL(p \| q)$



Variational inference

Probabilistic model: $p(x, \theta) = p(x \mid \theta)p(\theta)$

Main idea: find posterior approximation $p(\theta \mid x) \approx q(\theta) \in \mathcal{Q}$, using the following criterion function:

$$F(q) := KL(q(\theta) \| \, p(\theta \mid x)) \to \min_{q(\theta) \in \mathcal{Q}}$$

Variational inference

Probabilistic model: $p(x, \theta) = p(x \mid \theta)p(\theta)$

Main idea: find posterior approximation $p(\theta \mid x) \approx q(\theta) \in \mathcal{Q}$, using the following criterion function:

$$F(q) := KL(q(\theta) \| p(\theta \mid x)) \to \min_{q(\theta) \in \mathcal{Q}}$$

We could not compute the posterior in the first place

How to perform an optimization w.r.t. a distribution?

Mathematical magic

$$\log p(x)$$

Mathematical magic

$$\log p(x) = \int q(\theta) \log p(x) d\theta$$

Mathematical magic

$$\log p(x) = \int q(\theta) \log p(x) d\theta = \int q(\theta) \log \frac{p(x, \theta)}{p(\theta \mid x)} d\theta =$$

Mathematical magic

$$\log p(x) = \int q(\theta) \log p(x) d\theta = \int q(\theta) \log \frac{p(x, \theta)}{p(\theta \mid x)} d\theta =$$

$$= \int q(\theta) \log \frac{p(x, \theta) q(\theta)}{p(\theta \mid x) q(\theta)} d\theta =$$

Mathematical magic

$$
\begin{aligned}
\log p(x) &= \int q(\theta) \log p(x) d\theta = \int q(\theta) \log \frac{p(x, \theta)}{p(\theta \mid x)} d\theta = \\
&= \int q(\theta) \log \frac{p(x, \theta) q(\theta)}{p(\theta \mid x) q(\theta)} d\theta = \\
&= \int q(\theta) \log \frac{p(x, \theta)}{q(\theta)} d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta \mid x)} d\theta =
\end{aligned}
$$

Mathematical magic

$$
\begin{aligned}
\log p(x) &= \int q(\theta) \log p(x) d\theta = \int q(\theta) \log \frac{p(x,\theta)}{p(\theta \mid x)} d\theta = \\
&= \int q(\theta) \log \frac{p(x,\theta) q(\theta)}{p(\theta \mid x) q(\theta)} d\theta = \\
&= \int q(\theta) \log \frac{p(x,\theta)}{q(\theta)} d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta \mid x)} d\theta = \\
&= \mathcal{L}(q(\theta)) + KL(q(\theta) \| \, p(\theta \mid x))
\end{aligned}
$$

Mathematical magic

$$\log p(x) = \int q(\theta) \log p(x) d\theta = \int q(\theta) \log \frac{p(x,\theta)}{p(\theta \mid x)} d\theta =$$

$$= \int q(\theta) \log \frac{p(x,\theta) q(\theta)}{p(\theta \mid x) q(\theta)} d\theta =$$

$$= \int q(\theta) \log \frac{p(x,\theta)}{q(\theta)} d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta \mid x)} d\theta =$$

$$= \boxed{\mathcal{L}(q(\theta))} + \boxed{KL(q(\theta) \| p(\theta \mid x))}$$

Evidence lower bound (ELBO)

KL-divergence we need for VI

ELBO = Evidence Lower Bound

$$\log p(x) = \mathcal{L}(q(\theta)) + KL(q(\theta) \| p(\theta \mid x))$$

Evidence:

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{p(x)} = \frac{p(x \mid \theta)p(\theta)}{\int p(x \mid \theta)p(\theta)d\theta} = \frac{\text{Likelihood} \;\times\; \text{Prior}}{\text{Evidence}}$$

Evidence of the probabilistic model shows the total probability of observing the data.

Lower Bound: KL is non-negative $\longrightarrow$ $\log p(x) \geq \mathcal{L}(q(\theta))$

Variational inference

Optimization problem with intractable posterior distribution:

$$F(q) := KL(q(\theta) \,\|\, p(\theta \mid x)) \to \min_{q(\theta) \in \mathcal{Q}}$$

Variational inference

Optimization problem with intractable posterior distribution:

$$F(q) := KL(q(\theta) \| p(\theta \mid x)) \to \min_{q(\theta) \in \mathcal{Q}}$$

Let's use our magic:

$$\log p(x) = \mathcal{L}(q(\theta)) + KL(q(\theta) \| p(\theta \mid x))$$

Variational inference

Optimization problem with intractable posterior distribution:

$$F(q) := KL(q(\theta) \| p(\theta \mid x)) \to \min_{q(\theta) \in \mathcal{Q}}$$

Let's use our magic:

$$\log p(x) = \mathcal{L}(q(\theta)) + KL(q(\theta) \| p(\theta \mid x))$$

does not depend on q

depend on q

Variational inference

Optimization problem with intractable posterior distribution:

$$F(q) := KL(q(\theta) \| p(\theta \mid x)) \to \min_{q(\theta) \in \mathcal{Q}}$$

Let's use our magic:

$$\log p(x) = \mathcal{L}(q(\theta)) + KL(q(\theta) \| p(\theta \mid x))$$

does not depend on q

depend on q

$$KL(q(\theta) \| p(\theta \mid x)) \to \min_{q(\theta) \in \mathcal{Q}} \quad \Leftrightarrow \quad \mathcal{L}(q(\theta)) \to \max_{q(\theta) \in \mathcal{Q}}$$

Variational inference

Final optimisation problem:

$$\mathcal{L}(q(\theta)) = \int q(\theta) \log \frac{p(x, \theta)}{q(\theta)} d\theta \to \max_{q(\theta) \in \mathcal{Q}}$$

Variational inference: ELBO interpretation

Final optimisation problem:

$$\mathcal{L}(q(\theta)) = \int q(\theta) \log \frac{p(x, \theta)}{q(\theta)} d\theta = \int q(\theta) \log \frac{p(x \mid \theta) p(\theta)}{q(\theta)} d\theta =$$

Variational inference: ELBO interpretation

Final optimisation problem:

$$\mathcal{L}(q(\theta)) = \int q(\theta) \log \frac{p(x, \theta)}{q(\theta)} d\theta = \int q(\theta) \log \frac{p(x \mid \theta) p(\theta)}{q(\theta)} d\theta =$$

$$= \int q(\theta) \log p(x \mid \theta) d\theta + \int q(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta =$$

Variational inference: ELBO interpretation

Final optimisation problem:

$$\mathcal{L}(q(\theta)) = \int q(\theta) \log \frac{p(x,\theta)}{q(\theta)} d\theta = \int q(\theta) \log \frac{p(x \mid \theta) p(\theta)}{q(\theta)} d\theta =$$

$$= \int q(\theta) \log p(x \mid \theta) d\theta + \int q(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta =$$

$$= \mathbb{E}_{q(\theta)} \log p(x \mid \theta) - KL(q(\theta) \| p(\theta))$$

Variational inference: ELBO interpretation

Final optimisation problem:

$$\mathcal{L}(q(\theta)) = \int q(\theta) \log \frac{p(x,\theta)}{q(\theta)} d\theta = \int q(\theta) \log \frac{p(x \mid \theta) p(\theta)}{q(\theta)} d\theta =$$

$$= \int q(\theta) \log p(x \mid \theta) d\theta + \int q(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta =$$

$$= \underbrace{\boxed{\mathbb{E}_{q(\theta)} \log p(x \mid \theta)}}_{\text{data term}} - \underbrace{\boxed{KL(q(\theta) \| p(\theta))}}_{\text{regularizer}}$$

Variational inference

Final optimisation problem:

$$\mathcal{L}(q(\theta)) = \int q(\theta) \log \frac{p(x,\theta)}{q(\theta)} d\theta \to \max_{q(\theta) \in \mathcal{Q}}$$

How to perform an optimization w.r.t. a distribution?

Variational inference

Final optimisation problem:

$$\mathcal{L}(q(\theta)) = \int q(\theta) \log \frac{p(x, \theta)}{q(\theta)} d\theta \to \max_{q(\theta) \in \mathcal{Q}}$$

How to perform an optimization w.r.t. a distribution?

Mean field approximation

Factorized family

$$q(\theta) = \prod_{j=1}^{m} q_j(\theta_j), \quad \theta = [\theta_1, \ldots, \theta_m]$$

Parametric approximation

Parametric family

$$q(\theta) = q(\theta \mid \lambda)$$

Mean Field Approximation

Factorized family of variational distributions:

$$q(\theta) = \prod_{j=1}^{m} q_j\left(\theta_j\right), \quad \theta = [\theta_1, \ldots, \theta_m]$$

Why is it a restriction?

Mean Field Approximation

Factorized family of variational distributions:

$$q(\theta) = \prod_{j=1}^{m} q_j\left(\theta_j\right), \quad \theta = [\theta_1, \ldots, \theta_m]$$

Why is it a restriction? From product rule:

$$q(\theta) = \prod_{j=1}^{m} q_j\left(\theta_j \mid \theta_{<j}\right)$$

We assume that $\theta_1, \ldots, \theta_m$ are independent $\longrightarrow$ simpler approximation

Mean Field Approximation

Optimization problem:

$$\mathcal{L}(q(\theta)) = \int q(\theta) \log \frac{p(x,\theta)}{q(\theta)} d\theta \to \max_{q(\theta)=q_1(\theta_1)\cdot\ldots\cdot q_m(\theta_m)}$$

Block coordinate assent:

At each step fix all factors $\{q_i(\theta_i)\}_{i\neq j}$ except one and optimise w.r.t. to it:

$$\mathcal{L}(q(\theta)) \to \max_{q_j(\theta_j)}$$

Mean Field Approximation

$$\mathcal{L}(q(\theta)) = \mathbb{E}_{q(\theta)} \log p(x, \theta) - \mathbb{E}_{q(\theta)} \log q(\theta) =$$

Mean Field Approximation

$$\mathcal{L}(q(\theta)) = \mathbb{E}_{q(\theta)} \log p(x, \theta) - \mathbb{E}_{q(\theta)} \log q(\theta) =$$

$$= \mathbb{E}_{q(\theta)} \log p(x, \theta) - \sum_{k=1}^{m} \mathbb{E}_{q_k(\theta_k)} \log q_k(\theta_k) =$$

Mean Field Approximation

$$\begin{aligned}\mathcal{L}(q(\theta)) &= \mathbb{E}_{q(\theta)} \log p(x,\theta) - \mathbb{E}_{q(\theta)}\log q(\theta) = \\ &= \mathbb{E}_{q(\theta)} \log p(x,\theta) - \sum_{k=1}^{m} \mathbb{E}_{q_k(\theta_k)}\log q_k(\theta_k) = \\ &= \mathbb{E}_{q_j(\theta_j)} \left[\mathbb{E}_{q_{i\neq j}} \log p(x,\theta)\right] - \mathbb{E}_{q_j(\theta_j)}\log q_j(\theta_j) + Const =\end{aligned}$$

Mean Field Approximation

$$
\begin{aligned}
\mathcal{L}(q(\theta)) &= \mathbb{E}_{q(\theta)} \log p(x,\theta) - \mathbb{E}_{q(\theta)} \log q(\theta) = \\
&= \mathbb{E}_{q(\theta)} \log p(x,\theta) - \sum_{k=1}^{m} \mathbb{E}_{q_k(\theta_k)} \log q_k(\theta_k) = \\
&= \mathbb{E}_{q_j(\theta_j)} \left[\mathbb{E}_{q_{i\neq j}} \log p(x,\theta)\right] - \mathbb{E}_{q_j(\theta_j)} \log q_j(\theta_j) + Const = \\
&= \left\{ r_j(\theta_j) = \frac{1}{Z_j} \exp\left(\mathbb{E}_{q_{i\neq j}} \log p(x,\theta)\right) \right\} =
\end{aligned}
$$

Mean Field Approximation

$$\begin{aligned}
\mathcal{L}(q(\theta)) &= \mathbb{E}_{q(\theta)} \log p(x,\theta) - \mathbb{E}_{q(\theta)} \log q(\theta) = \\
&= \mathbb{E}_{q(\theta)} \log p(x,\theta) - \sum_{k=1}^{m} \mathbb{E}_{q_k(\theta_k)} \log q_k(\theta_k) = \\
&= \mathbb{E}_{q_j(\theta_j)} \left[\mathbb{E}_{q_{i\neq j}} \log p(x,\theta) \right] - \mathbb{E}_{q_j(\theta_j)} \log q_j(\theta_j) + Const = \\
&= \left\{ r_j(\theta_j) = \frac{1}{Z_j} \exp\left(\mathbb{E}_{q_{i\neq j}} \log p(x,\theta) \right) \right\} = \\
&= \mathbb{E}_{q_j(\theta_j)} \log \frac{r_j(\theta_j)}{q_j(\theta_j)} + Const = -KL\left(q_j(\theta_j) \| r_j(\theta_j) \right) + Const
\end{aligned}$$

Mean Field Approximation

Optimization problem at each step of the block coordinate assent:

$$\mathcal{L}(q(\theta)) = -KL\left(q_j(\theta_j) \| \, r_j(\theta_j)\right) + Const \to \max_{q_j(\theta_j)}$$

Mean Field Approximation

Optimization problem at each step of the block coordinate assent:

$$\mathcal{L}(q(\theta)) = -KL\left(q_j(\theta_j) \| \, r_j(\theta_j)\right) + Const \to \max_{q_j(\theta_j)}$$

Solution:

$$q_j\left(\theta_j\right) = r_j(\theta_j) = \frac{1}{Z_j} \exp\left(\mathbb{E}_{q_{i \neq j}} \log p(x, \theta)\right)$$

Mean Field Variational Inference

Algorithm:

Initialize $q(\theta) = \prod_{j=1}^{m} q_j(\theta_j)$

Iterations:

- Update each factor $q_1, \ldots, q_m$:

$$q_j(\theta_j) = \frac{1}{Z_j} \exp\left(\mathbb{E}_{q_{i \neq j}} \log p(x, \theta)\right)$$

- Compute ELBO $\mathcal{L}(q(\theta))$

Repeat until convergence of ELBO

Mean Field Variational Inference

Algorithm:

Initialize $q(\theta) = \prod_{j=1}^{m} q_j(\theta_j)$

Iterations:

- Update each factor $q_1, \ldots, q_m$:

$$q_j(\theta_j) = \frac{1}{Z_j} \exp\left(\mathbb{E}_{q_{i \neq j}} \log p(x, \theta)\right)$$

- Compute ELBO $\mathcal{L}(q(\theta))$

Repeat until convergence of ELBO

Assumption:
we can compute the update analytically

Mean Field Variational Inference

Probabilistic model: $p(x, \theta) = p(x \mid \theta)p(\theta), \quad \theta = [\theta_1, \ldots, \theta_m]$

When applicable?

Conditional conjugacy of likelihood and prior on each θ_j conditioned on all other $\{\theta_i\}_{i \neq j}$:

$$p(\theta_j \mid \theta_{i \neq j}) \in \mathcal{A}(\alpha), \quad p(x \mid \theta_j, \theta_{i \neq j}) \in \mathcal{B}(\theta_j) \longrightarrow p(\theta_j \mid x, \theta_{i \neq j}) \in \mathcal{A}(\alpha')$$

Mean Field Variational Inference

Probabilistic model: $p(x, \theta) = p(x \mid \theta)p(\theta), \quad \theta = [\theta_1, \ldots, \theta_m]$

When applicable?

Conditional conjugacy of likelihood and prior on each θ_j conditional on all other $\{\theta_i\}_{i \neq j}$:

$$p(\theta_j \mid \theta_{i \neq j}) \in \mathcal{A}(\alpha), \quad p(x \mid \theta_j, \theta_{i \neq j}) \in \mathcal{B}(\theta_j) \longrightarrow p(\theta_j \mid x, \theta_{i \neq j}) \in \mathcal{A}(\alpha')$$

How to check in practice?

For each θ_j:
- Fix all other $\{\theta_i\}_{i \neq j}$ (look at them as some constants)
- Check whether $p(x \mid \theta)$ and $p(\theta)$ are conjugate w.r.t. θ_j

Mean Field Variational Inference

In practice:

$$q_j\left(\theta_j\right) = \frac{1}{Z_j}\exp\left(\mathbb{E}_{q_{i\neq j}}\log p(x,\theta)\right)$$

$$\Downarrow$$

$$\log q_j\left(\theta_j\right) = \mathbb{E}_{q_{i\neq j}}\log p(x,\theta) + Const$$

Parametric approximation

Parametric family of variational distributions:

$$q(\theta) = q(\theta \mid \lambda), \qquad \lambda \text{ — some parameters}$$

Why is it a restriction? We choose a family of some fixed form:

- It may be too simple and insufficient to model the data
- If it is complex enough then there is no guarantee we can train it well to fit the data

Parametric approximation

Parametric family of variational distributions:

$$q(\theta) = q(\theta \mid \lambda), \qquad \lambda \text{ — some parameters}$$

Variational inference transforms to parametric optimization problem:

$$\mathcal{L}(q(\theta \mid \lambda)) = \int q(\theta \mid \lambda) \log \frac{p(x, \theta)}{q(\theta \mid \lambda)} d\theta \to \max_{\lambda}$$

If we're able to calculate derivatives of ELBO w.r.t. θ then we can solve this problem using some numerical optimization solver.

Inference methods: summary

Full Bayesian inference: $p(\theta \mid x)$

MP inference: $p(\theta \mid x) \approx \delta(\theta - \theta_{MP})$

Mean field variational inference: $p(\theta \mid x) \approx q(\theta) = \prod_{j=1}^{m} q_j\left(\theta_j\right)$

Parametric variational inference: $p(\theta \mid x) \approx q(\theta) = q(\theta \mid \lambda)$