# Introduction to Variational Inference

*Dmitry Kropotov*
*Lomonosov Moscow State University*

DeepBayes

# Contents

# The problem: finding posterior distribution

In Bayesian Inference we are interested in finding posterior distributions.

**Bayesian Ensembling:**

Probabilistic model: $p(\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, $\boldsymbol{x} = [\boldsymbol{x}_{train}, \boldsymbol{x}_{test}]$.

Prediction: $p(\boldsymbol{x}_{test}|\boldsymbol{x}_{train}) = \int p(\boldsymbol{x}_{test}|\boldsymbol{\theta}){\color{red}p(\boldsymbol{\theta}|\boldsymbol{x}_{train})}d\boldsymbol{\theta}$.

**EM-algorithm:**

Latent variable model: $p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})$.

Training: $\log p(\boldsymbol{x}|\boldsymbol{\theta}) \geq \underbrace{\mathbb{E}_{q(\boldsymbol{z})} \log p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) - \mathbb{E}_{q(\boldsymbol{z})} \log q(\boldsymbol{z})}_{ELBO(q, \boldsymbol{\theta})} \to \max_{q(\boldsymbol{z}), \boldsymbol{\theta}}$.

E-step: $q(\boldsymbol{z}) = {\color{red}p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})}$.

# The problem: finding posterior distribution

In Bayesian Inference we are interested in finding posterior distributions.

**Bayesian Ensembling:**

Probabilistic model: $p(\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, $\boldsymbol{x} = [\boldsymbol{x}_{train}, \boldsymbol{x}_{test}]$.

Prediction: $p(\boldsymbol{x}_{test}|\boldsymbol{x}_{train}) = \int p(\boldsymbol{x}_{test}|\boldsymbol{\theta}){\color{red}p(\boldsymbol{\theta}|\boldsymbol{x}_{train})}d\boldsymbol{\theta}$.

**EM-algorithm:**

Latent variable model: $p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})$.

Training: $\log p(\boldsymbol{x}|\boldsymbol{\theta}) \geq \underbrace{\mathbb{E}_{q(\boldsymbol{z})} \log p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) - \mathbb{E}_{q(\boldsymbol{z})} \log q(\boldsymbol{z})}_{ELBO(q, \boldsymbol{\theta})} \rightarrow \max_{q(\boldsymbol{z}), \boldsymbol{\theta}}$.

E-step: $q(\boldsymbol{z}) = {\color{red}p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})}$.

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

# Variational Inference

Probabilistic model: $p(\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$.

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

$$F(q) := \mathrm{KL}(q(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta}|\boldsymbol{x})) \to \min_{q(\boldsymbol{\theta}) \in \mathcal{Q}}. \tag{1}$$

# Variational Inference

Probabilistic model: $p(\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$.

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

$$F(q) := \mathrm{KL}(q(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta}|\boldsymbol{x})) \to \min_{q(\boldsymbol{\theta}) \in \mathcal{Q}}. \tag{1}$$

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

$$= \int q(\boldsymbol{\theta}) \log q(\boldsymbol{\theta}) d\boldsymbol{\theta} + \underbrace{\int q(\boldsymbol{\theta}) \log p(\boldsymbol{x}) d\boldsymbol{\theta}}_{\log p(\boldsymbol{x})} - \int q(\boldsymbol{\theta}) \log p(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{\theta} =$$

$$= \mathbb{E}_{q(\boldsymbol{\theta})}[\log q(\boldsymbol{\theta}) - \log p(\boldsymbol{x}, \boldsymbol{\theta})] + \mathrm{const} \to \min_{q \in \mathcal{Q}} \tag{2}$$

# Variational Inference

Probabilistic model: $p(\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$.

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

$$F(q) := \mathrm{KL}(q(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta}|\boldsymbol{x})) \to \min_{q(\boldsymbol{\theta}) \in \mathcal{Q}} . \tag{1}$$

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

$$= \int q(\boldsymbol{\theta}) \log q(\boldsymbol{\theta}) d\boldsymbol{\theta} + \underbrace{\int q(\boldsymbol{\theta}) \log p(\boldsymbol{x}) d\boldsymbol{\theta}}_{\log p(\boldsymbol{x})} - \int q(\boldsymbol{\theta}) \log p(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{\theta} =$$

$$= \mathbb{E}_{q(\boldsymbol{\theta})}[\log q(\boldsymbol{\theta}) - \log p(\boldsymbol{x}, \boldsymbol{\theta})] + \mathrm{const} \to \min_{q \in \mathcal{Q}} \tag{2}$$

The problem (2) is equivalent to maximizing ELBO:

$$\log p(\boldsymbol{x}) \geq \mathrm{ELBO(q)} = \mathbb{E}_{q(\boldsymbol{\theta})} \log p(\boldsymbol{x}, \boldsymbol{\theta}) - \mathbb{E}_{q(\boldsymbol{\theta})} \log q(\boldsymbol{\theta}) \to \max_{q \in \mathcal{Q}} .$$

# Variational EM-algorithm

Latent variable model: $p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})$.

**Conventional EM-algorithm:**

$$\log p(\boldsymbol{x}|\boldsymbol{\theta}) \geq \underbrace{\mathbb{E}_{q(\boldsymbol{z})} \log p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) - \mathbb{E}_{q(\boldsymbol{z})} \log q(\boldsymbol{z})}_{ELBO(q,\boldsymbol{\theta})} \to \max_{q(\boldsymbol{z}),\boldsymbol{\theta}},$$

$$q(\boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}).$$

**Variational EM-algorithm:**

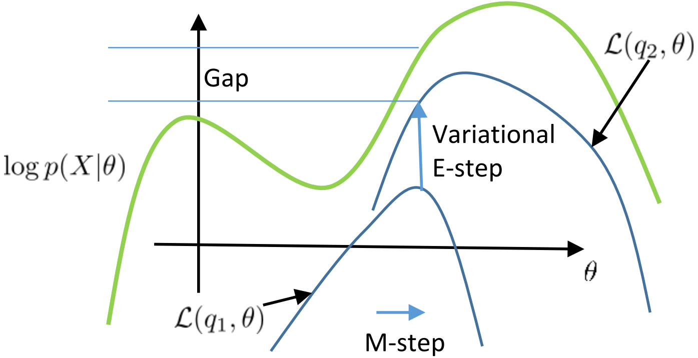
$$\log p(\boldsymbol{x}|\boldsymbol{\theta}) \geq \underbrace{\mathbb{E}_{q(\boldsymbol{z})} \log p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) - \mathbb{E}_{q(\boldsymbol{z})} \log q(\boldsymbol{z})}_{ELBO(q,\boldsymbol{\theta})} \to \max_{q(\boldsymbol{z}) \in \mathcal{Q},\boldsymbol{\theta}},$$

$q(\boldsymbol{z}) \in \mathcal{Q}$ – variational approximation to $p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})$.

# Variational EM-algorithm

$$\log p(\boldsymbol{x}|\boldsymbol{\theta}) \geq \mathcal{L}(q, \boldsymbol{\theta}) = \mathbb{E}_{q(\boldsymbol{z})} \log p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) - \mathbb{E}_{q(\boldsymbol{z})} \log q(\boldsymbol{z}) \to \max_{q \in \mathcal{Q}, \boldsymbol{\theta}}.$$

Even in the case of inexact E-step in many situations we are able to find good model parameters.



# ELBO interpretation

Probabilistic model: $p(\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$.

Finding variational approximation $p(\boldsymbol{\theta}|\boldsymbol{x}) \approx q(\boldsymbol{\theta}) \in \mathcal{Q}$:

$$\mathrm{ELBO(q)} = \mathbb{E}_{q(\boldsymbol{\theta})} \log p(\boldsymbol{x}, \boldsymbol{\theta}) - \mathbb{E}_{q(\boldsymbol{\theta})} \log q(\boldsymbol{\theta}) = \mathbb{E}_{q(\boldsymbol{\theta})}[\log p(\boldsymbol{x}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log q(\boldsymbol{\theta})] =$$

$$= \mathbb{E}_{q(\boldsymbol{\theta})} \log p(\boldsymbol{x}|\boldsymbol{\theta}) + \mathbb{E}_{q(\boldsymbol{\theta})} \log \frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} = \underbrace{\mathbb{E}_{q(\boldsymbol{\theta})} \log p(\boldsymbol{x}|\boldsymbol{\theta})}_{\text{data term}} - \underbrace{KL(q(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta}))}_{\text{KL term}} \to \max_{q \in \mathcal{Q}}.$$

Maximum likelihood approach:

$$\mathbb{E}_{q(\boldsymbol{\theta})} \log p(\boldsymbol{x}|\boldsymbol{\theta}) \to \max_{q} \;\Leftrightarrow\; q(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{ML}),\; \boldsymbol{\theta}_{ML} = \arg\max_{\theta} \log p(\boldsymbol{x}|\boldsymbol{\theta})$$

# Mean Field Approximation

Factorized family of variational distributions:

$$q(\boldsymbol{\theta}) = \prod_{j=1}^{m} q_j(\boldsymbol{\theta}_j), \quad \boldsymbol{\theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_m]. \tag{3}$$

Let's fix all factors $\{q_j(\boldsymbol{\theta}_j)\}_{j \neq i}$ except one in the factorized family (3) and find best variational approximation for $q_i(\boldsymbol{\theta}_i)$:

$$\begin{aligned}
\mathrm{ELBO}(q) &= \mathbb{E}_{q(\boldsymbol{\theta})}[\log p(\boldsymbol{x}, \boldsymbol{\theta}) - \log q(\boldsymbol{\theta})] = \\
&= \int \prod_{j=1}^{m} q_j(\boldsymbol{\theta}_j) \log p(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{\theta} - \int \prod_{j=1}^{m} q_j(\boldsymbol{\theta}_j) \left( \sum_{k=1}^{m} \log q_k(\boldsymbol{\theta}_k) \right) d\boldsymbol{\theta} = \\
&= \int \prod_{j=1}^{m} q_j(\boldsymbol{\theta}_j) \log p(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{\theta} - \sum_{k=1}^{m} \int \prod_{j=1}^{m} q_j(\boldsymbol{\theta}_j) \log q_k(\boldsymbol{\theta}_k) d\boldsymbol{\theta} = \\
&= \int \prod_{j=1}^{m} q_j(\boldsymbol{\theta}_j) \log p(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{\theta} - \sum_{k=1}^{m} \int q_k(\boldsymbol{\theta}_k) \log q_k(\boldsymbol{\theta}_k) d\boldsymbol{\theta}_k = \dots
\end{aligned}$$

# Mean Field Approximation

$$\dots = \int q_i(\boldsymbol{\theta}_i) \underbrace{\left[ \int \log p(\boldsymbol{x}, \boldsymbol{\theta}) \prod_{j \neq i} q_j(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j \right]}_{r_i(\boldsymbol{\theta}_i, \boldsymbol{x})} d\boldsymbol{\theta}_i - \int q_i(\boldsymbol{\theta}_i) \log q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i + \mathrm{const} =$$

$$= \int q_i(\boldsymbol{\theta}_i) \log \frac{\exp(r_i(\boldsymbol{\theta}_i, \boldsymbol{x}))}{q_i(\boldsymbol{\theta}_i)} d\boldsymbol{\theta}_i + \mathrm{const} = [Z_i \text{ – normalizing constant for distribution } r_i] =$$

$$= \int q_i(\boldsymbol{\theta}_i) \log \underbrace{\frac{\exp(r_i(\boldsymbol{\theta}_i, \boldsymbol{x}))}{Z_i}}_{\hat{r}_i(\boldsymbol{\theta}_i, \boldsymbol{x})} \cdot \frac{Z_i}{q_i(\boldsymbol{\theta}_i)} d\boldsymbol{\theta}_i + \mathrm{const} =$$

$$= \int q_i(\boldsymbol{\theta}_i) \log \frac{\hat{r}_i(\boldsymbol{\theta}_i, \boldsymbol{x})}{q_i(\boldsymbol{\theta}_i)} d\boldsymbol{\theta}_i + \log Z_i + \mathrm{const} = \mathrm{const} - KL(q_i(\boldsymbol{\theta}_i) \,||\, \hat{r}_i(\boldsymbol{\theta}_i, \boldsymbol{x})) \to \max_{q_i(\boldsymbol{\theta}_i)} .$$

Solution: $q_i(\boldsymbol{\theta}_i) = \hat{r}_i(\boldsymbol{\theta}_i, \boldsymbol{x}) = \frac{1}{Z_i} \exp(\int \log p(\boldsymbol{x}, \boldsymbol{\theta}) \prod_{j \neq i} q_j(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j)$

# Mean Field Approximation

**General scheme for Mean Field Variational Inference:**

1. Initialize $q(\boldsymbol{\theta}) = \prod_{j=1}^{m} q_j(\boldsymbol{\theta}_j)$;
2. For each factor $q_i(\boldsymbol{\theta}_i),\ i = 1, \ldots, m$ do:
   - Calculate $q_i(\boldsymbol{\theta}_i) = \frac{1}{Z_i} \exp(\int \log p(\boldsymbol{x}, \boldsymbol{\theta}) \prod_{j \neq i} q_j(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j)$, where $Z_i = \int \exp(\int \log p(\boldsymbol{x}, \boldsymbol{\theta}) \prod_{j \neq i} q_j(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j) d\boldsymbol{\theta}_i$;
3. Calculate $\mathrm{ELBO}(q) = \int \log p(\boldsymbol{x}, \boldsymbol{\theta}) \prod_{j=1}^{m} q_j(\boldsymbol{\theta}_j) d\boldsymbol{\theta} - \sum_{j=1}^{m} \int q_j(\boldsymbol{\theta}_j) \log q_j(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j$;
4. Repeat until convergence of $\mathrm{ELBO}(q)$.

# Mean Field Approximation

**General scheme for Mean Field Variational Inference:**

1. Initialize $q(\boldsymbol{\theta}) = \prod_{j=1}^{m} q_j(\boldsymbol{\theta}_j)$;
2. For each factor $q_i(\boldsymbol{\theta}_i)$, $i = 1, \dots, m$ do:
   - Calculate $q_i(\boldsymbol{\theta}_i) = \frac{1}{Z_i} \exp(\int \log p(\boldsymbol{x}, \boldsymbol{\theta}) \prod_{j \neq i} q_j(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j)$, where $Z_i = \int \exp(\int \log p(\boldsymbol{x}, \boldsymbol{\theta}) \prod_{j \neq i} q_j(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j) d\boldsymbol{\theta}_i$;
3. Calculate $\mathrm{ELBO}(q) = \int \log p(\boldsymbol{x}, \boldsymbol{\theta}) \prod_{j=1}^{m} q_j(\boldsymbol{\theta}_j) d\boldsymbol{\theta} - \sum_{j=1}^{m} \int q_j(\boldsymbol{\theta}_j) \log q_j(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j$;
4. Repeat until convergence of $\mathrm{ELBO}(q)$.

Main assumption: we are able to calculate $\int \log p(\boldsymbol{x}, \boldsymbol{\theta}) \prod_{j \neq i} q_j(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j$ and $Z_i$ analytically.

# Mean Field Approximation

**General scheme for Mean Field Variational Inference:**

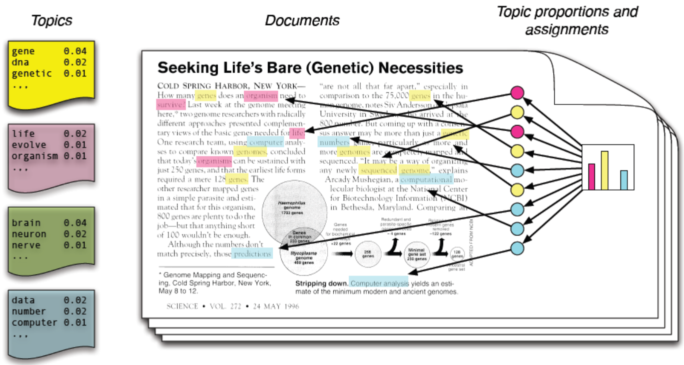
1. Initialize $q(\boldsymbol{\theta}) = \prod_{j=1}^{m} q_j(\boldsymbol{\theta}_j)$;
2. For each factor $q_i(\boldsymbol{\theta}_i)$, $i = 1, \dots, m$ do:
   - Calculate $q_i(\boldsymbol{\theta}_i) = \frac{1}{Z_i} \exp(\int \log p(\boldsymbol{x}, \boldsymbol{\theta}) \prod_{j \neq i} q_j(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j)$, where $Z_i = \int \exp(\int \log p(\boldsymbol{x}, \boldsymbol{\theta}) \prod_{j \neq i} q_j(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j) d\boldsymbol{\theta}_i$;
3. Calculate $\mathrm{ELBO}(q) = \int \log p(\boldsymbol{x}, \boldsymbol{\theta}) \prod_{j=1}^{m} q_j(\boldsymbol{\theta}_j) d\boldsymbol{\theta} - \sum_{j=1}^{m} \int q_j(\boldsymbol{\theta}_j) \log q_j(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j$;
4. Repeat until convergence of $\mathrm{ELBO}(q)$.

Main assumption: we are able to calculate $\int \log p(\boldsymbol{x}, \boldsymbol{\theta}) \prod_{j \neq i} q_j(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j$ and $Z_i$ analytically. For many probabilistic models this calculation is much simpler than $p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{\theta}$.

# Example: Latent Dirichlet Allocation (LDA) [Blei, Ng & Jordan, 2003]

LDA is a tool for topic modelling. LDA assumptions:

- Each document is a collection of words regardless their order;
- Each topic is a probability distribution over set of words;
- Each document is a discrete probability mixture of several topics.



# Example: Latent Dirichlet Allocation

Generative process for one document with $n$ words:

1. Choose topic probabilities $\boldsymbol{\theta}$ from prior distribution;
2. For each position $1, \ldots, n$ in the document:
   - Choose topic for current position $z_i$ using topic probabilities $\boldsymbol{\theta}$;
   - Choose current word $w_i$ using word probabilities $\Phi$ for current topic $z_i$.

# Example: Latent Dirichlet Allocation

Probabilistic model for one document:

$$p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta} | \Phi, \alpha) = p(\boldsymbol{\theta}|\alpha) \prod_{i=1}^{n} p(z_i|\boldsymbol{\theta}) p(w_i|z_i, \Phi),$$

$$p(\boldsymbol{\theta}|\alpha) = \mathrm{Dir}(\boldsymbol{\theta}|\alpha) \propto \prod_{t=1}^{T} \theta_t^{\alpha-1},$$

$$p(z_i|\boldsymbol{\theta}) = \prod_{t=1}^{T} \theta_t^{[z_i=t]},$$

$$p(w_i|z_i, \Phi) = \prod_{t=1}^{T} \prod_{w=1}^{W} \Phi_{t,w}^{[w_i=w][z_i=t]}.$$

# Variational EM-algorithm for LDA

LDA training: $\log p(\boldsymbol{w}|\Phi,\alpha) \geq \mathrm{ELBO}(q,\Phi,\alpha) = \mathbb{E}_{q(\boldsymbol{z},\boldsymbol{\theta})} \log \frac{p(\boldsymbol{w},\boldsymbol{z},\boldsymbol{\theta}|\Phi,\alpha)}{q(\boldsymbol{z},\boldsymbol{\theta})} \to \max_{q(\boldsymbol{z},\boldsymbol{\theta}),\Phi,\alpha}.$

Mean field approximation: $q(\boldsymbol{z},\boldsymbol{\theta}) = q(\boldsymbol{z})q(\boldsymbol{\theta})$.

Calculation of $q(\boldsymbol{\theta})$:

$$\log q(\boldsymbol{\theta}) = \int \log p(\boldsymbol{w},\boldsymbol{z},\boldsymbol{\theta}|\Phi,\alpha)q(\boldsymbol{z})d\boldsymbol{z} + \mathrm{const} =$$

$$= \log \mathrm{Dir}(\boldsymbol{\theta}|\alpha) + \sum_{i=1}^{n}\sum_{z_i} \log p(z_i|\boldsymbol{\theta})q(z_i) + \mathrm{const} =$$

$$= \sum_{t=1}^{T}(\alpha-1)\log\theta_t + \sum_{i=1}^{n}\sum_{z_i}[z_i = t]\log\theta_t q(z_i) + \mathrm{const} =$$

$$= \sum_{t=1}^{T}\log\theta_t \underbrace{\left[\alpha - 1 + \sum_{i=1}^{n} q(z_i = t)\right]}_{\gamma_t - 1} + \mathrm{const} = \log \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\gamma}).$$

# Variational Parametric Approximation

Alternative to mean field variational approximation is to use parametric variational approximation: $q(\boldsymbol{\theta}) = q(\boldsymbol{\theta}|\boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ are some parameters.

In this case variational inference transforms to parametric optimization problem:

$$\mathrm{ELBO}(q, \boldsymbol{\theta}) = \mathrm{ELBO}(\boldsymbol{\lambda}, \boldsymbol{\theta}) = \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\lambda})} \log \frac{p(\mathbf{x}, \boldsymbol{\theta})}{q(\boldsymbol{\theta}|\boldsymbol{\lambda})} \to \max_{\boldsymbol{\lambda}, \boldsymbol{\theta}}. \tag{4}$$

If we're able to calculate derivatives of $\mathrm{ELBO}$ w.r.t. $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$, we can solve problem (4) using some numerical optimization solver.

# Example: Bayesian Logistic Regression

Consider a 2-class classification problem. We have a dataset $(\boldsymbol{y}, X) = \{y_i, \boldsymbol{x}_i\}_{i=1}^n$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ – feature vectors and $y_i \in \{-1, +1\}$ – class labels and want to train a linear classifier:

$$\hat{y}(\boldsymbol{x}) = \mathrm{sign}\Big(\sum_{j=1}^{d} w_j x_j\Big) = \mathrm{sign}(\boldsymbol{w}^T \boldsymbol{x}).$$

Probabilistic model:

$$p(\boldsymbol{y}, \boldsymbol{w}|X, \alpha) = p(\boldsymbol{y}|\boldsymbol{w}, X)p(\boldsymbol{w}|\alpha) = p(\boldsymbol{w}|\alpha)\prod_{i=1}^{n} p(y_i|\boldsymbol{w}, \boldsymbol{x}_i),$$

$$p(y_i|\boldsymbol{w}, \boldsymbol{x}_i) = \sigma(y_i \boldsymbol{w}^T \boldsymbol{x}_i) = \frac{1}{1 + \exp(-y_i \boldsymbol{w}^T \boldsymbol{x}_i)},$$

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

Use variational EM-algorithm for training:

$$\log p(\boldsymbol{y}|X, \alpha) \geq \mathrm{ELBO}(q(\boldsymbol{w}), \alpha) = \mathbb{E}_{q(\boldsymbol{w})} \log \frac{p(\boldsymbol{y}, \boldsymbol{w}|X, \alpha)}{q(\boldsymbol{w})} \to \max_{q(\boldsymbol{w}) \in \mathcal{Q}, \alpha}.$$

# Example: Bayesian Logistic Regression

Parametric variational family: $q(\boldsymbol{w}|\boldsymbol{\mu}, \Sigma) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \Sigma)$.

$$\log p(\boldsymbol{y}|X, \alpha) \geq \mathrm{ELBO}(\boldsymbol{\mu}, \Sigma, \alpha) = \mathbb{E}_{q(\boldsymbol{w}|\boldsymbol{\mu},\Sigma)} \log p(\boldsymbol{y}|\boldsymbol{w}, X) - KL(q(\boldsymbol{w}|\boldsymbol{\mu}, \Sigma) \,||\, p(\boldsymbol{w}|\alpha)) =$$

$$= \mathbb{E}_{q(\boldsymbol{w}|\boldsymbol{\mu},\Sigma)} \sum_{i=1}^{n} \log p(y_i|\boldsymbol{w}, \boldsymbol{x}_i) - \underbrace{KL(\mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \Sigma) \,||\, \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha I))}_{\text{analytical expression}} =$$

$$= \sum_{i=1}^{n} \mathbb{E}_{\mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu},\Sigma)} \log \sigma(y_i \boldsymbol{w}^T \boldsymbol{x}_i) - KL(\mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \Sigma) \,||\, \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha I)) \to \max_{\boldsymbol{\mu},\Sigma,\alpha} .$$

If $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \Sigma)$, then $u_i = \boldsymbol{w}^T \boldsymbol{x}_i \sim \mathcal{N}(u_i|m_i, s_i^2)$, where $m_i = \boldsymbol{\mu}^T \boldsymbol{x}_i, s_i^2 = \boldsymbol{x}_i^T \Sigma \boldsymbol{x}_i$.
Hence, $\mathbb{E}_{\mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu},\Sigma)} \log \sigma(y_i \boldsymbol{w}^T \boldsymbol{x}_i) = \mathbb{E}_{\mathcal{N}(u_i|m_i,s_i^2)} \log \sigma(y_i u_i)$.

Finally, if $u_i \sim \mathcal{N}(u_i|m_i, s_i^2)$, then $\xi_i \sim \mathcal{N}(\xi_i|0, 1)$ and $u_i = \xi_i s_i + m_i$. Hence,
$\mathbb{E}_{\mathcal{N}(u_i|m_i,s_i^2)} \log \sigma(y_i u_i) = \mathbb{E}_{\mathcal{N}(\xi_i|0,1)} \log \sigma(y_i(\xi_i s_i + m_i))$.

# Example: Bayesian Logistic Regression

Final expression for ELBO:

$$\mathrm{ELBO}(\boldsymbol{\mu}, \Sigma, \alpha) = \sum_{i=1}^{n} \mathbb{E}_{\mathcal{N}(\xi_i|0,1)} \log \sigma(y_i(\xi_i s_i + m_i)) - KL(\mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \Sigma) \,||\, \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \alpha I)) \to \max_{\boldsymbol{\mu}, \Sigma, \alpha}$$

$$m_i = \boldsymbol{\mu}^T \boldsymbol{x}_i, s_i^2 = \boldsymbol{x}_i^T \Sigma \boldsymbol{x}_i.$$

For solving optimization problem we need to know derivatives of ELBO w.r.t. parameters $\boldsymbol{\mu}, \Sigma, \alpha$. These expressions can be calculated in the following way:

$$\nabla_{\boldsymbol{\mu}} \mathrm{ELBO}(\boldsymbol{\mu}, \Sigma, \alpha) = \sum_{i=1}^{n} \mathbb{E}_{\mathcal{N}(\xi_i|0,1)} \nabla_{\boldsymbol{\mu}} \log \sigma(y_i(\xi_i s_i + m_i)) - \nabla_{\boldsymbol{\mu}} KL(\mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \Sigma) \,||\, \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \alpha I))$$

Expectation w.r.t. one-dimensional standard normal distribution can be calculated using Gauss-Hermite quadrature.

# Conclusions

- Variational Inference approach transforms Bayesian inference problem to a certain type of optimization problem. In this sense it is usually much faster than MCMC sampling approach;
- Variational Inference requires a careful choice of variational approximation family. Depending of properties of probabilistic model this could be mean field or parametric approximation or mixture of them;
- Variational Inference can give poor posterior approximation due to restrictions in variational approximation family. Variational EM-algorithm usually gives good results for model parameters;
- Scalability issues of Variational Inference are not covered. See the next lecture!