Introduction to Variational Inference

Lili Mou
moull12@sei.pku.edu.cn

Outline

Outline

Functional

Functional: a mapping that takes a function as the input and returns a value as output

E.g.

$$H[p] = \int p(x) \ln p(x) \, \mathrm{d}x$$

Calculus

Derivative of a univariate function:

$$y(x+\epsilon) = y(x) + \frac{\mathrm{d}y}{\mathrm{d}x}\epsilon + \mathcal{O}(\epsilon^2)$$

Derivative of a multivariate function:

$$y(x_1+\epsilon_1+\cdots x_D+\epsilon_D) = y(x_1,\cdots x_D) + \sum_{i=1}^{D} \frac{\mathrm{d}y}{\mathrm{d}x_i}\epsilon_i + \mathcal{O}(\epsilon^2)$$

Derivative of a functional

$$F[y(x)+\epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\delta F}{\delta y(x)}\eta(x)\,\mathrm{d}x + \mathcal{O}(\epsilon^2)$$

Stationary condition:

$$\int \frac{\delta F}{\delta y(x)}\eta(x)\,\mathrm{d}x = 0, \qquad \forall\eta$$

$\Rightarrow$ Functional derivatives must vanish for all values of x.

Variational Inference

Big idea: find functions with limited forms

- Parametric form $\Rightarrow$ standard optimization
- Restricted but non-parametric distributions, e.g., factorization

Model

$$\boldsymbol{Z} = \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n\}$$

$$\downarrow$$

$$\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$$

Variational Bound

$$\ln p(\boldsymbol{X}) = \mathcal{L}(q) + KL(q||p)$$

wherex

$$\mathcal{L}(q) = \int q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} \right\} \mathrm{d}\boldsymbol{Z}$$

$$KL(q||p) = - \int q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})} \right\} \mathrm{d}\boldsymbol{Z}$$

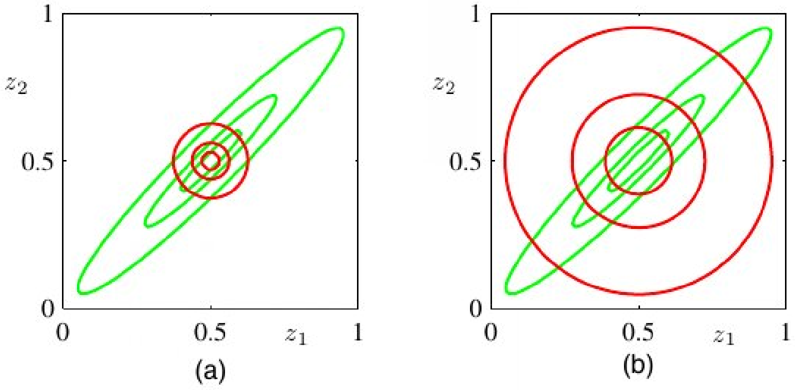
- $KL(\cdot||\cdot) \geq 0$
- Variational lower bound $\mathcal{L}(p)$

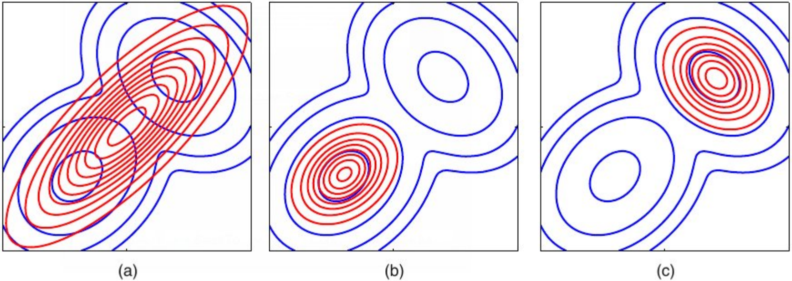
Maximize the lower bound $\mathcal{L}(q)$
$\Leftrightarrow$ Minimize the KL divergence
$\Leftrightarrow$ $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X})$ (usually intractable)

$KL(q||p)$ versus $KL(p||q)$

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

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





Discussion:

- $KL(q||p)$ versus $KL(p||q)$?
- Which one is better?

Outline

Factorized Distribution

Assumption

$$q(\boldsymbol{Z}) = \prod_{i=1}^{M} q_i(\boldsymbol{Z}_i)$$

Optimize $\mathcal{L}(q)$ w.r.t a group $\boldsymbol{Z}_j$ at a time

Lower Bound

$$
\begin{aligned}
\mathcal{L}(q) &= \int q(\boldsymbol{Z}) \left\{ \ln p(\boldsymbol{X}, \boldsymbol{Z}) - \ln q(\boldsymbol{Z}) \right\} \mathrm{d}\boldsymbol{Z} \\
&= \int \prod_i q_i \left\{ \ln p(\boldsymbol{X}, \boldsymbol{Z}) - \sum_i \ln q_i \right\} \mathrm{d}\boldsymbol{Z} \\
&= \int q_i \left\{ \int \ln p(\boldsymbol{X}, \boldsymbol{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\boldsymbol{Z}_i \right\} \mathrm{d}\boldsymbol{Z}_j \\
&\quad - \int q_j \ln q_j \, \mathrm{d}\boldsymbol{Z}_j + \text{const} \\
&\triangleq q_j \ln \tilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j) \, \mathrm{d}\boldsymbol{Z}_j - \int q_j \ln q_j \, \mathrm{d}\boldsymbol{Z}_j + \text{const} \\
&= -KL(q_i || \tilde{p})
\end{aligned}
$$

Notes

$$
\begin{aligned}
&\int \prod_i q_i \ln p(\boldsymbol{X}, \boldsymbol{Z}) \, \mathrm{d}\boldsymbol{Z} \\
&= \int \cdots \int q_i \cdots q_M \ln p(\boldsymbol{X}, \boldsymbol{Z}) \, \mathrm{d}Z_i \cdots \mathrm{d}\boldsymbol{Z}_M \\
&= \int q_i \left\{ \int \ln p(\boldsymbol{X}, \boldsymbol{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\boldsymbol{Z}_1 \cdots \mathrm{d}\boldsymbol{Z}_{j-1} \, \mathrm{d}\boldsymbol{Z}_{j+1} \cdots \mathrm{d}\boldsymbol{Z}_M \right\} \mathrm{d}\boldsymbol{Z}_j
\end{aligned}
$$

Notes (Cont.)

$$
\begin{aligned}
\int \prod_i q_i \sum_l \ln q_l \,\mathrm{d}\boldsymbol{Z} &= \sum_l \int \prod_i q_i \ln q_l \,\mathrm{d}\boldsymbol{Z} \\
&= \int \prod_i q_i \ln q_j \,\mathrm{d}\boldsymbol{Z} + \mathrm{const} \\
&= \int q_j \ln q_j \,\mathrm{d}\boldsymbol{Z}_j + \mathrm{const}
\end{aligned}
$$

Notes (Cont.)

$$\begin{aligned}\ln \tilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j) &\triangleq \mathbb{E}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \text{const} \\ &\triangleq \int \ln p(\boldsymbol{X}, \boldsymbol{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\boldsymbol{Z}_i + \text{const}\end{aligned}$$

$\mathbb{E}_{i \neq j}[\cdot]$ denotes an expectation w.r.t. the q distributions over all variables $\boldsymbol{Z}_i$ for $i \neq j$.

maximize $\mathcal{L}(q)$ w.r.t all possible forms of q_i
$\Leftrightarrow$ minimize $KL(q_i||\tilde{p})$
$\Leftrightarrow q_j^*(\boldsymbol{Z}_j) = \tilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j)$

$$\begin{aligned} \ln q_j^*(\boldsymbol{Z}_j) &= \mathbb{E}_{i\neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \text{const} \\ &= \int \ln p(\boldsymbol{X}, \boldsymbol{Z}) \prod_{i\neq j} q_i \,\mathrm{d}\boldsymbol{Z}_i + \text{const} \end{aligned}$$

$$q_j^*(\boldsymbol{Z}_j) = \frac{\exp\left\{\mathbb{E}_{i\neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})]\right\}}{\int \exp\{\mathbb{E}_{i\neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})]\} \,\mathrm{d}\boldsymbol{Z}_j}$$

Outline

Univariate Gaussian

Problem definition

Assume $\mathcal{D} = \{x_1, \cdots, \boldsymbol{x}_N\}$ i.i.d from a Gaussian

$$p(\mathcal{D}|\mu, \tau) = (\frac{\tau}{2\pi})^{N/2} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{N}(x_n - \mu)^2\right\}$$

Assume conjugate prior distributions for μ and τ

$$\begin{aligned} p(\mu|\tau) &= \mathcal{N}\left(\mu|\mu_0, (\lambda_0\tau)^{-1}\right) \\ p(\tau) &= \mathrm{Gam}(\tau|a_0, b_0) \\ &= \frac{1}{\Gamma(a_0)} b_0^{a_0} \tau^{a_0 - 1} \exp(-b_0\tau) \end{aligned}$$

Exact inference: Gaussian-gamma distribution
Variational inference (factorized distribution):

Assume $q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau)$

Compute $q_\mu(\mu)$

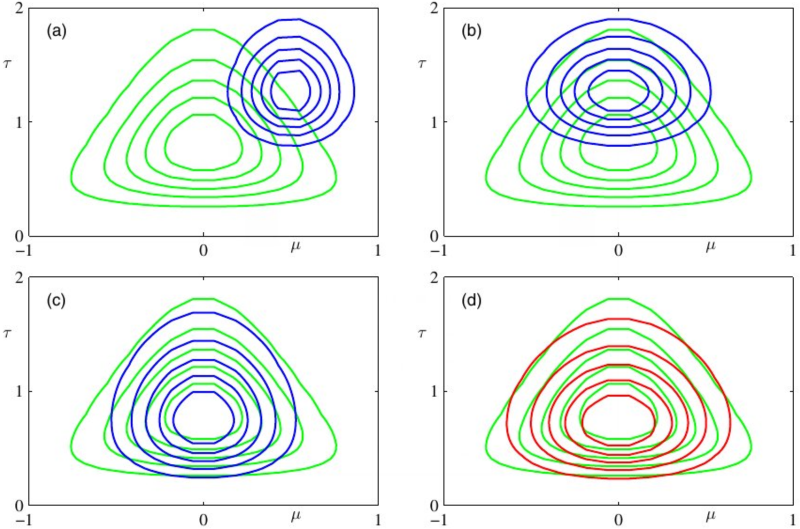
$$
\begin{aligned}
\ln q_\mu^*(\mu) &= \mathbb{E}_\tau[\ln p(\mathcal{D}, \mu, \tau)] + \text{const} \\
&= \mathbb{E}_\tau[\ln p(\mathcal{D}|\mu, \tau) + \ln p(\mu|\tau)] + \text{const} \\
&= \mathbb{E}_\tau\left[-\frac{\tau}{2}\sum_{n=1}^{N}(x_n - \mu)^2 + \frac{\lambda_0 \tau}{2}(\mu - \mu_0)^2\right] + const \\
&= -\frac{\mathbb{E}_\tau[\tau]}{2}\left\{\lambda_0(\mu - \mu_0)^2 + \sum_{i=1}^{N}(x_n - \mu)\right\} + \text{const}
\end{aligned}
$$

$$
\begin{aligned}
q_\mu(\mu) &= \mathcal{N}(\mu|\mu_N, \lambda_N^{-1}), \ \text{where} \\
\mu_N &= \frac{\lambda_0\mu_0 + N\bar{x}}{\lambda_0 + N} \\
\lambda_N &= (\lambda_0 + N)\,\mathbb{E}[\tau]
\end{aligned}
$$

Compute $q_\tau(\tau)$

$$
\begin{aligned}
\ln q_\tau^*(\tau) =& \mathbb{E}_\mu[\ln p(\mathcal{D}, \mu, \tau)] + \text{const} \\
=& \mathbb{E}_\mu[\ln p(\mathcal{D}|\mu, \tau) + \ln p(\mu|\tau)] + \ln p(\tau) + \text{const} \\
=& (a_0 - 1)\ln\tau - b_0\tau + \frac{N}{2}\ln\tau \\
& - \frac{\tau}{2}\,\mathbb{E}_\mu\left[\sum_{n=1}^{N}(x_n - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right] + \text{const}
\end{aligned}
$$

$$
\begin{aligned}
q_\tau(\tau) =& \mathrm{Gam}(\tau|a_N, b_N), \text{ where} \\
a_N =& a_0 + \frac{N}{2} \\
b_N =& b_0 + \frac{1}{2}\,\mathbb{E}_\mu\left[\sum_{n=1}^{N}(x_n - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right]
\end{aligned}
$$


(a)
(b)
(c)
(d)
τ
μ
2
1
0
−1
0
1

Thank you for listening!

Reference

[1] Christopher M. Bishop *Pattern Recognition and Machine Learning*, Springer, 2006.