变分推断

1. 背景 (Background)

1.1 频率派 vs 贝叶斯派 (Frequentist vs Bayesian)

核心区别在于对参数的看法以及随之而来的问题类型：

- **频率派视角 (Frequentist) $\rightarrow$ 优化问题 (Optimization Problem)**
    - 参数 w 被视为未知的常量。
    - **例子 1：回归 (Regression)**
        - 模型 (Model)：$f(w) = w^T x$
        - 损失函数 (Loss Function)：$L(w) = \sum_{i=1}^{N} \|w^T x_i - y_i\|^2$
        - 算法 (Algorithm)：$\hat{w} = \arg\min L(w)$
        - 求解：
            1. **解析解 (Analytic Solution)**：$\frac{\partial L(w)}{\partial w} = 0 \Rightarrow w^* = (X^T X)^{-1} X^T Y$
            2. **数值解 (Numerical Solution)**：梯度下降 (GD), 随机梯度下降 (SGD)
    - **例子 2：SVM (支持向量机)**
        - 模型：$f(w) = sign(w^T x + b)$
        - 损失函数：$\min \frac{1}{2} w^T w$ s.t. $y_i(w^T x_i + b) \geq 1$ (凸优化 Convex Optimization)
        - 算法：二次规划 (QP), 拉格朗日对偶 (Lagrange Duality)
- **贝叶斯视角 (Bayesian) $\rightarrow$ 积分问题 (Integration Problem)**
    - 参数 θ 被视为随机变量。
    - **贝叶斯定理 (Bayes' Theorem)**：
      $P(\theta|X) = \frac{P(X|\theta) \cdot P(\theta)}{P(X)}$
        - $P(\theta|X)$：后验 (Posterior)
        - $P(X|\theta)$：似然 (Likelihood)
        - $P(\theta)$：先验 (Prior)
        - $P(X) = \int P(X|\theta)P(\theta)d\theta$：证据 (Evidence, 归一化常数)
    - **贝叶斯推断** 的核心在于求解后验分布。

1.2 贝叶斯推断 (Bayesian Inference)

给定数据集 $X = \{x_1, \ldots, x_N\}$。
对于新样本 $\hat{x}$，我们希望求出 $P(\hat{x}|X)$ (后验预测分布 Posterior Predictive Distribution)。

$$
\begin{aligned}
P(\hat{x}|X) &= \int_\theta P(\hat{x}, \theta|X) d\theta \\
&= \int_\theta P(\hat{x}|\theta) \cdot P(\theta|X) d\theta \\
&= E_{\theta|X}[P(\hat{x}|\theta)]
\end{aligned} \tag{1}
$$

即通过对参数 θ 的积分（求期望）来获得预测。

1.3 推断方法 (Inference Methods)

推断任务（求后验或期望）通常分为两类：

1. **精确推断 (Exact Inference)**：可以得到精确的后验分布（例如共轭先验的情况）。
2. **近似推断 (Approximate Inference)**：当积分不可积或难以计算时使用。

- 确定性近似 (**Deterministic Approximation**) $\rightarrow$ **变分推断 (Variational Inference, VI)** - 本章重点
- **随机近似 (Stochastic Approximation)** $\rightarrow$ **MCMC (Markov Chain Monte Carlo)** (例如 MH 算法, Gibbs 采样)

1.4 期望最大化 (EM Algorithm)

EM 是一种基于优化的方法，常用于含有隐变量 Z 的模型参数估计。

- **目标**：$\hat{\theta} = \arg\max \log P(X|\theta)$
- **更新步骤 (E-step & M-step 结合)**：
  $\theta^{(t+1)} = \arg\max_\theta \int_z \log P(X, z|\theta) \cdot P(z|X, \theta^{(t)}) dz$

2. 变分推断 (Variational Inference)

变分推断的核心思想是寻找一个分布 $q(Z)$ 来近似后验分布 $P(Z|X)$。

- X: Observed Data (观测数据)
- Z: Latent Variable + Parameter (隐变量 + 参数)
- (X, Z): Complete Data (完整数据)

2.1 公式推导 (Formula Deduction)

我们从边缘似然 (Marginal Likelihood) $\log P(X)$ 出发。
由于涉及到隐变量 Z，我们引入分布 $q(Z)$：

$$\log P(X) = \log P(X, Z) - \log P(Z|X) \tag{2}$$

> **推导 (Derivation):**
> 根据条件概率公式 (Conditional Probability):
> $P(Z|X) = \frac{P(X,Z)}{P(X)}$
> 移项得:
> $P(X) = \frac{P(X,Z)}{P(Z|X)}$
> 两边取对数:
> $\log P(X) = \log\left(\frac{P(X,Z)}{P(Z|X)}\right) = \log P(X, Z) - \log P(Z|X)$

两边同时对 $q(Z)$ 求期望（积分）：

$$\begin{aligned}
\text{左边} &= \int_Z q(Z) \log P(X) dZ = \log P(X) \cdot \int_Z q(Z) dZ = \log P(X) \\
\text{右边} &= \int_Z q(Z) \log \frac{P(X, Z)}{P(Z|X)} dZ \\
&= \int_Z q(Z) \log \frac{P(X, Z)}{q(Z)} \cdot \frac{q(Z)}{P(Z|X)} dZ \\
&= \underbrace{\int_Z q(Z) \log \frac{P(X, Z)}{q(Z)} dZ}_{\text{ELBO}} - \underbrace{\int_Z q(Z) \log \frac{P(Z|X)}{q(Z)} dZ}_{-\text{KL}(q||p)} \\
&= \mathcal{L}(q) + \text{KL}(q||P(Z|X))
\end{aligned} \tag{3}$$

其中：

1. **ELBO (Evidence Lower Bound, 证据下界)**: $\mathcal{L}(q) = \int_Z q(Z) \log \frac{P(X,Z)}{q(Z)} dZ$
2. **KL 散度 (Kullback-Leibler Divergence)**: $\text{KL}(q||P) = \int_Z q(Z) \log \frac{q(Z)}{P(Z|X)} dZ \geq 0$

因为 $\log P(X)$ 是常数 (相对于 q 而言)，且 $\text{KL} \geq 0$，所以：

$$\log P(X) \geq \mathcal{L}(q) \tag{4}$$

要使 $q(Z) \approx P(Z|X)$，即最小化 $\mathrm{KL}(q||P)$，等价于 **最大化 ELBO** $\mathcal{L}(q)$。

$$\hat{q}(Z) = \arg\max_{q(Z)} \mathcal{L}(q) \tag{5}$$

2.2 平均场理论 (Mean Field Theory)

为了求解 $q(Z)$，我们需要对其形式做假设。常用的假设是 **平均场假设 (Mean Field Assumption)**：
假设 Z 可以划分为 M 个独立的组 $Z_1, Z_2, \dots, Z_M$，且它们之间相互独立：

$$q(Z) = \prod_{i=1}^{M} q_i(Z_i) \tag{6}$$

我们的目标是求解每一个 $q_j(Z_j)$。

将 $\mathcal{L}(q)$ 展开：

$$\begin{aligned}\mathcal{L}(q) &= \int_Z q(Z)\log P(X,Z)dZ - \int_Z q(Z)\log q(Z)dZ \\ &= \underbrace{\int_Z \left(\prod_{i=1}^{M} q_i(Z_i)\right)\log P(X,Z)dZ}_{①} - \underbrace{\int_Z \left(\prod_{i=1}^{M} q_i(Z_i)\right)\sum_{i=1}^{M}\log q_i(Z_i)dZ}_{②}\end{aligned} \tag{7}$$

分析第一项 ① (Term 1)：
我们将积分分解为 Z_j 和 Z_{-j} (除 Z_j 以外的变量)：

$$\begin{aligned}① &= \int_Z \left(\prod_{i=1}^{M} q_i(Z_i)\right)\log P(X,Z)dZ_1 dZ_2 \dots dZ_M \\ &= \int_{Z_j} q_j(Z_j)\left(\int_{Z_{-j}} \prod_{i\neq j} q_i(Z_i)\log P(X,Z)dZ_{\text{others}}\right)dZ_j \\ &= \int_{Z_j} q_j(Z_j)\underbrace{\left(\int_{Z_{-j}} \log P(X,Z)\cdot\prod_{i\neq j} q_i(Z_i)dZ_i\right)}_{E_{q_{-j}}[\log P(X,Z)]}dZ_j \\ &= \int_{Z_j} q_j(Z_j)\cdot E_{q_{-j}}[\log P(X,Z)]dZ_j\end{aligned} \tag{8}$$

这里 $E_{q_{-j}}$ 表示关于除 Z_j 以外所有变量的期望。

分析第二项 ② (Term 2)：
这一项是熵 (Entropy) 形式。

$$\begin{aligned}② &= \int_Z \left(\prod_{i=1}^{M} q_i(Z_i)\right)\log q(Z)dZ \\ &= \int_Z \prod_{i=1}^{M} q_i(Z_i)\cdot\sum_{i=1}^{M}\log q_i(Z_i)dZ \\ &= \int_Z \prod_{i=1}^{M} q_i(Z_i)[\log q_1(Z_1)+\dots+\log q_M(Z_M)]dZ\end{aligned} \tag{9}$$

考虑其中某一项（例如 $\log q_1$）：

$$\begin{aligned}&\int_{Z_1 \dots Z_M} q_1 q_2 \dots q_M \log q_1 dZ_1 \dots dZ_M \\ &= \int_{Z_1} q_1 \log q_1 dZ_1 \cdot \underbrace{\int_{Z_2} q_2 dZ_2}_{1} \cdot \dots \cdot \underbrace{\int_{Z_M} q_M dZ_M}_{1} \\ &= \int_{Z_1} q_1 \log q_1 dZ_1\end{aligned} \tag{10}$$

因此，总和可以保留 q_j 项，其余项相对于 q_j 是常数：

$$
② = \int_{Z_j} q_j(Z_j) \log q_j(Z_j) dZ_j + C \tag{11}
$$

合并 ① - ② 关于 $q_j(Z_j)$ 的部分：

$$
\mathcal{L}(q_j) = \int_{Z_j} q_j(Z_j) E_{q_{-j}}[\log P(X, Z)] dZ_j - \int_{Z_j} q_j(Z_j) \log q_j(Z_j) dZ_j + C \tag{12}
$$

令 $\log \hat{P}(X, Z_j) = E_{q_{-j}}[\log P(X, Z)]$，则：

$$
\begin{aligned}
\mathcal{L}(q_j) &= \int_{Z_j} q_j(Z_j) \log \hat{P}(X, Z_j) dZ_j - \int_{Z_j} q_j(Z_j) \log q_j(Z_j) dZ_j + C \\
&= \int_{Z_j} q_j(Z_j) \log \frac{\hat{P}(X, Z_j)}{q_j(Z_j)} dZ_j + C \\
&= -\mathrm{KL}(q_j || \hat{P}(X, Z_j)) + C
\end{aligned} \tag{13}
$$

因为 $\mathrm{KL} \geq 0$，所以要最大化 $\mathcal{L}(q_j)$，必须要最小化 KL 散度，即 $\mathrm{KL} = 0$。
这发生在两个分布相等时：

$$
\log q_j^*(Z_j) = \log \hat{P}(X, Z_j) = E_{q_{-j}}[\log P(X, Z)] \tag{14}
$$

因此：

$$
q_j^*(Z_j) = \hat{P}(X, Z_j) \tag{15}
$$

或者写成指数形式：

$$
q_j^*(Z_j) \propto \exp\left\{E_{Z_{-j}}[\log P(X, Z)]\right\} \tag{16}
$$

这就是 **坐标上升 (Coordinate Ascent)** 算法的更新公式。
我们固定其它 $q_{i \neq j}$，更新 q_j，迭代进行直到收敛。

具体的迭代过程（以 M 个变量为例）：

$$
\begin{aligned}
\log \hat{q}_1(Z_1) &= E_{Z_2, \ldots, Z_M}[\log P(X, Z)] \\
&= \int_{Z_2 \ldots Z_M} \hat{q}_2(Z_2) \ldots \hat{q}_M(Z_M) [\log P(X, Z)] dZ_2 \ldots dZ_M \\
\log \hat{q}_2(Z_2) &= E_{Z_1, Z_3, \ldots, Z_M}[\log P(X, Z)] \\
&= \int_{Z_1 Z_3 \ldots Z_M} \hat{q}_1(Z_1) \hat{q}_3(Z_3) \ldots \hat{q}_M(Z_M) [\log P(X, Z)] dZ_1 dZ_3 \ldots dZ_M \\
&\vdots \\
\log \hat{q}_M(Z_M) &= E_{Z_1, \ldots, Z_{M-1}}[\log P(X, Z)] \\
&= \int_{Z_1 \ldots Z_{M-1}} \hat{q}_1(Z_1) \ldots \hat{q}_{M-1}(Z_{M-1}) [\log P(X, Z)] dZ_1 \ldots dZ_{M-1}
\end{aligned} \tag{17}
$$

每一次更新 q_j 时，都使用最新的其它 $q_{i \neq j}$ 分布。

3. 随机梯度变分推断 (Stochastic Gradient Variational Inference, SGVI)

回顾 ELBO 的定义：

$$
\mathrm{ELBO} = \mathcal{L}(\phi) = E_{q_\phi(z)}\left[\log \frac{P_\theta(x, z)}{q_\phi(z)}\right] \tag{18}
$$

其中 x 是观测变量，z 是隐变量，ϕ 是变分分布 q 的参数，θ 是模型参数。

我们的目标是找到最优的 ϕ 使得 ELBO 最大化：
$\hat{\phi} = \arg\max_\phi \mathcal{L}(\phi)$

直接求导 $\nabla_\phi \mathcal{L}(\phi)$ 比较困难，因为期望分布 $q_\phi(z)$ 本身也包含参数 ϕ。

3.1 梯度推导 (Gradient Deduction)

我们可以利用 **Log-Derivative Trick (对数导数技巧)** 来交换梯度和积分的顺序。

$$\begin{aligned}\nabla_\phi \mathcal{L}(\phi) &= \nabla_\phi \int q_\phi(z)\left[\log P_\theta(x,z) - \log q_\phi(z)\right]dz \\ &= \int \nabla_\phi q_\phi(z)\left[\log P_\theta(x,z) - \log q_\phi(z)\right]dz + \int q_\phi(z)\nabla_\phi\left[\log P_\theta(x,z) - \log q_\phi(z)\right]dz\end{aligned} \tag{19}$$

第一项 (Term 1):
利用 $\nabla_\phi q_\phi(z) = q_\phi(z)\nabla_\phi \log q_\phi(z)$：

$$\text{Term 1} = \int q_\phi(z)\nabla_\phi \log q_\phi(z)\left[\log P_\theta(x,z) - \log q_\phi(z)\right]dz \tag{20}$$

这可以写成期望形式：$E_{q_\phi(z)}\left[\nabla_\phi \log q_\phi(z)(\log P_\theta(x,z) - \log q_\phi(z))\right]$。

第二项 (Term 2):

$$\begin{aligned}\text{Term 2} &= \int q_\phi(z)\nabla_\phi\left[\log P_\theta(x,z) - \log q_\phi(z)\right]dz \\ &= \int q_\phi(z)\left(0 - \frac{1}{q_\phi(z)}\nabla_\phi q_\phi(z)\right)dz \\ &= -\int \nabla_\phi q_\phi(z)dz = -\nabla_\phi \int q_\phi(z)dz = -\nabla_\phi 1 = 0\end{aligned} \tag{21}$$

(注意：$\log P_\theta(x,z)$ 对 ϕ 求导为 0)

结论:

$$\nabla_\phi \mathcal{L}(\phi) = E_{q_\phi(z)}\left[\nabla_\phi \log q_\phi(z)\left(\log P_\theta(x,z) - \log q_\phi(z)\right)\right] \tag{22}$$

这被称为 **Score Function Estimator** (或是 REINFORCE 梯度)。

3.2 蒙特卡洛近似 (Monte Carlo Approximation)

通过从 $q_\phi(z)$ 中采样 L 个样本 $z^{(l)} \sim q_\phi(z), l = 1, \ldots, L$，我们可以近似上述期望：

$$\nabla_\phi \mathcal{L}(\phi) \approx \frac{1}{L}\sum_{l=1}^{L} \nabla_\phi \log q_\phi(z^{(l)})\left(\log P_\theta(x, z^{(l)}) - \log q_\phi(z^{(l)})\right) \tag{23}$$

这样我们就可以使用随机梯度上升 (Stochastic Gradient Ascent) 来优化 ϕ。

3.3 重参数化技巧 (Reparameterization Trick)

问题 (Problem):
前面提到的 Score Function Estimator (REINFORCE) 虽然是无偏估计，但是方差很大 (High Variance)，导致训练不稳定。

解决方案 (Solution):
通过重参数化技巧 (Reparameterization Trick) 来降低方差。

假设隐变量 z 可以表示为一个无参数分布的辅助变量 ϵ 的确定性变换：

$z = g_\phi(\epsilon, x^{(i)}) \quad$ 其中 $\quad \epsilon \sim p(\epsilon)$

例如：

- 如果 $q_\phi(z|x) = \mathcal{N}(z; \mu, \sigma^2)$，则 $z = \mu + \sigma \cdot \epsilon$，其中 $\epsilon \sim \mathcal{N}(0, I)$。

梯度推导:
利用重参数化，期望中的分布不再依赖于 ϕ，我们可以将梯度算子直接移入期望内部：

$$
\begin{aligned}
\nabla_\phi \mathcal{L}(\phi) &= \nabla_\phi E_{q_\phi(z)}[\log P_\theta(x^{(i)}, z) - \log q_\phi(z|x^{(i)})] \\
&= \nabla_\phi E_{p(\epsilon)}[\log P_\theta(x^{(i)}, g_\phi(\epsilon, x^{(i)})) - \log q_\phi(g_\phi(\epsilon, x^{(i)})|x^{(i)})] \\
&= E_{p(\epsilon)}[\nabla_\phi(\log P_\theta(x^{(i)}, z) - \log q_\phi(z|x^{(i)}))] \\
&= E_{p(\epsilon)}[\nabla_z(\log P_\theta(x^{(i)}, z) - \log q_\phi(z|x^{(i)})) \cdot \nabla_\phi g_\phi(\epsilon, x^{(i)})]
\end{aligned}
\tag{24}
$$

蒙特卡洛估计:
采样 L 个噪声样本 $\epsilon^{(l)} \sim p(\epsilon)$，计算梯度：

$$
\nabla_\phi \mathcal{L}(\phi) \approx \frac{1}{L} \sum_{l=1}^{L} \nabla_z \left(\log P_\theta(x^{(i)}, z^{(l)}) - \log q_\phi(z^{(l)}|x^{(i)}) \right) \cdot \nabla_\phi g_\phi(\epsilon^{(l)}, x^{(i)}) \tag{25}
$$

其中 $z^{(l)} = g_\phi(\epsilon^{(l)}, x^{(i)})$。

3.4 SGVI 算法

基于随机梯度的变分推断算法 (Stochastic Gradient Variational Inference):

$$
\phi^{(t+1)} \leftarrow \phi^{(t)} + \lambda^{(t)} \cdot \nabla_\phi \mathcal{L}(\phi) \tag{26}
$$

其中 $\lambda^{(t)}$ 是学习率 (Learning Rate) 或步长 (Step Size)。