参数化变分推断教程

Jens Sjölund [1] ✉

[1]乌普萨拉大学信息技术系，瑞典

arXiv:2301.01236v1 [stat.ML] 3 Jan 2023

✉ **联系邮箱：**
jens.sjolund@it.uu.se

资助： 本研究部分得到了由克努特和爱丽丝·瓦伦堡基金会资助的瓦伦堡人工智能、自主系统与软件计划（WASP）的支持。

摘要

在贝叶斯模型中，变分推断通过优化而非积分来近似边际似然，进而近似后验分布。得益于过去十年计算可扩展性的进步，变分推断如今已成为许多高维模型和大型数据集的首选方法。本教程从参数化视角介绍变分推断；与此相对，其他入门读物通常采用平均场视角，而后者并非近年来发展的主流视角。

引言

在贝叶斯机器学习和统计学中，核心关注对象是通过贝叶斯推断得到的后验分布——即根据贝叶斯定理，将先验信念与观测值结合起来。在一些简单情形下，例如共轭模型中，可以精确完成这一过程。但一般的（非共轭）模型需要采用蒙特卡洛或变分推断等近似推断技术。这些方法各有优缺点，因此最合适的选择取决于具体应用。我们重点讨论变分推断：它一方面并非保证渐近精确，另一方面在计算上高效，并且能够扩展到高维模型和大型数据集。

记号

我们用单个观测变量 $\boldsymbol{x}$ 同时表示观测到的输入和输出。不过，我们主要关注的是潜变量 $\boldsymbol{z}$。由于我们遵循贝叶斯框架，模型中被赋予先验分布的“参数”（例如线性回归模型中的斜率和截距）实际上是潜变量。我们用 $\boldsymbol{\theta}$ 表示其余感兴趣的参数。

变分推断

那么，我们为什么需要变分推断？首先，回想一下，要根据观测值推断潜变量的任何信息，我们需要后验分布：

$$p_\theta(\boldsymbol{z} \mid \boldsymbol{x}) = \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{p_\theta(\boldsymbol{x})}. \tag{1}$$

分母中的表达式 $p_\theta(\boldsymbol{x})$ 称为 $\boldsymbol{x}$ 的边际似然，因为它可以改写为对潜变量的边缘化：

$$p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{x}, \boldsymbol{z})\, d\boldsymbol{z}. \tag{2}$$

问题在于，实际应用中这个积分往往难以求解，也就是说，无法用闭式计算。由于 $\boldsymbol{x}$是我们的观测值，边际似然是一个（归一化）常数。然而，如果不知道这个常数，后验分布的实用性就会受到限制。因此，我们需要近似推断。

核心思想

变分推断的核心思想是用一个易于处理的下界替代难以求解的边际似然，然后对该下界进行最大化。建模主要包括选择一个性质良好且表达能力足够强的概率分布族 $\mathcal{Q}$。更具体地说，我们希望存在一个分布 $q \in \mathcal{Q}$，称为变分后验，可用来直接替代真实后验。因而，变分后验应当与真实后验$p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$“接近”，同时又相对容易求得。搜索过程本质上是数学优化，这也是为什么变分推断有时被描述为：将一个困难的积分问题转化为一个更容易的优化问题。

证据下界（ELBO）

在变分推断中，真实后验 $p(\boldsymbol{z} \mid \boldsymbol{x})$ 与变分后验 $q(\boldsymbol{z})$ 之间的距离用库尔贝克–莱布勒（KL）散度来衡量，

$$\mathrm{KL}(q(\boldsymbol{z}) \parallel p(\boldsymbol{z} \mid \boldsymbol{x})) = -\int q(\boldsymbol{z}) \log \left(\frac{p(\boldsymbol{z} \mid \boldsymbol{x})}{q(\boldsymbol{z})} \right) d\boldsymbol{z}. \tag{3}$$

也可以使用其他距离度量，使变分后验接近真实后验，但 KL 散度有一个独特的优点：通过一个巧妙的技巧，我们可以同时估计边际似然，并避免计算方程 3中的后验分布。为了解其原理，我们先来看以下两个数学恒等式：

$$\int q(\boldsymbol{z})\, d\boldsymbol{z} = 1, \tag{4}$$

$$p(\boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{z})}{p(\boldsymbol{z} \mid \boldsymbol{x})} = \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} \left(\frac{p(\boldsymbol{z} \mid \boldsymbol{x})}{q(\boldsymbol{z})} \right)^{-1}. \tag{5}$$

利用这两个恒等式，我们可以将边际似然改写如下：

$$\begin{aligned} \log p(\boldsymbol{x}) &\underset{(4)}{=} \log p(\boldsymbol{x}) \cdot \int q(\boldsymbol{z})\, d\boldsymbol{z} = \int q(\boldsymbol{z}) \log p(\boldsymbol{x})\, d\boldsymbol{z} \\ &\underset{(5)}{=} \int q(\boldsymbol{z}) \log \left(\frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} \right) d\boldsymbol{z} - \int q(\boldsymbol{z}) \log \left(\frac{p(\boldsymbol{z} \mid \boldsymbol{x})}{q(\boldsymbol{z})} \right) d\boldsymbol{z} \\ &= \int q(\boldsymbol{z}) \log \left(\frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} \right) d\boldsymbol{z} + \mathrm{KL}(q(\boldsymbol{z}) \parallel p(\boldsymbol{z} \mid \boldsymbol{x})). \end{aligned} \tag{6}$$

由于 KL 散度始终非负，因此对于任意 q，第一项都是对数边缘似然（也称为证据）的下界，因此称为证据下界（ELBO）：

$$\mathrm{ELBO}(q(\boldsymbol{z})) = \int q(\boldsymbol{z}) \log \left(\frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} \right) d\boldsymbol{z} = \mathbb{E}_{q(\boldsymbol{z})} \left[\log p(\boldsymbol{x}, \boldsymbol{z}) - \log q(\boldsymbol{z}) \right]. \tag{7}$$

方程 6 因此可以更简洁地写作

$$\log p(\boldsymbol{x}) = \mathrm{ELBO}(q(\boldsymbol{z})) + \mathrm{KL}(q(\boldsymbol{z}) \parallel p(\boldsymbol{z} \mid \boldsymbol{x})). \tag{8}$$

对于一个固定的模型 $p(\boldsymbol{x}, \boldsymbol{z})$，证据的（对数）值是一个常数。因此，回想起 KL 散度非负，我们可以得出结论：最大化证据下界等价于最小化 KL散度。这一点很有利，因为要计算 KL 散度，我们必须对一个包含同一难以处理的后验分布的函数求边缘化，而这个后验分布正是我们想要估计的。相比之下，模型只通过联合分布 $p(\boldsymbol{x}, \boldsymbol{z})$影响证据下界，这意味着，首先，我们不需要计算方程2 中的棘手积分；其次，我们可以对联合分布进行因子分解，例如，按照有向图模型所表示的方式（Wainwright 和 Jordan，2008）。

示例 1

假设我们有一个来自 Exp(λ) 似然的单个观测值 x，其率参数 λ的先验分布为 Gamma(α, β)。假设 α 和 β 已知，那么唯一感兴趣的潜变量是 $z = \{\lambda\}$。具体来说，

$$p(x \mid \lambda) = \lambda e^{-\lambda x},$$
$$p(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda},$$

其中 $\Gamma(\alpha)$ 是伽马函数。由于伽马分布是 λ 的共轭先验分布，我们知道后验分布也是伽马分布。应用贝叶斯定理，并忽略所有不含 λ 的因子，可得

$$p(\lambda \mid x) \propto p(x \mid \lambda)p(\lambda) \propto \lambda^\alpha e^{-\lambda(\beta+x)}.$$

因此，我们将后验分布确定为 $p(\lambda \mid x) =$ 伽马分布$(\alpha + 1, \beta + x)$。
But，让我们假设自己并不知道这一点，而是希望用对数正态分布(μ, σ^2)来拟合 t o th后验分布，通过变分推断来完成，即

$$q(\lambda) = \frac{1}{\lambda\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log\lambda - \mu)^2}{2\sigma^2}\right).$$

根据方程 7可得

$$\begin{aligned}
&\text{ELBO}(q(\lambda)) = \mathbb{E}_{q(\lambda)}\left[\log\left(p(x \mid \lambda)p(\lambda)\right) - \log q(\lambda)\right] \\
&= \mathbb{E}_{q(\lambda)}\left[\log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right) + \alpha\log\lambda - \lambda(\beta + x) + \log\sqrt{2\pi} + \log\sigma + \log\lambda + \frac{(\log\lambda - \mu)^2}{2\sigma^2}\right] \\
&= \log\left(\frac{\beta^\alpha\sqrt{2\pi}}{\Gamma(\alpha)}\right) + (\alpha + 1)\mathbb{E}_{q(\lambda)}\left[\log\lambda\right] - (\beta + x)\mathbb{E}_{q(\lambda)}\left[\lambda\right] + \log\sigma + \frac{1}{2\sigma^2}\mathbb{E}_{q(\lambda)}\left[(\log\lambda - \mu)^2\right].
\end{aligned}$$

期望 $\mathbb{E}_{q(\lambda)}[\lambda] = \exp\left(\mu + \frac{\sigma^2}{2}\right)$，因为根据定义，它是对数正态分布的均值。此外，积分变量的替换 $y = \log\lambda$ （将 $q(\lambda)$ 转换为 $q(y) = \mathcal{N}(\mu, \sigma^2)$ ）表明

$$\mathbb{E}_{q(\lambda)}\left[\log\lambda\right] = \mathbb{E}_{q(y)}\left[y\right] = \mu,$$
$$\mathbb{E}_{q(\lambda)}\left[(\log\lambda - \mu)^2\right] = \mathbb{E}_{q(y)}\left[(y - \mu)^2\right] = \sigma^2.$$

因此，证据下界（ELBO）的最终表达式为

$$\text{ELBO}(q(\lambda)) = \log\left(\frac{\beta^\alpha\sqrt{2\pi}}{\Gamma(\alpha)}\right) + (\alpha + 1)\mu - (\beta + x)e^{\mu + \frac{\sigma^2}{2}} + \log\sigma + \frac{1}{2}.$$

建模

那么，我们该如何选择变分族 $\mathcal{Q}$？从历史上看，主流做法是假设变分后验具有特定的因子分解形式，并运用变分法寻找符合该因子分解形式的分布。这种方法称为平均场变分推断（Blei、Kucukelbir 和 McAuliffe，2017年），至今仍是课堂上最常讲授的方法。然而，平均场变分推断仅适用于相当有限的一类模型。过去 10 至 15 年间，变分推断的大多数成功应用采用的都是参数化方法：通过深度神经网络等表达能力很强的模型对变分族进行参数化。随后，我们便可以使用“标准”优化技术来寻找能够 $\boldsymbol{\theta}^*$ 最大化证据下界（ELBO）的参数。鉴于上述情况，本教程将专门讨论参数化变分推断。

在示例 1中，我们确实采用了参数化方法，因为变分后验由具有参数 $\boldsymbol{\theta} = \{\mu, \sigma\}$ 的对数正态分布显式参数化。在示例 2中，我们将仔细考察该模型某个具体实例的证据下界。

为了准确近似真实后验分布，我们希望变分族 $\mathcal{Q}$ 尽可能丰富，同时保持可处理性——毕竟不可能过拟合！然而，示例 3 表明，有一点需要注意： $q(\boldsymbol{z})$ 在 $p(\boldsymbol{z} \mid \boldsymbol{x})$ 为零时也必须为零。

估计证据下界

在目前为止的示例中，期望都可以用闭式计算。但一般情况下（在非共轭模型中）很少能做到这一点。不过，我们可以使用蒙特卡洛估计，将期望替换为一个求和，

$$\begin{aligned}\text{ELBO}(q(\boldsymbol{z})) &= \int q(\boldsymbol{z}) \log\left(\frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})}\right) d\boldsymbol{z} = \mathbb{E}_{q(\boldsymbol{z})}\left[\log p(\boldsymbol{x}, \boldsymbol{z}) - \log q(\boldsymbol{z})\right] \\ &\approx \frac{1}{L}\sum_{i=1}^{L}\left(\log p(\boldsymbol{x}, \boldsymbol{z}^{(i)}) - \log q(\boldsymbol{z}^{(i)})\right). \end{aligned} \tag{9}$$

关键要求是能够从变分后验 $q(\boldsymbol{z})$中抽取样本 $\boldsymbol{z}^{(i)}$。但是，正如上一节所述，仅仅针对给定的 $q \in \mathcal{Q}$计算证据下界还不够——我们希望找到最优的 q！将变分后验 $q_{\boldsymbol{\theta}}(\boldsymbol{z})$ 用参数 $\boldsymbol{\theta}$参数化后，我们可以将其重新表述为：寻找使证据下界最大化的参数值。然而，要进行高效优化，我们需要同时计算目标函数（证据下界）及其梯度。

基于梯度的优化：证据下界

在优化中，通常考虑最小化问题。（由于最大化问题可以通过对目标函数取负转换为最小化问题，因此这样做不失一般性。）因此，我们将优化问题表示为：

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} -\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z})}\left[\log p(\boldsymbol{x}, \boldsymbol{z}) - \log q_{\boldsymbol{\theta}}(\boldsymbol{z})\right]. \tag{10}$$

例如，对这个问题应用梯度下降，对应的迭代过程为：

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \eta \nabla_{\boldsymbol{\theta}} \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z})}\left[\log p(\boldsymbol{x}, \boldsymbol{z}) - \log q_{\boldsymbol{\theta}}(\boldsymbol{z})\right], \qquad k = 0, 1, \ldots \tag{11}$$

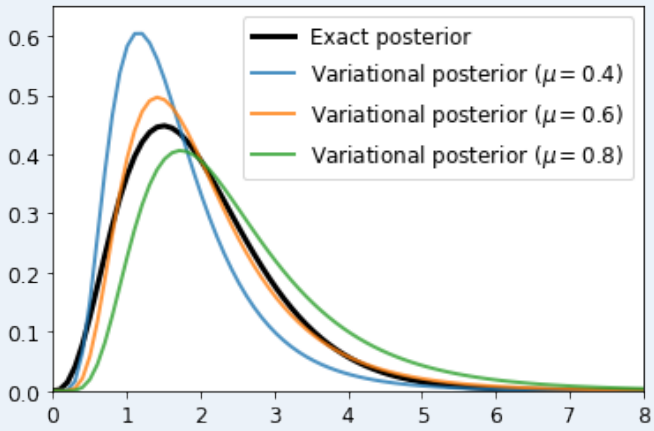
其中，超参数 $\eta > 0$ 是步长。但这揭示了一个难点：梯度作用于我们计算期望所依据的分布的参数。因此，我们不能简单地将梯度移入期望，也不能先用样本替代期望，再对这些样本求梯度，也就是不能使用蒙特卡洛技巧。不过，还有其他不那么直接的方法可以应用蒙特卡洛思想，而且确实有效（顺带一提，这会将梯度下降转变为随机梯度下降）。我们先重写证据下界的梯度

例2

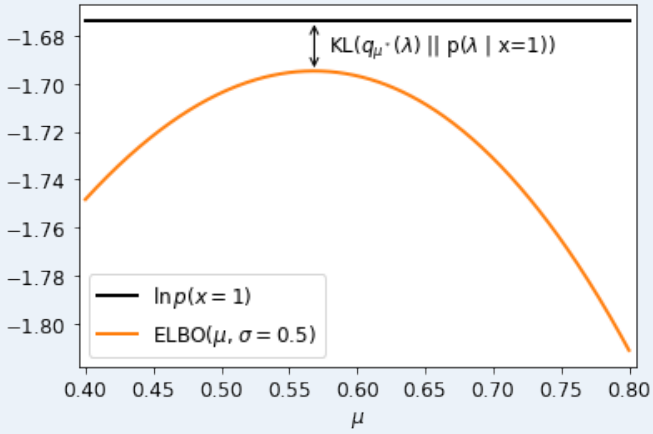
为了更具体地说明问题，我们继续采用示例 1 中的设定，并令 $\alpha = 3$，$\beta = 1$，以及 $x = 1$。证据 $p(x)$ 是此前忽略的比例常数，它联系着后验分布与联合分布，

$$p(x) = \frac{p(x, \lambda)}{p(\lambda \mid x)} = \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha)} \frac{\beta^\alpha}{(\beta + x)^{\alpha+1}} = \frac{\alpha \beta^\alpha}{(\beta + x)^{\alpha+1}}.$$

代入上述数值可得 $p(x = 1) = 3/16$。
为简单起见，我们在变分后验中固定 $\sigma = 0.5$ （这近似对应于通过矩匹配得到的值），并研究改变 μ 的影响。



例2——图1。不同 μ 取值下，对数正态分布($\mu, \sigma^2 = 0.25$)变分后验对伽马分布(4, 2)后验分布的拟合效果。



例2——图2。证据下界对数证据的近似效果取决于参数 μ。两者之间的差距恰好对应 KL散度，因此最大化证据下界等价于最小化 KL散度。

示例 3

让我们回到例子 1，看看如果我们尝试使用一个 $\mathcal{N}(\mu, \sigma^2)$ 分布作为变分后验，也就是：

$$q(\lambda) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\lambda-\mu)^2}{2\sigma^2}\right).$$

像之前一样推导证据下界，可得

$$\begin{aligned}
&\text{ELBO}(q(\lambda)) = \mathbb{E}_{q(\lambda)}\left[\log\left(p(x \mid \lambda)p(\lambda)\right) - \log q(\lambda)\right] \\
&= \mathbb{E}_{q(\lambda)}\left[\log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right) + \alpha\log\lambda - \lambda(\beta + x) + \log\sqrt{2\pi} + \log\sigma + \frac{(\lambda-\mu)^2}{2\sigma^2}\right] \\
&= \log\left(\frac{\beta^\alpha\sqrt{2\pi}}{\Gamma(\alpha)}\right) + \alpha\underbrace{\mathbb{E}_{q(\lambda)}\left[\log\lambda\right]}_{\text{undefined!}} - (\beta + x)\underbrace{\mathbb{E}_{q(\lambda)}\left[\lambda\right]}_{=\mu} + \log\sigma + \frac{1}{2\sigma^2}\underbrace{\mathbb{E}_{q(\lambda)}\left[(\lambda-\mu)^2\right]}_{=\sigma^2}
\end{aligned}$$

对数仅对正值有定义，因此 $\mathbb{E}_{q(\lambda)}\left[\log\lambda\right]$ 未定义。这说明在选择变分分布时有一个重要注意事项：$q(\boldsymbol{z})$ 必须在$p(\boldsymbol{z} \mid \boldsymbol{x})$为零时也为零。

如下所示（Ranganath、Gerrish 和 Blei，2014年）：

$$\begin{aligned}
\nabla_\theta \mathbb{E}_{q_\theta(\boldsymbol{z})}\left[\log p(\boldsymbol{x}, \boldsymbol{z}) - \log q_\theta(\boldsymbol{z})\right] &= \nabla_\theta \int \left(\log p(\boldsymbol{x}, \boldsymbol{z}) - \log q_\theta(\boldsymbol{z})\right) q_\theta(\boldsymbol{z})\, d\boldsymbol{z} \\
&= \int \left(\log p(\boldsymbol{x}, \boldsymbol{z}) - \log q_\theta(\boldsymbol{z})\right) \nabla_\theta\, q_\theta(\boldsymbol{z})\, d\boldsymbol{z} - \int \left(\nabla_\theta \log q_\theta(\boldsymbol{z})\right) q_\theta(\boldsymbol{z})\, d\boldsymbol{z} \qquad (12)
\end{aligned}$$

但该表达式中的第二项为零，

$$\int \left(\nabla_\theta \log q_\theta(\boldsymbol{z})\right) q_\theta(\boldsymbol{z})\, d\boldsymbol{z} = \int \frac{\nabla_\theta q_\theta(\boldsymbol{z})}{\cancel{q_\theta(\boldsymbol{z})}} \cancel{q_\theta(\boldsymbol{z})}\, d\boldsymbol{z} = \nabla_\theta \underbrace{\int q_\theta(\boldsymbol{z})\, d\boldsymbol{z}}_{=1} = 0. \qquad (13)$$

总之，可得

$$\nabla_\theta \mathbb{E}_{q_\theta}(\boldsymbol{z})\left[\log p(\boldsymbol{x}, \boldsymbol{z}) - \log q_\theta(\boldsymbol{z})\right] = \int \left(\log p(\boldsymbol{x}, \boldsymbol{z}) - \log q_\theta(\boldsymbol{z})\right) \nabla_\theta\, q_\theta(\boldsymbol{z}) d\boldsymbol{z}. \qquad (14)$$

有时，如示例 4所示，我们可以改写 $\nabla_\theta q_\theta(\boldsymbol{z})$（变分后验的梯度），从而直接使用蒙特卡罗方法来估计积分。稍后，我们将介绍另外两种更通用的基于蒙特卡罗的方法：重参数化（Kingma 和 Welling，2014年）和黑盒变分推断（Ranganath、Gerrish 和 Blei，2014年）。

重参数化

“重参数化技巧”因介绍变分自编码器的论文而广为人知（Kingma 和 Welling，2014年），但其一般原理的历史要悠久得多（Devroye，1996）。其思路是通过巧妙地重新表述随机变量 $\boldsymbol{z} \sim q_\theta(\boldsymbol{z})$，将随机性的来源与参数解耦，使其成为另一个易于采样的随机变量 $z = g_\theta(\epsilon)$ 的参数化变换 $\epsilon \sim p(\epsilon)$。实际上，这样就将随机性移到了模型“外部”，从而可以将梯度移入期望中，如下例所示。

当且仅当 $g(\epsilon, \boldsymbol{\theta})$ 是一个连续的关于 $\boldsymbol{\theta}$ 对于所有 ϵ(Schulman 等人，2015）。此外，与上面的示例相同，该技巧也适用于期望 $\mathbb{E}_{q_\theta(\boldsymbol{z})}[f(\boldsymbol{z})]$其中 $f(\boldsymbol{z})$是关于 $\boldsymbol{z}$的一般非线性函数，

$$\nabla_\theta \mathbb{E}_{q_\theta(\boldsymbol{z})}\left[f(\boldsymbol{z})\right] = \nabla_\theta \mathbb{E}_{p(\epsilon)}\left[f(g_\theta(\epsilon))\right] = \mathbb{E}_{p(\epsilon)}\left[\nabla_\theta f(g_\theta(\epsilon))\right]. \qquad (15)$$

通过设置 $f(\boldsymbol{z}) = \log p(\boldsymbol{x}, \boldsymbol{z}) - \log q_\theta(\boldsymbol{z})$，我们可以将证据下界作为一个特例推导出来。

示例 4

考虑一个由均值 μ 和标准差 σ 参数化的单变量正态变分后验，即

$$q_\theta(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right), \qquad \boldsymbol{\theta} = \{\mu, \sigma\}.$$

经过一些代数运算后，偏导数可以写成：

$$\begin{aligned}
\frac{\partial q_\theta}{\partial \mu} &= \frac{z-\mu}{\sigma^2} \cdot q_\theta(z), \\
\frac{\partial q_\theta}{\partial \sigma} &= \frac{1}{\sigma}\left(\frac{(z-\mu)^2}{\sigma^2} - 1\right) \cdot q_\theta(z).
\end{aligned}$$

注意， $q_\theta(z)$ 出现在这两个表达式中。将上述内容代入方程14，即可得到一个期望，我们可以用蒙特卡洛估计来替代它：

$$\begin{aligned}
&\nabla_\theta \mathbb{E}_{q_\theta(z)}\left[\log p(\boldsymbol{x}, z) - \log q_\theta(z)\right] \\
&= \int \left(\log p(\boldsymbol{x}, z) - \log q_\theta(z)\right) \left(\tfrac{z-\mu}{\sigma^2}, \tfrac{1}{\sigma}\left(\tfrac{(z-\mu)^2}{\sigma^2} - 1\right)\right)^\top q_\theta(z) dz \\
&\approx \frac{1}{L}\sum_{i=1}^{L} \left(\log p(\boldsymbol{x}, z^{(i)}) - \log q_\theta(z^{(i)})\right) \left(\tfrac{z^{(i)}-\mu}{\sigma^2}, \tfrac{1}{\sigma}\left(\tfrac{(z^{(i)}-\mu)^2}{\sigma^2} - 1\right)\right)^\top,
\end{aligned}$$

其中 $z^{(i)} \sim q_\theta(z)$。

示例 5

假设变分后验是一个以均值 μ 和标准差 σ 为参数的单变量正态分布，即 $q_\theta(z) = \mathcal{N}(z; \boldsymbol{\theta})$，其中 $\boldsymbol{\theta} = \{\mu, \sigma\}$。这可以重新参数化为 $z = g_\theta(\epsilon) = \mu + \sigma \cdot \epsilon$，其中 $\epsilon \sim \mathcal{N}(0, 1)$。L

让我们来看看这对期望 $\mathbb{E}_{q_\theta(z)}\left[\log z\right]$ 有何影响。

(i) 原始表达式：

$$\begin{aligned}
\mathbb{E}_{q_\theta(z)}\left[\log z\right] &= \frac{1}{\sqrt{2\pi\sigma^2}} \int \log z \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right) dz, \\
\nabla_\theta \mathbb{E}_{q_\theta(z)}\left[\log z\right] &= \int \log z \, \nabla_\theta q_\theta(z) dz \\
&= \mathbb{E}_{q_\theta(z)}\left[\log z \cdot \left(\left(\tfrac{z-\mu}{\sigma^2}\right), \tfrac{1}{\sigma}\left(\tfrac{(z-\mu)^2}{\sigma^2} - 1\right)\right)^\top\right]
\end{aligned}$$

其中我们使用了 $\nabla_\theta q_\theta(z)$ 示例4中的表达式。

（ii）重参数化表达式：

$$\begin{aligned}
\mathbb{E}_{q_\theta(z)}\left[\log z\right]\Big|_{z=\mu+\sigma\epsilon} &= \frac{1}{\sqrt{2\pi\sigma^2}} \int \log(\mu+\sigma\epsilon) \exp\left(-\frac{(\mu+\sigma\epsilon-\mu)^2}{2\sigma^2}\right) \sigma \, d\epsilon \\
&= \frac{1}{\sqrt{2\pi}} \int \log(\mu+\sigma\epsilon) \exp\left(-\frac{\epsilon^2}{2}\right) d\epsilon = \mathbb{E}_{p(\epsilon)}\left[\log(\mu+\sigma\epsilon)\right], \\
\nabla_\theta \mathbb{E}_{p(\epsilon)}\left[\log(\mu+\sigma\epsilon)\right] &= \mathbb{E}_{p(\epsilon)}\left[\nabla_\theta \log(\mu+\sigma\epsilon)\right] \\
&= \mathbb{E}_{p(\epsilon)}\left[\left(\tfrac{1}{\mu+\sigma\epsilon}, \tfrac{\epsilon}{\mu+\sigma\epsilon}\right)^\top\right].
\end{aligned}$$

摊销变分推断

许多概率模型都为每个数据点 $\boldsymbol{x}_i$关联了局部潜变量 $\boldsymbol{z}_i$ 。最简单的情况是，联合分布可分解为

$$p(\boldsymbol{x}, \boldsymbol{z}) = \prod_{i=1}^{N} p(\boldsymbol{x}_i \mid \boldsymbol{z}_i) p(\boldsymbol{z}_i). \tag{16}$$

假设我们使用一个按相应方式分解的变分后验，

$$q_{\boldsymbol{\theta}}(\boldsymbol{z}) = \prod_{i=1}^{N} q_{\boldsymbol{\theta}_i}(\boldsymbol{z}_i), \tag{17}$$

那么，式10 中的证据下界最大化问题可分解为一系列局部证据下界之和

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} - \sum_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\theta}_i}(\boldsymbol{z}_i)} \left[\log p(\boldsymbol{x}_i \mid \boldsymbol{z}_i) + \log p(\boldsymbol{z}_i) - \log q_{\boldsymbol{\theta}_i}(\boldsymbol{z}_i) \right]. \tag{18}$$

由于优化变量为 $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N\}$，大型数据集会带来规模庞大的优化问题，求解起来计算开销很大。这促成了摊销变分推断的提出（Rezende、Mohamed 和 Wierstra，2014年）：训练一个机器学习模型（通常是神经网络），使其能够直接预测该优化问题的解 $\boldsymbol{\theta}^*$ 。

具体来说，令 $\Lambda_{\boldsymbol{\phi}}$ 表示一个由 $\boldsymbol{\phi}$ 参数化的神经网络，它将单个数据点$\boldsymbol{x}_i$ 映射为局部变分后验 $\boldsymbol{\theta}_i$ 的相应参数$q_{\boldsymbol{\theta}_i}(\boldsymbol{z}_i)$。训练该模型时，使用式18 中的表达式作为损失函数，但将 $\boldsymbol{\theta}_i = \Lambda_{\boldsymbol{\phi}}(\boldsymbol{x}_i)$替换掉。请注意，尽管目标函数相同，这仍是一种摊销优化（Amos，2022年），因为现在我们使用 $\boldsymbol{\phi}$ 而非 $\boldsymbol{\theta}$作为优化变量。此外，损失函数是对数据点求和，因此可以使用训练神经网络的标准方法（如随机梯度下降）。在变分自编码器中，模型 $\Lambda_{\boldsymbol{\phi}}$ 称为编码器，并配有一个联合训练的解码器，对应于概率分布 $p(\boldsymbol{x} \mid \boldsymbol{z})$ （Kingma 和 Welling，2019年）。

黑盒变分推断

重参数化技巧可以通过自动微分计算精确梯度，这无疑十分方便。另一方面，许多模型无法进行重参数化。在这些情况下，可以改用黑盒变分推断（BBVI）（Ranganath、Gerrish 和 Blei，2014年）来估计梯度。这种方法适用范围更广，同时仍然方便。不过，BBVI 估计器的方差较高。

BBVI 基于这样的观测：

$$\nabla_{\boldsymbol{\theta}} \log q_{\boldsymbol{\theta}}(\boldsymbol{z}) = \frac{\nabla_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\boldsymbol{z})}{q_{\boldsymbol{\theta}}(\boldsymbol{z})}, \tag{19}$$

这有时也称为 REINFORCE 技巧（Williams，1992年）。利用这一点，可以将方程 14 改写为

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z})} \left[\log p(\boldsymbol{x}, \boldsymbol{z}) - \log q_{\boldsymbol{\theta}}(\boldsymbol{z}) \right] &= \int \left(\log p(\boldsymbol{x}, \boldsymbol{z}) - \log q_{\boldsymbol{\theta}}(\boldsymbol{z}) \right) \underbrace{q_{\boldsymbol{\theta}}(\boldsymbol{z}) \nabla_{\boldsymbol{\theta}} \log q_{\boldsymbol{\theta}}(\boldsymbol{z})}_{= \nabla_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\boldsymbol{z})} \, d\boldsymbol{z} \\ &= \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z})} \left[\left(\log p(\boldsymbol{x}, \boldsymbol{z}) - \log q_{\boldsymbol{\theta}}(\boldsymbol{z}) \right) \nabla_{\boldsymbol{\theta}} \log q_{\boldsymbol{\theta}}(\boldsymbol{z}) \right] \approx \frac{1}{L} \sum_{i=1}^{L} \left(\log p(\boldsymbol{x}, \boldsymbol{z}^{(i)}) - \log q_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)}) \right) \nabla_{\boldsymbol{\theta}} \log q_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)}). \end{aligned} \tag{20}$$

由于我们通常可以使用自动微分来评估得分函数 $\nabla_{\boldsymbol{\theta}} \log q_{\boldsymbol{\theta}}(\boldsymbol{z})$，因此这一重新表述似乎解决了从样本中估计证据下界梯度的问题。然而，这种估计量的方差通常过高，因而在实践中并不实用。可以说，黑盒变分推断的关键贡献在于将两种方差缩减技术——Rao-Blackwell化和控制变量——应用于期望中的估计量20。不过，详细介绍这些方差缩减技术将超出变分推断基础教程的范围。感兴趣的读者可参阅 Ranganath、Gerrish 和 Blei 的原始工作（2014年）。

致谢

这项工作得到了由 Knut and Alice Wallenberg Foundation 资助的 Wallenberg AI、自主系统与软件计划（WASP）的部分支持。本预印本使用 LaPreprint 模板创建（https://github.com/roaldarbol/lapreprint），该模板由 Mikkel Roald-Arbøl 制作。

参考文献

Amos, Brandon（2022年）。“用于学习在连续域上进行优化的摊销优化教程”。载于：*arXiv* 预印本 *2202.00665*。Blei, David M、Alp Kucukelbir 和 Jon D. McAuliffe（2017年）。“变分推断：统计学家综述”。载于：《美国统计学会杂志》112.518，第859–877页。Devroye, Luc（1996年）。“一行代码中的随机变量生成”。载于：冬季仿真会议论文集。IEEE，第265–272页。Kingma, Diederik P 和 Max Welling（2014年）。“自编码变分贝叶斯”。载于：第*2*届国际学习表征会议。—（2019年）。“变分自编码器简介”。载于：《机器学习基础与趋势》®《机器学习》12.4，第307–392页。Ranganath, Rajesh、Sean Gerrish 和 David M. Blei（2014年）。“黑盒变分推断”。载于：人工智能与统计学。PMLR，第814–822页。Rezende, Danilo Jimenez、Shakir Mohamed 和 Daan Wierstra（2014年）。“深度生成模型中的随机反向传播与近似推断”。载于：国际机器学习会议。PMLR，第1278–1286页。Schulman, John 等（2015年）。“使用随机计算图进行梯度估计”。载于：神经信息处理系统进展28。Wainwright, Martin J 和 Michael I. Jordan（2008年）。“图模型、指数族与变分推断”。载于：《机器学习基础与趋势》1.1–2，第1–305页。Williams, Ronald J（1992年）。“用于联结主义强化学习的简单统计梯度跟随算法”。载于：《机器学习》8.3，第229–256页。