强化学习和变分推理

0. 来源（鸣谢！！！）

https://mbernste.github.io/posts/elbo/

https://mbernste.github.io/posts/variational_inference/

https://mbernste.github.io/posts/reparameterization_vi/

https://andrew-cr.github.io/posts/RLandVI/

https://zhuanlan.zhihu.com/p/614237272

1. ELBO的由来

参考网站：https://mbernste.github.io/posts/elbo/

假设：存在两个随机变量 X 和 Z；存在随机变量 X 的观测值 x。

X 和 Z 满足联合概率分布： $p(X, Z; \theta)$，其中 θ 是表达这个分布的参数（例如：正态分布就是均值和方差）。

现在的问题是：只存在随机变量 X 的观测值 x，并没有随机变量 Z 的观测值，也就是 Z 没有被观测到（remain unobserved），可以称 Z 为潜变量（latent variable）。

对于只有**部分观测值**的联合概率分布 $p(X, Z; \theta)$，我们能解决两个问题：

1. θ 给定，也就是知道这个联合概率分布，在这种情况下求随机变量 Z 的后验分布：$p(Z|X; \theta)$；

   这个问题是变分推理（variational inference）的基础。
2. θ 未知，求解 θ 的极大似然估计。

$$
\begin{aligned}
&\arg\max_{\theta} l(\theta) \\
l(\theta) &= \log p(x; \theta) \\
&\left(\text{根据似然函数的定义}\right) \\
&= \log \int_z p(x, z; \theta)\, dz \\
&\left(\text{先假装添加Z的观测值z,然后对z做积分}\right)
\end{aligned} \tag{1}
$$

   这个问题是期望最大化（expectation-maximum）问题的基础。

“证据（evidence）”的概念是什么？

对于 $p(x;\theta)$，根据极大似然的思想，我们能拿到的观测值 x，说明在合适的模型 p 及其参数 θ 下，对应观测值 x 的边缘概率（marginal probability，此处不考虑隐变量 Z 何种情况）应该尽可能地大。

那么，我们就可以用 $p(x;\theta)$ 来量化我们选取的模型 p 和参数 θ “对不对” / “合适不合适”。

如果选取的模型和参数不合适，那么 $p(x;\theta)$ 就会很小，不符合极大似然的思想；反之，合适的话，$p(x;\theta)$ 就会很大，符合极大似然的思想。

由此可以下定义：$\text{evidence} = \log p(x;\theta)$，加对数是让后续更好算，不会影响本质性质的变化。

但是，如果我们对隐变量 Z 一点都不知道的话，这样的“证据（evidence）”会无从下手。

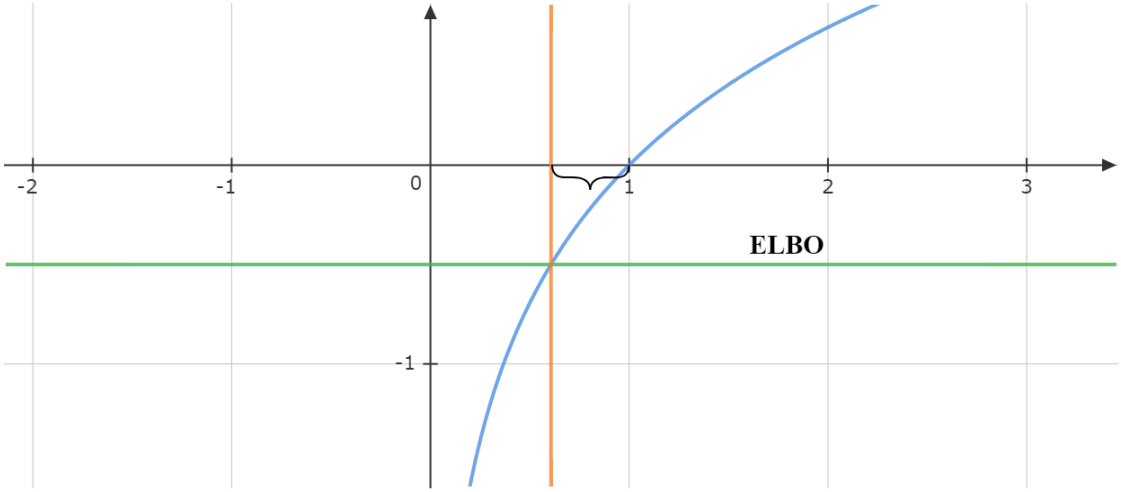
如果我们事先知道隐变量 Z 服从一个概率分布 q，那么 $p(x,z;\theta) = p(x|z;\theta)q(z)$，等式两边同时除以 $q(z)$，得到：$p(x|z;\theta) = \frac{p(x,z;\theta)}{q(z)}$。两边取对数，并对 z 求期望，就可以得到证据下界（ELBO）。

$$
\begin{aligned}
\log p(x;\theta) &= \log \int_z p(x,z;\theta)\, dz \\
&\left(\text{先假装添加Z的观测值z,然后对z做积分}\right) \\
&= \log \int_z p(x|z;\theta) q(z)\, dz \\
&\left(\text{条件概率公式展开}\right) \\
&= \log \int_z \frac{p(x,z;\theta)}{q(z)} q(z)\, dz \\
&\left(\text{把除以q(z)后的等式替换进来}\right) \\
&= \log E_{Z\sim q}\left[\frac{p(x,z;\theta)}{q(z)}\right] \\
&\left(\text{用期望符号表达}\right) \\
&\geq E_{Z\sim q}\left[\log \frac{p(x,z;\theta)}{q(z)}\right] \\
&\left(\text{根据琴声不等式,log是上凸函数}\right) = ELBO
\end{aligned}
\tag{2}
$$

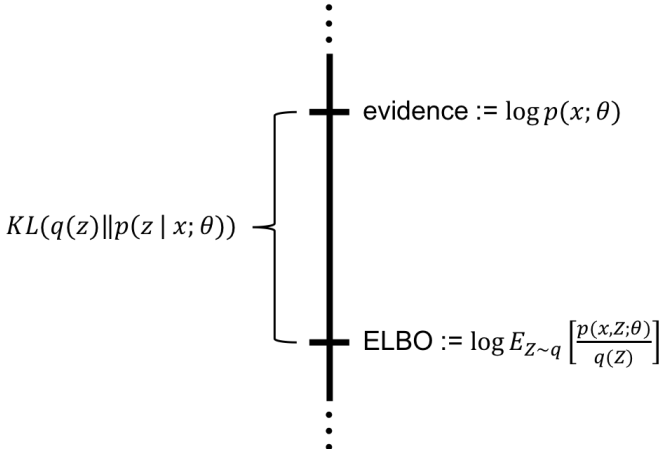
另外“证据下界（ELBO）”还能这样推导，其实也是大同小异：

$$
\begin{aligned}
\text{evidence} &= \log p(x;\theta) \\
&\left(\text{根据刚才的定义}\right) \\
&= \log \int_z p(x,z;\theta)\, dz \\
&\left(\text{先假装添加Z的观测值z,然后对z做积分}\right) \\
&= \log \int_z p(x,z;\theta) \frac{q(z)}{q(z)}\, dz \\
&\left(\text{分子分母同时加q(z)不改变等式}\right) \\
&= \log \int_z p(x,z;\theta) \frac{1}{q(z)} \cdot q(z)\, dz \\
&\left(\text{写成期望公式的连续型版本}\right) \\
&= \log E_{Z\sim q}\left[\frac{p(x,Z;\theta)}{q(z)}\right] \\
&\left(\text{用期望符号表达}\right) \\
&\geq E_{Z\sim q} log\left[\frac{p(x,Z;\theta)}{q(z)}\right] \\
&\left(\text{根据琴声不等式,log是上凸函数}\right) = ELBO
\end{aligned}
\tag{3}
$$

我的理解：从对数曲线图中可以看出，ELBO把对数概率限制在了 $[e^{ELBO}, 1]$ 之间。因此，如果要想极大化证据 $\log p(x;\theta)$，那么就要“抬高” ELBO，也就是极大化变分下界。



证据下界（ELBO）与 KL 散度之间是什么关系?



KL 散度是用来衡量证据（evidence）和证据下界（ELBO）之间的数值差。

$$
\begin{aligned}
KL(q(z) \ || \ p(z \mid x;\theta)) &= \sum_z q(z) \log \frac{q(z)}{p(z \mid x;\theta)} \\
&= E_{Z\sim q}\left[\log \frac{q(Z)}{p(Z \mid x;\theta)}\right] \\
&= E_{Z\sim q}\left[\log q(Z)\right] - E_{Z\sim q}\left[\log \frac{p(x,Z;\theta)}{p(x;\theta)}\right] \\
&= E_{Z\sim q}\left[\log q(Z)\right] - E_{Z\sim q}\left[\log p(x,Z;\theta)\right] + E_{Z\sim q}\left[\log p(x;\theta)\right] \\
&= \log p(x;\theta) - E_{Z\sim q}\left[\log \frac{p(x,Z;\theta)}{q(z)}\right] \\
&= \text{evidence} - \text{ELBO}
\end{aligned}
\tag{4}
$$

从这个推导式中可以得到等式：

$$
\text{evidence} = KL + ELBO \tag{5}
$$

这就意味着，极大化证据下界 $ELBO$ 等于极小化 KL 散度，这就是在强化学习中经常会出现的两个优化方向。

2. 变分推理

变分推断是用于，在计算后验分布明确不可行时，对后验分布的估计。

假设：存在两个随机变量 X 和 Z ，以及一个关于这两个变量的联合概率密度 $p(X,Z)$ 。

假设：随机变量 X 的观测数据可获得 x ， Z 是模型内部的隐藏随机变量。

问题：计算关于随机变量 Z 的后验分布 $p(Z|X)$ 。

解法：最理想的方式是使用贝叶斯定理。 x 和 z 分别是 X 和 Z 的观测值。 $p(\cdot)$ 是两个随机变量的边缘概率分布

$$p(z|x) = \frac{p(x,z)}{p(x)} = \frac{p(x|z)p(z)}{p(x)} \tag{6}$$

贝叶斯定理的方案太难算了，很难知道 $p(\cdot)$ 是什么样的形式。

因此我们会采用逼近/近似的方式来获得后验概率分布 $p(z|x)$ 。

我们的想法是：寻找一个新的分布 $q(z)$ ，通过某种度量手段计算 $q(z)$ 和 $p(z|x)$ 之间的差异，然后我们认为：如果 $q(z)$ 和 $p(z|x)$ 之间的差异很小很小，那么我们就可以用 $q(z)$ 的分布取代 $p(z|x)$ 。

找分布是一件很困难的事情，因为有很多分布（正态分布、均匀分布、指数分布，等等），此外每类分布都有自己的待定参数（对正态分布就是均值和方差）。我们的做法是：分布类型就固定为一类，比如论文中经常固定成（高维度）正态分布。假定正态分布的参数统一记为 ϕ ，那么不同 ϕ （变分参数，**variational parameter**）就有不同的分布，最后就是一族分布（变分分布族，**variational distribution family**）记为 Q 。

我们的任务就是：从变分分布族 Q 中找到一个最好的、由变分参数 $\hat{\phi}$ 刻画的具体分布 $q(z|\hat{\phi})$ ，使得 $q(z;\hat{\phi})$ 与理论上的后验分布 $p(z|x)$ 非常近似 / 最近似。

如何实现？

上面提到：“某种度量手段”，在这里指的是 KL 散度，这里写下表达式。

$$KL(q(z)||p(z|x)) = E_{Z\sim q}[\log \frac{q(Z)}{P(Z|x)}] \tag{7}$$

那么变分推理的目标是：

$$\hat{q} = \arg\min_q KL(q(z)||p(z|x)) \tag{8}$$

最后返回的是最匹配/最接近的分布 $\hat{q}(z|\hat{\phi}) = \hat{q}(z;\hat{\phi})$

根据等式 (5) ，最小化 KL 散度等价于最大化 $ELBO$ 证据下界。因此就得到了另一个优化式子：

$$\begin{aligned} \hat{q} &= \arg\max_q ELBO(q) \\ &= \arg\max_q E_{Z\sim q}[\log p(x,Z)] - E_{Z\sim q}[\log q(Z)] \end{aligned} \tag{9}$$

回忆一下 $\hat{q}$ ： 在 KL 散度度量下， $\hat{q}$ 与 $p(z|x)$ 是最接近的 / 差异最小的。

3. 基于重参数化梯度方法的黑盒变分推理

这一部分主要解决如何计算等式 (9) 的优化问题。

能实现的效果是：只需要输入 p 和 q 的分布表达式，算法就会自动执行变分推理。

重参数化梯度方法是一种用于在 $ELBO$ 上执行**随机梯度上升**的方法，它用重参数化技巧，重写 $ELBO$ 表达式。

如果选定了确定的变分分布族 Q ，那么分布 q 可以用它的参数 ϕ 刻画，重写等式 (9) 的优化问题。

$$\hat{\phi} = \arg\max_\phi ELBO(\phi) \tag{10}$$

对于这种优化问题，一个非常自然的想法是延梯度上升的方向最大化，那么就可以写出关于 ϕ 的表达式。

$$\phi_{t+1} \leftarrow \phi + \alpha \nabla_\phi ELBO(\phi)|_{\phi_t} \tag{11}$$

很明显， α 表示学习率，这样的迭代直到收敛为止。

现在，问题变成了，我们如何计算 $ELBO$ 的梯度?

这里的一个关键挑战是处理 $ELBO$ 中的期望（抑或是连续场景中的积分）。

重参数化梯度方法，通过计算**可行的随机梯度**，而不是计算**不可行的精确梯度**，来执行随机梯度上升，从而解决了这个挑战。

随机梯度上升的工作原理如下：我们不计算 $ELBO$ 对 ϕ 的精确梯度，而是制定一个随机变量 $V(\phi)$，其期望值是在 ϕ 处 $ELBO$ 的梯度，即：

$$E[V(\phi)] = \nabla_\phi ELBO(\phi) \tag{12}$$

那么，在第 t 次迭代中，我们从 $V(\phi_t)$ 中采样近似梯度，并朝着这个随机梯度的方向迈出一小步：

$$\begin{aligned} V &\sim V(\phi) \\ \phi_{t+1} &= \phi_t + \alpha \cdot V \end{aligned} \tag{13}$$

现在的问题是，我们如何制定一个分布 $V(\phi)$，其期望值是 $ELBO$ 的梯度 $\nabla_\phi ELBO(\phi)$？

重新参数化技巧将使得制定这样一个分布成为可能。

一种思考方式是，我们不直接从后验分布 $q_\phi(z)$ 中采样 z，而是“重新设计” z 的生成过程。

首先采样一个替代随机变量 ϵ，然后将 ϵ 转换为 z，同时确保最终 z 的分布仍然遵循 q_ϕ。至关重要的是， D 必须是我们可以轻松从中采样的东西，例如标准正态分布。

$$\epsilon \sim D \tag{14}$$

$$z = g_\phi(\epsilon) \tag{15}$$

重参数化 $q_\phi(z)$ 有时可能很困难，但是如果使用正确的变分分布族 Q，它可以变得很容易。例如，如果 $q_\phi(z)$ 是一个 “位置-尺度” 族分布，那么重新参数化就变得非常简单。

> “位置-尺度” 族是一类由**位置参数**和**非负尺度参数**进行参数化的概率分布族。
>
> - 对于**任何属于该族的**随机变量，它们的**分布函数**也属于该族。
> - 位置-尺度族通常限制在具有相同函数形式的分布上。
>
> 包括正态分布、柯西分布和 t 分布等常见的位置-尺度分布族。

经典的重参数化案例如下：

$$\begin{cases} q_\phi(z) = N(\mu, \sigma^2), \phi = \{\mu, \sigma\} \\ \epsilon \sim N(0,1) \rightarrow z = g_\phi(\epsilon) = \mu + \sigma\epsilon \end{cases} \tag{16}$$

那么，重参数化技巧和 $ELBO$ 如何产生联系呢?

基于重参数化技巧，可以将 $ELBO$ 重写出新的表达式：

$$\begin{aligned} ELBO(q_\phi) =& E_{Z\sim q_\phi}[\log p(x, Z)] - E_{Z\sim q_\phi}[\log q_\phi(Z)] \\ =& E_{\epsilon\sim D}[\log p(x, g_\phi(\epsilon))] - E_{\epsilon\sim D}[\log q_\phi(g_\phi(\epsilon))] \end{aligned} \tag{17}$$

就是把原本带 Z 的变量重新改成 $\epsilon \sim D, Z = g_\phi(\epsilon)$ 就行，不难的~

公式 (17) 使我们能够通过蒙特卡罗抽样来近似计算 $ELBO$ 。

首先，从分布 D 中采样 L 个代理随机变量：

$$\epsilon_1, \epsilon_2, \cdots, \epsilon_L \sim D \tag{18}$$

那么，我们就可以计算 $ELBO$ 的蒙特卡罗近似：

$$\begin{aligned} ELBO(q_\phi) &= E_{\epsilon \sim D}[\log p(x, g_\phi(\epsilon))] - E_{\epsilon \sim D}[\log q_\phi(g_\phi(\epsilon))] \\ \hat{ELBO}(q_\phi) &= \frac{1}{L} \sum_{i=1}^{L} \big[\log p(x, g_\phi(\epsilon_i)) - \log q_\phi(g_\phi(\epsilon_i)) \big] \end{aligned} \tag{19}$$

蒙特卡洛近似，在讲强化学习基础的时候也用过。我的理解，就是通过采用把期望变成样本的均值。

只要 p 是连续型概率分布，且 g_ϕ 对 ϕ 是连续函数，也就是保证都可导的情况下，可以求出基于蒙特卡罗抽样的 $ELBO$ 的梯度值：

$$\nabla_\phi \hat{ELBO}(q_\phi) = \nabla_\phi \frac{1}{L} \sum_{i=1}^{L} \big[\log p(x, g_\phi(\epsilon_i)) - \log q_\phi(g_\phi(\epsilon_i)) \tag{20}$$

基于蒙特卡罗抽样的 $ELBO$ 的梯度值，其本质是：自变量为 L 个代理的随机变量的函数，因此基于蒙特卡罗抽样的 $ELBO$ 的梯度值也是随机变量！

因此回到了最初的起点~

$$E[\nabla_\phi \hat{ELBO}(q_\phi)] = \nabla_\phi ELBO(\phi) \tag{21}$$

总之：从 D 中采样 $\epsilon_1, \cdots, \epsilon_L$ ，计算蒙特卡洛近似 $ELBO$ ，然后对这个近似计算梯度，等价于从随机梯度分布 $V(\phi)$ 中采样，其期望是 $ELBO$ 的梯度。

4. 变分参数和模型参数的共同优化

在许多情况下，我们不仅拥有具有潜在变量 z 的模型，还拥有模型参数 θ 。

也就是说，我们的联合分布 $p(x, z)$ 由一些参数 θ 参数化。因此，我们将完整的联合分布表示为 $p_\theta(x, z)$ 。

在这种情况下，如果我们不知道 θ 的真实值，我们如何估计后验分布 $p_\theta(z|x)$ ？

一种想法是在 θ 上放置一个先验分布，并将 θ 视为类似于 z 的潜变量（即，让 z 包括潜变量和模型参数）。

然而，这可能并不总是理想的。首先，我们可能不需要 θ 的完整后验分布。此外，估计后验分布是具有挑战性的！在同时估计 $p_\theta(z|x)$ 的情况下，是否有可能得出 θ 的点估计值?

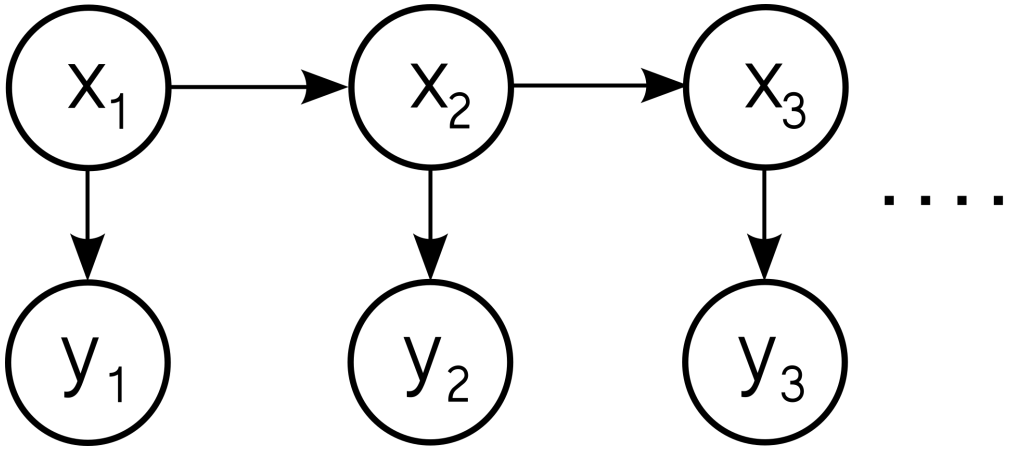
事实证明，可以通过同时针对变分参数 ϕ 和模型参数 θ 最大化 $ELBO$ 来推断 θ 。

$$
\begin{aligned}
\hat{\theta}, \hat{\phi} &= \arg\max_{\theta,\phi} \ ELBO(\theta, \phi) \\
&= \arg\max_{\theta,\phi} \ E_{Z\sim q_\phi}[\log p_\theta(x, Z) - \log q_\phi(Z)] \\
&\left(\text{将模型参数和变分参数带入可得}\right) \\
&= \arg\max_{\theta,\phi} \ E_{\epsilon\sim D}[\log p_\theta(x, g_\phi(\epsilon)) - \log q_\phi(g_\phi(\epsilon))] \\
&\left(\text{使用重参数化技巧}\right) \\
&= \arg\max_{\theta,\phi} \ \frac{1}{L}\sum_{i=1}^{L}\left[\log p_\theta(x, g_\phi(\epsilon_i)) - \log q_\phi(g_\phi(\epsilon_i))\right] \\
&\left(\text{使用蒙特卡洛抽样}\right)
\end{aligned}
\tag{22}
$$

最后一个等式与前面的内容呼应了：“只需要输入 p 和 q 的分布表达式，算法就会自动执行变分推理。”甚至在这里，不需要输入 p 和 q 的分布表达式，只要会采样就可以了。

5. 变分推理在强化学习中的应用

Part (1) 采用强化学习的价值函数做替代近似



根据这个图，我们可以先定义观测分布（纵向）：这个 $g(y_t|x_t)$ 的意思是，x_t 存在时 y_t 的观测值，可以从“以 x_t 为条件的 y_t 的概率分布 g”抽样出来。

$$
y_t|x_t \sim g(y_t|x_t) \tag{23}
$$

再定义转移分布（横向）：

$$
x_{t+1}|x_t \sim f(x_{t+1}|x_t) \tag{24}
$$

通过表达式 (23) 和表达式 (24) 可以得到“从过去到现在的”概率分布：

$$
\begin{aligned}
p(x_{1:T}, y_{1:T}) &= p(x_1|x_0)p(y_1|x_1)p(x_2|x_1)p(y_2|x_2)\cdots p(x_T|x_{T-1})p(y_T|x_T) \\
&= f(x_1|x_0)g(y_1|x_1)f(x_2|x_1)g(y_2|x_2)\cdots f(x_T|x_{T-1})g(y_T|x_T) \\
&= \prod_{i=1}^{T} f(x_i|x_{i-1})g(y_i|x_i)
\end{aligned}
\tag{25}
$$

问题是：我们希望在已观测到 $y_1, y_2, \cdots, y_T = y_{1:T}$ 的情况下推断隐藏状态。也就是说，我们希望通过后验分布进行推理，也就是：

$$
p(x_{1:T}|y_{1:T}) = p(Z|X) \tag{26}
$$

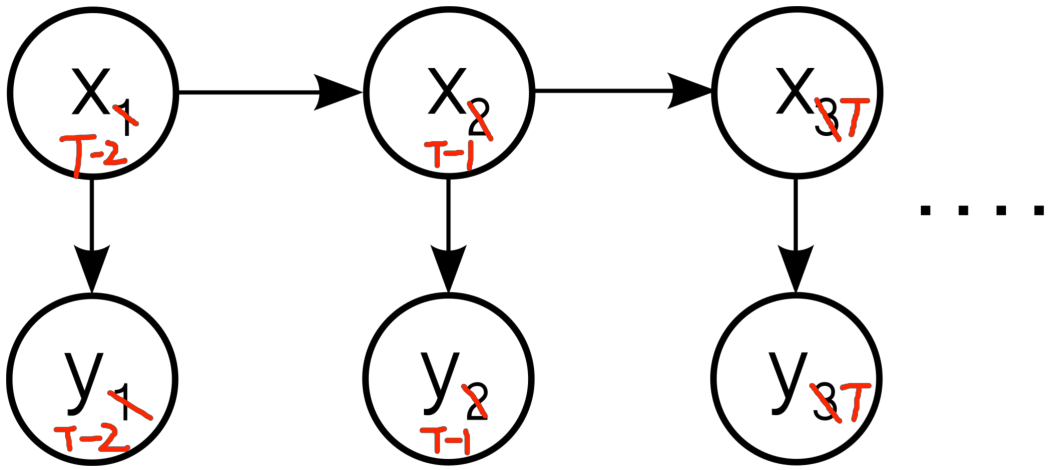
根据我们前面的理论：直接计算这个后验分布是困难的，那么我们采用变分推理的框架，拿来一个分布 $q(x_{1:T};\phi)$ 来近似我们的后验分布。那么，我们的目标就是最大化 $ELBO$ 。

$$\arg\max_{q_\phi} E_{x_{1:T}\sim q_\phi}[\log p(y_{1:T}, x_{1:T}) - \log q_\phi(x_{1:T})] \tag{27}$$

接着就可以采用我们之前提到的 “重参数化技巧 + 蒙特卡洛采样” 技术来求解梯度，甚至可以共同优化模型参数和变分参数。

但是，强化学习在环境中获得的序列长度不是固定的！

这就意味着：我们需要近似 $p(x_{1:T}|y_{1:T})$ 同时，近似 $p(x_{1:T+1}|y_{1:T+1})$ 和 $p(x_{1:T+2}|y_{1:T+2})$ 乃至后面的。



首先，我们需要对理论上的后验分布做一个纯粹的分解：

$$\begin{aligned}
p(x_{1:T}|y_{1:T}) &= p(x_1, x_2 \cdots x_T|y_{1:T}) \\
&\text{(展开来写)} \\
&= p(x_T|y_{1:T}) \cdot p(x_1, x_2 \cdots x_{T-1}|x_T, y_{1:T}) \\
&\text{(根据条件概率公式展开)} \\
&= p(x_T|y_{1:T}) \cdot p(x_1, x_2 \cdots x_{T-1}|x_T, y_{1:T-1}) \\
&\text{(与 } x_{1:T-1} \text{相关联的只有 } y_{1:T-1} \text{和 } x_T\text{,与 } y_T \text{无关)} \\
&= p(x_T|y_{1:T})p(x_{T-1}|x_T, y_{1:T-1}) \cdot p(x_1, x_2 \cdots x_{T-2}|x_T, x_{T-1}, y_{1:T-2}) \\
&\text{(根据条件概率公式展开)} \\
&= p(x_T|y_{1:T})p(x_{T-1}|x_T, y_{1:T-1})p(x_{T-2}|x_{T-1}, y_{1:T-2}) \\
&\cdot p(x_1, x_2 \cdots x_{T-3}|x_T, x_{T-1}, x_{T-2}, y_{1:T-3}) \\
&\text{(根据条件概率公式展开)} \\
&\cdots \\
&\text{(以此类推)} \\
&= p(x_T|y_{1:T})p(x_{T-1}|x_T, y_{1:T-1})p(x_{T-2}|x_{T-1}, y_{1:T-2}) \cdots p(x_1|x_2, y_1) \\
&\text{(得到最后的结果~)}
\end{aligned} \tag{28}$$

请注意，每个因子 $p(x_k|x_{k+1}, y_{1:k})$ 只取决于时间 k 之前的观测值，而不取决于未来观测值 $y_{k+1:T}$ 。

这是因为它被条件于 x_{k+1} ，而这个隐藏状态总结了所有未来的信息（未来的观测、未来的状态转移，根本都是由 x_{k+1} 产生的）。未来的观测值 $y_{k+1:T}$ 对于 x_k 的分布没有额外的信息（未来的观测根本都是由 x_{k+1} 产生的，也即是说， x_{k+1} 具有描绘未来观测的 “能力”）。

这很重要，因为我们可以在时间步 k 时用 $q_\phi(x_k|x_{k+1})$ 来近似计算 $p(x_k|x_{k+1}, y_{1:k})$ ：

$$q_\phi(x_k|x_{k+1}) \hat{=} p(x_k|x_{k+1}, y_{1:k}) \tag{29}$$

那么就可以化简得到：

$$q_\phi(x_{1:T}) = q_\phi(x_T)q_\phi(x_{T-1}|x_T) \cdots q_\phi(x_1|x_2) \tag{30}$$

当我们移动到下一个时间步时，我们将重复使用所有变分因子 $q_\phi(x_{T-1}|x_T)\cdots q_\phi(x_1|x_2)$ ，并在前面添加 $q_\phi(x_{T+1})q_\phi(x_T|x_{T+1})$ 。这样的好处就是可以一直做递推~~

我们现在已将计算工作负载减少到只需要优化两个变分因子 $q_\phi(x_{T+1})q_\phi(x_T|x_{T+1})$ ，但我们仍需要设置我们的目标。将上述变分分布简单地插入到 $ELBO$ 中，即可得到：

$$
\begin{aligned}
ELBO_T &= E_{q_\phi(x_{1:T})}[\log \frac{p(x_{1:T},y_{1:T})}{q_\phi(x_{1:T})}] \\
&\text{(按照定义展开)} \\
&= E_{q_\phi(x_{1:T})}\big[\log \frac{f(x_T|x_{T-1})g(y_T|x_T)f(x_{T-1}|x_{T-2})g(y_{T-1}|x_{T-1})\cdots f(x_1|x_0)g(y_1|x_1)}{q_\phi(x_T)q_\phi(x_{T-1}|x_T)\cdots q_\phi(x_1|x_2)}\big] \\
&\text{(分子按照公式(25)展开，分母按照公式(30)展开)} \\
&= E_{q_\phi(x_{1:T})}\big[\log \frac{f(x_T|x_{T-1})g(y_T|x_T)}{q_\phi(x_T)q_\phi(x_{T-1}|x_T)}\cdot\frac{p(x_{1:T-1},y_{1:T-1})}{q_\phi(x_{T-2}|x_{T-1})\cdots q_\phi(x_1|x_2)}\big] \\
&\text{(拆分这个分式，后面的余项又合并起来)} \\
&= E_{q_\phi(x_{1:T})}\big[\log \frac{f(x_T|x_{T-1})g(y_T|x_T)}{q_\phi(x_T)q_\phi(x_{T-1}|x_T)}\cdot\frac{p(x_{1:T-1},y_{1:T-1})}{q_\phi(x_{T-2}|x_{T-1})\cdots q_\phi(x_1|x_2)}\cdot\frac{q_\phi(x_{T-1})}{q_\phi(x_{T-1})}\big] \\
&\text{(分子分母同时添加一个}q_\phi(x_{T-1})\text{)} \\
&= E_{q_\phi(x_{1:T})}\big[\log \frac{f(x_T|x_{T-1})g(y_T|x_T)q_\phi(x_{T-1})}{q_\phi(x_T)q_\phi(x_{T-1}|x_T)}\cdot\frac{p(x_{1:T-1},y_{1:T-1})}{q_\phi(x_{T-1})q_\phi(x_{T-2}|x_{T-1})\cdots q_\phi(x_1|x_2)}\big] \\
&\text{(插入到适当的位置)} \\
&= E_{q_\phi(x_T)q_\phi(x_{T-1}|x_T)}\big[\log \frac{f(x_T|x_{T-1})g(y_T|x_T)q_\phi(x_{T-1})}{q_\phi(x_T)q_\phi(x_{T-1}|x_T)}+\log\frac{p(x_{1:T-1},y_{1:T-1})}{q_\phi(x_{T-1})q_\phi(x_{T-2}|x_{T-1})\cdots q_\phi(x_1|x_2)}\big] \\
&\text{(根据对数运算做拆分)} \\
&= E_{q_\phi(x_T)q_\phi(x_{T-1}|x_T)}\big[\log \frac{f(x_T|x_{T-1})g(y_T|x_T)q_\phi(x_{T-1})}{q_\phi(x_T)q_\phi(x_{T-1}|x_T)}\big]+E_{q_\phi(x_{1:T-1})}\big[\log\frac{p(x_{1:T-1},y_{1:T-1})}{q_\phi(x_{1:T-1})}\big] \\
&\text{(把期望拆开，并对分母做合并)} \\
&= E_{q_\phi(x_T)q_\phi(x_{T-1}|x_T)}\big[\log \frac{f(x_T|x_{T-1})g(y_T|x_T)q_\phi(x_{T-1})}{q_\phi(x_T)q_\phi(x_{T-1}|x_T)}\big]+E_{q_\phi(x_T)q_\phi(x_{T-1}|x_T)}\big[E_{q_\phi(x_{T-1})q_\phi(x_{T-2}|x_{T-1})\cdots q_\phi(x_1|x_2)}\log\frac{p(x_{1:T-1},y_{1:T-1})}{q_\phi(x_{T-1})q_\phi(x_{T-2}|x_{T-1})\cdots q_\phi(x_1|x_2)}\big] \\
&\text{(根据分母把第二项期望服从的分布展开)}
\end{aligned}
\tag{31}
$$

最后，我们得到了一个很重要的递推等式：

$$
\begin{aligned}
ELBO_T &= E_{q_\phi(x_{1:T})}[\log\frac{p(x_{1:T},y_{1:T})}{q_\phi(x_{1:T})}] \\
&= E_{q_\phi(x_T)q_\phi(x_{T-1}|x_T)}\big[\log\frac{f(x_T|x_{T-1})g(y_T|x_T)q_\phi(x_{T-1})}{q_\phi(x_T)q_\phi(x_{T-1}|x_T)}\big]+E_{q_\phi(x_{1:T-1})}\big[\log\frac{p(x_{1:T-1},y_{1:T-1})}{q_\phi(x_{1:T-1})}\big]
\end{aligned}
\tag{32}
$$

递推等式的关键就是 $E_{q_\phi(x_{1:T-1})}\big[\log\frac{p(x_{1:T-1},y_{1:T-1})}{q_\phi(x_{1:T-1})}\big]$ 的计算，我们的目标是要让这个式子不要包含那么多的信息。

变分分布 $q_\phi(x_{1:T-1})$ 只包含固定因子，因此期望内的项不会直接对新因子 $q_\phi(x_T)$ 和 $q_\phi(x_{T-1}|x_T)$ 的目标梯度产生贡献。然而，这个项取决于 x_{T-1} ，而 x_{T-1} 取决于新因子，因为 x_{T-1} 是从 $x_{T-1}\sim q_\phi(x_T)q_\phi(x_{T-1}|x_T)$ 中采样得到的。

因此，这个项将对目标的梯度产生贡献，不能忽略。我们需要一种避免直接计算这个项的方法，因为它需要每个时间步从 $T-1$ 回溯到 1 。换句话说，我们希望有一个只包含固定数量项且其评估成本不随时间增长的目标。

现在我们将看到如何使用强化学习的思想来实现这一点。

$$
V_t(s_t)=E_{s_{t+1},a_t\sim P(s_{t+1}|s_t,a_t)\pi\phi(a_t|s_t)}[r(s_t,a_t)+V_{t+1}(s_{t+1})] \tag{33}
$$

Bellman递归是 ”向后“ 操作的。也就是说，时间 t 的值函数取决于当前奖励和下一步 $t+1$ 的未来值函数。

但是我们的最终目的是，将这些想法应用于序列场景中的变分推理，其中我们只 ”向前“ 移动时间。‘

这需要一个向前递归，其中 “价值函数” 将取决于上一个时间步骤的过去价值函数。

接下来基于这个Bellman递归表达式做一些改写。

假如我们让智能体在环境中走了一个序列：

$$\tau = \{s_0, a_1, s_1, a_2, \cdots, s_{T-1}, a_T, s_T\} \tag{34}$$

那么我们可以知道最后状态 s_T 的概率 $p(s_T)$ ，那么整个序列的概率可以写成：

$$p_\phi(\tau) = p(s_T)\pi_\phi(a_T|s_T) \prod_{i=T-1}^{1} p(s_i|s_{i+1}, a_{i+1})\pi_\phi(a_i|s_i) \tag{35}$$

说白了就是把之前的序列反着写，状态转移和动作策略都是反着的。

$$V_t(s_t) = E\left[\sum_{k=1}^{r}(s_k, a_k)\right] \tag{36}$$

那么就可以得到前馈式的Bellman递归表达式：

$$V_t(s_t) = E_{\pi_\phi(a_t|s_t)P(s_{t-1}|s_t,a_t)}[r(s_t, a_t) + V_{t-1}(s_{t-1})] \tag{37}$$

我们现在终于准备好了，看看如何利用这个来解决我们之前的问题，其中我们的顺序变分推断目标中有一个项依赖于通过 x_{T-1} 的新变分因子，但从 $T-1$ 到 1 有大量项。

主要思想是定义一些“奖励”，使得从1到T的这些奖励的期望总和为 $ELBO_T$ 。然后我们可以复制我们的前向递归方程，得到一个“值函数”，它总结了先前 $ELBO$ 的一部分。因此，总体上，我们将获得一个目标，每个时间步只有2个项，而不是 T 个项。

我们首先将强化学习中的状态定义为我们模型中的隐藏状态 $s_t = x_t$。然后我们也将动作 a_t 定义为模型中前一个时间步的隐藏状态 $a_t = x_{t-1}$。

这似乎很奇怪，但它使我们的策略 $\pi_\phi(a_t|s_t)$ 对应于我们的变分因子之一 $\pi_\phi(a_t|s_t) = q_\phi(x_{t-1}|x_t)$。

由于状态和动作的奇怪定义，我们的状态转移函数变成了一个 δ 函数 $P(s_t|s_{t+1}, a_{t+1}) = \delta(s_t = a_{t+1})$ 。

$$\begin{aligned}
&P(s_t|s_{t+1}, a_{t+1}) \\
&= P(x_t|x_{t+1}, x_t) \\
&\left(\text{按照我们的定义改写一下}\right) \\
&= \delta(x_t = x_t) \\
&\left(\text{以}x_{t+1}, x_t\text{共同发生作为条件算}x_t\text{的概率，相当于}x_t\text{事件发生！}\right) \\
&= \delta(s_t = a_{t+1}) \\
&\left(\text{按照我们的定义再改写回来}\right)
\end{aligned} \tag{38}$$

现在我们可以定义我们的新“奖励”。

$$r(s_t, a_t) = r(x_t, x_{t-1}) = \log \frac{f(x_t|x_{t-1})g(y_t|x_t)q_\phi(x_{t-1})}{q_\phi(x_t)q_\phi(x_{t-1}|x_t)} \tag{39}$$

基于这个奖励，那么我们就可以将 $ELBO$ 改写成累计奖励的形式：

$$ELBO_T = E_{q_\phi(x_{1:T})}[r(x_T, x_{T-1}) + \cdots + r(x_2, x_1) + r(x_1, x_0))] \tag{40}$$

紧接着就可以定义我们的值函数：

$$
\begin{aligned}
V_{t-1}(x_{t-1}) &= E_{q_\phi(x_{1:t-2})q_\phi(x_{t-1})}[\sum_{k=1}^{t-1} r(x_k, x_{k-1})] \\
&\quad (\text{根据定义写出来}) \\
&= E_{q_\phi(x_{1:t-2})q_\phi(x_{t-1})}[\sum_{k=1}^{t-1} \log \frac{f(x_k|x_{k-1})g(y_k|x_k)q_\phi(x_{k-1})}{q_\phi(x_k)q_\phi(x_{k-1}|x_k)}] \\
&\quad (\text{代入公式}(39)) \\
&= E_{q_\phi(x_{1:t-2})q_\phi(x_{t-1})}[\log \frac{f(x_{t-1}|x_{t-2})g(y_{t-1}|x_{t-1})q_\phi(x_{t-2})\cdots f(x_1|x_0)g(y_1|x_1)q_\phi(x_1)}{q_\phi(x_{t-1})q_\phi(x_{t-2}|x_{t-1})\cdots q_\phi(x_1)q_\phi(x_0|x_1)}] \\
&\quad (\text{去掉累加符号，得到一个连乘表达式}) \\
&= E_{q_\phi(x_{1:t-2})q_\phi(x_{t-1})}[\log \frac{f(x_{t-1}|x_{t-2})g(y_{t-1}|x_{t-1})\cdots f(x_1|x_0)g(y_1|x_1)}{q_\phi(x_{t-1})q_\phi(x_{t-2}|x_{t-1})\cdots q_\phi(x_0|x_1)} \\
&\quad (\text{消去分子分母同时存在的} q\phi) \\
&= E_{q_\phi(x_{1:t-2})q_\phi(x_{t-1})}[\log \frac{p(x_{1:t-1}, y_{1:t-1})}{q_\phi(x_{1:t-1})}] \\
&\quad (\text{化简即可得到})
\end{aligned} \tag{41}
$$

最后我们就能得到关于我们定义的值函数的递推式：

$$
V_t(x_t) = E_{q_\phi(x_{t-1}|x_t)}[r(x_t, x_{t-1}) + V_{t-1}(x_{t-1})] \tag{42}
$$

最后就能将 $ELBO$ 化简成我们想要的形式：

$$
ELBO_T = E_{q_\phi(x_T)q_\phi(x_{T-1}|x_T)}[r(x_T, x_{T-1}) + V_{T-1}(x_{T-1})] \tag{43}
$$

我们可以立即看到，因为我们现在只有一个关于 x 和 $x-1$的目标。无论我们之前处理了多少观察结果，这将具有恒定的计算成本。我们需要做的就是跟踪我们的价值函数 $V_T(x_T)$，我们可以通过监督学习来实现。我们用 $\hat{V}_T(x_T)$ 来近似 $V_T(x_T)$，并创建的以下回归问题来学习 $\hat{V}_T(x_T)$。

$$
\min_{\hat{V}_T} E_{x_T, x_{T-1}} ||\hat{V}_T(x_T) - \big(r(x_T, x_{T-1}) + \hat{V}_{T-1}(x_{T-1})\big)||_2^2 \tag{44}
$$

Part (2) 元强化学习：隐变量任务推断

任务的推断通常使用变分推断或VAE来做。一个任务的信息用隐变量 z 来表示，而将这个任务的轨迹称为任务上下文 $\tau_{:N}$，目标是根据任务上下文推断出当前任务的信息，即求隐变量 z 的后验 $p(z|\tau_{:N})$。

$$
\begin{aligned}
ELBO(Z) &= E_{Z\sim q}[\log p(x, Z)] - E_{Z\sim q}[\log q(Z)] \\
&= E_{Z\sim q}[\log p(Z)] + E_{Z\sim q}[\log(x|Z)] - E_{Z\sim q}[\log q(Z)] \\
&= E_{Z\sim q}[\log(x|Z)] + KL(p(Z)||q(Z))
\end{aligned} \tag{45}
$$

由于p(x)的概率是固定的，即最大化 $ELBO$ 就等价于最小化 KL 散度。

由此，原问题就转化成了最大化 $ELBO$ 的问题，从 $ELBO$ 的式子来看，前一项是重建损失，即从隐变量 z 的假设分布 q 中抽样 Z 重新得到原来的 x 的似然；后一项则为正则损失，它把后验推向先验。

接着就可以采用我们之前提到的 “重参数化技巧 + 蒙特卡洛采样” 技术来求解梯度，甚至可以共同优化模型参数和变分参数。

用隐变量 z 来表示任务的信息，往往隐变量 z 和状态 s 一起输入到策略网络里：

$$
a \sim \pi(a|s, z) \tag{46}
$$

我们希望能从一部分的轨迹推断出一整个任务的信息，这样就能实现一边任务执行获得部分轨迹 $\tau_{:N}$ 的同时一边进行任务推断得到隐变量 z 的分布。我们采取的目标是：通过推断器 q_ϕ 推断出来的分布 $q_\phi(z|\tau_{:N})$ 要与完整历史轨迹 $\tau_{:H}$ 的分布 $p(z|\tau_{:H})$ 要接近，这里隐含 $N \leq H$ 的条件，可得：

$$
\begin{aligned}
&E_{\rho(Z,\tau_{:H})}\left[KL(q_\phi(z|\tau_{:N})||p(z|\tau_{:H}))\right] \\
&= E_{\rho(Z,\tau_{:H})}[E_q[\log q_\phi(z|\tau_{:N})] - E_q[\log p(z|\tau_{:H})]] \\
&\text{(根据KL散度的定义展开)} \\
&= E_{\rho(Z,\tau_{:H})}[E_q[\log q_\phi(z|\tau_{:N})] - E_q[\log \frac{p(z)p(\tau_{:H}|z)}{p(\tau_{:H})}]] \\
&\text{(根据条件概率公式展开)} \\
&= E_{\rho(Z,\tau_{:H})}[E_q[\log q_\phi(z|\tau_{:N})] - E_q[\log p(z)] - E_q[\log p(\tau_{:H}|z)] + E_q[\log p(\tau_{:H})]] \\
&\text{(根据对数公式展开)} \\
&= E_{\rho(Z,\tau_{:H})}[E_q[KL(q_\phi(z|\tau_{:N})||p(z)) + \log p(\tau_{:H}) - \log p(\tau_{:H}|z)] \\
&\text{(最后做简单的合并)}
\end{aligned} \tag{47}
$$

$\rho(Z, \tau_{:H})$ 表示一种联合分布。因为隐变量 z 是从分布 Z 中抽取出来的，轨迹 $\tau_{:H}$ 也是从分布中抽取出来的，因此两个可以组合成一种联合分布，才能对 (47) 求期望。

前一项是重建损失，它要求从一部分轨迹重建出MDP。对这部分损失我们可以用一些目标函数 $R(\rho, z)$ 来衡量，如累计回报，或者是SAC、PPO等算法里的损失函数的负数，因为从 $q_\phi(z|\tau_{:N})$ 中抽样得到隐变量 z 后，累计回报越大，损失函数越小越能表明隐变量 z 包含了更多任务的信息。

Part (3) 元强化学习：用VAE重建任务推断

最大化似然 $E_{\rho(Z,\tau_{:H})}[\log p_\phi(\tau_{:H})]$ 。

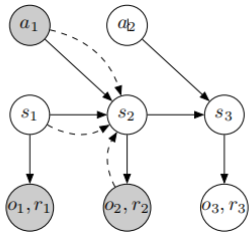
$$
\begin{aligned}
&E_{\rho(Z,\tau_{:H})}[\log p_\phi(\tau_{:H})] \\
&= E_{\rho(Z,\tau_{:H})}[\log \int_z p_\phi(\tau_{:H}, z) \frac{q_\phi(z|\tau_{:t})}{q_\phi(z|\tau_{:t})} dz] \\
&\text{(加入隐变量z,分子分母同时增加一个项} q_\phi(z|\tau_{:t})) \\
&= E_{\rho(Z,\tau_{:H})}[\log E_{q_\phi(z|\tau_{:t})}[\frac{p_\phi(\tau_{:H}, z)}{q_\phi(z|\tau_{:t})}]] \\
&\text{(写成期望的形式)} \\
&\geq E_{\rho(Z,\tau_{:H}),q_\phi}[\log \frac{p_\phi(\tau_{:H}, z)}{q_\phi(z|\tau_{:t})}] \\
&\text{(根据琴声不等式把期望符号提前)} \\
&= E_{\rho(Z,\tau_{:H}),q_\phi}[\log p_\phi(\tau_{:H}|z) + \log p_\phi(z) - \log q_\phi(z|\tau_{:t})] \\
&\text{(对数运算展开每一项)} \\
&= E_{\rho(Z,\tau_{:H}),q_\phi}[\log p_\phi(\tau_{:H}|z)] - E_{\rho(Z,\tau_{:H})}[KL(q_\phi(z|\tau_{:t})||p_\phi(z))] \\
&\text{(后面两项合并成KL散度)} \\
&= ELBO_t
\end{aligned} \tag{48}
$$

所以优化目标还是 $ELBO_t$ ，对 $ELBO_t$ 的第一项做进一步的计算：

$$
\begin{aligned}
&\log p_\phi(\tau_{:H}|z, a_{:H-1}) \\
&= \log p_\phi(s_0, r_0, s_1, r_1, \cdots, s_{t-1}, r_{t-1}, s_t|z, a_{:H-1}) \\
&\text{(把概率展开来写)} \\
&= \log p_\phi(s_0|z) + \\
&\text{(原始状态只与隐变量z有关)} \\
&\sum_{i=0}^{H-1}[\log p_\phi(s_{i+1}|s_i, a_i, z) + \log p_\phi(r_{i+1}|s_i, a_i, z, s_{i+1})] \\
&\text{(下一个状态} s_{i+1} \text{只与当前状态、动作和隐变量有关)} \\
&\text{(下一个奖励} r_{i+1} \text{只与当前状态、下一个状态、动作和隐变量有关)}
\end{aligned} \tag{49}
$$

Part (4) 多模态强化学习：Planet 变分公式推导

首先介绍一下 SSM 的公式。



先标注状态转移模型、观测模型和奖励模型：

$$\begin{aligned} s_t &\sim p(s_t|s_{t-1}, a_{t-1}) \\ o_t &\sim p(o_t|s_t) \\ r_t &\sim p(r_t|s_t) \end{aligned} \tag{50}$$

那么，我们可以推导出基于动作序列的状态动力学模型：

$$p(o_{1:T}, s_{1:T}|a_{1:T}) = \prod_{t=1}^{T} p(s_t|s_{t-1}, a_{t-1}) p(o_t|s_t) \tag{51}$$

我们之所以注意“以 $a_{1:T}$ 条件”，是因为我们的隐藏信息 $s_{1:T}$ 和观测信息 $o_{1:T}$ 能发生变化的原因就是 $a_{1:T}$ 的发生！那么我们的后验分布就是想知道隐藏信息 $s_{1:T}$ 的情况，“以动作信息 $a_{1:T}$ 和观测信息 $o_{1:T}$ 为条件（我们能记录的信息）”，求隐藏信息 $s_{1:T}$ 的分布。

我们在这里设置一个后验分布：

$$\begin{aligned} q(s_{1:T}|o_{1:T}, a_{1:T}) &= \prod_{t=1}^{T} q(s_t|o_{1:t}, a_{1:t}) \\ &= q(s_1|o_1, a_1) \cdot q(s_2|o_1, a_1, o_2, a_2) \cdots q(s_t|o_1, a_1 \cdots o_t, a_t) \end{aligned} \tag{52}$$

对于这个分布 $p(o_{1:T}, s_{1:T}|a_{1:T})$，暂时不看动作信息 $a_{1:T}$，记观测信息 $o_{1:T}$ 为 X，隐藏信息 $s_{1:T}$ 为 Z；那么分布变成了 $p(X, Z)$。那么 $p(Z|X)$ 就是我们的后验分布，就可以用我们之前的 $ELBO$！参考公式 $(9)(26)$。

$$
\begin{aligned}
\ln p(o_{1:T}|a_{1:T}) &= \ln \int_S p(o_{1:T}, s_{1:T}|a_{1:T}) ds_{1:T} \\
&\left(\text{补充一项然后做积分}\right) \\
&= \ln \int_S p(o_{1:T}, s_{1:T}|a_{1:T}) \frac{q(s_t|o_{1:t}, a_{1:t-1})}{q(s_t|o_{1:t}, a_{1:t-1})} ds_{1:T} \\
&\left(\text{重要性采样法}\right) \\
&= \ln E_{q(s_t|o_{1:t}, a_{1:t-1})} \left[\frac{p(o_{1:T}, s_{1:T}|a_{1:T})}{q(s_t|o_{1:t}, a_{1:t-1})} \right] \\
&\left(\text{写成带有期望符号的形式}\right) \\
&= \ln E_{q(s_t|o_{1:t}, a_{1:t-1})} \left[\frac{\prod_{t=1}^{T} p(s_t|s_{t-1}, a_{t-1}) p(o_t|s_t)}{q(s_t|o_{1:t}, a_{1:t-1})} \right] \\
&\left(\text{分子用公式}(51)\text{带入}\right) \\
&\geq E_{q(s_t|o_{1:t}, a_{1:t-1})} \left[\ln \frac{\prod_{t=1}^{T} p(s_t|s_{t-1}, a_{t-1}) p(o_t|s_t)}{q(s_t|o_{1:t}, a_{1:t-1})} \right] \\
&\left(\text{采用琴声不等式}\right) \\
&= E_{q(s_t|o_{1:t}, a_{1:t-1})} \left[\sum_{t=1}^{T} \ln p(o_t|s_t) + \ln p(s_t|s_{t-1}, a_{t-1}) - \ln q(s_t|o_{1:t}, a_{1:t-1}) \right] \\
&\left(\text{采用对数变形}\right) \\
&= \sum_{t=1}^{T} \left(E[\ln p(o_t|s_t)] - KL(q(s_t|o_{1:t}, a_{1:t-1}) || p(s_t|s_{t-1}, a_{t-1})) \right) \\
&\left(\text{合并得到结果}\right)
\end{aligned} \tag{53}
$$

这个论文还有一个多步预测：

$$
\begin{aligned}
p(s_t|s_{t-d}) &= \int_S \int_A p(s_t|s_{t-1}, a_{1:t-1}) p(s_{t-1}|s_{t-2}, a_{1:t-2}) \cdots p(s_{t-d+1}|s_{t-d}, a_{1:t-d}) ds_{t-d+1:t} d_{a_{1:i-1}} \\
&\left(\text{根据状态转移模型展开}\right) \\
&= \int_S \int_A p(s_t|s_{t-1}, a_{1:t-1}) \cdot p(s_{t-1}|s_{t-2}, a_{1:t-2}) \cdots p(s_{t-d+1}|s_{t-d}, a_{1:t-d}) ds_{t-d+1:t} d_{a_{1:i-1}} \\
&\left(\text{保留第一项,合并后面的若干项}\right) \\
&= \int_S \int_A p(s_t|s_{t-1}, a_{1:t-1}) \cdot p(s_{t-1}|s_{t-d}) ds_{t-d+1:t} d_{a_{1:i-1}} \\
&\left(\text{保留第一项,合并后面的若干项}\right) \\
&= E_{p(s_{t-1}|s_{t-d})} [p(s_t|s_{t-1}, a_{1:t-1})] \\
&\left(\text{写成带有期望符号的形式}\right)
\end{aligned} \tag{54}
$$

最后得到多步预测的 $ELBO$：

$$
\begin{aligned}
\ln p(o_{1:T}|a_{1:T}) &= \ln \int_S p(o_{1:T}, s_{1:T}|a_{1:T}) ds_{1:T} \\
&\text{(补充一项然后做积分)} \\
&= \ln \int_S p(o_{1:T}, s_{1:T}|a_{1:T}) \frac{q(s_t|o_{1:t}, a_{1:t-1})}{q(s_t|o_{1:t}, a_{1:t-1})} ds_{1:T} \\
&\text{(重要性采样法)} \\
&= \ln E_{q(s_t|o_{1:t}, a_{1:t-1})} \Big[\frac{p(o_{1:T}, s_{1:T}|a_{1:T})}{q(s_t|o_{1:t}, a_{1:t-1})} \Big] \\
&\text{(写成带有期望符号的形式)} \\
&= \ln E_{q(s_t|o_{1:t}, a_{1:t-1})} \Big[\frac{\prod_{t=1}^{T} p(s_t|s_{t-d}, a_{t-d-1:t-1}) p(o_t|s_t)}{q(s_t|o_{1:t}, a_{1:t-1})} \Big] \\
&\text{(分子用公式(54)带入)} \\
&\geq E_{q(s_t|o_{1:t}, a_{1:t-1})} \Big[\ln \frac{\prod_{t=1}^{T} p(s_t|s_{t-d}, a_{t-d-1:t-1}) p(o_t|s_t)}{q(s_t|o_{1:t}, a_{1:t-1})} \Big] \\
&\text{(采用琴声不等式)} \\
&= E_{q(s_t|o_{1:t}, a_{1:t-1})} \Big[\sum_{t=1}^{T} \ln p(o_t|s_t) + E_{p(s_{t-1}|s_{t-d}, a_{t-d-1:t-2})} \ln p(s_t|s_{t-d}, a_{t-d-1:t-1}) \\
&- \ln q(s_t|o_{1:t}, a_{1:t-1}) \Big] \\
&\text{(采用对数变形)} \\
&= \sum_{t=1}^{T} \big(E[\ln p(o_t|s_t)] - KL(q(s_t|o_{1:t}, a_{1:t-1}) || p(s_t|s_{t-d}, a_{t-d-1:t-1})) \big) \\
&\text{(合并得到结果)}
\end{aligned} \tag{53}
$$

2023.07.21