Importance Sampling

- $p(x)$ is a P.D.F.

$$\mathbb{E}_{x\sim p}[f(x)] \approx \frac{1}{N}\sum_{i=1}^{N} f(x^i)$$

Problem: We do not know the distribution $p(x)$

We can sample it from $q(x)$, a P.D.F.

$$\begin{aligned}
\mathbb{E}_{x\sim p}[f(x)] &= \int f(x)p(x)dx \\
&= \int f(x)\frac{p(x)}{q(x)}q(x)dx \\
&= \mathbb{E}_{x\sim q}[f(x)\frac{p(x)}{q(x)}]
\end{aligned}$$

Question:

Note: $\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}(x))^2$

$$\begin{aligned}
\mathbb{E}_{x\sim p}[f(x)] &= \mathbb{E}_{x\sim q}[f(x)\frac{p(x)}{q(x)}] \\
\mathrm{Var}_{x\sim p}[f(x)] &\ ?\ \mathrm{Var}_{x\sim q}[f(x)\frac{p(x)}{q(x)}]
\end{aligned}$$

See:

$$\mathrm{Var}_{x\sim p}[f(x)] = \mathbb{E}_{x\sim p}[f(x)^2] - (\mathbb{E}_{x\sim p}[f(x)])^2$$

$$\begin{aligned}
\mathrm{Var}_{x\sim q}[f(x)\frac{p(x)}{q(x)}] &= \mathbb{E}_{x\sim q}\left[\left(f(x)\frac{p(x)}{q(x)}\right)^2\right] - \left(\mathbb{E}_{x\sim q}\left[f(x)\frac{p(x)}{q(x)}\right]\right)^2 \\
&= \mathbb{E}_{x\sim p}\left[f(x)^2\frac{p(x)}{q(x)}\right] - (\mathbb{E}_{x\sim p}[f(x)])^2
\end{aligned}$$

这个过程的关键在于利用了变量替换和期望的定义。我们知道，对于随机变量 X，其期望 $\mathbb{E}[X]$ 的定义是 $\int xp(x)dx$，其中 $p(x)$ 是 X 的概率分布函数。因此，我们可以将期望从 q 分布转换到 p 分布。

对于第一个期望 $\mathbb{E}_{x\sim q}\left[\left(f(x)\frac{p(x)}{q(x)}\right)^2\right]$，我们可以写成

$$\mathbb{E}x \sim q\left[\left(f(x)\frac{p(x)}{q(x)}\right)^2\right] = \int\left(f(x)\frac{p(x)}{q(x)}\right)^2 q(x)dx = \int f(x)^2\frac{p(x)}{q(x)}p(x)dx = \mathbb{E}x$$

对于第二个期望 $\mathbb{E}_{x\sim q}\left[f(x)\frac{p(x)}{q(x)}\right]$，我们可以写成

$$\mathbb{E}x \sim q\left[f(x)\frac{p(x)}{q(x)}\right] = \int f(x)\frac{p(x)}{q(x)}q(x)dx = \int f(x)p(x)dx = \mathbb{E}x \sim p[f(x)]$$

因此，我们得到了

$$\mathrm{Var}x \sim q[f(x)\frac{p(x)}{q(x)}] = \mathbb{E}x \sim p\left[f(x)^2\frac{p(x)}{q(x)}\right] - (\mathbb{E}_{x\sim p}[f(x)])^2$$

这就是推导的过程。

比较发现, 第一项不同. $p(x)$和$q(x)$的差距不能很大.

On-policy to Off-policy

$$\nabla\bar{R}_\theta = \mathbb{E}_{\tau\sim p_\theta(\tau)}[R(\tau)\nabla\log p_\theta(\tau)]$$

- Use π_θ to collect data, when θ is updated, we have to sample training data agian.
- Goal: Using the sample from $\pi_{\theta'}$ to train θ. θ' is fixed, so we can re-use the sample data.

$$\nabla\bar{R}_\theta = \mathbb{E}_{\tau\sim p_{\theta'}(\tau)}\left[\frac{p_\theta(\tau)}{p_{\theta'}(\tau)}R(\tau)\nabla\log p_\theta(\tau)\right]$$

θ'是负责和环境做互动, 示范给θ

- Sample the data from θ'.
- Use the data to train θ many times.

Gradient for Update

Using agent θ to sample state and action pair.

> $A^{\theta}(s_t, a_t)$ 估算出来, 在现在的state, 采取动作a_t是好的还是不好的.

$$
\begin{aligned}
&= \mathbb{E}_{(s_t,a_t)\sim\pi_\theta}\left[A^{\theta}(s_t, a_t)\nabla \log p_\theta(a_t^n \mid s_t^n)\right] \\
&\text{Using important sampling} \\
&= \mathbb{E}_{(s_t,a_t)\sim\pi_{\theta'}}\left[\frac{p_\theta(s_t, a_t)}{p_{\theta'}(s_t, a_t)} A^{\theta'}(s_t, a_t)\nabla \log p_\theta(a_t^n \mid s_t^n)\right] \\
&\text{Some math} \\
&= \mathbb{E}_{(s_t,a_t)\sim\pi_{\theta'}}\left[\frac{p_\theta(a_t \mid s_t)}{p_{\theta'}(a_t \mid s_t)} A^{\theta'}(s_t, a_t)\nabla \log p_\theta(a_t^n \mid s_t^n)\right]
\end{aligned}
$$

Thus, the objective function:

$$
\mathcal{J}^{\theta'}(\theta) = \mathbb{E}_{(s_t,a_t)\sim\pi_{\theta'}}\left[\frac{p_\theta(a_t \mid s_t)}{p_{\theta'}(a_t \mid s_t)} A^{\theta'}(s_t, a_t)\right]
$$

如何避免$p(x)$和$q(x)$相差太多?

PPO (adding the constrains)

$$
\mathcal{J}^{\theta'}(\theta) = \mathcal{J}^{\theta'}(\theta) - \beta \mathrm{KL}(\theta, \theta')
$$

Process of PPO

- Initial policy parameters θ^0.
- In each iteration k
  - Using θ^k to interact with the environment to collect $\{s_t, a_t\}$ and compute advantage $A^{\theta^k}(s_t, a_t)$
  - Find θ optimizing $\mathcal{J}_{\mathrm{PPO}}^{\theta^k}(\theta) = \mathcal{J}^{\theta^k}(\theta) - \beta \mathrm{KL}(\theta, \theta^k)$