本科毕设：

Mc Gradient Estimation in ML

1. Overview：

ϕ : structural parameters　θ : distributional parameters

☆ Central Question: computing $\mathcal{F}(\theta) = \int \underbrace{p(x;\theta)}_{\text{measure}} \underbrace{f(x;\phi)}_{\text{cost}} dx = \mathbb{E}_{p(x;\theta)}[f(x;\phi)]$ ①

$p(x;\theta)$: probability distribution that continuous in its domain and differentiable with θ

first we want to learn more about $\eta := \nabla_\theta \mathcal{F}(\theta) = \nabla_\theta \mathbb{E}_{p(x;\theta)}[f(x;\phi)]$ [Sensitivity Analysis] ②

$= \nabla_\theta \mathcal{F}(\theta) = \left[\frac{\partial \mathcal{F}(\theta)}{\partial \theta_1}, \ldots\ldots, \frac{\partial \mathcal{F}(\theta)}{\partial \theta_D}\right]$

Section 2: general principles and considerations for Monte Carlo Methods

Section 4-6: develop three classes of gradient estimators
- score function estimator
- pathwise estimator
- measure valued gradient estimator

Section 7: methods to control the variance of the estimators

2. Section 2: MC Methods and Stochastic Optimization

2.1 Monte Carlo Estimators

MC 方法： ① 从分布 $p(x;\theta)$ 中抽取 $\hat{x}^{(1)}, \ldots\ldots, \hat{x}^{(N)}$

② 计算 $\bar{\mathcal{F}}_N = \frac{1}{N}\sum_{n=1}^{N} f(\hat{x}^{(n)})$ ⓪ where $\hat{x}^{(n)} \sim p(x;\theta)$ for $n=1,\ldots\ldots,N$

$\bar{\mathcal{F}}_N$ 为随机变量　→ Monte Carlo Estimator

MC 估计量的四个性质

① Consistency (一致性)：$N \uparrow$ $\bar{\mathcal{F}}_N$ converge to true value

② 无偏性：多次重复估计所得的均值的期望为真实值

$\mathbb{E}_{p(x;\theta)}(\bar{\mathcal{F}}_N) = \mathbb{E}_{p(x;\theta)}\left[\frac{1}{N}\sum_{n=1}^{N} f(\hat{x}^{(n)})\right] = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{p(x;\theta)} f(\hat{x}^{(n)}) = \mathbb{E}_{p(x;\theta)}[f(x)]$

③ 最小方差

④ Computational efficiency (计算效率)：使用最少样本数量来计算期望

2.2 随机优化

gradient ② 可以用来表征 θ 变化时 cost 的敏感性 $\Rightarrow$ 用于 optimisation of the distribution parameter θ

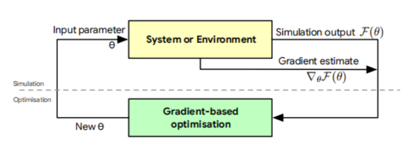

Figure 1: Stochastic optimisation loop comprising a simulation phase and an optimisation phase. The simulation phase produces a simulation of the stochastic system or interaction with the environment, as well as unbiased estimators of the gradient.

⟶ 有点类似梯度下降

2.3 Gradient Estimation 的五个应用领域

(1) Variational Inference : x 由 $p(x|z)p(z)$ 生成 后验 $p(z|x)$ 未知

利用 $q(z|x,\theta)$ 来近似

θ: variational paramete e.g: 可以给一个 Gauss 分布

$$\eta = \nabla_\theta E_{q(z|x;\theta)}\left[\underbrace{\boxed{\log p(x|z) - \log \frac{q(z|x;\theta)}{p(z)}}}_{\text{cost } f(z;x)}\right]$$

(x: 观测得到 ; z 观测不到)

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

(2) Reinforcement Learning

(3) Sensitivity Analysis

(4) Discrete Event Systems and Queuing Theory

(5) Experimental design

3. Two ways to compute the gradients $\nabla_\theta E_{p(x;\theta)}[f(x)]$

- Derivatives of Measure (测度的导数) (对 measure $p(x;\theta)$ 求微分)
  - score function estimator (Section 4)
  - measure-valued gradient (Section 6)
- Derivatives of Path (路径的导数): 计算 cost $f(x)$ 的微分
  - path wise gradient (Section 5)

4. Score Function Gradient Estimators

4.1 score function : $\nabla_\theta \log p(x;\theta) = \frac{\nabla_\theta p(x;\theta)}{p(x;\theta)}$ key in maximum likelyhood estimation

A properety of score function : $E_{p(x;\theta)}[\nabla_\theta \log p(x;\theta)] = \int p(x;\theta)\cdot\frac{\nabla_\theta p(x;\theta)}{p(x;\theta)}dx = \int \nabla_\theta p(x;\theta)dx$

$\boxed{\text{LDC}}$ $\overset{?}{=} \nabla_\theta \int p(x;\theta)dx = \nabla_\theta 1 = 0$

4.2 $\eta = \nabla_\theta E_{p(x;\theta)}[f(x)] = \underbrace{\nabla_\theta \int p(x;\theta)f(x)dx \overset{?}{=} \int f(x)\nabla_\theta p(x;\theta)dx}_{(4.2a)} = \int f(x)(\nabla_\theta \log p(x;\theta))p(x;\theta)dx$

$= E_{p(x;\theta)}[f(x)\nabla_\theta \log p(x;\theta)]$

$\bar{\eta} = \frac{1}{N}\sum_{n=1}^{N}[f(\hat{x}^{(n)})\nabla_\theta \log p(\hat{x}^{(n)};\theta)]$ $\quad \hat{x}^{(n)} \sim p(x;\theta)$ MC

4.3.1 (4.2 a)操作的可行性

When the interchange between differentiation and integration in (13a) is valid, we will obtain an unbiased estimator of the gradient (L'Ecuyer, 1995). Intuitively, since differentiation is a process of limits, the validity of the interchange will relate to the conditions for which it is possible to exchange limits and integrals, in such cases most often relying on the use of the dominated convergence theorem or the Leibniz integral rule (Flanders, 1973; Grimmett and Stirzaker, 2001). The interchange will be valid if the following conditions are satisfied:

- The measure $p(\mathbf{x};\boldsymbol{\theta})$ is continuously differentiable in its parameters $\boldsymbol{\theta}$.
- The product $f(\mathbf{x})p(\mathbf{x};\boldsymbol{\theta})$ is both integrable and differentiable for all parameters $\boldsymbol{\theta}$.
- There exists an integrable function $g(\mathbf{x})$ such that $\sup_{\boldsymbol{\theta}} \|f(\mathbf{x})\nabla_{\boldsymbol{\theta}} p(\mathbf{x};\boldsymbol{\theta})\|_1 \le g(\mathbf{x})\ \forall \mathbf{x}$.

4.3.3 估计量方差

denote the estimator mean as $\mu(\theta) := E_{p(x;\theta)}[\bar{\eta}_N]$ 先假设 $N=1$ 情况

故估计量方差可写作 $V_{p(x;\theta)}[\bar{\eta}_N] = V_{p(x;\theta)}[\underline{f(x)\nabla_\theta \log p(x;\theta)}]$ → $\bar{\eta}_N, N=1$

$= E_{p(x;\theta)}[(f(x)\nabla_\theta \log p(x;\theta))^2] - (\mu(\theta))^2$

$V_{p(x;\theta)}[\bar{\eta}_{N=1}] = \lim_{h\to 0}\frac{1}{h^2}E_{p(x;\theta)}[(w(\theta,h)-1)^2 f(x)^2] - (\mu(\theta))^2$ 其中 $w(\theta,h) := \frac{p(x,\theta+h)}{p(x,\theta)}$ → importance ratio

$f(x)\nabla_\theta \log p(x;\theta) = f(x)\cdot\frac{\nabla_\theta p(x;\theta)}{p(x;\theta)} = f(x)\left(\lim_{h\to 0}\frac{p(x,\theta+h)-p(x,\theta)}{h\cdot p(x;\theta)}\right)$

$= \lim_{h\to 0} f(x)(w(\theta,h)-1)\cdot\frac{1}{h}$

故 $E_{p(x;\theta)}[(f(x)\nabla_\theta \log p(x;\theta))^2] = \int p(x;\theta)\left(\lim_{h\to 0} f(x)\cdot\frac{1}{h}(w(\theta,h)-1)\right)^2 dx = \int p(x;\theta)\lim_{h\to 0}\frac{1}{h^2}(w(\theta,h)-1)^2 f(x)^2 dx$

$= \lim_{h\to 0}\frac{1}{h^2}E_{p(x;\theta)}[(w(\theta,h)-1)^2 f(x)^2]$

which, for a fixed h, exposes the dependency of the variance on the importance weight ω. Although we will not explore it further, we find it instructive to connect these variance expressions to the variance bound for the estimator given by the Hammersley-Chapman-Robbins bound (Lehmann and Casella, 2006, ch 2.5)

$$\mathbb{V}_{p(\mathbf{x};\theta)}[\bar{\eta}_{N=1}] \ge \sup_h \frac{(\mu(\theta+h)-\mu(\theta))^2}{\mathbb{E}_{p(\mathbf{x};\theta)}\left[\frac{p(\mathbf{x};\theta+h)}{p(\mathbf{x};\theta)}-1\right]^2} = \sup_h \frac{(\mu(\theta+h)-\mu(\theta))^2}{\mathbb{E}_{p(\mathbf{x};\theta)}[\omega(\theta,h)-1]^2}, \tag{19}$$ → 未证

which is a generalisation of the more widely-known Cramer-Rao bound and describes the minimal variance achievable by the estimator.

5. Pathwise Gradient Estimators

(losing generality, but lower variance and ease of implementation)

5.1 Sampling Paths

$\hat{x}$ can be sampled directly from $p(x;\theta)$

can also be sampled from indirect way: $p(x;\theta) = p(\varepsilon)\left|\nabla_\varepsilon g(\varepsilon;\theta)\right|^{-1}$

Lotus: $\mathbb{E}_{p(x;\theta)}[f(x)] = \mathbb{E}_{p(\varepsilon)}[f(g(\varepsilon;\theta))] \Rightarrow$ 可以不知道x的分布情况下计算$f(x)$的期望

e.g [One liners]: $p(x;\theta) = \mathcal{N}(x|\mu,\Sigma)$ 先从 $p(\varepsilon) = \mathcal{N}(0,I)$ 中抽取 然后进行转换 $g(\varepsilon,\theta) = \mu + L\varepsilon$

$\Sigma = LL^T$

e.g [Polar transformations]: Box-Muller

5.2 生成估计量

$$\eta = \nabla_\theta \mathbb{E}_{p(x;\theta)}[f(x)] \overset{\text{lotus}}{=} \nabla_\theta \mathbb{E}_{p(\varepsilon)}[f(g(\varepsilon;\theta))] = \nabla_\theta \int p(\varepsilon) f(g(\varepsilon;\theta))\, d\varepsilon$$

由 MC Methods 可得 $\bar{\eta}_N = \frac{1}{N}\sum_{n=1}^{N} \nabla_\theta f(g(\hat{\varepsilon}^{(n)};\theta))$; $\quad \hat{\varepsilon}^{(n)} \sim p(\varepsilon)$

pathwise estimator can be rewritten as a more general form:

$$\eta = \nabla_\theta \mathbb{E}_{p(x;\theta)}[f(x)] = \mathbb{E}_{p(\varepsilon)}\left[\nabla_\theta f(x)\big|_{x=g(\varepsilon;\theta)}\right] = \int p(\varepsilon)\, \nabla_x f(x)\big|_{x=g(\varepsilon;\theta)} \nabla_\theta g(\varepsilon;\theta)\, d\varepsilon \quad (5.2a)$$

$$= \int p(x;\theta)\, \nabla_x f(x)\, \nabla_\theta x\, dx = \mathbb{E}_{p(x;\theta)}\left[\nabla_x f(x)\, \nabla_\theta x\right] \quad (5.2b)$$

6. Measure-valued Gradients

6.1 Weak Derivatives

对于 D-dimensional parameters $\vec{\theta}$　θ_i 表示其第 i 个分量

$\nabla_{\theta_i} p(x;\theta)$ 可能存在负部，故其本身不是概率密度，故对 $\nabla_{\theta_i} p(x;\theta)$ 进行如下分解：

$\nabla_{\theta_i} p(x;\theta) = c^+_{\theta_i} p^+_i(x;\theta) - c^-_{\theta_i} p^-_i(x;\theta)$ 其中 $p^+_i(x;\theta)$、$p^-_i(x;\theta)$ 是概率密度

LHS: $\int \nabla_{\theta_i} p(x;\theta)dx = \nabla_{\theta_i}\int p(x;\theta)dx = 0$; RHS: $\int (c^+_{\theta_i} p^+_i(x;\theta) - c^-_{\theta_i} p^-_i(x;\theta))dx = c^+_{\theta_i} - c^-_{\theta_i} = 0 \Rightarrow c^+_{\theta_i} = c^-_{\theta_i} = c_{\theta_i}$

故 $\nabla_{\theta_i} p(x;\theta) = c_{\theta_i}(p^+_i(x;\theta) - p^-_i(x;\theta))$

Deriving the Estimator:
$$
\begin{aligned}
\eta_i = \nabla_{\theta_i} \mathbb{E}_{p(x;\theta)}[f(x)] &= \nabla_{\theta_i}\int p(x;\theta) f(x)\,dx \\
&= \int \nabla_{\theta_i} p(x;\theta) f(x)\,dx \\
&= \int c_{\theta_i}(p^+_i(x;\theta) - p^-_i(x;\theta)) f(x)\,dx \\
&= c_{\theta_i}\left(\int p^+_i(x;\theta) f(x)dx - \int p^-_i(x;\theta) f(x)dx\right) \\
&= c_{\theta_i}\left(\mathbb{E}_{p^+_i(x;\theta)}[f(x)] - \mathbb{E}_{p^-_i(x;\theta)}[f(x)]\right)
\end{aligned}
$$

(MC Methods)

故 $\bar{\eta}_{i,N} = \frac{c_{\theta_i}}{N}\left(\sum_{n=1}^{N} f(\dot{x}^{(n)}) - \sum_{n=1}^{N} f(\ddot{x}^{(n)})\right)$；其中 $\dot{x}^{(n)} \sim p^+_i(x;\theta)$　$\ddot{x}^{(n)} \sim p^-_i(x;\theta)$