STA 4273: Minimizing Expectations

Lecture 3 - Gradient Estimation I

Chris J. Maddison

University of Toronto

Announcements

- Presentation assignments are out.
- Project handout(s) are out.

Presentation assignments

- Quercus: People $\rightarrow$ Groups $\rightarrow$ look for yourself in the Week N Presentation Assignments groups.
- Please let me know ASAP, if you cannot do your week.
- If you have not gotten an assignment, either I made a mistake or you didn't send in your rankings. **Email me**!

Project

- Handouts are up. Apologies for the delay. Come to office hours or email me for help!
- I've moved the due date of the Proposal back to Feb 22.
  - Reduce conflict Prof. Grosse's course.
  - I was late on getting the handout up.
- The proposal is to get you started. You do not have to end up working on the same project that you propose!
- Can I work alone? Yes, but standards will be just as high as for groups of 4.

Gradient estimation

Assuming it exists, today and next week we will consider the problem of gradient estimation, i.e. computing

$$\nabla_\theta \mathbb{E}_{X \sim q_\theta}[f(X, \theta)]$$

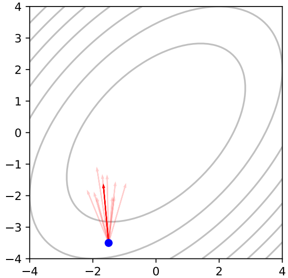
- Same old beloved assumptions.
- X is a random variable taking values in $\mathcal{X}$ with a prob. density q_θ in a parametric family of densities parameterized by $\theta \in \mathbb{R}^D$.
- $f : \mathcal{X} \times \mathbb{R}^D \to \mathbb{R}$ is a function.

Gradient estimation

- A gradient estimator is a random variable $G(\theta)$ such that

$$\mathbb{E}[G(\theta)] = \nabla_\theta \mathbb{E}_{X \sim q_\theta}[f(X, \theta)]$$

- Will briefly introduce two basic approaches.
  - Score function estimator (we've actually seen this).
  - Pathwise gradient estimator, also called reparameterization estimator.

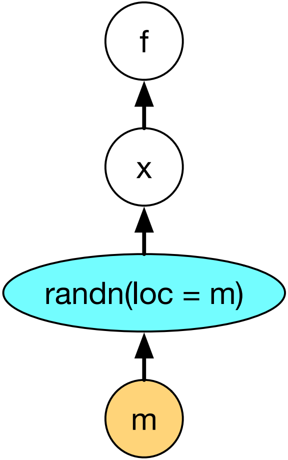


Pathwise gradient

- Let's start with pathwise gradient.
- Example: Let $f : \mathbb{R} \to \mathbb{R}$ be continuously differentiable, $X \sim \mathcal{N}(m, 1)$ be a Gaussian with mean $m \in \mathbb{R}$. We want to compute:

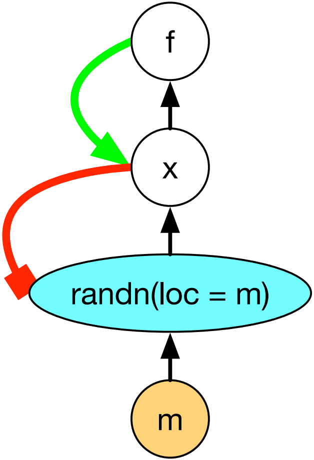
$$\nabla_m \mathbb{E}[f(X)]$$

- Imagine the flow of computation required to compute a sample $f(X)$ using numpy.
- Can we use the state of this computation to compute an estimator?



Pathwise gradient

- Naive idea: compute $\nabla_m f(X)$ using the chain rule given a realization like the one to the right.
- What is the partial derivative of `randn(loc=m)`??
- Intuitively, it should be 0.



Pathwise gradient

- One idea that works: reparameterize the sampling process of X:

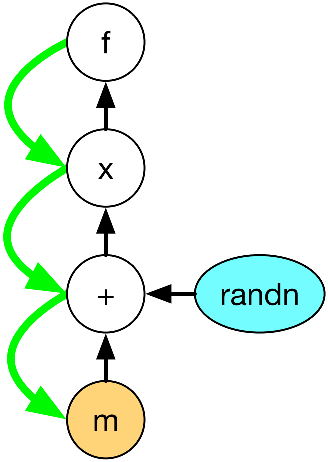
$$\epsilon \sim \mathcal{N}(0,1) \quad X \overset{d}{=} \epsilon + m$$
$$\mathbb{E}[f(X)] = \mathbb{E}[f(\epsilon + m)]$$

- Assuming we can exchange the derivative and the integral, we now get

$$\nabla_m \mathbb{E}[f(X)] = \mathbb{E}[\nabla_m f(\epsilon + m)]$$

Suggesting the estimator $\nabla_m f(\epsilon + m)$

- Key idea: ϵ does not depend on m.



Pathwise gradient

- The pathwise gradient estimator is based on a change of variables.
- More generally, suppose there exists a random variable $\epsilon \in \mathcal{E}$ with density $p(\epsilon)$ and a function $y : \mathcal{E} \times \mathbb{R}^d \to \mathcal{X}$ such that

$$X \stackrel{d}{=} y(\epsilon, \theta)$$

- Then, assuming we can exchange the derivative and integral operation:

$$\begin{aligned} \nabla_\theta \mathbb{E}_{X \sim q_\theta}[f(X, \theta)] &= \nabla_\theta \mathbb{E}_{\epsilon \sim p(\epsilon)}[f(y(\epsilon, \theta), \theta)] \\ &= \mathbb{E}_{\epsilon \sim p(\epsilon)}[\nabla_\theta f(y(\epsilon, \theta), \theta)] \end{aligned}$$

- Suggesting the pathwise gradient estimator:

$$\nabla_\theta f(y(\epsilon, \theta), \theta)$$

Pathwise gradient

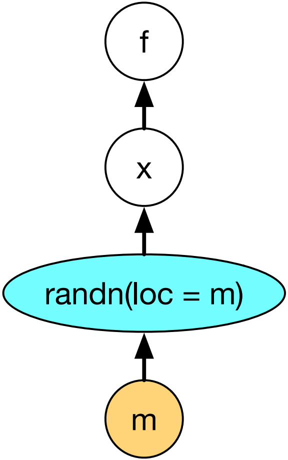
- When can we exchange derivative and integral operators?
- Asmussen and Glynn (2007) give some simple conditions in Chap. 7.2, prop. 2.3.
- Rule-of-thumb:
  - Typically valid when $Z(\theta) := f(y(\epsilon, \theta), \theta)$ is continuous and differentiable except at finitely many points.
  - Does NOT hold for reparameterizations of discrete X!

Score function gradient

- Let's move to the score function gradient.
- Example: Let $f : \mathbb{R} \to \mathbb{R}$ be continuously differentiable, $X \sim \mathcal{N}(m, 1)$ be a Gaussian with mean $m \in \mathbb{R}$. We want to compute:

$$\nabla_m \mathbb{E}[f(X)]$$

- Let's see if we can attack this directly.



Score function gradient

Assuming we can exchange derivative and integral,

$$\begin{aligned}
\nabla_m \mathbb{E}[f(X)] &= \nabla_m \int_x f(x) \frac{\exp\left(-\frac{(x-m)^2}{2}\right)}{\sqrt{2\pi}}\, dx \\
&= \int_x f(x) \nabla_m \frac{\exp\left(-\frac{(x-m)^2}{2}\right)}{\sqrt{2\pi}}\, dx \\
&= \int_x f(x) \frac{\exp\left(-\frac{(x-m)^2}{2}\right)}{\sqrt{2\pi}} \nabla_m (-(x-m)^2/2)\, dx \\
&= \int_x f(x) \frac{\exp\left(-\frac{(x-m)^2}{2}\right)}{\sqrt{2\pi}} (x-m) \\
&= \mathbb{E}[f(X)(X-m)]
\end{aligned}$$

i.e., weight the function value by the distance to the mean, very intuitive!

Score function gradient

- More generally, the score function gradient is based on the following identity
$$\nabla_\theta q_\theta(x) = q_\theta(X)\nabla_\theta \log q_\theta(x)$$
- Assuming we can exchange derivative and integral:
$$\nabla_\theta \mathbb{E}_{X\sim q_\theta}[f(X)] = \mathbb{E}_{X\sim q_\theta}[f(X)\nabla_\theta \log q_\theta(X)]$$
- Suggesting the score function gradient gradient estimator:
$$f(X)\nabla_\theta \log q_\theta(X)$$
- Note, this is for the case in which f does not depend on θ. To get a gradient with dependence on θ, just add $\partial_\theta f(X,\theta)$, where ∂_θ is the vector of partial derivatives of f w.r.t. θ.

Control variates

- The score function has an important property that makes it easy to design control variates.
- Let C be *any* real-valued random variable that is uncorrelated to X

$$\begin{aligned}
\mathbb{E}_{X\sim q_\theta}[C\nabla_\theta \log q_\theta(X)] &= \mathbb{E}[C]\mathbb{E}_{X\sim q_\theta}[\nabla_\theta \log q_\theta(X)] \\
&= \mathbb{E}[C]\nabla_\theta \mathbb{E}_{X\sim q_\theta}[1] \\
&= 0
\end{aligned}$$

- We can use this to reduce the variance of score function estimators!

Control variates–RL

Recall our gradient estimator for the finite-horizon MDP:

$$\nabla J(\theta) = \mathbb{E}_{\tau \sim p}\left[\sum_{t=0}^{T} r(\tau) \nabla \log \pi_\theta(a_t|s_t)\right]$$

- Simulate a random trajectory $\tau \sim p$.
- $\sum_{t=0}^{T} r(\tau) \nabla \log \pi_\theta(a_t|s_t)$ is a score function estimator! We can reduce the variance using our new knowledge.

Control variates–RL

Given s_t, $r(s_{t'}, a_{t'})$ is independent of a_t for $t' < t$, so,

$$\begin{aligned}
\nabla J(\theta) &= \mathbb{E}_{\tau\sim p}\left[\sum\nolimits_{t=0}^{T} r(\tau)\nabla \log \pi_\theta(a_t|s_t)\right] \\
&= \mathbb{E}_{\tau\sim p}\left[\sum\nolimits_{t=0}^{T} \left(\sum_{t'=0}^{T} r(s_{t'}, a_{t'})\right) \nabla \log \pi_\theta(a_t|s_t)\right] \\
&= \mathbb{E}_{\tau\sim p}\left[\sum\nolimits_{t=0}^{T} \left(\sum_{t'<t} r(s_{t'}, a_{t'}) + \sum_{t'\geq t} r(s_{t'}, a_{t'})\right) \nabla \log \pi_\theta(a_t|s_t)\right] \\
&= \mathbb{E}_{\tau\sim p}\left[\sum\nolimits_{t=0}^{T} \left(\sum_{t'\geq t} r(s_{t'}, a_{t'})\right) \nabla \log \pi_\theta(a_t|s_t)\right]
\end{aligned}$$

This is lower variance. But we can do more...

Control variates–RL

Given s_t, the value $V_t^\pi(s_t)$ is independent of a_t, so,

$$\begin{aligned}
\nabla J(\theta) &= \mathbb{E}_{\tau\sim p}\left[\sum\nolimits_{t=0}^{T}\left(\sum_{t'\geq t} r(s_{t'},a_{t'})\right)\nabla\log\pi_\theta(a_t|s_t)\right] \\
&= \mathbb{E}_{\tau\sim p}\left[\sum\nolimits_{t=0}^{T}\left(\sum_{t'\geq t} r(s_{t'},a_{t'}) - V_t^\pi(s_t)\right)\nabla\log\pi_\theta(a_t|s_t)\right] \\
&= \mathbb{E}_{\tau\sim p}\left[\sum\nolimits_{t=0}^{T} A_t^\pi(s_t,a_t)\nabla\log\pi_\theta(a_t|s_t)\right]
\end{aligned}$$

This quantity

$$A_t^\pi(s_t,a_t) = \sum_{t'\geq t} r(s_{t'},a_{t'}) - V_t^\pi(s_t)$$

is an example of an advantage, i.e., how much better is it to take action a_t at time t than the average value.

Control variates–RL

(4) much lower variance gradient estimator than (3):

$$\sum_{t=0}^{T} \left(\sum_{t'=0}^{T} r(s_{t'}, a_{t'}) \right) \nabla \log \pi_\theta(a_t | s_t) \tag{1}$$

$$\sum_{t=0}^{T} \left(\sum_{t' \geq t} r(s_{t'}, a_{t'}) - V_t^\pi(s_t) \right) \nabla \log \pi_\theta(a_t | s_t) \tag{2}$$

Talks today

- Can we use pathwise gradients for discrete random variables?
- Can we reduce the variance of the score function estimator for discrete random varibles using clever subset structure?
- Can we reduce the variance of RL gradients by estimating the advantage in more clever ways?