Intro to the research

Lecture Note:

in current off-policy gradient, we have the policy gradient (without causality form) as a expectation

$$\nabla_{\theta'} J(\theta') = E_{\tau\sim\pi_\theta(\tau)}\left[\underbrace{\left(\prod_{t=1}^{T}\frac{\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t)}{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)}\right)}_{\text{exp term with T}}\underbrace{\left(\sum_{t=1}^{T}\nabla_{\theta'}\log\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t)\right)\left(\sum_{t=1}^{T}r(\mathbf{s}_t,\mathbf{a}_t)\right)}_{\text{policy gradient with high variance itself}}\right] \tag{1}$$

with following strategies:

- causality
  From common sense, future cannot affect the past, therefore we can no longer consider the past reward. By which we reduced the magnitude of $\sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t)$ so the variance is controlled
- baseline
  "normalize" the reward of the policy, by which restrict the unexpected behaviour from "shifted reward"

Solution from Infinite paper 1

1. curse of horizon: the exponential product w.r.t. T which have the variance grow exponentially with T(in infinite-term it will be badly-defined).
2. significant decrease in estimation variance is possible when we apply IS on the state space rather than the trajectory.
   Definition:
   (a) $d_{\pi,t}(\cdot)$: the distribution of state s_t at time t when using policy π and start from s_0 from initial distribution $d_0(\cdot)$
   (b) $d_\pi(s)$: the average visitation distribution, based on $d_{\pi,t}(\cdot)$
   $$d_\pi(s) = \lim_{T\to\infty}\frac{\sum_{t=0}^{T}\gamma^t d_{\pi,t}(s)}{\sum_{t=0}^{T}\gamma^t} \tag{2}$$
      i. if $\gamma \in (0,1)$ (discounted reward case), by geometric series we have
      $$d_\pi(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t d_{\pi,t}(s) \tag{3}$$
      ii. if $\gamma = 1$(average reward case)
      $$d_\pi(s) = \lim_{T\to\infty}\frac{\sum_{t=0}^{T} d_{\pi,t}(s)}{T+1} = \lim_{t\to\infty} d_{\pi,t}(s) \tag{4}$$
      which is the stationary distribution of s_t
   (c) R_π: the expected reward of π
      i. $p_\pi(\cdot)$: the distribution of trajectory τ under policy π

ii. $R^T(\tau)$: the reward of trajectory τ up to time T, defined as

$$R^T(\tau) = \frac{\sum_{t=0}^{T} \gamma^t r_t}{\sum_{t=0}^{T} \gamma^t}, \text{where } r_t \text{ is defined in } \tau \tag{5}$$

iii. R_π: the expected reward of policy π

$$R_\pi = \lim_{T\to\infty} E_{\tau\sim p_\pi}[R^T(\tau)] \tag{6}$$

(d) with the distribution in (b), we have the R_π in (c) in state-space version

$$R_\pi = \sum_{s,a} d_\pi(s)\pi(a|s)r(s,a) = E_{(s,a)\sim d_\pi(s)\pi(a|s)}[r(s,a)] \tag{7}$$

focus on state-action pair, the IS will be on both state and action

$$R_\pi = E_{(s,a)\sim d_{\pi_0}}\left[\underbrace{\frac{\pi(a|s)}{\pi_0(a|s)}}_{\substack{\text{policy}\\ \text{importance}\\ \text{ratio}}} \underbrace{\frac{d_\pi(s)}{d_{\pi_0}(s)}}_{\substack{\text{visitation}\\ \text{importance}\\ \text{ratio}}} r(s,a)\right] \tag{8}$$

where the visitation importance ratio is not known but can be estimated. Denote the following:

i. $\beta_{\pi/\pi_0}(a,s) = \frac{\pi(a|s)}{\pi_0(a|s)}$ policy importance ratio.

ii. $w_{\pi/\pi_0}(s) = \frac{d_\pi(s)}{d_{\pi_0}(s)}$ visitation importance ratio.

(e) from (7), a weighted-IS can be generated as following

i. run π_0 (original policy), generate the data,

ii. for any new policy π, the policy is evaluated as:

$$\hat{R_\pi} = \sum_{i=1}^{m}\sum_{t=0}^{T} \frac{\gamma^t w_{\pi/\pi_0}(s_t^i)\beta_{\pi/\pi_0}(a_t^i, s_t^i)}{\text{normed sum}} r_t^i \tag{9}$$

by which we restrict the space in station-action pair(s,a) rather than trajectory.

(f) when policy π and π_0 is given, the policy importance ratio $\beta_{\pi/\pi_0}(a,s)$ is obtained, and we need to obtained visitation ratio $w_{\pi/\pi_0}(s)$

- in average reward case:

  i. $T_\pi(s'|s) = \sum_a T(s'|s,a)\pi(a|s)$ the transition probability from s to s', following policy π. Considering the stationary case, there's

  $$d_\pi(s') = \sum_s T_\pi(s'|s)d_\pi(s) \tag{10}$$

  ii. Theorem 1: in average reward case, assume d_π is the unique invariant distribution of T_π and $d_{\pi_0}(s) > 0$(irreducibility). Then there's a function $w(s)$ equals visitation importance ratio $w_{\pi/\pi_0}(s)$ iff:

  $$E_{(s,a)|s'\sim d_{\pi_0}}[w(s)\beta_{\pi/\pi_0}(a,s) - w(s')|s'] = 0 \tag{11}$$

i.e. when given the REVERSED transition distribution, for any next state, a correct estimator w(s) of visitation ratio should have the distribution of prior state-space pair $w(s)\beta_{\pi/\pi_0}(a,s) = w(s')$ on expectation.

Problem: What is the policy gradient?

Copy from (8), denote the θ as the parameter of new policy π_θ

$$J(\theta) = E_{(s,a)\sim d_{\pi_0}}\left[\frac{\pi_\theta(a|s)}{\pi_0(a|s)}\frac{d_{\pi_\theta}(s)}{d_{\pi_0}(s)}r(s,a)\right] \tag{12}$$

$$\nabla_\theta J(\theta) = E_{(s,a)\sim d_{\pi_0}}\left[\frac{\nabla_\theta\pi_\theta(a|s)d_{\pi_\theta}(s)}{\pi_0(a|s)d_{\pi_0}(s)}r(s,a)\right] \tag{13}$$

We know that, by product rule

$$\nabla_\theta\pi_\theta(a|s)d_{\pi_\theta}(s) = \nabla_\theta\pi_\theta(a|s)\cdot d_{\pi_\theta}(s) + \pi_\theta(a|s)\cdot\nabla_\theta d_{\pi_\theta}(s)$$

from log identity:

$$\nabla_\theta\pi(a|s) = \pi_\theta(a|s)\nabla_\theta\log\pi_\theta(a|s)$$
$$\nabla_\theta d_{\pi_\theta}(s) = d_{\pi_\theta}(s)\nabla_\theta\log d_{\pi_\theta}(s)$$

Thus we have

$$\nabla_\theta\pi_\theta(a|s)d_{\pi_\theta}(s) = \pi_\theta(a|s)d_{\pi_\theta}(s)\left(\nabla_\theta\log\pi_\theta(a|s) + \nabla_\theta\log d_{\pi_\theta}(s)\right) \tag{14}$$

from (13), we have

$$\nabla_\theta J(\theta) = E_{(s,a)\sim d_{\pi_0}}\left[\frac{\pi_\theta(a|s)d_{\pi_\theta}(s)}{\pi_0(a|s)d_{\pi_0}(s)}\left(\nabla_\theta\log\pi_\theta(a|s) + \nabla_\theta\log d_{\pi_\theta}(s)\right)r(s,a)\right] \tag{15}$$

$\nabla_\theta\log\pi_\theta(a|s)$ can be solved by auto-diff, we need to solve $\nabla_\theta\log d_{\pi_\theta}(s)$

(Result in Jierong's email, I simply LaTeX-lized it)

Try 1: Given a policy π_θ (should be a NN), the forward pass of NN should be instant, roll out from initial distribution d_0, after a large amount of steps, save the respective partial derivative for each step, then calculate the mean, it should be able to be solved by auto-diff?

This approach will suffer a environment not derivable(where auto-diff won't work).

Try 2: in average reward case: $T_\pi(s'|s) = \sum_a T(s'|s,a)\pi(a|s)$ the transition probability from s to s', following policy π. Considering the stationary case, there's

$$
\begin{aligned}
d_{\pi_\theta}(s') &= \sum_s \sum_a T(s'|s,a)\pi_\theta(a|s)d_{\pi_\theta}(s) \\
\nabla_\theta d_\pi(s') &= \nabla_\theta \sum_s \sum_a \underbrace{T(s'|s,a)}_{environment} \pi_\theta(a|s)d_{\pi_\theta}(s) \\
&= \sum_s \sum_a T(s'|s,a)\nabla_\theta \left(\pi_\theta(a|s)d_{\pi_\theta}(s)\right) \\
\text{by product rule} &= \sum_s \sum_a T(s'|s,a)d_{\pi_\theta}(s)\nabla_\theta\pi_\theta(a|s) + \sum_s \sum_a T(s'|s,a)\pi_\theta(a|s)\nabla_\theta d_{\pi_\theta}(s) \\
\nabla_\theta d_{\pi_\theta}(s) \text{ is on s, so} &= \cdots + \sum_s \nabla_\theta d_{\pi_\theta}(s) \sum_a T(s'|s,a)\pi_\theta(a|s) \\
T_{\pi_\theta} \text{ for simplicity} &= \cdots + \sum_s \nabla_\theta d_{\pi_\theta}(s) T_{\pi_\theta}(s'|s) \\
&= \cdots + \sum_{s\backslash s'} \nabla_\theta d_{\pi_\theta}(s) T_{\pi_\theta}(s'|s) + \nabla_\theta d_{\pi_\theta}(s') T_{\pi_\theta}(s'|s') \\
(1 - T_{\pi_\theta}(s'|s'))\nabla_\theta d_{\pi_\theta}(s') &= \sum_s \sum_a T(s'|s,a)\nabla_\theta\pi_\theta(a|s)\cdot d_{\pi_\theta}(s) + \sum_{s\backslash s'} \nabla_\theta d_{\pi_\theta}(s) T_{\pi_\theta}(s'|s) \\
\nabla_\theta d_{\pi_\theta}(s') &= \frac{\sum_s \sum_a T(s'|s,a)\nabla_\theta\pi_\theta(a|s)\cdot d_{\pi_\theta}(s)}{1 - T_{\pi_\theta}(s'|s')} + \frac{\sum_{s\backslash s'} T_{\pi_\theta}(s'|s)}{1 - T_{\pi_\theta}(s'|s')}\nabla_\theta d_{\pi_\theta}(s)
\end{aligned}
$$

where: $T(s'|s,a)$ from environment
$\nabla_\theta \pi_\theta(a|s)$ from auto-diff
$d_{\pi_\theta}(s)$ from policy
$T_{\pi_\theta}(s'|s')$,$T_{\pi_\theta}(s'|s)$ from policy
Thus we can solve a linear system

$$
x_i = k_i + \sum_{j\neq i} a_{ij} * x_j
$$

to obtain the derivative. However this solution only works for discrete scenario.

For continuous state space problem, $t_\pi(s'|s)$ becomes the transition density function and we ought to use integral instead of summation:

$$
\begin{aligned}
d_{\pi_\theta}(s') &= \int_s \sum_a t(s'|s,a)\pi_\theta(a|s)d_{\pi_\theta}(s)ds \\
\nabla_\theta d_\pi(s') &= \nabla_\theta \int_s \sum_a \underbrace{t(s'|s,a)}_{environment} \pi_\theta(a|s)d_{\pi_\theta}(s)ds \\
&= \int_s \sum_a t(s'|s,a)\nabla_\theta \left(\pi_\theta(a|s)d_{\pi_\theta}(s)\right) ds \\
&= \int_s \sum_a t(s'|s,a)d_{\pi_\theta}(s)\nabla_\theta\pi_\theta(a|s)ds + \int_s \sum_a t(s'|s,a)\pi_\theta(a|s)\nabla_\theta d_{\pi_\theta}(s)ds \\
&= \int_s \sum_a t(s'|s,a)d_{\pi_\theta}(s)\nabla_\theta\pi_\theta(a|s)ds + \int_s \nabla_\theta d_{\pi_\theta}(s) \sum_a t(s'|s,a)\pi_\theta(a|s)ds \\
&= \cdots + \int_s \nabla_\theta d_{\pi_\theta}(s) t_{\pi_\theta}(s'|s)ds \\
\nabla_\theta d_\pi(s') &= \int_s \sum_a t(s'|s,a)d_{\pi_\theta}(s)\nabla_\theta\pi_\theta(a|s)ds + \int_s \nabla_\theta d_{\pi_\theta}(s) t_{\pi_\theta}(s'|s)ds
\end{aligned}
$$

We find that the LHS of equation is not an integral, and we want to make it a an integral w.r.t s. To do this, first notice that LHS is independent of s so we can move it inside or out side of integral. Second, the integral of density function is 1 by definition. Therefore:

$$\int_s \nabla_\theta d_\pi(s')t_{\pi_\theta}(s|s')ds = \int_s \sum_a t(s'|s,a)d_{\pi_\theta}(s)\nabla_\theta \pi_\theta(a|s)ds + \int_s \nabla_\theta d_{\pi_\theta}(s)t_{\pi_\theta}(s'|s)ds$$

By looking at the first term in the RHS of equation, we find that $\sum_a t(s'|s,a)\nabla_\theta \pi_\theta(a|s) = \nabla_\theta t_{\pi_\theta}(s'|s)$ So

$$\int_s \nabla_\theta d_\pi(s')t_{\pi_\theta}(s|s')ds = \int_s \nabla_\theta t_{\pi_\theta}(s'|s)d_{\pi_\theta}(s)ds + \int_s \nabla_\theta d_{\pi_\theta}(s)t_{\pi_\theta}(s'|s)ds$$

Try 3: Another idea is that we can take derivative of the Expectation term in Theorem 1 and assume expectation and taking gradient can be interchanged. Assume $e(s)$ is any test function that is integrable

$$\because d_\theta(s') = \iint_{(s,a)} P(s'|s,a)d\pi_\theta(a|s)dd_\theta(s)$$
$$\therefore \int_{s'} e(s')dd_\theta(s') = \iiint_{(s,a,s')} e(s')dd_\theta(s)d\pi_\theta(a|s)dP(s'|s,a)$$

Differentiating both size w.r.t θ

$$\begin{aligned}\int_{s'} e(s')\nabla_\theta log(f_{d_\theta}(s'))dd_\theta(s') = &\iiint_{(s,a,s')} e(s')\nabla_\theta log(f_{d_\theta}(s))dd_\theta(s)d\pi_\theta(a|s)dP(s'|s,a)\\ &+\iiint_{(s,a,s')} e(s')\nabla_\theta log(f_{d_{\pi_\theta}}(s))dd_\theta(s)d\pi_\theta(a|s)dP(s'|s,a)\end{aligned}$$

Denote $w(s) = \nabla_\theta log(f_{d_\theta}(s))$, and since $e(s)$ is arbitrary, the function inside the integral must be equal almost everywhere, therefore

$$w(s') = \mathbb{E}_{(s,a)|s'\sim d_{\pi_0}}[w(s) + \nabla_\theta log(f_{d_{\pi_\theta}}(s))](*)$$

So what we should do is to find a w which satisfies (*)

In order to obtain $w(s)$ from the previous expression(*), we have the following strategies:

Approach 1: We solve

$$\sum_{i=1}^{n}(w(s_{i+1}) - w(s_i) - \nabla_\theta logf(a_i|s_i))^2$$

Where

1.In the small state-space case, we directly solve $w(s)$ for all states s.

2.In the large state-space case, we approximate $w(s)$ by $\sum q_j\phi_j(s)$, where ϕ_j is a set of basis functions. Then here we solve for the q_j.

Ideas on how to calculate $\nabla_\theta logf(a_i|s_i)$ in discrete setting:
Suppose the there are n state space, denoted by $s_1, s_2 \cdot\cdot\cdot, s_n$
Suppose for each state $s = s_i$, there are totally m_i actions to be chosen, so the conditional distribution function is:

$$p(a = x|s = si) = \prod_{j=1}^{m_i} \theta_{ij}^{x_j} \text{①}$$

Where $x = (x_1, x_2, \cdot\cdot\cdot, x_{m_i}), x_i \in \{0, 1\}$, and $\sum_{j=1}^{m_i} x_j = 1$, $x_j = 1$ denote takes action j, otherwise 0. And $\sum_{j=1}^{m_i} \theta_{ij} = 1, \forall i$

$$\therefore p(a = x|s) = \sum_{i=1}^{n} [\mathbb{1}_{\{s=s_i\}} \prod_{j=1}^{m_i} \theta_{ij}^{x_j}] \triangleq f_{d_{\pi_\theta}}(s)$$

$$\therefore log(f_{d_{\pi_\theta}}(s)) = \mathbb{1}_{\{s=s_i\}} log \prod_{j=1}^{m_i} \theta_{ij}^{x_j}$$

We can assume further that $m_1 = m_2 \cdot\cdot\cdot = m_n$, i.e. every state has the same action. Therefore we can derive that:

$$log(f_{d_{\pi_\theta}}(s_i)) = \sum_{j=1}^{m} x_j log\theta_{ij}$$

$$\forall i' \neq i, \forall j, \frac{\partial log(f_{d_{\pi_\theta}}(s_i))}{\partial\theta_{i'j}} = 0$$

$$\frac{\partial log(f_{d_{\pi_\theta}}(s_i))}{\partial\theta_{ij}} = \frac{x_j}{\theta_{ij}} - \frac{x_m}{1 - \sum_{j\neq m} \theta_{ij}} = \frac{x_j}{\theta_{ij}} - \frac{x_m}{\theta_{im}}$$

$$\therefore \nabla_\theta(og(f_{d_{\pi_\theta}}(s_i))) = (\frac{\partial log(f_{d_{\pi_\theta}}(s_i))}{\partial\theta_{ij}})_{i,j,i=1,2\cdots n,j=1,2\cdots m}$$

When both action space and state space is small, we can enumerate all situation and solve $w(s)$ exactly.
When either of them is large, if we generate the Markov Chain for several steps. Then if we cansider consider a single steps from s to s', i.e. s_i to s_j, the previous state and the action we take is all known and fixed, so we can plug in plug everything to the expression to obtain $\nabla_\theta(og(f_{d_{\pi_\theta}}(s_i)))$, which is a vector of length n*m. Then we plug it into the cost function and use similar method as linear regression, we can get a least square approximation of $w(s)$