Assumptions and derivation of graduate research initiatives

In order that our new method is correct and valid, there are a few technical conditions that should be satisfied. Therefore, we have made some assumptions of our settings, which is inevitable when applying our new method.

Assumptions:

Part 1

On-policy

In our derivation in the later chapter, we have to take gradient of expect reward function and move it inside the integral (Note that expectation is an integral). So it is required that the operator of taking gradient and taking integral can be interchanged. So we have:

Assumption 1(Interchangeable of $\nabla and \int$):

$\nabla_\theta \int \pi_\theta(\tau) r(\tau) d\tau = \int \nabla_\theta \pi_\theta(\tau) r(\tau) d\tau$

To figure out the when assumption 1 will be satisfied, first we should notice that the variable that we are taking gradient with respect to is not the variable that has been integrated. So it is different from the situation of fundamental theorem of calculus, in which the two variables I mentioned above are the same. Generally our problem is whether:

$\frac{d}{dt}\int_a^b f(x,t)dx = \int_a^b \frac{\partial}{\partial t} f(x,t)dx (1)$

holds. A weak condition I could come up with is to use dominated convergence theorem. The basic idea is to write LHS of (1) in the form of difference quotient, and use related theorem to change the order of differentiation and integral.

Theorem1 . Suppose that $f : X \times [a,b] \to \mathbb{R}$ and $f(\cdot,t) : X \to \mathbb{R}$ is integrable for each $t \in [a,b]$. Let $F(t) = \int f(x,t)dx$. Suppose that $\frac{\partial f}{\partial t}$ exists and there is a $g \in L^1$ such that $|(\frac{\partial f}{\partial t})(x,t)| \leq g(x), \forall x,t$. Then F is differentiable and $F'(x) = \int \frac{\partial f}{\partial t}(x,t)dx$

Proof: $\forall t_0$, Let $h_n(x) = \frac{f(x,t_n)-f(x,t_0)}{t_n - t_0}$, where $\{t_n\}$ is any sequence that converges to t_0. By the definition of derivative

$\frac{\partial f}{\partial t}(x,t_0) = \lim h_n(x)$

Since f is measurable, so h_n is measurable, $\frac{\partial f}{\partial t}$ is limit of h_n thus measurable. By the mean value theorem,

$|h_n(x)| \leq \sup_{t\in[a,b]} |\frac{\partial f}{\partial t}(x,t)| \leq g(x)$

Finally using dominated convergence theorem

$F'(t_0) = \lim \frac{F(t_n)-F(t_0)}{t_n - t_0} = \lim \int h_n(x)dx = \int \frac{\partial f}{\partial t}(x,t_0)dx$ ■

Notice that dominated theorem applies to any measure, thus the measure induced by the probability distribution of environment parametrized by θ is also applicable here. And notice that when dealing with taking gradient, we just have to check the condition entry by entry thus the problem is reduced to 1-D case proved above.

In our method, we make use of stationary distribution of the markov chain parametrized by θ. More precisely, we require it to be a limiting distribution, under which assumption the chain has the unique stationary distribution equals to the limiting distribution. Furthermore, when doing optimization θ has been updated, we require there exists a limiting distribution for every fixed θ. Formally we have:

Assumption 2(Limiting distribution):
$\forall\theta \in \Theta$, the markov chain generated by θ has a limiting distribution.
A sufficient but not necessary condition of having a limiting distribution is that the chain is aperiodic and irreducible.
Part 2
Off-policy
In the setting of off-policy, there are two policies, and we have to use importance sampling. A basic requirement is that the density appeared at denominator be nonzero whenever nominator is nonzero. So we have:
Assumption 3(Importance sampling):
$\forall t \in [0, T]$,when π_θ is not zero,$\pi_{\theta'}$ should be nonzero.

Basic formula
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
- control variate
  "centralize" the reward of the policy by subtracting the average reward from reward obtained at each step.

Shortcomings of existing method and instinction

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:

1. $d_{\pi,t}(\cdot)$: the distribution of state s_t at time t when using policy π and start from s_0 from initial distribution $d_0(\cdot)$
2. $d_\pi(s)$: the average visitation distribution, based on $d_{\pi,t}(\cdot)$

$$d_\pi(s) = \lim_{T\to\infty}\frac{\sum_{t=0}^{T}\gamma^t d_{\pi,t}(s)}{\sum_{t=0}^{T}\gamma^t} \tag{2}$$

(a) if $\gamma \in (0,1)$ (discounted reward case), by geometric series we have

$$d_\pi(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t d_{\pi,t}(s) \tag{3}$$

(b) 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

3. R_π: the expected reward of π

(a) $p_\pi(\cdot)$: the distribution of trajectory τ under policy π

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

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

(c) R_π: the expected reward of policy π

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

4. 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:

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

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

Steps

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

(a) run π_0 (original policy), generate the data,

(b) 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.

2. 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 $w_{\pi/\pi_0}(s)$

3. in average reward case:

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

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

   (b) 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: How to obtain $\nabla_\theta \log d_{\pi_\theta}(s)$ instead of $\log d_{\pi_\theta}(s)$?
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 $\nabla_\theta \log d_{\pi_\theta}(s)$

Method of solving the problem previously stated

Here for simplicity, we constrained our derivation to on-policy setting. Notice that the result of off-policy differs the one of on-policy just by a factor of cumulative production of density ratio.

1. 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

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

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

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

2. 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

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

Differentiating both size w.r.t θ, by log-identity

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

where f_{d_θ} and f_{π_θ} denote the densities of d_θ and π_θ

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

$$w(s') = \mathbb{E}_{(s,a)|s' \sim d_{\pi_0}}[w(s) + \nabla_\theta \log(f_{\pi_\theta}(a|s))] \tag{16}$$

So what we should do is to find a w which satisfies (16)

Here we use the latter method to do implementations, and all the following discussions from now on are based on method 2

The approach of estimating $w(s)$

In order to obtain $w(s)$ from the previous expression (16), we have the following strategies:

Approach 1: (Least Square) We solve

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

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.

Note: In least square method, we find that the solution of $w(s)$ is invariant to parallel shift. That is, if $w^*(s)$ is a solution, then $w^*(s)+c$ is also a solution for any constant c. To handle this problem, we notice that w itself is gradient of log-likelihood function. Under the assumption that taking gradient and taking integration can be interchanged, the expectation of log-likelihood function is always zero. Using this proposition, we could calculate a solution first and deducted its mean to obtain the real solution of w.

We would like to calculate the minimum value of square sum, for which we need $\nabla_\theta \log f(a_i|s_i)$. The approach to calculate it is as follows.

Calculate $\nabla_\theta \log f(a_i|s_i)$

in discrete setting:

Suppose the there are state space $s_1, s_2, \cdots, s_n$, for each state s_i, there are m_i actions to be chosen, so the empirical conditional distribution function is:

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

Where

- x_j is the indicator function of taking action j
- $\sum_{j=1}^{m_i} \theta_{ij} = 1, \forall i$ as the selection probability.
- $x = (x_1, x_2, \cdots, x_{m_i}) \in \{0,1\}^{m_i}$ as the binary notation of selecting action j

Write the conditional distribution with an indicator function to write the complete distribution of a

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

$f_{d_{\pi_\theta}}(a|s) = \pi_\theta(a|s)$: the distribution of action a given state$s$ and policy π_θ
For each individual s_i

$$\log f_{d_{\pi_\theta}}(a|s_i) = \log \prod_{j=1}^{m_i} \theta_{ij}^{x_j}$$

We can assume that every state has the same action space, which is common for infinite horizon problem. from log, we expand the product

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

The selection probability from another state should not affect the selection probability of state i, i.e.

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

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

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

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, we generate the Markov Chain for several steps. Then if we 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, we can plug in everything to the expression to obtain $\nabla_\theta(log(f_{d_{\pi_\theta}}(s_i)))$, which is a vector of length n*m. Then we plug it into the cost function and use similar method as linear regression, thus we can get a least square approximation of $w(s)$.