APPENDICES: Taylor Expansions of Discount Factors

A. Proofs

Proposition 3.1. The following holds for all $K \geq 0$,

$$V_{\gamma'}^{\pi} = \sum_{k=0}^{K} \left((\gamma' - \gamma)(I - \gamma P^{\pi})^{-1} P^{\pi}\right)^{k} V_{\gamma}^{\pi} + \underbrace{\left((\gamma' - \gamma)(I - \gamma P^{\pi})^{-1} P^{\pi}\right)^{K+1} V_{\gamma'}^{\pi}}_{\text{residual}}. \tag{9}$$

When $\gamma < \gamma' < 1$, the residual norm converges to 0, which implies

$$V_{\gamma'}^{\pi} = \sum_{k=0}^{\infty} \left((\gamma' - \gamma)(I - \gamma P^{\pi})^{-1} P^{\pi}\right)^{k} V_{\gamma}^{\pi}. \tag{10}$$

Proof. Recall the Woodbury matrix identity

$$(I - \gamma' P^{\pi})^{-1} = (I - \gamma P^{\pi})^{-1} + (\gamma' - \gamma)(I - \gamma P^{\pi})^{-1} P^{\pi} (I - \gamma' P^{\pi})^{-1}.$$

Recall the equality $V_{\gamma'}^{\pi} = (I - \gamma' P^{\pi})^{-1} r^{\pi}$. By plugging in the Woodbury matrix identity, this immediate shows

$$\begin{aligned} V_{\gamma'}^{\pi} &= (I - \gamma P^{\pi})^{-1} r^{\pi} + (\gamma' - \gamma)(I - \gamma P^{\pi})^{-1} P^{\pi} (I - \gamma' P^{\pi})^{-1} r^{\pi} \\ &= V_{\gamma}^{\pi} + (\gamma' - \gamma)(I - \gamma P^{\pi})^{-1} P^{\pi} V_{\gamma'}^{\pi}. \end{aligned}$$

Now, observe that the second term involves $V_{\gamma'}^{\pi}$. We can plug in the definition of $V_{\gamma'}^{\pi} = (I - \gamma' P^{\pi})^{-1} r^{\pi}$ and invoke the Woodbury matrix identity again. This produces

$$V_{\gamma'}^{\pi} = V_{\gamma}^{\pi} + (\gamma' - \gamma)(I - \gamma P^{\pi})^{-1} P^{\pi} V_{\gamma}^{\pi} + \left((\gamma' - \gamma)(I - \gamma P^{\pi})^{-1} P^{\pi}\right)^{2} V_{\gamma'}^{\pi}.$$

By induction, it is straightforward to show that iterating the above procedure $K \geq 0$ times produces the following equalities

$$V_{\gamma'}^{\pi} = \sum_{k=0}^{K} \left((\gamma' - \gamma)(I - \gamma P^{\pi})^{-1} P^{\pi}\right)^{k} V_{\gamma}^{\pi} + + \underbrace{\left((\gamma' - \gamma)(I - \gamma P^{\pi})^{-1} P^{\pi}\right)^{K+1} V_{\gamma'}^{\pi}}_{\text{residual}}.$$

Consider the norm of the residual term. Since P^{π} is a transition matrix, $\|P^{\pi}\|_{\infty} < 1$. As a result, $\left\|(I - \gamma P^{\pi})^{-1}\right\|_{\infty} = \|\sum_{t=0}^{\infty} \gamma^{t} (P^{\pi})^{t}\|_{\infty} < (1 - \gamma)^{-1}$. This implies

$$\left\|\left((\gamma' - \gamma)(I - \gamma P^{\pi})^{-1} P^{\pi}\right)^{K+1} V_{\gamma'}^{\pi}\right\|_{\infty} < \left(\frac{\gamma' - \gamma}{1 - \gamma}\right)^{K+1} \cdot \frac{R_{\max}}{1 - \gamma'}.$$

When $\gamma < \gamma' < 1$, the residual norm decays exponentially and $\to 0$ as $K \to \infty$. This implies that the infinite series converges,

$$V_{\gamma'}^{\pi} = \sum_{k=0}^{\infty} \left((\gamma' - \gamma)(I - \gamma P^{\pi})^{-1} P^{\pi}\right)^{k} V_{\gamma}^{\pi}.$$

Additional consideration when $\gamma' = 1$. When $\gamma' = 1$, in order to ensure finiteness of $V_{\gamma'=1}^{\pi}$, we assume the following two conditions: **(1)** The Markov chain induced by π is absorbing; **(2)** for any absorbing state x, $r^{\pi}(x) = 0$. Without loss of generality, assume there exists a single absorbing state. In general, the transition matrix P^{π} can be decomposed as follows (Grinstead and Snell, 2012; Ross, 2014),

$$P^{\pi} = \begin{pmatrix} \widetilde{P}^{\pi} & \widetilde{p}^{\pi} \\ 0 & 1 \end{pmatrix},$$

where $\widetilde{P}^{\pi} \in \mathbb{R}^{(|\mathcal{X}|-1)|\mathcal{A}| \times (|\mathcal{X}|-1)|\mathcal{A}|}$ and $\widetilde{p}^{\pi} \in \mathbb{R}^{\mathcal{X}-1)}$. Here, the first $\mathcal{X} - 1$ states are transient and the last state is absorbing. For convenience, define $\widetilde{r}^{\pi}$ as the reward vector r^{π} constrained on the first $\mathcal{X} - 1$ transient states. We provide a few lemmas below.

Lemma A.1. The matrix $(I - \widetilde{P})^\pi$ is invertible and its inverse is $(I - \widetilde{P})^\pi = \sum_{k=0}^{\infty} (\widetilde{P})^k$.

Proof. Define a matrix $N = \sum_{k=0}^{\infty} (\widetilde{P}^\pi)^k$, then $N[x, y]$ defines the expected number of times it takes to transition from x to y before absorption. By definition of the absorbing chain, N is finite. This further shows that $(I - \widetilde{P}^\pi)$ is invertible, because

$$N(I - \widetilde{P}^\pi) = (I - \widetilde{P}^\pi)N = I.$$

□

Lemma A.2. Let $f(A, B)$ be a matrix polynomial function of matrix A and B. Then

$$f\left(P^\pi, (I - \gamma P^\pi)^{-1}\right) = \begin{pmatrix} f\left(\widetilde{P}^\pi, \left(I - \gamma \widetilde{P}^\pi\right)^{-1}\right) & B \\ 0 & 1 \end{pmatrix},$$

where B is some matrix.

Proof. The intuition for the above result is that polynomial transformation preserves the *block triangle* property of P^π and $(I - \gamma P^\pi)^{-1}$. In general, we can assume

$$f(A, B) = \sum_{m,n \le K} c_{m,n} A^m B^n,$$

for some $K \ge 0$ and $c_{m,n} \in \mathbb{R}$ are scalar coefficients. First, note that $(P^\pi)^k, k \ge 0$ is of the form

$$(P^\pi)^k = \begin{pmatrix} \left(\widetilde{P}^\pi\right)^k & C \\ 0 & 1 \end{pmatrix},$$

for some matrix C. Since $(I - \gamma A)^{-1} = \sum_{k=0}^{\infty} A^k$ for $A \in \{P^\pi, \widetilde{P}^\pi\}$, this implies that

$$(I - \gamma P^\pi)^{-1} = \sum_{k=0}^{\infty} (\gamma P^\pi)^k = \begin{pmatrix} \left(\widetilde{P}^\pi\right)^k & D \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \left(I - \gamma \widetilde{P}^\pi\right)^{-1} & D \\ 0 & 1 \end{pmatrix},$$

for some matrix D. The above two results show that both polynomials of P^π and $(I - \gamma P^\pi)^{-1}$ are *block upper triangle* matrices. It is then straightforward that

$$(P^\pi)^m \left((I - \gamma P^\pi)^{-1}\right)^n = \begin{pmatrix} \left(\widetilde{P}^\pi\right)^m \left(\left(I - \gamma \widetilde{P}^\pi\right)^{-1}\right)^n & E \\ 0 & 1 \end{pmatrix},$$

for some matrix E. Finally, since $f(P^\pi, (I - \gamma P^\pi)^{-1})$ is a linear combination of $(P^\pi)^m \left((I - \gamma P^\pi)^{-1}\right)^n$, we conclude the proof. □

Lemma A.3. Under assumption **(1)** and **(2)**, one could write the value function $V^\pi_{\gamma'=1}$ as

$$V^\pi_{\gamma'=1} = \sum_{k=0}^{\infty} (P^\pi)^t r^\pi,$$

where the infinite series on the RHS converges. In addition, for any transient state x, $V^\pi_{\gamma'=1}(x) = \left[\sum_{k=0}^{\infty} (\widetilde{P}^\pi)^k \widetilde{r}^\pi\right](x)$.

Proof. Recall that $V^\pi_\gamma(x) \coloneqq \mathbb{E}\left[\sum_{t=0}^\infty r_t \mid x_0 = x\right]$. Under assumption **(2)**, for any absorbing state x, $V^\pi_\gamma(x) = 0 = \left[\sum_{k=0}^\infty (P^\pi)^k r^\pi\right](x)$. We can instead constrain the Markov chain to the transient states. For any transient state x, recall the definition of N from Lemma A.1, it follows that

$$V^\pi_{\gamma'=1}(x) = \sum_y N(\text{expected number of times in } y | x_0 = x) r^\pi(y) = \left[N r^\pi\right](x) = \left[\left(I - \widetilde{P}^\pi\right)^{-1} \widetilde{r}^\pi\right](x) = \left[\sum_{k=0}^\infty (\widetilde{P}^\pi)^k \widetilde{r}^\pi\right](x).$$

By Lemma A.2, this is equivalent to $\left[\sum_{k=0}^\infty (P^\pi)^t r^\pi\right](x)$. We thus complete the proof. □

Lemma A.4. The following holds for any $\gamma < 1$,

$$\begin{aligned}
\left(I - \widetilde{P}^\pi\right)^{-1} &= \left(I - \gamma\widetilde{P}^\pi - (1-\gamma)\widetilde{P}^\pi\right)^{-1} \\
&= \sum_{k=0}^K \left((1-\gamma)\left(I - \gamma\widetilde{P}^\pi\right)^{-1}\widetilde{P}^\pi\right)^k \left(I - \gamma\widetilde{P}^\pi\right)^{-1} + \left((1-\gamma)\left(I - \gamma\widetilde{P}^\pi\right)^{-1}\widetilde{P}^\pi\right)^{K+1}\left(I - \widetilde{P}^\pi\right)^{-1} \\
&= \sum_{k=0}^\infty \left((1-\gamma)\left(I - \gamma\widetilde{P}^\pi\right)^{-1}\widetilde{P}^\pi\right)^k \left(I - \gamma\widetilde{P}^\pi\right)^{-1}. \qquad (20)
\end{aligned}$$

Proof. The first two lines derive from a straightforward application of Woodbury matrix identity to $(I - \widetilde{P}^\pi)^{-1}$. This is valid because by Lemma A.1, $(I - \widetilde{P}^\pi)$ is invertible. The convergence of the infinite series is guaranteed for all $\gamma < 1$. To see why, recall that the finiteness of $N = \sum_{k=0}^\infty (\widetilde{P}^\pi)^k$ implies $(\widetilde{P}^\pi)^{K+1} \to 0$. We can bound the residual,

$$\left\|\left((1-\gamma)\left(I - \gamma\widetilde{P}^\pi\right)^{-1}\widetilde{P}^\pi\right)^{K+1}\left(I - \widetilde{P}^\pi\right)^{-1}\right\|_\infty \le \left\|(\widetilde{P}^\pi)^{K+1}\left(I - \widetilde{P}^\pi\right)^{-1}\right\|_\infty \to 0.$$

□

Finally, we combine results from the above to prove the main claim. First, consider the absorbing state x. Due to Assumption **(2)**, $V^\pi_\gamma(x) = 0$ for any $\gamma \in [0, 1]$. The matrix equalities in Proposition 3.2 holds in this case.

In the following, we consider any transient states x. By Lemma A.3 and Lemma A.4

$$\begin{aligned}
\widetilde{V}^\pi_{\gamma'=1}(x) &= \left[\sum_{k=0}^\infty (\widetilde{P}^\pi)^k \widetilde{r}^\pi\right](x) \\
&= \left[\sum_{k=0}^K \left((1-\gamma)\left(I - \gamma\widetilde{P}^\pi\right)^{-1}\widetilde{P}^\pi\right)^k \left(I - \gamma\widetilde{P}^\pi\right)^{-1}\widetilde{r}^\pi + \left((1-\gamma)\left(I - \gamma\widetilde{P}^\pi\right)^{-1}\widetilde{P}^\pi\right)^{K+1}\left(I - \widetilde{P}^\pi\right)^{-1}\widetilde{r}^\pi\right](x)
\end{aligned}$$

Now, notice that because the last entries of $r^\pi, V^\pi_\gamma, V^\pi_{\gamma'=1}$ are zero (for the absorbing state),

$$\left[\left(I - \gamma\widetilde{P}^\pi\right)^{-1}\widetilde{r}^\pi\right](x) = \left[\left(I - \gamma P^\pi\right)^{-1} r^\pi\right](x).$$

Combining with Lemma A.2,

$$\begin{aligned}
\widetilde{V}^\pi_{\gamma'=1}(x) &= \left[\sum_{k=0}^K \left((1-\gamma)\left(I - \gamma P^\pi\right)^{-1} P^\pi\right)^k \left(I - \gamma P^\pi\right)^{-1}\widetilde{r}^\pi + \left((1-\gamma)\left(I - \gamma P^\pi\right)^{-1} P^\pi\right)^{K+1} V^\pi_{\gamma'=1}\right](x) \\
&= \left[\underbrace{\sum_{k=0}^K \left((1-\gamma)\left(I - \gamma P^\pi\right)^{-1} P^\pi\right)^k V^\pi_\gamma}_{\text{K-th order expansion}} + \underbrace{\left((1-\gamma)\left(I - \gamma P^\pi\right)^{-1} P^\pi\right)^{K+1} V^\pi_{\gamma'=1}}_{\text{residual}}\right](x).
\end{aligned}$$

The residual term $\to 0$ as $K \to 0$ with similar arguments used for Lemma A.4. We hence conclude the proof. □

Proposition 3.2. The following bound holds for all $K \geq 0$,

$$\left|V^\pi_{\gamma'}(x) - V^\pi_{K,\gamma,\gamma'}(x)\right| \leq \left(\frac{\gamma' - \gamma}{1-\gamma}\right)^{K+1} \frac{R_{\max}}{1-\gamma'}. \tag{12}$$

Proof. The proof follows directly from the residual term in Proposition 3.1. Recall that the residual term takes the form

$$V^\pi_{\gamma'} - V^\pi_{K,\gamma,\gamma'} = \left((\gamma' - \gamma)(I - \gamma P^\pi)^{-1} P^\pi\right)^{K+1} V^\pi_{\gamma'}.$$

Its infinity norm can be bounded as $(\frac{\gamma'-\gamma}{1-\gamma})^{K+1} \frac{R_{\max}}{1-\gamma'}$ □

Lemma 4.1. Assume $\gamma < \gamma' < 1$. We can write $V^\pi_{\gamma'}(x) = (\rho^\pi_{x,\gamma,\gamma'})^T V^\pi_\gamma$, where the weight vector $\rho^\pi_{x,\gamma,\gamma'} \in \mathbb{R}^{\mathcal{X}}$ is

$$\left(I - \gamma (P^\pi)^T\right)\left(I - \gamma' (P^\pi)^T\right)^{-1} \delta_x.$$

Also we can rewrite $V^\pi_{\gamma'}(x)$, using an expectation, as:

$$V^\pi_\gamma(x) + \mathbb{E}_\pi\left[\sum_{t=1}^{\infty} (\gamma' - \gamma)(\gamma')^{t-1} V^\pi_\gamma(x_t) \,\middle|\, x_0 = x\right]. \tag{16}$$

When $\gamma' = 1$, $\rho^\pi_{x,\gamma,\gamma'}$ might be undefined. However, Eqn (16) still holds if assumptions **A.1** and **A.2** are satisfied.

Proof. We will derive the above result with the matrix form. Recall by applying Woodbury inversion identity to $(I - \gamma' P^\pi)^{-1} = (I - (\gamma' - \gamma)P^\pi - \gamma P^\pi)^{-1}$, we get

$$\begin{aligned}
(I - \gamma' P^\pi)^{-1} &= \sum_{k=0}^{\infty} \left((\gamma' - \gamma)(I - \gamma P^\pi)^{-1} P^\pi\right)^k (I - \gamma P^\pi)^{-1} \\
&= (I - \gamma P^\pi)^{-1} + \sum_{k=1}^{\infty} \left((\gamma' - \gamma)(I - \gamma P^\pi)^{-1} P^\pi\right)^k (I - \gamma P^\pi)^{-1} \\
&= (I - \gamma P^\pi)^{-1} + (\gamma' - \gamma) \sum_{k=1}^{\infty} \left((\gamma' - \gamma)(I - \gamma P^\pi)^{-1} P^\pi\right)^k \cdot (I - \gamma P^\pi)^{-1} \cdot P^\pi (I - \gamma P^\pi)^{-1} \\
&= (I - \gamma P^\pi)^{-1} + (\gamma' - \gamma)(I - \gamma' P^\pi)^{-1} \cdot P^\pi \cdot (I - \gamma P^\pi)^{-1}.
\end{aligned}$$

Then, right multiply the above equation by r^π,

$$\begin{aligned}
V^\pi_{\gamma'} &= V^\pi_\gamma + (\gamma' - \gamma)(I - \gamma' P^\pi)^{-1} P^\pi V^\pi_\gamma \\
&= V^\pi_\gamma + (\gamma' - \gamma) \sum_{t=1}^{\infty} (\gamma')^{t-1} (P^\pi)^t V^\pi_\gamma.
\end{aligned}$$

By indexing both sides at state x, we recover the following equality,

$$V^\pi_{\gamma'}(x) = V^\pi_\gamma(x) + \mathbb{E}_\pi\left[\sum_{t=1}^{\infty} (\gamma' - \gamma)(\gamma')^{t-1} V^\pi_\gamma(x_t) \,\middle|\, x_0 = x\right].$$

To derive the expression for $\rho^\pi_{x,\gamma,\gamma}$, note that also

$$V^\pi_{\gamma'} = (I - \gamma' P^\pi)^{-1} r^\pi = (I - \gamma P^\pi)(I - \gamma' P^\pi)^{-1}(I - \gamma P^\pi)^{-1} r^\pi = \underbrace{(I - \gamma P^\pi)(I - \gamma' P^\pi)^{-1}}_{\text{weight matrix } W} V^\pi_\gamma,$$

where we use the fact that $(I - \gamma P^\pi)$ commutes with $(I - \gamma' P^\pi)^{-1}$. Since we define $\rho^\pi_{x,\gamma,\gamma'}$ as such that $V^\pi_{\gamma'}(x) = \left(\rho^\pi_{x,\gamma,\gamma'}\right)^T V^\pi_\gamma$, we can derive the matrix form of $\rho^\pi_{x,\gamma,\gamma'}$ by indexing the x-th row of weight matrix W. This directly leads to the desired result

$$\rho^\pi_{x,\gamma,\gamma'} = \left(I - \gamma (P^\pi)^T\right)\left(I - \gamma' (P^\pi)^T\right)^{-1} \delta_x.$$

□

Proposition 4.3. For any $\gamma < \gamma' < 1$, the first partial gradient $(\partial_V F(V_\gamma^{\pi_\theta}, \rho_{x,\gamma,\gamma'}^{\pi_\theta}))^T \nabla_\theta V_\gamma^{\pi_\theta}$ can be expressed as

$$\mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty}(\gamma')^t Q_\gamma^{\pi_\theta}(x_t,a_t)\nabla_\theta \log \pi_\theta(a_t|x_t) \;\middle|\; x_0 = x\right]. \tag{17}$$

When $\gamma' = 1$, under assumptions **A.1** and **A.2**, the first partial gradient exists and is expressed as

$$\mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T} Q_\gamma^{\pi_\theta}(x_t,a_t)\nabla_\theta \log \pi_\theta(a_t|x_t) \;\middle|\; x_0 = x\right]. \tag{18}$$

Proof. First we assume $\gamma' < 1$, we will consider the extension to $\gamma' = 1$ at the end of the proof. Recall that the policy gradient takes the following form,

$$\nabla_\theta V_\gamma^{\pi_\theta}(x) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t Q_\gamma^{\pi_\theta}(x_t,a_t)\nabla_\theta \log \pi_\theta(a_t|x_t) \;\middle|\; x_0 = x.\right]$$

We plug in the above, the partial derivative $\left(\partial_V F(V_\gamma^{\pi_\theta}, \rho_{x,\gamma,\gamma'}^{\pi_\theta})\right)^T \nabla_\theta V_\gamma^{\pi_\theta}$ evaluates to the following

$$\begin{aligned}
&\nabla_\theta V_\gamma^{\pi_\theta}(x) + \mathbb{E}_{\pi_\theta}\left[(\gamma'-\gamma)\sum_{t=1}^{\infty}(\gamma')^{t-1}\nabla_\theta V_\gamma^{\pi_\theta}(x_t)\right] \\
&= \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t Q_\gamma^{\pi_\theta}(x_t,a_t)\nabla_\theta \log \pi_\theta(a_t|x_t) \;\middle|\; x_0 = x\right] \\
&+ \mathbb{E}_{\pi_\theta}\left[(\gamma'-\gamma)(\gamma')^{t-1}\sum_{t=1}^{\infty}\sum_{s=0}^{\infty}\gamma^s Q_\gamma^{\pi_\theta}(x_{t+s},a_{t+s})\nabla_\theta \log \pi_\theta(a_{t+s}|x_{t+s}) \;\middle|\; x_0 = x\right] \\
&= \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty}\left(\underbrace{\gamma^t + \sum_{u=1}^{t}(\gamma'-\gamma)(\gamma')^{u-1}\gamma^{t-u}}_{\text{coefficient at time t}}\right) Q_\gamma^{\pi_\theta}(x_t,a_t)\nabla_\theta \log \pi_\theta(a_t|x_t) \;\middle|\; x_0 = x.\right]
\end{aligned}$$

In the above, the coefficient term at time t can be calculated by carefully grouping terms across different time steps. It can be shown that the coefficient term evaluates to $(\gamma')^t$ for all $t \geq 0$. This concludes the proof.

Alternative proof based on matrix notations. We introduce an alternative proof based on matrix notations as it will make the extension to $\gamma' = 1$ simpler. First, note that

$$V_{\gamma'}^{\pi} = \left(I - \gamma' P^{\pi}\right)^{-1} r^{\pi} = \left(I - \gamma' P^{\pi}\right)^{-1}\left(I - \gamma P^{\pi}\right)\left(I - \gamma P^{\pi}\right)^{-1} r^{\pi} = \left(I - \gamma P^{\pi}\right)\left(I - \gamma' P^{\pi}\right)^{-1}\left(I - \gamma P^{\pi}\right)^{-1} r^{\pi},$$

where for the second equality we exploit the fact that $(I - \gamma P^{\pi})$ commutes with $(I - \gamma' P^{\pi})^{-1}$. Now, notice that the above rewrites as

$$V_{\gamma'}^{\pi} = \underbrace{\left(I - \gamma P^{\pi}\right)\left(I - \gamma' P^{\pi}\right)^{-1}}_{W_{\gamma,\gamma'}} V_{\gamma}^{\pi}$$

where $W_{\gamma,\gamma'}$ is the weight matrix. This matrix is equivalent to the weighting distribution $\rho_{x,\gamma,\gamma'}^{\pi}$ by $W_{\gamma,\gamma'}[x] = \rho_{x,\gamma,\gamma'}^{\pi}$ where $A[x]$ is the x-th row of matrix A. The first partial gradient corresponds to differentiating $V_{\gamma'}^{\pi_\theta}$ only through $V_{\gamma}^{\pi_\theta}$. To make the derivation clear in matrix notations, let θ_i be the i-th component of the parameter θ. Define $\nabla_{\theta_i} V_\gamma^{\pi_\theta} \in \mathbb{R}^{\mathcal{X}}$ such that $\nabla_{\theta_i} V_\gamma^{\pi_\theta}(x) = \nabla_{\theta_i} V_\gamma^{\pi_\theta}(x)$, This means the i-th component of the first partial gradient across all states is

$$W_{\gamma,\gamma'} \nabla_{\theta_i} V_\gamma^{\pi_\theta} \in \mathbb{R}^{\mathcal{X}}.$$

Let $G_\gamma^{\theta_i} \in \mathbb{R}^{\mathcal{X}}$ to be the vector of local gradient (for parameter θ_i) such that $G_\gamma^{\theta_i}(x) = \sum_a \nabla_{\theta_i}\pi_\theta(a|x)Q_\gamma^{\pi_\theta}(x,a)$. Vanilla PG (Sutton et al., 2000) can be expressed as

$$\nabla_{\theta_i} V_\gamma^{\pi_\theta} = (I - \gamma P^\pi)^{-1} G_\gamma^{\theta_i}.$$

We can finally derive the following,

$$\begin{aligned} W_{\gamma,\gamma'}\nabla_{\theta_i}V_\gamma^{\pi_\theta} &= (I-\gamma P^\pi)(I-\gamma' P^\pi)^{-1}G_\gamma^{\theta_i} \\ &= (I-\gamma P^\pi)(I-\gamma' P^\pi)^{-1}(I-\gamma P^\pi)^{-1}G_\gamma^{\theta_i} \\ &= (I-\gamma P^\pi)(I-\gamma P^\pi)^{-1}(I-\gamma' P^\pi)^{-1}G_\gamma^{\theta_i} \\ &= (I-\gamma' P^\pi)^{-1}G_\gamma^{\theta_i} \end{aligned}$$

Now, consider the x-th component of the above vector. We have $\nabla_\theta[J(\pi_\theta,\pi_t)]_{\pi_t=\pi_\theta}$ is equal to

$$\mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty}(\gamma')^t\sum_a \nabla_{\theta_i}\pi_\theta(a|x_t)Q_\gamma^{\pi_\theta}(x_t,a)\,\middle|\, x_0=x\right] = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty}(\gamma')^t\nabla_{\theta_i}\log\pi_\theta(a_t|x_t)Q_\gamma^{\pi_\theta}(x_t,a)\,\middle|\, x_0=x\right]$$

When concatenating the gradient for all component $\theta_i i$ of θ, we conclude the proof.

Extensions to the case $\gamma' = 1$. Similar to the arguments made in the proof of Proposition 3.2, under assumptions **A.1** and **A.2**, we can decompose the transition matrix P^π as

$$P^\pi = \begin{pmatrix} \widetilde{P} & \widetilde{p} \\ 0 & 1 \end{pmatrix},$$

where the last state is assumed to be absorbing. Though $(I-\gamma' P^\pi)^{-1}$ for $\gamma'=1$ is in general not necessarily invertible, the matrix $(I-\widetilde{P})^{-1}$ is invertible. Since $r^\pi(x)$ for the absorbing state x, we have deduced that $Q_\gamma^\pi(x.a) = V_\gamma^\pi(x) = 0$, and accordingly $G_\gamma^{\theta_i}(x) = 0$. As such, though $(I-\gamma' P^\pi)^{-1}$ for $\gamma'=1$ might be undefined, the multiplication $(I-\gamma' P^\pi)^{-1}G_\gamma^{\theta_i}$ is defined, with the last entry being 0. Since at time $t=T$, the chain enters the absorbing states, all local gradient terms that come after T are zero. As a result, the x-th component of $(I-\gamma' P^\pi)^{-1}G_\gamma^{\theta_i}$ is

$$\mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T}(\gamma')^t\nabla_{\theta_i}\log\pi_\theta(a_t|x_t)Q_\gamma^{\pi_\theta}(x_t,a)\,\middle|\, x_0=x\right]$$

□

Proposition 4.4. Assume $\gamma < \gamma' < 1$. For any $x \in \mathcal{X}$, define the K^{th} Taylor expansion to $\rho^\pi_{x,\gamma,\gamma'}$ as

$$\rho^\pi_{x,K,\gamma,\gamma'} = \sum_{k=0}^{K}\left((\gamma'-\gamma)\left(I-\gamma(P^\pi)^T\right)^{-1}(P^\pi)^T\right)^k\delta_x.$$

It can be shown that $V^\pi_{K,\gamma,\gamma'}(x) = (\rho^\pi_{x,K,\gamma,\gamma'})^T V_\gamma^\pi$ and $\left\|\rho^\pi_{x,K,\gamma,\gamma'} - \rho^\pi_{K,\gamma,\gamma'}\right\|_\infty = O\left(\left(\frac{\gamma'-\gamma}{1-\gamma}\right)^{K+1}\right)$.

Proof. Recall from Lemma 4.1, by construction,

$$\rho^\pi_{x,\gamma,\gamma'} = \left(I-\gamma'(P^\pi)^T\right)^{-1}\left(I-\gamma(P^\pi)^T\right)\delta_x.$$

Simialr to the case of primal space expansions in Section 3.1, we construct the K^{th} order expansion to $\rho^\pi_{x,\gamma,\gamma'}$ via the expansion of the matrix $\left(I-\gamma(P^\pi)^T\right)^{-1}$. Recall that

$$\left(I-\gamma'(P^\pi)^T\right)^{-1} = \sum_{k=0}^{\infty}\left((\gamma'-\gamma)(P^\pi)^T\left(I-\gamma(P^\pi)^T\right)^{-1}\right)^k\left(I-\gamma(P^\pi)^T\right)^{-1}.$$

□

When truncating the infinite series to the first $K+1$ terms, we derive the K^{th} order expansion $\rho^{\pi}_{x,K,\gamma,\gamma'}$,

$$\left((\gamma'-\gamma)(P^\pi)^T\left(I-\gamma(P^\pi)^T\right)^{-1}\right)^k\left(I-\gamma(P^\pi)^T\right)^{-1}\left(I-\gamma(P^\pi)^T\right)\delta_x=\sum_{k=0}^{K}\left((\gamma'-\gamma)(P^\pi)^T\left(I-\gamma(P^\pi)^T\right)^{-1}\right)^k\delta_x.$$

Note that since

$$\left\|\sum_{k=K+1}^{\infty}\left((\gamma'-\gamma)(P^\pi)^T\left(I-\gamma(P^\pi)^T\right)^{-1}\right)^k\left(I-\gamma(P^\pi)^T\right)^{-1}\right\|_\infty\leq\left(\frac{\gamma'-\gamma}{1-\gamma}\right)^{K+1}\frac{1}{1-\gamma'}.$$

This concludes the proof.

B. Further results on Taylor expansions in the dual space

The dual representation of value function $V^\pi_{\gamma'}(x)$ in Eqn (6) is $V^\pi_{\gamma'}(x)=(1-\gamma')^{-1}(r^\pi)^T d^\pi_{x,\gamma'}$ where $r^\pi, d^\pi_{x,\gamma'}\in\mathbb{R}^{\mathcal{X}}$ are vector rewards and visitation distribution starting at state x. Here, we abuse the notation $d^\pi_{x,\gamma}$ to denote both a function and a vector, i.e., $d^\pi_{x,\gamma}(x')$ can be interpreted as both a function evaluation and a vector indexing. Given such a dual representation, one natural question is whether the K^{th} expansion in the primal space corresponds to some approximations of the discounted visitation distribution $d^\pi_{K,\gamma,\gamma'}\approx d^\pi_{x,\gamma'}$. Below, we answer in the affirmative.

Let $\delta_x\in\mathbb{R}^{\mathcal{X}}$ be the one-hot distribution such that $[\delta_x]_{x'}=1$ only when $x'=x$. The visitation distribution satisfies the following balance equation in matrix form

$$d^\pi_{x,\gamma'}=(1-\gamma')\delta_x+\gamma'(P^\pi)^T d^\pi_{x,\gamma'}. \tag{21}$$

Inverting the equation, we obtain an explicit expression for the visitation distribution $d^\pi_{x,\gamma'}=(1-\gamma')(I-\gamma' P^\pi)^{-1}\delta_x$. Following techniques used in the derivation of Propo 3.1, we can derive similar approximation results for dual variables. See Appendix B.

Proposition B.1. The following holds for all $K\geq 0$,

$$\begin{aligned}d^\pi_{x,\gamma'}=&\frac{1-\gamma'}{1-\gamma}\sum_{k=0}^{K}\left((\gamma'-\gamma)\left(I-\gamma(P^\pi)^T\right)^{-1}(P^\pi)^T\right)^k d^\pi_{x,\gamma}.\\&+\underbrace{\left((\gamma'-\gamma)\left(I-\gamma(P^\pi)^T\right)^{-1}(P^\pi)^T\right)^{K+1}d^\pi_{x,\gamma'}}_{\text{residual}}.\end{aligned} \tag{22}$$

When $\gamma<\gamma'<1$, the residual norm $\to 0$, which implies that the following holds

$$d^\pi_{x,\gamma'}=\frac{1-\gamma'}{1-\gamma}\sum_{k=0}^{\infty}((\gamma'-\gamma)\left(I-\gamma(P^\pi)^T\right)^{-1}(P^\pi)^T)^k d^\pi_{x,\gamma}. \tag{23}$$

Proof. Starting from the fixed point equation satisfied by $d^\pi_{\gamma'}$, we can apply Woodbury inversion indentity

$$\begin{aligned}d^\pi_{\gamma'}&=(1-\gamma')\left(I-\gamma'(P^\pi)^T\right)^{-1}\delta_x\\&=(1-\gamma')\sum_{k=0}^{K}\left((\gamma'-\gamma)(I-\gamma P^\pi)^{-1}(P^\pi)^T\right)^k\delta_x+(1-\gamma')\left((\gamma'-\gamma)\left(I-\gamma(P^\pi)^T\right)^{-1}(P^\pi)^T\right)^K\left(I-\gamma'(P^\pi)^T\right)^{-1}\delta_x\\&=\frac{1-\gamma'}{1-\gamma}\sum_{k=0}^{K}\left((\gamma'-\gamma)\left(I-\gamma(P^\pi)\right)^{-1}P^\pi\right)^k d^\pi_{\gamma}+(1-\gamma')\left((\gamma'-\gamma)(I-\gamma P^\pi)^{-1}(P^\pi)^T\right)^K d^\pi_{\gamma'}\end{aligned}$$

The norm of the residual term could be bounded as

$$\left\|(1-\gamma')\left((\gamma'-\gamma)(I-\gamma P^\pi)^{-1}P^\pi\right)^K d^\pi_{\gamma'}\right\|_\infty\leq(1-\gamma')\left(\frac{\gamma'-\gamma}{1-\gamma}\right)^{K+1}\to 0.$$

□

With a similar motivation as expansions in the primal space, we define the K^{th} order expansion by truncating to first $K+1$ terms,

$$d^{\pi}_{x,K,\gamma,\gamma'} \coloneqq \frac{1-\gamma'}{1-\gamma}\sum_{k=0}^{K}((\gamma'-\gamma)(I-\gamma P^{\pi})^{-1}P^{\pi})^{k}d^{\pi}_{x,\gamma} \tag{24}$$

The following result formalizes the connection between the K^{th} order dual approximation to the visitation distribution $d^{\pi}_{K,\gamma,\gamma'}$ and the primal approximation to the value function at state x, $V^{\pi}_{K,\gamma,\gamma'}(x)$.

Proposition B.2. The K^{th} order primal and dual approximations are related by the following equality for any $K \geq 0$,

$$V^{\pi}_{K,\gamma,\gamma'}(x) = (1-\gamma')^{-1}\left(d^{\pi}_{x,K,\gamma,\gamma'}\right)^{T} r^{\pi} \tag{25}$$

Proof. The proof follows by expanding out the RHS of the equation. Recall the definition of $d^{\pi}_{K,\gamma,\gamma'}$,

$$\begin{aligned}
(d^{\pi}_{K,\gamma,\gamma'})^{T} &= \frac{1-\gamma'}{1-\gamma}\sum_{k=0}^{K}(d^{\pi}_{\gamma})^{T}\left((\gamma'-\gamma)\left(I-\gamma P^{\pi}\right)^{-1}\right)^{k} \\
&= (1-\gamma')\sum_{k=0}^{K}\delta_x^{T}(I-\gamma P^{\pi})^{-1}\left((\gamma'-\gamma)\left(I-\gamma P^{\pi}\right)^{-1}\right)^{k} \\
&= (1-\gamma')\delta_x^{T}\left[\sum_{k=0}^{K}\left((\gamma'-\gamma)(I-\gamma P^{\pi})^{-1}P^{\pi}\right)^{k}\right]\cdot(I-\gamma P^{\pi})^{-1}.
\end{aligned}$$

Now multiply the RHS by r^{π} and recall that $V^{\pi}_{\gamma} = (I-\gamma P^{\pi})^{-1}r^{\pi}$, we conclude the proof,

$$\text{RHS} = \frac{1-\gamma'}{1-\gamma}\delta_x^{T}\left[\sum_{k=0}^{K}\left((\gamma'-\gamma)(I-\gamma P^{\pi})^{-1}P^{\pi}\right)^{k}\right]V^{\pi}_{\gamma} = (1-\gamma')\delta_x^{T}V^{\pi}_{K,\gamma,\gamma'} = (1-\gamma')V^{\pi}_{K,\gamma,\gamma'}(x).$$

□

Proposition B.2 shows that indeed, the K^{th} order approximation of the value function is equivalent to the K^{th} order approximation of the visitation distribution in the dual space. It is instructive to consider the special case $K=1$.

C. Details on Taylor expansion Q-function advantage estimation

Proposition C.1. Let $Q^{\pi}_{\gamma} \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ be the vector advantage functions. Let $\overline{P}^{\pi} \in \mathbb{R}^{(\mathcal{X}\times\mathcal{A})\times(\mathcal{X}\times\mathcal{A})}$ be the transition matrix such that $\overline{P}^{\pi}(x,a,x',a') = \pi(x'|x')p(x'|x,a)$. Define the K^{th} order Taylor expansion of advantage as $Q^{\pi}_{K,\gamma,\gamma'} \coloneqq \sum_{k=0}^{K}((\gamma'-\gamma)(I-\gamma\overline{P}^{\pi})^{-1}\overline{P}^{\pi})^{k}Q^{\pi}_{\gamma}$. Then $\lim_{K\to\infty}Q^{\pi}_{K,\gamma,\gamma'} = Q^{\pi}_{\gamma'}$ for any $\gamma < \gamma' < 1$.

Algorithm 4 Estimating the K^{th} term of the expansion (Q-function)

Require: A trajectory $(x_t,a_t,r_t)_{t=0}^{\infty} \sim \pi$ and discount factors $\gamma < \gamma' < 1$

1. Compute advantage function estimates $\widehat{Q}^{\pi}_{\gamma}(x_t,a_t)$ for states on the trajectory. For example, $\widehat{Q}^{\pi}_{\gamma}(x_t,a_t) = \sum_{t'\geq t}\gamma^{t'-t}r_{t'}$. One could also apply other alternatives (e.g., (Schulman et al., 2015b)) which potentiall reduce the variance of $\widehat{Q}^{\pi}_{\gamma}(x_t,a_t)$.
2. Sample K random time $\tau_i, 1 \leq i \leq K$, all i.i.d. geometrically distributed $\tau_i \sim \text{Geometric}(1-\gamma)$.
3. Return $\frac{(\gamma'-\gamma)^{K}}{(1-\gamma)^{K}}\widehat{Q}^{\pi}_{\gamma}(x_\tau,a_\tau)$, where $\tau = \sum_{i=1}^{K}\tau_i$.

Proof. The proof follows closely that of Taylor expansion based approximation to value functions in Proposition 3.2. Importantly, notice that here we define $\overline{P}^{\pi}$, which differs from P^{π} used in the derivation of value functions. In particular,

$\overline{P}^\pi(x,a,y,b) = p(y|x,a)\pi(b|y)$ for any $x,y \in \mathcal{X}, a,b \in \mathcal{A}$. Let r be the vector reward function. The Bellmen equation for Q-function is

$$Q^\pi_{\gamma'} = r + \gamma' \overline{P}^\pi Q^\pi_{\gamma'}.$$

Inverting the equation and applying the Woodbury inversion identity,

$$Q^\pi_{\gamma'} = (I - \gamma'\overline{P}^\pi)^{-1} r = \sum_{k=0}^{\infty} \left((\gamma' - \gamma) \left(I - \gamma \overline{P}^\pi \right)^{-1} \overline{P}^\pi \right)^k Q^\pi_\gamma$$

The above equality holds for all $\gamma < \gamma' < 1$ due to similar convergence argument as in Proposition 3.2. Truncating the infinite series at step K, we arrive at the K^{th} order expansion $Q^\pi_{K,\gamma,\gamma'}$. By construction, $\lim_{K\to\infty} Q^\pi_{K,\gamma,\gamma'} = Q^\pi_{\gamma'}$. □

D. Details on Taylor expansion update weighting

Proposition D.1. The following is true for all $K \geq 0$,

$$\rho_{x,K,\gamma,\gamma'}(x') = \mathbb{I}[x' = x] + \mathbb{E}_\pi \left[\sum_{t=1}^{\infty} f(K,t,\gamma,\gamma') \mathbb{I}[x_t = x'] \,\middle|\, x_0 = x \right],$$

Equivalently, the K^{th} order Taylor expansion of $V^\pi_{\gamma'}(x)$ is

$$V^\pi_{K,\gamma,\gamma'}(x) = V_\gamma(x) + \mathbb{E}_\pi \left[\sum_{t=1}^{\infty} f(K,t,\gamma,\gamma') V^\pi_\gamma(x_t) \,\middle|\, x_0 = x \right], \tag{26}$$

where $f(K,t,\gamma,\gamma') = \sum_{u=1}^{\min(K,t)} (\gamma' - \gamma)^u \gamma^{t-u} \binom{t-1}{t-u}$ is a weight function.

Proof. We start with a few lemmas.

Lemma D.2. For any $n \geq 0, k \geq 1$, define a set of k-dimensional vector $\left\{ x_1, ...x_k | x_i \in \mathbb{Z}_{\geq 0}, \sum_{i=1}^k x_i = n \right\}$ and let $F(n,k)$ be the size of this set. Then

$$F(n,k) = \binom{n+k-1}{k-1}.$$

Proof. By construction, the above set can be decomposed into smaller sets by fixing the value of x_k, i.e.,

$$\left\{ x_1, ...x_k | x_i \in \mathbb{Z}_{\geq 0}, \sum_{i=1}^k x_i = n \right\} = \cup_{s=0}^n \left\{ x_1, ...x_{k-1}, x_k | x_i \in \mathbb{Z}_{\geq 0}, \sum_{i=1}^k x_i = n, x_k = s \right\}$$

Since these sets do not overlap, we have a recursive formula, $F(n,k) = \sum_{s=0}^n F(n-s, k-1)$. Starting from the base case $F(n,1) = 1, \forall n \geq 0$, it is straightforward to prove by induction that for all $n \geq 0, k \geq 1$

$$F(n,k) = \binom{n+k-1}{k-1}.$$

□

Lemma D.3. Consider $V^\pi_{K+1,\gamma,\gamma'} - V^\pi_{K,\gamma,\gamma'}$ for $K \geq 0$. It can be shown that

$$V^\pi_{K+1,\gamma,\gamma'} - V^\pi_{K,\gamma,\gamma'} = (\gamma' - \gamma)^{K+1} \left(\sum_{t=0}^{\infty} F(t, K+1) \left(P^\pi\right)^t \right) \left(P^\pi\right)^{K+1} V^\pi_\gamma.$$

Proof. Starting with the definition,

$$\begin{aligned} V^\pi_{K+1,\gamma,\gamma'} - V^\pi_{K,\gamma,\gamma'} &= \left((\gamma'-\gamma)\left(I-\gamma P^\pi\right)^{-1} P^\pi\right)^{K+1} V^\pi_\gamma \\ &= \left((\gamma'-\gamma)\left(I-\gamma P^\pi\right)^{-1}\right)^{K+1} \left(P^\pi\right)^{K+1} V^\pi_\gamma, \end{aligned}$$

where for the second equality we use the fact that P^π commutes with $(I-\gamma P^\pi)^{-1}$. Then consider $\left((I-\gamma P^\pi)^{-1}\right)^{K+1}$,

$$\left((I-\gamma P^\pi)^{-1}\right)^{K+1} = \left(\sum_{t=0}^\infty \left(\gamma P^\pi\right)^t\right)^{K+1} = \sum_{s_1\geq 0} \ldots \sum_{s_{K+1}\geq 0} \left(\gamma P^\pi\right)^{\sum_{i=1}^{K+1} s_i} = \sum_{s=0}^\infty F(s,K+1)\left(\gamma P^\pi\right)^s.$$

Note that the last equality corresponds to a regrouping of terms in the infinite summation – instead of summing over $s_1, \ldots s_{K+1}$ sequentially, we count the number of examples such that $\sum_{i=1}^{K+1} s_i = s$ and then sum over s. This count is exactly $F(s, K+1)$ as defined in Lemma D.2. Hence the proof is completed. □

With the above lemmas, we are ready to prove the final result. We start by summing up all the differences of expansions,

$$\begin{aligned} V^\pi_{K,\gamma,\gamma'} &= V^\pi_{0,\gamma,\gamma'} + \sum_{k=0}^{K-1} \left(V^\pi_{k+1,\gamma,\gamma'} - V^\pi_{k,\gamma,\gamma'}\right) \\ &= V^\pi_\gamma + \sum_{k=0}^{K-1} (\gamma'-\gamma)^{k+1} \left(\sum_{t=0}^\infty F(t,k+1)\left(\gamma P^\pi\right)^t\right) \left(P^\pi\right)^{k+1} V^\pi_\gamma \\ &= V^\pi_\gamma + \sum_{t=0}^\infty \sum_{k=0}^{K-1} (\gamma'-\gamma)^{k+1} \gamma^{-k-1} F(t,k+1)\left(\gamma P^\pi\right)^{t+k+1} V^\pi_\gamma \\ &= V^\pi_\gamma + \sum_{t=0}^\infty \sum_{u=1}^{K} (\gamma'-\gamma)^{u} \gamma^{-u} F(t,u)\left(\gamma P^\pi\right)^{t+u} V^\pi_\gamma \\ &= V^\pi_\gamma + \sum_{s=0}^\infty \sum_{u=1}^{K} (\gamma'-\gamma)^{u} \gamma^{-u} F(s-u,u)\left(\gamma P^\pi\right)^{s} V^\pi_\gamma \\ &= V^\pi_\gamma + \sum_{s=1}^\infty \sum_{u=1}^{K} (\gamma'-\gamma)^{u} \gamma^{-u} F(s-u,u)\left(\gamma P^\pi\right)^{s} V^\pi_\gamma \\ &= V^\pi_\gamma + \sum_{s=1}^\infty \sum_{u=1}^{K} (\gamma'-\gamma)^{u} \gamma^{s-u} \binom{s-1}{u-1} \left(P^\pi\right)^{s} V^\pi_\gamma \\ &= V^\pi_\gamma + \sum_{s=1}^\infty \sum_{u=1}^{\min(K,s)} (\gamma'-\gamma)^{u} \gamma^{s-u} \binom{s-1}{u-1} \left(P^\pi\right)^{s} V^\pi_\gamma \end{aligned}$$

In the above derivation, we have applied the transformation $u = k+1, s = t+u$. Then we have modified the bound of the summation with the definition of $F(s-u,u)$ (in particular, if $s<u$, $F(s-u,u)=0$). If we index the x-th component of the vector, we recover the desired result.

□

D.1. Further discussions on the objectives

Recall that the full gradient $\nabla_\theta V^{\pi_\theta}_{\gamma'}(x)$ is

$$\nabla_\theta V^{\pi_\theta}_{\gamma'}(x) = \mathbb{E}_{x'\sim\rho^{\pi_\theta}_{\gamma,\gamma'}(\cdot;x)}\left[\nabla_\theta V^{\pi_\theta}_{\gamma}(x')\right] + \underbrace{\mathbb{E}_{x'\sim\rho^{\pi_\theta}_{\gamma,\gamma'}(\cdot;x)}\left[V^{\pi_\theta}_{\gamma}(x')\nabla_\theta \log \rho^{\pi_\theta}_{\gamma,\gamma'}(x';x).\right]}_{\text{second term}}$$

Consider the second term. Now, we derive this term in an alternative way which imparts more intuitions on why its estimation is challenging. Note that

$$V_{\gamma'}^{\pi_\theta}(x) = V_{\gamma}^{\pi_\theta}(x) + (\gamma' - \gamma)\mathbb{E}_\pi\left[\sum_{t=1}^{\infty}(\gamma')^{t-1}V_{\gamma}^{\pi_\theta}(x_t)\right]$$

The second term of the full gradient is equivalent to differentiating through the above expression, while keeping all $V_{\gamma}^{\pi_\theta}(x_t)$ fixed. This leads the following gradient

$$\text{second term} = (\gamma' - \gamma)(\gamma')^{-1}\mathbb{E}_\pi\left[\sum_{t=1}^{\infty}(\gamma')^{t}W_{\gamma,\gamma'}^{\pi_\theta}(x_t)\nabla_\theta \log \pi_\theta(a_t|x_t)\right].$$

Here, we introduce $W_{\gamma,\gamma'}^{\pi_\theta}(x_t) = \mathbb{E}_\pi\left[\sum_{s=0}^{\infty}(\gamma')^{s}V_{\gamma}^{\pi_\theta}(x_{t+s})\right]$, which is equivalent to a value function that treats $V_{\gamma}^{\pi_\theta}(x)$ as rewards and with discount factor γ'. Naturally, constructing an unbiased estimator of the second term of the full gradient requires estimating $W_{\gamma,\gamma'}^{\pi_\theta}$, which is difficult in at least two aspects: **(1)** in practice, value functions are already estimated, which could introduce additional bias and variance; **(2)** as a premise of our work, estimating discounted values with discount factor γ' is challenging potentially due to high variance.

E. Details on approximation errors with finite samples

Intuitively, as K increases, the K^{th} order expansion $V_{K,\gamma,\gamma'}^{\pi}$ approximates $V_{\gamma'}^{K}$ more accurately in expectation. However, in practice where all constituent terms of the approximation are built from the same batch of data, the variance might negatively impact the accuracy of the estimate.

To formalize such intuitions, we characterize the bias and variance trade-off under the phased TD-learning framework (Kearns and Singh, 2000). Consider estimating the value function $V_{\gamma}^{\pi}(x)$ under discount γ, with estimator $\widehat{V}_{\gamma}^{\pi}(x)$. At each iteration t, let $\Delta_t^\gamma := \max_{x\in\mathcal{X}}|V_{\gamma}^{\pi}(x) - \widehat{V}_{\gamma}^{\pi}(x)|$ be the absolute error of value function estimates $\widehat{V}_{\gamma}^{\pi}$. Assume from each state x, there are independent n trajectories generated under π, (Kearns and Singh, 2000) shows that commonly used TD-learning methods (e.g. TD(λ)) have error bounds of the following form with probability $1-\delta$,

$$\Delta_t^\gamma \le A(\gamma,\delta) + B(\gamma)\Delta_{t-1}^\gamma. \tag{27}$$

Here, the factor $A(\gamma,\delta)$ is an error term which characterizes the errors arising from the finite sample size n. As $n\to\infty$, $A(\gamma,\delta)\to 0$; the constant $B(\gamma)$ is a contraction coefficient that shows how fast the error decays in expectation. See Appendix E for details.

With the calculations of estimators $\widehat{V}_{\gamma}^{\pi}(x)$ as a subroutine, we construct the n-sample K^{th} order estimator $\widehat{V}_{K,\gamma,\gamma'}^{\pi}(x)$,

$$\widehat{V}_{K,\gamma}(x_0) = \sum_{k=0}^{K}\frac{1}{n}\sum_{i=1}^{n}(\gamma'-\gamma)^k\widehat{V}_{\gamma}^{\pi}(x_{i,k}), \tag{28}$$

where $x_{i,k}$ is sampled from $(P^\pi\cdot d_\gamma^\mu)^k(\cdot;x)$. Note that if $K=0$, Eqn (28) reduces to $\frac{1}{n}\sum_{i=1}^{n}\widehat{V}_{\gamma}^{\pi}(x_0)$, the estimator analyzed by (Kearns and Singh, 2000). We are interested in the error $\Delta_{K,t}^\gamma := \max_{x\in\mathcal{X}}|V_{\gamma'}^{\pi}(x) - \widehat{V}_{K,\gamma,\gamma'}^{\pi}(x)|$, measured against the value function of discount γ'. The following summarizes how errors propagate across iterations,

Proposition E.1. Assume all samples $x_{i,k}$ are generated independently. Define a factor $\varepsilon := \frac{1-(\gamma'-\gamma)^{K+1}}{1-(\gamma'-\gamma)}$. Then with probability at least $1-2\delta$ if $K\ge 1$ and probability $1-\delta$ if $K=0$, the following holds[1],

$$\Delta_{K,t}^\gamma \le \underbrace{\varepsilon(A(\gamma,\delta)+U)}_{\text{finite sample error}} + \underbrace{E(\gamma,\gamma',K)}_{\text{expected gap error}} + \underbrace{\varepsilon B(\gamma)}_{\text{contraction coeff}}\ \Delta_t^\gamma, \tag{29}$$

where $U = \sqrt{2\log\frac{2(K+1)}{\delta}/n}$ for $K\ge 1$ and $U=0$ if $K=0$. The expected gap error $E(\gamma,\gamma',K) = \left(\frac{\gamma'-\gamma}{1-\gamma}\right)^{K+1}\frac{R_{\max}}{1-\gamma}$ is defined in Proposition 3.2.

[1]The error bounds could be further improved, e.g., by adapting the concentration bounds at different steps $1\le k\le K$. Note that its purpose is to illustrate the bias and variance trade-off induced by the Taylor expansion order K.

Proof. Recall the results from (Kearns and Singh, 2000): Let $\Delta_t^\gamma \coloneqq \max_{x\in\mathcal{X}} |V_\gamma^\pi(x) - \widehat{V}_\gamma^\pi(x)|$. Then with probability at least $1-\delta$, the following holds

$$\Delta_t^\gamma \le A(\gamma,\delta) + B(\gamma)\Delta_{t-1}^\gamma.$$

In the following, we condition all analysis on the event set that the above inequality holds. Now, using $\widehat{V}_\gamma^\pi(x)$ as a subroutine, define the estimator for the K^{th} Taylor expansion as in Eqn (28),

$$\widehat{V}_{K,\gamma}(x_0) = \sum_{k=0}^{K} \frac{1}{n} \sum_{i=1}^{n} (\gamma' - \gamma)^k \widehat{V}_\gamma^\pi(x_{i,k}).$$

Define the error $\Delta_{K,t}^\gamma \coloneqq \max_{x\in\mathcal{X}} |V_{\gamma'}^\pi(x) - \widehat{V}_{K,\gamma}^\pi(x)|$, which is measured against the value function $V_{\gamma'}^\pi(x)$ with a higher discount factor γ'. Consider for a given starting state x_0,

$$\begin{aligned}
|V_{\gamma'}^\pi(x_0) - \widehat{V}_{K,\gamma}^\pi(x_0)| &= V_{\gamma'}^\pi(x_0) - V_{K,\gamma}^\pi(x_0) + V_{K,\gamma}^\pi(x_0) - \widehat{V}_{K,\gamma}^\pi(x) \\
&\le |V_{\gamma'}^\pi(x_0) - V_{K,\gamma}^\pi(x_0)| + V_{K,\gamma}^\pi(x_0) - \widehat{V}_{K,\gamma}^\pi(x_0) \\
&\le E(\gamma,\gamma',K) + \underbrace{V_{K,\gamma}^\pi(x_0) - \mathbb{E}\left[\widehat{V}_{K,\gamma}^\pi(x_0)\right]}_{\text{second term}} + \underbrace{\mathbb{E}\left[\widehat{V}_{K,\gamma}^\pi(x_0)\right] - \widehat{V}_{K,\gamma}^\pi(x_0)}_{\text{third term}}.
\end{aligned}$$

Now, we bound each term in the equation above. Recall $\varepsilon \coloneqq \sum_{k=0}^{K}(\gamma'-\gamma)^k = \frac{1-(\gamma'-\gamma)^{K+1}}{1-\gamma'+\gamma}$. The second term is bounded as follows

$$V_{K,\gamma}^\pi(x_0) - \mathbb{E}\left[\widehat{V}_{K,\gamma}^\pi(x_0)\right] \le \varepsilon\Delta_t^\gamma.$$

The third term is bounded by applying concentration bounds. Recall that the estimator $\widehat{V}_{K,\gamma}^\pi(x_0) \coloneqq \sum_{k=0}^{K}\frac{1}{n}\sum_{i=1}^{n}(\gamma'-\gamma)^k\widehat{V}_\gamma^\pi(x_{i,k})$ decomposes into $K+1$ estimators, each being an average over n i.i.d. samples drawn from the K^{th} step visitation distribution $(P^\pi\cdot d_\gamma^\pi)^k, 0\le k\le K$. Applying similarly naive techniques in (Kearns and Singh, 2000), we bound each of the $K+1$ terms individually and then take a union bound over all $K+1$ terms. This implies that, with probability at least $1-\delta$, the following holds

$$\mathbb{E}\left[\widehat{V}_{K,\gamma}^\pi(x_0)\right] - \widehat{V}_{K,\gamma}^\pi(x_0) \le \varepsilon U = \varepsilon\sqrt{2\log\frac{2(K+1)}{\delta}/n}.$$

Aggregating all results, we have

$$\begin{aligned}
|V_{\gamma'}^\pi(x_0) - \widehat{V}_{K,\gamma}^\pi(x_0)| &\le E(\gamma,\gamma',K) + \varepsilon\Delta_t^\gamma + \varepsilon U \\
&\le \varepsilon(A(\gamma,\delta)+U) + E(\gamma,\gamma',K) + \varepsilon B(\gamma)\Delta_{t-1}^\gamma.
\end{aligned}$$

This holds with probability at least $(1-\delta)^2 \ge 1-2\delta$.

□

Bias-variance trade-off via K. The error terms come from two parts: the first term contains errors $A(\gamma,\delta)$ in the subroutine estimator $\widehat{V}_\gamma^\pi(x)$, and its propagated errors through the sampling of K^{th} order approximations for $1\le k\le K$ (shown via the multiplier ε). This first term also contains U, a concentration bound that scales with $O(\sqrt{\log K})$, which shows that the variance of the overall estimator grows with K. This first error term scales with $\sqrt{n}$ and vanishes as the number of samples increases. The second term is due to the gap between the expected K^{th} order Taylor expansion and $V_{\gamma'}^\pi(x_0)$, which decreases with K and does not depend on sample size n. The new contraction coefficient is $\varepsilon B(\gamma)$, where it can be shown that $\varepsilon \in [1, \frac{1}{1-\gamma'+\gamma}]$. Since typical estimators have $B(\gamma)\le\gamma$, in general $\varepsilon B(\gamma) < 1$ and the error contracts with respect to Δ_t. In general, the contraction becomes slower as K increases. For example, for TD(λ), $B(\gamma) = \frac{(1-\lambda)\gamma}{1-\gamma\lambda}$.

F. Further experiment details

Below, we provide further details on experiment setups along with additional results.

F.1. Further details on the toy example

We presented a toy example that highlighted the trade-off between bias and variance, mediated by the order parameter K. Here, we provide further details of the experiments.

Toy MDP. We consider tabular MDPs with $|\mathcal{X}| = 10$ states and $|\mathcal{A}| = 2$ actions. The transition table $p(y|x, a)$ is drawn from a Dirichlet distribution $(\alpha, \ldots, \alpha)$ for $\alpha = 0.01$. Here, α is chosen such that the MDP is not very communicative (i.e., the distribution $p(\cdot|x, a)$ concentrates only on a few states). The rewards are random $r(x, a) = \overline{r}(x, a)(1 + \varepsilon)$ where $\varepsilon \sim \mathcal{N}(0, 0.2^2)$ and mean rewards $\overline{r}(x, a)$ are drawn from Uniform$(0, 1)$ and fixed for the problem.

F.2. Deep RL algorithms

Proximal policy optimization (PPO). PPO (Schulman et al., 2017) implements a stochastic actor $\pi_\theta(a|x)$ as a Gaussian distribution $a \sim \mathcal{N}(\mu_\theta(x), \sigma^2\mathbb{I})$ with state-conditional mean $\mu_\theta(x)$ and a global standard deviation $\sigma^2\mathbb{I}$; and a value function $V_\phi(x)$. The behavior policy μ is the previous policy iterate. The policy is updated as $\widehat{A}^\mu_\gamma(x, a)\nabla_\theta \text{clip}(\frac{\pi_\theta(a|x)}{\mu(a|x)}, 1 - \varepsilon, 1 + \varepsilon)$ with $\varepsilon = 0.2$[2]. The advantages $\widehat{A}^\mu_\gamma(x, a)$ estimated using generalized advantage estimation (GAE, (Schulman et al., 2015b)) with $\gamma = 0.99, \lambda = 0.95$. Value functions are trained by minimizing $(V_\phi(x) - R(x))^2$ with returns $R(x) = V_{\phi'}(x) + \widehat{A}^\mu_\gamma(x, a)$ with ϕ' being a prior parameter. Both parameters θ, ϕ are trained with the Adam optimizer (Kingma and Ba, 2014) with learning rate $\alpha = 3 \cdot 10^{-4}$. We adopt other default hyper-parameters in (Dhariwal et al., 2017), for details, please refer to the code base.

Trust region policy optimization (TRPO). TRPO (Schulman et al., 2015b) implements the same actor-critic pipeline as PPO, the difference is in the updates. Instead of enforcing a *soft* clipping constraint, TRPO enforces a strict KL-divegergence constraint $\mathbb{E}_{x\sim\mu}\left[\text{KL}(\pi_\theta(\cdot|x), \mu(\cdot|x)\right] \leq \varepsilon$ with $\varepsilon = 0.01$. The policy gradient is computed as $\widehat{A}^\mu_\gamma(x, a)\nabla_\theta \log \pi_\theta(a|x)$, and then the final update is constructed by approximately solving a constrained optimization problem, see (Schulman et al., 2015a) for details. The scale of the final update is found through a line search, to ensure that the KL-divergence constraint is satisfied. The implementations are based in (Achiam and OpenAI, 2018).

F.3. Deep RL architecture

Across all algorithms, the policy $\pi_\theta(a|x) = \mathcal{N}(\mu_\theta(x), \sigma^2\mathbb{I})$ has a parameterized mean $\mu_\theta(x)$ and a single standard deviation σ^2. The mean $\mu_\theta(x)$ is a 2-layer neural network with hidden units $h = 64$, and $f(x) = \tanh(x)$ activation functions. The output layer does not have any activation functions; The value function $V_\phi(x)$ is a 2-layer neural network with hidden units $h = 64$ and $f(x) = \tanh(x)$ as activation functions. The output layer does not have any activation functions.

F.4. Additional deep RL experiment results

F.4.1. Taylor expansion Q-function estimation: ablation study on η

Recall that throughout the experiments, we choose $K = 1$ and construct the new Q-function estimator as a mixture of the default estimator and Taylor expansion Q-function estimator. In particular, the final Q-function estimator is

$$\widehat{Q}(x, a) = (1 - \eta)\widehat{Q}^\pi_\gamma(x, a) + \eta\widehat{Q}^\pi_{K,\gamma,\gamma'}(x, a).$$

We choose $\eta \in [0, 1]$ such that it balances the numerical scales of the two combining estimators. In our implementation, we find that the algorithm performs more stably when η is small in the absolute scale. In Figure 5(a)-(b), we show the ablation study on the effect of η, where we vary $\eta \in [0.01, 0.03]$. The y-axis shows the normalized performance against PPO baselines (which is equivalent to $\eta = 0$), such that the PPO baseline achieves a normalized performance of 1.

Overall, we see on different tasks, η impacts the performance differently. For example: on HalfCheetah(B), better performance is achieved with larger values of η, this is consistent with the observation that PPO with $\gamma = 0.999$ also achieves better performance; on Ant(B), however, as η increases from zero, the performance increases marginally before degrading. In Figure 5, we show the median and mean performance across all tasks. Note that in general, the average performance increases as η increases from zero, but later starts to decay a bit. When accounting for the effect of performance variance across all tasks, we chose $\eta = 0.01$ as the fixed hyper-parameter throughout experiments in the main paper.

[2]The exact PPO update is more complicated than this. Refer to (Schulman et al., 2017) for the exact formula.

Further details on computing $\widehat{Q}^{\pi}_{K,\gamma,\gamma'}(x,a)$. Below we assume $K=1$. In Algorithm 4, we showed we can construct unbiased estimates of $Q^{\pi}_{K,\gamma,\gamma'}(x,a)$ using $\widehat{Q}^{\pi}_{\gamma}(x,a)$ as building blocks. With a random time $\tau \sim \text{Geometric}(1-\gamma)$, the estimator takes the following form

$$\widehat{Q}_{K,\gamma,\gamma'}(x_t,a_t) = \widehat{Q}^{\pi}_{\gamma}(x_t,a_t) + \frac{\gamma'-\gamma}{1-\gamma} Q^{\pi}_{\gamma}(x_{t+\tau},a_{t+\tau}).$$

However, since the estimator is based on a single random time, it can have high variance. To reduce variance, we propose the following procedure: let (x_t,a_t) be the target state-action pair, we can compute the estimate as

$$\widehat{Q}_{K,\gamma,\gamma'}(x_t,a_t) = \widehat{Q}^{\pi}_{\gamma}(x_t,a_t) + \frac{\gamma'-\gamma}{1-\gamma} \sum_{s=1}^{H} \frac{\gamma^s}{\sum_{s'=1}^{H}\gamma^{s'}} Q^{\pi}_{\gamma}(x_{t+s},a_{t+s}).$$

When $H=\infty$, the above estimator corresponds to an estimator which marginalizes over the random time. This should achieve variance reduction compared to the random time based estimate in Algorithm 4. However, then the estimate requires computing cumulative sums over an infinite horizon (or in general a horizon of T), which might be computationally expensive. To mitigate this, we propose to truncate the above summation up to $H=10$ steps. This choice of H aims to achieve a trade-off between computation efficiency and variance. Note that this estimator was previously introduced in (Tang et al., 2020) for off-policy learning.

F.4.2. Taylor expansion update weighting: ablation on K

In Figure 5(c)-(d), we carry out ablation study on the effect of K for the update weighting. Recall that K interpolates two extremes: when $K=0$, it recovers the vanilla PG (Sutton et al., 2000) while when $K=\infty$, it recovers the deep RL heuristic update. We expect an intermediate value of K to achieve some trade-off between bias and variance of the overall update.

In Figure 5(c), we see the effect on individual environments. The effect is case dependent. For HalfCheetah(G), larger K improves the performance; however, for Walker(G), the improvement is less prominent over a large range of K. When aggregating the performance metric in Figure 5(d), we see that intermediate values of K indeed peak in performance. We see that on average, both $K=10$ and $K=100$ achieve locally optimal mean performance, while $K=10$ also achieves the locally optimal median performance.

Note on how the practical updates impact the effect of K. Based on our theoretical analysis, when $K=0$ the update should recover the vanilla PG (Sutton et al., 2000), which is generally considered too conservative for the undiscounted objective in Eqn (1). However, in practice, as shown in Figure 5(d), the algorithm does not severely underperform even when $K=0$. We speculate that this is because practical implementations of PG updates use batches of data instead of the full trajectories. This means that the relative weights $w(t)$ of the local gradients $\widehat{Q}_t \nabla_\theta \log \pi_\theta(a_t|x_t)$ are effectively self-normalized: $\widetilde{w}(t) \leftarrow \frac{w(t)}{\sum w(t')}$ where the summation is over the time steps in a sampled mini-batch. The self-normalized weights $\widetilde{w}(t)$ are increased in the absolute scale relative to $w(t)$ and partly offset the effect of an initially aggressive discount $w(t)=\gamma^t$.

F.4.3. Comparison to results in (Romoff et al., 2019)

Recently, Romoff et al. (2019) derived a recursive relations between differences value functions defined with different discount factors. This was shown in Lemma 4.1. Given a sequence of discount factors $\gamma_1 < \gamma_2 < \ldots < \gamma_N < \gamma'$, they derived a value function estimator to $V^{\pi}_{\gamma'}(x)$ based on recursive bootstraps of value function differences $V^{\pi}_{\gamma_i}(x) - V^{\pi}_{\gamma_{i-1}}(x')$. Because they aim at recovering the exact value functions, this estimator could be interpreted as similar to Taylor expansions but with $K=\infty$.

Different from their motives, we focus on the trade-off achieved by intermediate values of K. We argued that by using $K=0$, the estimate might be too conservative; however, using $K=\infty$ might be challenging due to the variance induced in the recursive bootstrapping procedure. Though it is not straightforward to theoretically show, we conjecture that using the Taylor expansion Q-function estimator with $K=\infty$ is as difficult as directly estimating $V^{\pi}_{\gamma'}(x)$.

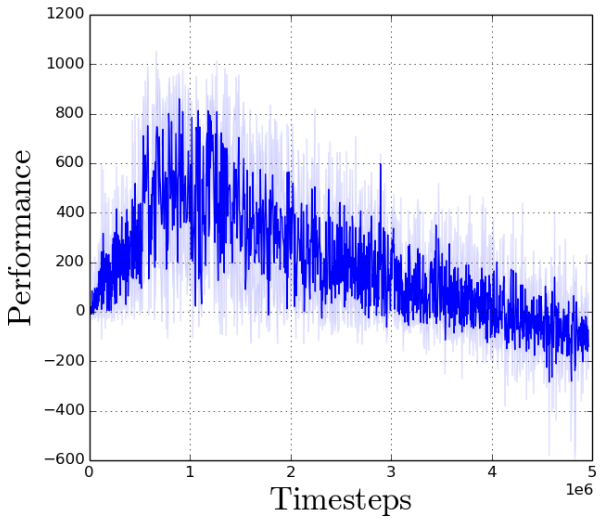
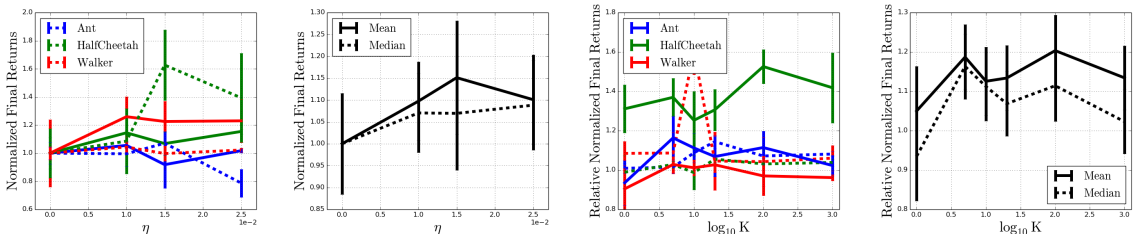


Figure 4. Learning curves generated by running the open source implementation of (Romoff et al., 2019) on Walker2d(G), averaged across 5 runs. There is little progress of learning for the algorithm on other benchmark tasks.



(a) Ablation on η (individual) (b) Ablation on η (average) (c) Ablation on K (individual) (d) Ablation on K (average)

Figure 5. Ablation study of hyper-parameters. We study two hyper-parameters: (a) η (b) K. In both cases, we calculate the task-dependent normalized final returns after training for 10^6 steps. See Appendix F for how such normalized returns are computed. In (a), normalized returns are computed with respect to $\eta = 0$ (i.e, the PPO baseline), such that when $\eta = 0$, the normalized returns are ones; in (b), normalized returns are computed with respect to the default PPO baseline, such that values of ones imply that the baseline performs the same as the default PPO baseline. Dashed curves (bullet tasks) and solid curves (gym tasks) are both mean scores averaged over 5 seeds.

Empirical comparison. The base algorithm of (Romoff et al., 2019) is PPO(Schulman et al., 2017). Their algorithm uses the recursive bootstraps to estimate Q-functions and advantage functions. The new estimate is used as a direct plug-in replacement to $\widehat{Q}^{\pi}_{\gamma}(x,a)$ and $\widehat{A}^{\pi}_{\gamma}(x,a)$ adopted in the PPO algorithm. We run experiments with the open source implementation of (Romoff et al., 2019) from the original authors[3]. We evaluate the algorithm's performance over continuous control benchmark tasks. We applied the default configurations from the code base with minimum changes to run on continuous problems (note that (Romoff et al., 2019) focused on a few discrete control problems). Overall, we find that the algorithm does not learn stably (see Figure 4).

G. Extensions of update weighting techniques to off-policy algorithms

Below, we show that techniques developed in this paper could be extended to off-policy learning algorithms. We provide both details in theoretical derivations, algorithms, as well as experimental results.

[3]See https://github.com/facebookresearch/td-delta.

G.1. Off-policy actor-critic algorithms

Off-policy actor-critics (Mnih et al., 2015; Lillicrap et al., 2015) maintain a deterministic policy $\pi_\theta(x)$ and a Q-function critic $Q_\phi(x,a)$. The agent takes exploratory actions under the environment, and saves data (x_t, a_t, r_t) into a common replay buffer $\mathcal{D}$. At training time, the algorithm samples data from the replay to update parameters. The policy is updated via the deterministic policy gradient (Silver et al., 2014), $\theta \leftarrow \theta + \alpha \nabla_\theta \mathbb{E}_\mu \left[Q_\phi(x, \pi_\theta(x))\right]$, where μ is implicitly defined by the past behavior policy.

Deep deterministic policy gradient (DDPG). DDPG (Lillicrap et al., 2015) maintains a deterministic policy network $\pi_\theta(a|x) \equiv \pi_\theta(x)$ and a Q-function critic $Q_\phi(x,a)$. The algorithm explores by executing a perturbed policy $a = \varepsilon + \pi_\theta(x)$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ for $\sigma = 0.1$, and then saves the data (x, a, r, x') into a replay buffer $\mathcal{D}$. At training time, the behavior data is sampled uniformly from the replay buffer $(x_i, a_i, r_i, x_i')_{i=0}^{B-1} \sim \mathcal{U}(\mathcal{D})$ with $B = 100$. The critic is updated via TD(0), by minimizing: $\frac{1}{B}\sum_{i=0}^{B-1}(Q_\phi(x_i,a_i) - Q_{\text{target}}(x_i,a_i))^2$ where $Q_{\text{target}}(x_i,a_i) = r_i + \gamma Q_{\phi'}(x_i', \pi_{\theta'}(x_i'))$, where θ', ϕ' are delayed versions of θ, ϕ respectively (Mnih et al., 2015). The policy is updated by maximizing $\frac{1}{B}\sum_{i=0}^{B-1} Q_\phi(x_i, \pi_\theta(x_i))$ with respect to θ. Both parameters θ, ϕ are trained with the Adam optimizer (Kingma and Ba, 2014) with learning rate $\alpha = 10^{-4}$. We adopt other default hyper-parameters in (Achiam and OpenAI, 2018), for details, please refer to the code base.

Twin-delayed DDPG (TD3). TD3 (Fujimoto et al., 2018) adopts the same training pipeline and architectures as DDPG. TD3 also adopts two critic networks $Q_{\phi_1}(x,a), Q_{\phi_2}(x,a)$ with parameters ϕ_1, ϕ_2, in order to minimize the over-estimation bias (Hasselt, 2010; Van Hasselt et al., 2016).

Soft actor-critic. SAC (Haarnoja et al., 2018) adopts a similar training pipeline and architectures as TD3. A major conceptual difference is that SAC is based on the maximum-entropy formulation of RL (Ziebart et al., 2008; Fox et al., 2016). The Q-function is augmented by entropy regularization bonus and the policy is optimized such that it does not collapse to a deterministic policy.

G.2. Architecture

All algorithms share the same architecture. The policy network $\pi_\theta(x)$ takes as input the state x, and is a 2-layer neural network with hidden units $h = 256$ and $f(x) = \text{relu}(x)$ activation functions. The output is squashed by $f(x) = \tanh(x)$ to comply with the action space boundaries; The critic $Q_\phi(x,a)$ takes a concatenated vector $[x,a]$ as inputs, is 2-layer neural network with hidden units $h = 256$ and $f(x) = \text{relu}(x)$ activation functions. The output does not have any activation functions.

For stochastic policies, the policy network parameterizes a Gaussian also parameterizes a log standard deviation vector $\log\sigma(x)$, which is a neural network with the same architecture above. The stochastic output is a reparameterized function $a = \pi_\theta(x) + \exp(\log\sigma(x)) \cdot \varepsilon$ where the noise $\varepsilon \sim \mathcal{N}(0,1)$. Finally, the action output is squashed by $\tanh(x)$ to comply with the action boundary (Haarnoja et al., 2018).

G.3. Algorithm details for update weighting

To derive an update based on update weighting, we start with the undiscounted on-policy objective $V_{\gamma'}(x) = \mathbb{E}_{x' \sim \rho^{\pi_\theta}_{x,\gamma,\gamma'}}\left[V_\gamma^{\pi_\theta}(x')\right]$. Given behavior data generated under μ, we abuse the notation and also use μ to denote the state distribution under μ (usually implicitly defined by sampling from a replay buffer $\mathcal{D}$). By rewriting the objective with importance sampling (IS),

$$V_{\gamma'}^{\pi_\theta}(x) = \mathbb{E}_{x' \sim \rho^{\pi_\theta}_{x,\gamma,\gamma'}}\left[V_\gamma^{\pi_\theta}(x')\right] = \mathbb{E}_{x' \sim \mu}\left[\frac{\rho^{\pi_\theta}_{x,\gamma,\gamma'}(x')}{\mu(x')} V_\gamma^{\pi_\theta}(x')\right], \tag{30}$$

we derive an off-policy learning objective. By dropping a certain terms (see (Degris et al., 2012) for details about the justifications for dropping such terms), we can derive the IS-based gradient update

$$\mathbb{E}_{x' \sim \mu}\left[\frac{\rho^{\pi_\theta}_{x,\gamma,\gamma'}(x')}{\mu(x')} \nabla_\theta V_\gamma^{\pi_\theta}(x')\right] \approx \mathbb{E}_{x' \sim \mu}\left[\frac{\rho^{\pi_\theta}_{x,\gamma,\gamma'}(x')}{\mu(x')} \nabla_\theta Q_\phi\left(x', \pi_\theta(x')\right)\right]$$

To render the update feasible, we need to estimate the ratio $\frac{\rho^{\pi_\theta}_{x,\gamma,\gamma'}(x')}{\mu(x')}$. Inspired by (Sinha et al., 2020), we propose to maintain a *fast replay buffer* $\mathcal{D}_f$ which contains the most recent sampled data (which implicitly defines $\rho^{\pi_\theta}_{x,\gamma,\gamma'}$), then the

estimator w_ψ is trained to estimate the density ratio between $\mathcal{D}$ (which implicitly defines μ) and $\mathcal{D}_f$. See Appendix F for further details. The full off-policy actor-critic algorithm is summarized in Algorithm G.3. In practice, we implement a undiscounted uniform distribution instead of $\rho^\pi_{x,\gamma,\gamma'}(x')$ with $\gamma' = 1$. The main motivation is that this distribution is much easier to specify as it corresponds to sampling from the replay buffer uniformly without discounts, as explained below.

As an important observation for practical implementations, note that

$$\rho^\pi_{x,\gamma,\gamma'}(x') = \frac{\gamma}{\gamma'}\mathbb{I}[x_0 = x'] + (\gamma' - \gamma)\mathbb{E}_\pi\left[\sum_{t\geq 1}(\gamma')^{t-1}\mathbb{I}[x_t = x'] \,\middle|\, x_0 = x\right]$$

when setting $\gamma' = 1$, we see that the second term of the distribution is proportional to $\mathbb{E}_\pi\left[\sum_{t\geq 1}\mathbb{I}[x_t = x'] \,\middle|\, x_0 = x\right]$, which corresponds to a uniform distribution over states on sampled trajectories, *without* discounting. This will make implementations much simpler. We will see that this could also lead to performance gains. We leave Taylor expansion based extension of this method for future work.

Details on training the density estimator $w_\psi(x)$. The density estimator $w_\psi(x)$ is parameterized with exactly the same architecture as the policy network $\pi_\theta(x)$, except that its output activation is replaced by $\log(1 + \exp(x))$ to ensure that $w_\psi(x) > 0$. The off-policy actor-critic algorithm maintains an original buffer $\mathcal{D}$ of size $|\mathcal{D}| = 10^6$; in addition, we maintain a fast replay buffer $\mathcal{D}_f$ with $|\mathcal{D}_f| = 10^4$, which is used for saving the most recently generated data points. For ease of analysis, assume that the data sampled from $\mathcal{D}_f$ come from π_θ, while the data sampled from $\mathcal{D}$ come from μ.

To learn the ratio $\frac{\rho^{\pi_\theta}_{x,\gamma,\gamma'}(x')}{\mu(x')}$, we adopt a simple discriminative loss function as follows

$$L(\psi) = -\mathbb{E}_{x'\sim\rho^{\pi_\theta}_{x,\gamma,\gamma'}}\left[\log\frac{w_\psi(x')}{1+w_\psi(x')}\right] - \mathbb{E}_{x'\sim\mu}\left[\log\frac{1}{1+w_\psi(x')}\right] \approx -\mathbb{E}_{x\sim\mathcal{D}_f}\left[\log\frac{w_\psi(x')}{1+w_\psi(x')}\right] - \mathbb{E}_{x\sim\mathcal{D}}\left[\log\frac{1}{1+w_\psi(x')}\right].$$

The optimal solution to $\psi^* = \arg\min_\psi L(\psi)$ is $w_{\psi^*}(x') = \frac{\rho^{\pi_\theta}_{x,\gamma,\gamma'}(x')}{\mu(x')}$(assuming enough expressiveness). Then, the density estimator is used for weighting the policy update: when sampling a batch of B data from the buffer, the weight $w_\psi(x_i), 1 \leq i \leq B$ is computed for each data point x_i. Then the weights are normalized across batch $\widetilde{w}_i = \frac{w_\psi(x_i)^\tau}{\sum_{j=1}^B w_\psi(x_j)^\tau}$ where the inverse temperature is $\tau = 0.1$. Then $\widetilde{w}_i$ is used for weighting the such that the policy is updated as $\theta \leftarrow \theta + \alpha\frac{1}{B}\sum_{i=1}^B \widetilde{w}_i\nabla_\theta Q_\phi(x_i, \pi_\theta(x_i))$.

Algorithm 5 Update weighting Off-policy actor-critic

Require: policy $\pi_\theta(x)$, Q-function critic $Q_\phi(x, a)$, density estimator $w_\psi(x)$ and learning rate $\alpha \geq 0$

while not converged **do**

1. Collect data $(x_t, a_t, r_t) \sim \mu$ and save to the buffer $\mathcal{D}$ and the fast buffer $\mathcal{D}_f$
2. Estimate the density by the discriminative loss between $\mathcal{D}, \mathcal{D}_f$, such that $w_\psi(x') \approx \rho^{\pi_\theta}_{x,\gamma,\gamma'}(x')/\mu(x')$, where x is the initial state of the MDP.
3. Sample data from $(x_i, a_i, r_i)_{i=1}^B \sim \mathcal{D}$.

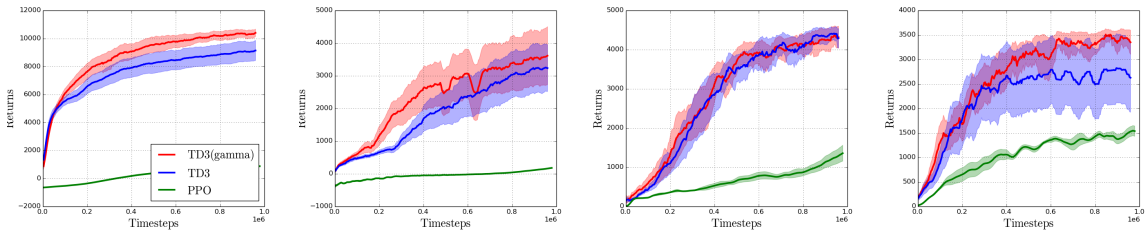
3(a). Update the Q-function critic $Q_\phi(x, a)$ via TD-learning, such that $Q_\phi(x, a) \approx Q^{\pi_\theta}_\gamma(x, a)$.

3(b). Update the policy parameter with the gradient $\theta \leftarrow \theta + \alpha\sum_{i=1}^B w_\psi(x_i)\nabla_\theta Q_\phi(x_i, \pi_\theta(x_i))$.

end while

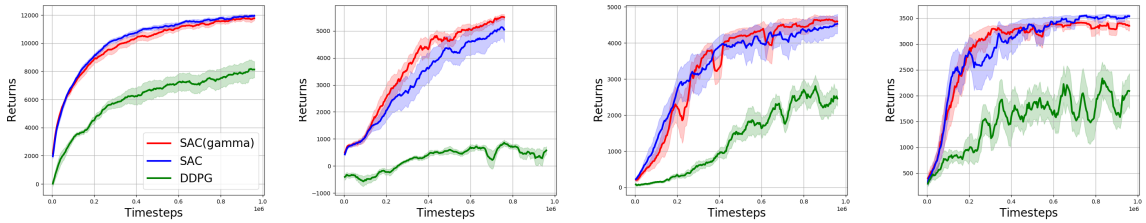
We carry out the update in Algorithm 2, where the density estimator $w_\psi(x)$ is trained based on a discriminative loss between $\mathcal{D}$ and $\mathcal{D}_f$. For any given batch of data $\{x_i\}_{i=1}^B$, we normalize the prediction $\widetilde{w}_i = w_\psi(x_i)^\tau / \sum_{j=1}^B w_\psi(x_j)^\tau$ with hyper-parameter $\tau = 0.1$ as similarly implemented in (Sinha et al., 2020). The temperature annealing moves $\widetilde{w}_i$ closer to a uniform distribution and tends to stabilize the algorithm. See Appendix F for further details.

Discussion on relations to other algorithms. Previous work focuses on re-weighting transitions to stabilize the training of critics. For example, prioritized replay (Schaul et al., 2015) prioritizes samples with high Bellman errors. Instead, Algorithm 2 reweighs samples to speed up the training of the policy. Our observation above also implies that when sampling from $\mathcal{D}, \mathcal{D}_f$ for training the estimates $w_\psi \approx \frac{\rho^{\pi_\theta}_{x,\gamma,\gamma'}(x')}{\mu(x')}$, it is not necessary to discount the transitions. This is in clear contrast



(a) HalfCheetah(G) (b) Ant(G) (c) Walker2d(G) (d) Hopper(G)

Figure 6. Evaluation of near off-policy actor-critic algorithms over continuous control domains. Each curve corresponds to a baseline algorithm averaged over 5 random seeds. TD3(γ) (red curve) consistently outperforms or performs similarly as other baselines.



(a) HalfCheetah(G) (b) Ant(G) (c) Walker2d(G) (d) Hopper(G)

Figure 7. Evaluation of near off-policy actor-critic algorithms over continuous control domains. Each curve corresponds to a baseline algorithm averaged over 5 random seeds. SAC(γ) (red curve) consistently outperforms or performs similarly as other baselines.

to prior work, such as (Sinha et al., 2020), where they propose to train $w_\psi(x') \approx d^{\pi_\theta}_{x,\gamma}(x')/d^{\mu}_{x,\gamma}(x')$, which is the fully discounted visitation distribution under γ based on the derivation of optimizing a discounted objective $V^{\pi_\theta}_\gamma(x)$.

Results. We build the algorithmic improvements based on TD3 (Fujimoto et al., 2018) and SAC (Haarnoja et al., 2018), and name the correspinding algorithms TD3(γ) and SAC(γ) respectively. We compare with TD3, SAC, and DDPG (Lillicrap et al., 2015), all of which are off-policy algorithms.

We first compare TD3(γ) with TD3 in Figure 6. To highlight the default sample efficiency of off-policy methods, we include PPO as a baseline as well. Across all four presented tasks, we see that TD(γ) performs similarly or marginally outperforms the TD3 baseline. To make concrete the comparison between final performance, we report the final score mean $\pm$ 0.5std of each algorithm in Table 1. As a default baseline, we also show the results of DDPG reported in (Achiam and OpenAI, 2018). Overall, TD3(γ) provides a modest yet consistent boost over baseline TD3.

Then we compare SAC(γ) with SAC in Figure 7 and Table 1. We see that SAC(γ) provides marginal performance gains over Walker2d and Ant, while it is slightly overperformed by baseline SAC for HalfCheetah and Hopper. We speculate that this is partly because the hyper-parameters of baseline SAC are well tuned on HalfCheetah, and it is difficult to achieve further significant gains without exhaustive hyper-parameter search. Overall, SAC(γ) is competitive compared to SAC.

Tasks	TD3(γ)	TD3	DDPG-v1
ANT(G)	$\mathbf{3601 \pm 879}$	$\mathbf{3269 \pm 686}$	≈ 1000
HALFCHEETAH(G)	$\mathbf{10350 \pm 279}$	9156 ± 718	≈ 8500
WALKER2D(G)	$\mathbf{4090 \pm 440}$	$\mathbf{4233 \pm 314}$	≈ 2000
HOPPER(G)	$\mathbf{3340 \pm 262}$	2626 ± 677	≈ 1800
Tasks	**SAC(γ)**	**SAC**	**DDPG-v2**
ANT(G)	$\mathbf{5572 \pm 115}$	4886 ± 530	706 ± 123
HALFCHEETAH(G)	11774 ± 96	$\mathbf{12059 \pm 91}$	7957 ± 527
WALKER2D(G)	$\mathbf{4626 \pm 165}$	$\mathbf{4522 \pm 269}$	2261 ± 147
HOPPER(G)	3384 ± 81	$\mathbf{3557 \pm 20}$	2024 ± 297

Table 1. Final performance of baseline algorithms over benchmark tasks. The final performance is computed as the mean scores over the last 10 iterations of each algorithm, averaged over 5 seeds. When compared with TD3, the performance of DDPG-v1 is taken from (Achiam and OpenAI, 2018); when compared with SAC, the performance is based on re-runs of the DDPG-v2 baselines with (Achiam and OpenAI, 2018). For each task, the best algorithms are highlighted in bold fonts (potentially with ties).