Off-Policy Actor–Critic Errata

The current theoretical results only apply to tabular representations for the policy π and not necessarily to function approximation for the policy. Thanks to Hamid Reza Maei for pointing out this issue. We are working on correcting this issue, both by re-examining the current theoretical results and working on modifications to the algorithm. Note, however, that even theoretical results restricted to tabular representations indicate that the algorithm has a principled design.

The are two mistakes in the theoretical analysis that cause us to scale back the claims. The first problem is about the existence of stable minima for our approximate gradient. Because the approximate gradient is not the gradient of any objective function, it is not clear if any stable minima exist. To justify the existence of these minima, we need to prove that when \( g(u) = 0 \), perturbations to \( u \) push \( u \) back to this minimum (rather than away from it). To do so, we will need to consider linearizations around the minimum using the Hartman-Grobman theorem. For a tabular representation, this is not a problem because each iteration strictly improves the value via a local change to \( \pi(a|s) \) for a specific state and action, without aliasing causing ripple effects (as we showed in Theorem 1).

The second error is in the proof of the claim in Theorem 2 that \( Z \subseteq \tilde{Z} \). It remains true that \( Z = \tilde{Z} \) for tabular representations. In the Policy Improvement claim in Theorem 1, we can say that the policy update overall improves the policy, allotting more importance to states more highly weighted by the stationary behavior distribution, \( d_b \). Therefore, it is still true that

\[
J_\gamma(u_t) \leq \sum_{s \in S} \sum_{a \in A} \pi_u(a|s) \sum_{s', a'} P(s, a, s') [R(s, a, s') + \gamma V^{\pi_u, \gamma}(s')] 
\]

Expanding this further, we get

\[
J_\gamma(u_t) \leq \sum_{s \in S} \sum_{a \in A} \pi_u(a|s) \sum_{s', a'} P(s, a, s', a') [R(s, a, s', a') + \gamma V^{\pi_u, \gamma}(s')] 
\]

We know that

\[
\sum_{s' \in S} \pi_u(a'|s') \sum_{s'' \in S} P(s', a', s'', a'') [R(s', a', s'', a'') + \gamma V^{\pi_u, \gamma}(s'')] 
\]

Therefore, it is still true that

\[
J_\gamma(u_t) \leq \sum_{s \in S} \sum_{a \in A} \pi_u(a|s) \sum_{s', a'} P(s, a, s', a') [R(s, a, s', a') + \gamma V^{\pi_u, \gamma}(s')] 
\]

Expanding this further, we get

\[
J_\gamma(u_t) \leq \sum_{s \in S} \sum_{a \in A} \pi_u(a|s) \sum_{s', a'} P(s, a, s', a') [R(s, a, s', a') + \gamma V^{\pi_u, \gamma}(s')] 
\]

We know that

\[
\sum_{s' \in S} \pi_u(a'|s') \sum_{s'' \in S} P(s', a', s'', a'') [R(s', a', s'', a'') + \gamma V^{\pi_u, \gamma}(s'')] 
\]

Therefore, it is still true that

\[
J_\gamma(u_t) \leq \sum_{s \in S} \sum_{a \in A} \pi_u(a|s) \sum_{s', a'} P(s, a, s', a') [R(s, a, s', a') + \gamma V^{\pi_u, \gamma}(s')] 
\]
We see that unless $P(s, a, \cdot)$ and $d^b$ are similar, then the next important inequality, in the middle of the proof of Theorem 1, with actions selected according to $\pi_u$, might not hold:

$$\sum_a \pi_u'(a|s) \sum_{s,a,s_{t+1}} P(s, a, s_{t+1}) \sum_{a_{t+1}} \pi_u(a_{t+1}|s_{t+1}) \sum_{s_{t+2}} P(s_{t+1}, a_{t+1}, s_{t+2}) \left[ R(s_{t+1}, a_{t+1}, s_{t+2} + \gamma_{t+2}V^\pi \cdot \gamma(s_{t+2}) \right]$$

$$\leq \sum_a \pi_u'(a|s) \sum_{s,a,s_{t+1}} P(s, a, s_{t+1}) \sum_{a_{t+1}} \pi_u'(a_{t+1}|s_{t+1}) \sum_{s_{t+2}} P(s_{t+1}, a_{t+1}, s_{t+2}) \left[ R(s_{t+1}, a_{t+1}, s_{t+2} + \gamma_{t+2}V^\pi \cdot \gamma(s_{t+2}) \right]$$

For tabular representations, this weighting is not relevant because the policy is updated individually for each state. As the representation becomes less local, the weighting by $d^b$ becomes more relevant for trading off error in different states. Moreover, for $b = \pi$, i.e. on-policy, the policy improvement claim is true. As the stationary distributions induced by $b$ and $\pi$ become more different, and there is significant aliasing in the policy function approximator, then it is unclear if updates that improve $\pi$ according to weighting states by $d^b$ improve $\pi$ according to weighting states by $d^\pi$. 