Decision-Aware Actor-Critic with Function Approximation & Theoretical Guarantees

Sharan Vaswani, Amirreza Kazemi, Reza Babanezhad, Nicolas Le Roux

   

NeurIPS 2023

Motivation

- ✓ Policy Gradient (PG) methods can easily handle function approximation and structured state-action spaces, making them widely used for RL tasks.
- × PG methods require estimating a policy's return. Using Monte-Carlo sampling results in high variance in the estimated return, leading to higher sample-complexity.

Motivation

✓ Policy Gradient (PG) methods can easily handle function approximation and structured state-action spaces, making them widely used for RL tasks.

× PG methods require estimating a policy's return. Using Monte-Carlo sampling results in high variance in the estimated return, leading to higher sample-complexity.

✓ Actor-critic (AC) methods alleviate this problem by using a value-based method (critic) to approximate a policy's estimated value, while a PG method (actor) uses this estimate to improve the policy towards obtaining higher returns.

Motivation

- ✓ Policy Gradient (PG) methods can easily handle function approximation and structured state-action spaces, making them widely used for RL tasks.
- × PG methods require estimating a policy's return. Using Monte-Carlo sampling results in high variance in the estimated return, leading to higher sample-complexity.
- ✓ Actor-critic (AC) methods alleviate this problem by using a value-based method (critic) to approximate a policy's estimated value, while a PG method (actor) uses this estimate to improve the policy towards obtaining higher returns.
- × Unclear how to train the actor and critic components jointly in order to learn good policies. E.g, the critic is typically trained by minimizing the TD error, an objective that is potentially decorrelated with the actor's objective of learning a good policy.

Motivation

- ✓ Policy Gradient (PG) methods can easily handle function approximation and structured state-action spaces, making them widely used for RL tasks.
- × PG methods require estimating a policy's return. Using Monte-Carlo sampling results in high variance in the estimated return, leading to higher sample-complexity.
- ✓ Actor-critic (AC) methods alleviate this problem by using a value-based method (critic) to approximate a policy's estimated value, while a PG method (actor) uses this estimate to improve the policy towards obtaining higher returns.
- × Unclear how to train the actor and critic components jointly in order to learn good policies. E.g, the critic is typically trained by minimizing the TD error, an objective that is potentially decorrelated with the actor's objective of learning a good policy.
- Want the critic to use its model capacity to correctly estimate the state-action values that are useful for improving the actor's policy.

Motivation

- ✓ Policy Gradient (PG) methods can easily handle function approximation and structured state-action spaces, making them widely used for RL tasks.
- × PG methods require estimating a policy's return. Using Monte-Carlo sampling results in high variance in the estimated return, leading to higher sample-complexity.
- ✓ Actor-critic (AC) methods alleviate this problem by using a value-based method (critic) to approximate a policy's estimated value, while a PG method (actor) uses this estimate to improve the policy towards obtaining higher returns.
- × Unclear how to train the actor and critic components jointly in order to learn good policies. E.g, the critic is typically trained by minimizing the TD error, an objective that is potentially decorrelated with the actor's objective of learning a good policy.
- Want the critic to use its model capacity to correctly estimate the state-action values that are useful for improving the actor's policy.

Contribution: Theoretically principled objective to jointly train the actor and critic.

Functional Mirror Ascent for PG Framework (VBTMGGMCR'22)

- **Problem Formulation**: Given an infinite-horizon discounted MDP: $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, p, r, \rho, \gamma \rangle$, and a set of feasible policies Π, $\max_{\pi \in \Pi} J(\pi) := \mathbb{E}_{s_0, a_0, \ldots}[\sum_{\tau=0}^{\infty} \gamma^{\tau} r(s_{\tau}, a_{\tau})]$.

Functional Mirror Ascent for PG Framework (VBTMGGMCR'22)

- **Problem Formulation**: Given an infinite-horizon discounted MDP: $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, p, r, \rho, \gamma \rangle$, and a set of feasible policies Π, $\max_{\pi \in \Pi} J(\pi) := \mathbb{E}_{s_0, a_0, \ldots}[\sum_{\tau=0}^{\infty} \gamma^\tau r(s_\tau, a_\tau)]$.
- **Functional representation vs Policy Parameterization**
  - Functional representation: Specifies a policy's sufficient statistics. Examples:
    - Direct functional representation: Conditional distribution over actions $p^\pi(\cdot|s)$ for $s \in \mathcal{S}$.
    - Softmax functional representation: Logits $z^\pi(s,a)$ such that $p^\pi(a|s) \propto \exp(z^\pi(s,a))$.
  - Policy parameterization: Realization of the sufficient statistics. Determines Π. Examples:
    - Tabular parameterization for the direct functional representation: $p^\pi(a|s,\theta) = \theta(s,a)$.
    - Linear parameterization for the softmax functional representation: $z^\pi(a,s) = \langle \theta, \Psi(s,a) \rangle$.

Functional Mirror Ascent for PG Framework (VBTMGGMCR'22)

- **Problem Formulation**: Given an infinite-horizon discounted MDP: $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, p, r, \rho, \gamma \rangle$, and a set of feasible policies Π, $\max_{\pi \in \Pi} J(\pi) := \mathbb{E}_{s_0, a_0, \ldots}[\sum_{\tau=0}^{\infty} \gamma^{\tau} r(s_\tau, a_\tau)]$.
- **Functional representation vs Policy Parameterization**
  - Functional representation: Specifies a policy's sufficient statistics. Examples:
    - Direct functional representation: Conditional distribution over actions $p^{\pi}(\cdot|s)$ for $s \in \mathcal{S}$.
    - Softmax functional representation: Logits $z^{\pi}(s, a)$ such that $p^{\pi}(a|s) \propto \exp(z^{\pi}(s, a))$.
  - Policy parameterization: Realization of the sufficient statistics. Determines Π. Examples:
    - Tabular parameterization for the direct functional representation: $p^{\pi}(a|s, \theta) = \theta(s, a)$.
    - Linear parameterization for the softmax functional representation: $z^{\pi}(a, s) = \langle \theta, \Psi(s, a) \rangle$.
- **FMA-PG Algorithm**: Iteratively form and approximately maximize the surrogate function:

$$\ell_t(\theta) := J(\pi_t) + \langle \pi(\theta), \nabla_{\pi} J(\pi(\theta_t)) \rangle - \frac{1}{\eta} D_{\Phi}(\pi(\theta), \pi(\theta_t)) \; ; \; \pi_{t+1} = \pi(\theta_{t+1}).$$

Functional Mirror Ascent for PG Framework (VBTMGGMCR'22)

- **Problem Formulation**: Given an infinite-horizon discounted MDP: $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, p, r, \rho, \gamma \rangle$, and a set of feasible policies Π, $\max_{\pi \in \Pi} J(\pi) := \mathbb{E}_{s_0, a_0, \ldots}[\sum_{\tau=0}^{\infty} \gamma^{\tau} r(s_\tau, a_\tau)]$.
- **Functional representation vs Policy Parameterization**
  - Functional representation: Specifies a policy's sufficient statistics. Examples:
    - Direct functional representation: Conditional distribution over actions $p^{\pi}(\cdot|s)$ for $s \in \mathcal{S}$.
    - Softmax functional representation: Logits $z^{\pi}(s, a)$ such that $p^{\pi}(a|s) \propto \exp(z^{\pi}(s, a))$.
  - Policy parameterization: Realization of the sufficient statistics. Determines Π. Examples:
    - Tabular parameterization for the direct functional representation: $p^{\pi}(a|s, \theta) = \theta(s, a)$.
    - Linear parameterization for the softmax functional representation: $z^{\pi}(a, s) = \langle \theta, \Psi(s, a) \rangle$.
- **FMA-PG Algorithm**: Iteratively form and approximately maximize the surrogate function:

$$\ell_t(\theta) := J(\pi_t) + \langle \pi(\theta), \nabla_{\pi} J(\pi(\theta_t)) \rangle - \frac{1}{\eta} D_{\Phi}(\pi(\theta), \pi(\theta_t)) \; ; \; \pi_{t+1} = \pi(\theta_{t+1}).$$

✓ Maximizing $\ell_t(\theta)$ does not require computing $\nabla_{\pi} J(\pi)$ and results in off-policy updates.

✓ Monotonic policy improvement for any complex parameterization.

× Forming $\ell_t(\theta)$ requires knowledge of $\nabla_{\pi} J(\pi)$, which involves either Q^{π} or A^{π} functions.

Joint Objective for Actor & Critic

Generic lower-bound

For any gradient estimator $\hat{g}_t$ at iteration t of FMA-PG, for $c > 0$ and η such that $J + \frac{1}{\eta}\Phi$ is convex in π, if $\Phi^*(y) := \max_\pi[\langle y, \pi \rangle - \Phi(\pi)]$ is the Fenchel conjugate of Φ, we have

$$J(\pi) - J(\pi_t) \geq \underbrace{\langle \hat{g}_t, \pi(\theta) - \pi_t \rangle - \left(\frac{1}{\eta} + \frac{1}{c}\right) D_\Phi(\pi(\theta), \pi_t)}_{\text{Surrogate function to be maximized by the actor}}$$

$$- \underbrace{\frac{1}{c} D_{\Phi^*}\left(\nabla\Phi(\pi_t) - c[\nabla J(\pi_t) - \hat{g}_t], \nabla\Phi(\pi_t)\right)}_{\text{Error in } Q^\pi \text{ or } A^\pi \text{ estimation. Can be minimized by training a critic}}$$

Joint Objective for Actor & Critic

Generic lower-bound

For any gradient estimator $\hat{g}_t$ at iteration t of FMA-PG, for $c > 0$ and η such that $J + \frac{1}{\eta}\Phi$ is convex in π, if $\Phi^*(y) := \max_\pi [\langle y, \pi \rangle - \Phi(\pi)]$ is the Fenchel conjugate of Φ, we have

$$J(\pi) - J(\pi_t) \geq \underbrace{\langle \hat{g}_t, \pi(\theta) - \pi_t \rangle - \left(\frac{1}{\eta} + \frac{1}{c}\right) D_\Phi(\pi(\theta), \pi_t)}_{\text{Surrogate function to be maximized by the actor}}$$

$$- \underbrace{\frac{1}{c} D_{\Phi^*}\left(\nabla \Phi(\pi_t) - c[\nabla J(\pi_t) - \hat{g}_t], \nabla \Phi(\pi_t)\right)}_{\text{Error in } Q^\pi \text{ or } A^\pi \text{ estimation. Can be minimized by training a critic}}$$

- To maximize policy improvement, an algorithm should (i) learn $\hat{g}_t$ to minimize the blue term (critic objective) and (ii) compute $\pi \in \Pi$ that maximizes the green term (actor objective).

Joint Objective for Actor & Critic

Generic lower-bound

For any gradient estimator $\hat{g}_t$ at iteration t of FMA-PG, for $c > 0$ and η such that $J + \frac{1}{\eta}\Phi$ is convex in π, if $\Phi^*(y) := \max_\pi[\langle y, \pi \rangle - \Phi(\pi)]$ is the Fenchel conjugate of Φ, we have

$$J(\pi) - J(\pi_t) \geq \underbrace{\langle \hat{g}_t, \pi(\theta) - \pi_t \rangle - \left(\frac{1}{\eta} + \frac{1}{c}\right) D_\Phi(\pi(\theta), \pi_t)}_{\text{Surrogate function to be maximized by the actor}}$$
$$- \underbrace{\frac{1}{c} D_{\Phi^*}\left(\nabla\Phi(\pi_t) - c[\nabla J(\pi_t) - \hat{g}_t], \nabla\Phi(\pi_t)\right)}_{\text{Error in } Q^\pi \text{ or } A^\pi \text{ estimation. Can be minimized by training a critic}}$$

- To maximize policy improvement, an algorithm should (i) learn $\hat{g}_t$ to minimize the blue term (critic objective) and (ii) compute $\pi \in \Pi$ that maximizes the green term (actor objective).
- c is a parameter relating the critic error to the permissible movement in the actor update.

Instantiating Decision-aware Actor-Critic – Direct functional representation

Lower-bound for direct representation

For the direct representation and negative entropy mirror map, $c > 0$, $\eta \leq \frac{(1-\gamma)^3}{2\gamma\,|A|}$,

$$J(\pi) - J(\pi_t) \geq C + \mathbb{E}_{s \sim d^{\pi_t}} \left[\mathbb{E}_{a \sim p^{\pi_t}(\cdot|s)} \left[\frac{p^{\pi}(a|s)}{p^{\pi_t}(a|s)} \left(\hat{Q}^{\pi_t}(s,a) - \left(\frac{1}{\eta} + \frac{1}{c} \right) \log \left(\frac{p^{\pi}(a|s)}{p^{\pi_t}(a|s)} \right) \right) \right] \right]$$

$$- \mathbb{E}_{s \sim d^{\pi_t}} \left[\mathbb{E}_{a \sim p^{\pi_t}(\cdot|s)} [Q^{\pi_t}(s,a) - \hat{Q}^{\pi_t}(s,a)] + \frac{1}{c} \log \left(\mathbb{E}_{a \sim p^{\pi_t}(\cdot|s)} \left[\exp \left(-c \left[Q^{\pi_t}(s,a) - \hat{Q}^{\pi_t}(s,a) \right] \right) \right] \right) \right]$$

Instantiating Decision-aware Actor-Critic – Direct functional representation

Lower-bound for direct representation

For the direct representation and negative entropy mirror map, $c > 0$, $\eta \leq \frac{(1-\gamma)^3}{2\gamma\,|A|}$,

$$J(\pi) - J(\pi_t) \geq C + \mathbb{E}_{s\sim d^{\pi_t}}\left[\mathbb{E}_{a\sim p^{\pi_t}(\cdot|s)}\left[\frac{p^{\pi}(a|s)}{p^{\pi_t}(a|s)}\left(\hat{Q}^{\pi_t}(s,a) - \left(\frac{1}{\eta}+\frac{1}{c}\right)\log\left(\frac{p^{\pi}(a|s)}{p^{\pi_t}(a|s)}\right)\right)\right]\right]$$
$$- \mathbb{E}_{s\sim d^{\pi_t}}\left[\mathbb{E}_{a\sim p^{\pi_t}(\cdot|s)}[Q^{\pi_t}(s,a) - \hat{Q}^{\pi_t}(s,a)] + \frac{1}{c}\log\left(\mathbb{E}_{a\sim p^{\pi_t}(\cdot|s)}\left[\exp\left(-c\,[Q^{\pi_t}(s,a) - \hat{Q}^{\pi_t}(s,a)]\right)\right]\right)\right]$$

- Lower-bound holds for any policy or critic parameterization i.e. $p^{\pi}(\cdot|s) = p^{\pi}(\cdot|s,\theta)$, $\hat{Q}^{\pi}(s,a) = Q^{\pi}(s,a|\omega)$, and instantiates the actor and critic objectives at iteration t.

Instantiating Decision-aware Actor-Critic – Direct functional representation

Lower-bound for direct representation

For the direct representation and negative entropy mirror map, $c > 0$, $\eta \leq \frac{(1-\gamma)^3}{2\gamma\,|A|}$,

$$J(\pi) - J(\pi_t) \geq C + \mathbb{E}_{s\sim d^{\pi_t}}\left[\mathbb{E}_{a\sim p^{\pi_t}(\cdot|s)}\left[\frac{p^{\pi}(a|s)}{p^{\pi_t}(a|s)}\left(\hat{Q}^{\pi_t}(s,a) - \left(\frac{1}{\eta} + \frac{1}{c}\right)\log\left(\frac{p^{\pi}(a|s)}{p^{\pi_t}(a|s)}\right)\right)\right]\right]$$
$$- \mathbb{E}_{s\sim d^{\pi_t}}\left[\mathbb{E}_{a\sim p^{\pi_t}(\cdot|s)}[Q^{\pi_t}(s,a) - \hat{Q}^{\pi_t}(s,a)] + \frac{1}{c}\log\left(\mathbb{E}_{a\sim p^{\pi_t}(\cdot|s)}\left[\exp\left(-c\,[Q^{\pi_t}(s,a) - \hat{Q}^{\pi_t}(s,a)]\right)\right]\right)\right]$$

- Lower-bound holds for any policy or critic parameterization i.e. $p^{\pi}(\cdot|s) = p^{\pi}(\cdot|s,\theta)$, $\hat{Q}^{\pi}(s,a) = Q^{\pi}(s,a|\omega)$, and instantiates the actor and critic objectives at iteration t.
- The decision-aware critic loss is asymmetric and penalizes the under/over-estimation of the Q^{π} function differently.

Instantiating Decision-aware Actor-Critic – Direct functional representation

Lower-bound for direct representation

For the direct representation and negative entropy mirror map, $c > 0$, $\eta \leq \frac{(1-\gamma)^3}{2\gamma\,|A|}$,

$$J(\pi) - J(\pi_t) \geq C + \mathbb{E}_{s\sim d^{\pi_t}} \left[\mathbb{E}_{a\sim p^{\pi_t}(\cdot|s)} \left[\frac{p^{\pi}(a|s)}{p^{\pi_t}(a|s)} \left(\hat{Q}^{\pi_t}(s,a) - \left(\frac{1}{\eta} + \frac{1}{c} \right) \log \left(\frac{p^{\pi}(a|s)}{p^{\pi_t}(a|s)} \right) \right) \right] \right]$$

$$- \mathbb{E}_{s\sim d^{\pi_t}} \left[\mathbb{E}_{a\sim p^{\pi_t}(\cdot|s)} [Q^{\pi_t}(s,a) - \hat{Q}^{\pi_t}(s,a)] + \frac{1}{c} \log \left(\mathbb{E}_{a\sim p^{\pi_t}(\cdot|s)} \left[\exp\left(-c \left[Q^{\pi_t}(s,a) - \hat{Q}^{\pi_t}(s,a) \right] \right) \right] \right) \right]$$

- Lower-bound holds for any policy or critic parameterization i.e. $p^{\pi}(\cdot|s) = p^{\pi}(\cdot|s,\theta)$, $\hat{Q}^{\pi}(s,a) = Q^{\pi}(s,a|\omega)$, and instantiates the actor and critic objectives at iteration t.
- The decision-aware critic loss is asymmetric and penalizes the under/over-estimation of the Q^{π} function differently.
- Can construct two-armed bandit examples where minimizing the squared loss results in convergence to the sub-optimal action, while minimizing the decision-aware loss above results in convergence to the optimal action.

Instantiating Decision-aware Actor-Critic – Direct functional representation

Lower-bound for direct representation

For the direct representation and negative entropy mirror map, $c > 0$, $\eta \leq \frac{(1-\gamma)^3}{2\gamma\,|A|}$,

$$J(\pi) - J(\pi_t) \geq C + \mathbb{E}_{s \sim d^{\pi_t}} \left[\mathbb{E}_{a \sim p^{\pi_t}(\cdot|s)} \left[\frac{p^\pi(a|s)}{p^{\pi_t}(a|s)} \left(\hat{Q}^{\pi_t}(s,a) - \left(\frac{1}{\eta} + \frac{1}{c} \right) \log \left(\frac{p^\pi(a|s)}{p^{\pi_t}(a|s)} \right) \right) \right] \right]$$

$$- \mathbb{E}_{s \sim d^{\pi_t}} \left[\mathbb{E}_{a \sim p^{\pi_t}(\cdot|s)} [Q^{\pi_t}(s,a) - \hat{Q}^{\pi_t}(s,a)] + \frac{1}{c} \log \left(\mathbb{E}_{a \sim p^{\pi_t}(\cdot|s)} \left[\exp \left(-c\, [Q^{\pi_t}(s,a) - \hat{Q}^{\pi_t}(s,a)] \right) \right] \right) \right]$$

- Lower-bound holds for any policy or critic parameterization i.e. $p^\pi(\cdot|s) = p^\pi(\cdot|s,\theta)$, $\hat{Q}^\pi(s,a) = Q^\pi(s,a|\omega)$, and instantiates the actor and critic objectives at iteration t.
- The decision-aware critic loss is asymmetric and penalizes the under/over-estimation of the Q^π function differently.
- Can construct two-armed bandit examples where minimizing the squared loss results in convergence to the sub-optimal action, while minimizing the decision-aware loss above results in convergence to the optimal action.
- Similar results for the softmax functional representation.

Contributions

- Lower-bound (on the return of an arbitrary policy) depends on both the actor and critic.
  - ✓ The lower-bound is valid for any policy representation, and actor or critic parameterization.
- Generic AC algorithm and its instantiation for the direct and softmax policy representations.
  - ✓ The actor supports off-policy updates like in PPO, whereas the critic minimizes a decision-aware loss.
  - ✓ Examples to demonstrate that minimizing the proposed critic loss results in convergence to the optimal policy, whereas minimizing the standard squared loss does not.
- Conditions for the AC algorithm to guarantee monotonic policy improvement
  - ✓ Improvement guarantees hold regardless of the policy or critic parameterization.
- Simple experiments that demonstrate the importance of being decision-aware

Contributions

- Lower-bound (on the return of an arbitrary policy) depends on both the actor and critic.
  - ✓ The lower-bound is valid for any policy representation, and actor or critic parameterization.
- Generic AC algorithm and its instantiation for the direct and softmax policy representations.
  - ✓ The actor supports off-policy updates like in PPO, whereas the critic minimizes a decision-aware loss.
  - ✓ Examples to demonstrate that minimizing the proposed critic loss results in convergence to the optimal policy, whereas minimizing the standard squared loss does not.
- Conditions for the AC algorithm to guarantee monotonic policy improvement
  - ✓ Improvement guarantees hold regardless of the policy or critic parameterization.
- Simple experiments that demonstrate the importance of being decision-aware

Poster # 1904, Poster Session 2, Tue Dec 12, 5:15 p.m

Paper: `https://arxiv.org/abs/2305.15249`

Contact: `vaswani.sharan@gmail.com`