Of Moments and Matching: A Game-Theoretic Framework for Closing the Imitation Gap

G. Swammy, S. Choudhury, J. Bagnell, Z. Wu

presented by: Bao Do

Applied Math Program
University of Arizona

November 2021

Table of Contents

Problem Definition

MDP : $(\mathcal{S}, \mathcal{A}, \mathcal{T}, r, T, P_0)$

State space – $\mathcal{S}$

Action space – $\mathcal{A}$

Transition operator – $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$

Reward function – $r : \mathcal{S} \times \mathcal{A} \rightarrow [-1, 1]$

Horizon – T

initial state distribution – P_0

Problem Definition

- Policy class: $\Pi = \{\pi : \mathcal{S} \to \Delta(\mathcal{A})\}$
- Trajectory: $\tau = (s_t, a_t)_{t=1}^{T}$
  $\tau \sim \pi$ means that τ is generated by $s_1 \sim P_0$, $a_t \sim \pi(s_t)$ and $s_{t+1} = \mathcal{T}(s_t, a_t)$
- Value function : $V_t^\pi(s) = \mathbb{E}\left[\sum_{t'=t}^{T} r(s_{t'}, a_{t'}) \middle| \tau \sim \pi, s_t = s\right]$
- Q-value function: $Q_t^\pi(s) = \mathbb{E}\left[\sum_{t'=t}^{T} r(s_{t'}, a_{t'}) \middle| \tau \sim \pi, s_t = s, a_t = a\right]$
- Advantage function: $A_t^\pi(s, a) = Q_t^\pi(s, a) - V_t^\pi(s)$
- Performance: $J(\pi) = \mathbb{E}\left[\sum_{t'=t}^{T} r(s_{t'}, a_{t'}) \middle| \tau \sim \pi\right]$
- Imitation Gap: $J(\pi_E) - J(\pi)$

Problem Definition

- $\mathcal{F}_r = \{f : \mathcal{S} \times \mathcal{A} \to [-1, 1]\}$: class of reward functions
- $\mathcal{F}_Q = \{f : \mathcal{S} \times \mathcal{A} \to [-T, T]\}$: set of Q functions induced by sampling actions from some π
- $\mathcal{F}_{Q_E} = \{f : \mathcal{S} \times \mathcal{A} \to [-1, 1]\}$: set of Q functions induced by sampling actions from π_E
- $\mathcal{F} = \{f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$ is convex, compact, closed under negation, and finite dimensional

Reward:

$$
\begin{aligned}
J(\pi_E) - J(\pi) &= \mathbb{E}_{\tau \sim \pi_E} \sum_{t=1}^{T} r(s_t, a_t) - \mathbb{E}_{\tau \sim \pi} \sum_{t=1}^{T} r(s_t, a_t) \\
&= \mathbb{E}_{\tau \sim \pi} \sum_{t=1}^{T} -r(s_t, a_t) - \mathbb{E}_{\tau \sim \pi_E} \sum_{t=1}^{T} -r(s_t, a_t) \\
&\leq \sup_{f \in \mathcal{F}_r} \left[\mathbb{E}_{\tau \sim \pi_E} \sum_{t=1}^{T} f(s_t, a_t) - \mathbb{E}_{\tau \sim \pi_E} \sum_{t=1}^{T} f(s_t, a_t) \right]
\end{aligned}
$$

Off-policy Q:

$$J(\pi_E) - J(\pi) = \mathbb{E}_{\tau\sim\pi_E}\left[\sum_{t=1}^{T} Q_t^{\pi}(s_t, a_t) - \mathbb{E}_{\tau\sim\pi(s_t)} Q_t^{\pi}(s_t, a_t)\right]$$

$$\leq \sup_{f\in\mathcal{F}_Q} \mathbb{E}_{\tau\sim\pi_E}\left[\sum_{t=1}^{T} \mathbb{E}_{\tau\sim\pi(s_t)} Q_t^{\pi}(s_t, a) - Q_t^{\pi}(s_t, a_t)\right]$$

$$(Q_t^{\pi} \in \mathcal{F}_Q \forall\pi \in \Pi, r \in \mathcal{F}_r)$$

Need to justify using the performance diff lemma

Problem Definition
Moments

On-policy Q:

$$J(\pi_E) - J(\pi) = -\mathbb{E}_{\tau\sim\pi}\left[\sum_{t=1}^{T} Q_t^{\pi_E}(s_t,a_t) - \mathbb{E}_{\tau\sim\pi_E(s_t)} Q_t^{\pi_E}(s_t,a_t)\right]$$
$$\leq \sup_{f\in\mathcal{F}_{Q_E}} \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=1}^{T} Q_t^{\pi}(s_t,a_t) - \mathbb{E}_{\tau\sim\pi(s_t)}\left[Q_t^{\pi}(s_t,a)\right]\right]$$
$$(Q_t^{\pi_E} \in \mathcal{F}_{Q_E} \forall r \in \mathcal{F}_r)$$

In realizable setting, $\pi_E \in \Pi, \mathcal{F}_{Q_e} \subseteq \mathcal{F}_Q$.

2 player minimax game:

1. Learner (min player): select policy $\pi \in \Pi$
2. Discriminator (max player): select function $f \in \mathcal{F}$
   $\mathcal{F} = \{f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$ is convex, compact, closed under negation, and finite dimensional.

Payoff Functions:

1. On-policy reward:

$$U_1(\pi, f) = \frac{1}{T}\left(\mathbb{E}_{\tau\sim\pi} \sum_{t=1}^{T} f(s_t, a_t) - \mathbb{E}_{\tau\sim\pi_E} \sum_{t=1}^{T} f(s_t, a_t)\right)$$

2. On-policy Q:

$$U_1(\pi, f) = \frac{1}{T}\left(\mathbb{E}_{\substack{\tau\sim\pi \\ a\sim\pi(s_t)}} \sum_{t=1}^{T} f(s_t, a) - \mathbb{E}_{\tau\sim\pi_E} \sum_{t=1}^{T} f(s_t, a_t)\right)$$

3. Off-policy Q:

$$U_1(\pi, f) = \frac{1}{T}\left(\mathbb{E}_{\tau\sim\pi} \sum_{t=1}^{T} f(s_t, a_t) - \mathbb{E}_{\substack{\tau\sim\pi \\ a\sim\pi_E(s_t)}} \sum_{t=1}^{T} f(s_t, a_t)\right)$$

Approximate Equilibria

A pair $(\hat{\pi} \in \Pi, \hat{f} \in \mathcal{F})$ is a δ-**approximate equilibrium solution** if

$$\sup_{f \in \mathcal{F}} U_j(f, \hat{\pi}) - \frac{\delta}{2} \leq U_j(\hat{f}, \hat{\pi}) \leq \inf_{\pi \in \Pi} U_j(\hat{f}, \pi) + \frac{\delta}{2}.$$

An **imitation game** δ-**oracle** $\phi\{\delta\}(\cdot)$ takes payoff function U and return $(k\delta)$-approximate equilibrium strategy for the policy player.

MDP examples

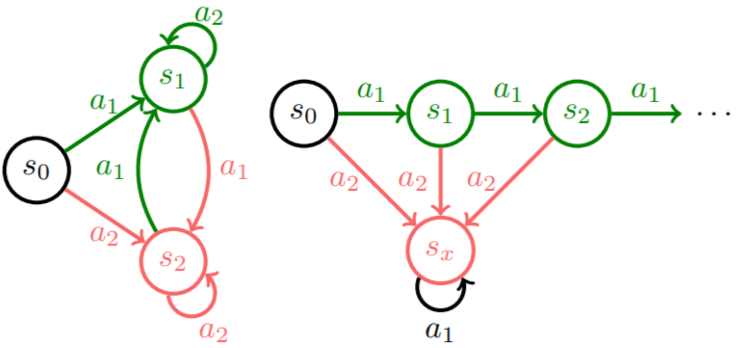

Figure 2. Left: Borrowed from (Ross et al. 2011), the goal of LOOP is to spend time in s_1. Right: a folklore MDP CLIFF, where the goal is to not "fall off the cliff" and end up in s_x evermore.

Performance Bounds

Moment Matched	Upper Bound	Lower Bound
Reward	$O(\epsilon T)$	$\Omega(\epsilon T)$
Off-Policy Q	$O(\epsilon T^2)$	$\Omega(\epsilon T^2)$
On-Policy Q	$O(\epsilon HT)$	$\Omega(\epsilon T)$

Table 2. An overview of the difference in bounds between the three types of moment matching. All bounds are on imitation gap (1).

Performance Bounds

Lemma 1. ***Reward Upper Bound:*** *If $\mathcal{F}_r$ spans $\mathcal{F}$, then for all MDPs, π_E, and $\pi \leftarrow \Psi\{\epsilon\}(U_1)$, $J(\pi_E) - J(\pi) \leq O(\epsilon T)$.*

Proof.

$$
\begin{aligned}
J(\pi_E) - J(\pi) &\leq \sup_{f \in \mathcal{F}_r} \left(\mathbb{E}_{\tau \in \pi} \sum_{t=1}^{T} f(s_t, a_t) - \mathbb{E}_{\tau \in \pi_E} \sum_{t=1}^{T} f(s_t, a_t) \right) \\
&\leq \sup_{f \in \mathcal{F}} \left(\mathbb{E}_{\tau \in \pi} \sum_{t=1}^{T} 2f(s_t, a_t) - \mathbb{E}_{\tau \in \pi_E} \sum_{t=1}^{T} 2f(s_t, a_t) \right) \\
&= 2T \sup_{f \in \mathcal{F}} U_1(\pi, f) \leq 2T\varepsilon.
\end{aligned}
$$

Performance Bounds

Lemma 2. ***Reward Lower Bound:*** *There exists an MDP, π_E, and $\pi \leftarrow \Psi\{\epsilon\}(U_1)$ such that $J(\pi_E) - J(\pi) \geq \Omega(\epsilon T)$.*

Proof. Consider CLIFF example with $r(s,a) = -\mathbb{1}_{s_x} - \mathbb{1}_{a_2}$ and a perfect expert that never takes a_2. If $P(a_2|s_0) = \varepsilon$, the optimal discriminator would be able to penalize the learner on average ε per step for T steps. Therefore, $J(\pi_E) - J(\pi) = \varepsilon T \leq \Omega(\varepsilon T)$.

Recoverability

considering skipping

Goal: Construct the oracle.

Assumptions:

1. State is finite
2. Policy class is complete

Approach:

1. Outer player follows a *no regret* strategy
2. Inner player follows a *best response* strategy

An **efficient no-regret algorithm** over a class $\mathcal{X}$ produces $x^1, \ldots, x^H \in \mathcal{X}$ that satisfy the following property for any sequence of loss functions $l^1, \ldots, l^H$:

$$\mathrm{Regret}(H) = \sum_t^H l^t(x^t) - \min_{x \in \mathcal{X}} \sum_t^H l^t(x) \leq \beta_{\mathcal{X}}(H)$$

where $\frac{\beta_{\mathcal{X}}(H)}{H} \leq \varepsilon$ holds for H that are $\mathcal{O}(\mathrm{poly}(1/\varepsilon))$

Primal. We execute a no-regret algorithm on the policy representation, while a maximization oracle over the space $\mathcal{F}$ computes the best response to those policies.

Dual. We execute a no-regret algorithm on the space $\mathcal{F}$, while a minimization oracle over policies computes *entropy regularized* best response policies.

	Outer player	Inner player	Application
Primal	Learner	Discriminator	Off-Q, On-Q
Dual	Discriminator	Learner	Reward

Theorem 1. *Given access to the no-regret and maximization oracles in either* ***primal*** *or* ***dual*** *above, for all three imitation games we are able to compute a* δ*-approximate equilibrium strategy for the policy player in* $poly(\frac{1}{\delta}, T, \ln|\mathcal{S}|, \ln|\mathcal{A}|)$ *iterations of the outer player optimization.*

Goal: Find $\hat{\pi}$ such that $\max_{f \in \mathcal{F}} U_j(\hat{\pi}, f) \leq \delta$

Procedure:

1. For $t = 1, \ldots, N$ do:
    - No-regret algorithm to find π^t
    - Set f^t to be the best response to π^t
2. Return $\hat{\pi} = \pi^{t^*}, t^* = \text{argmin}_t U_j(\pi^t, f^t)$

By the no-regret assumption with $l^t(\pi) = U_j(\pi, f^t)$:

$$\frac{1}{N}\sum_{t}^{N} U_j(\pi^t, f^t) - \frac{1}{N}\min_{\pi\in\Pi} U_j(\pi^t, f^t) \leq \frac{\beta_\Pi(N)}{N} \leq \delta$$

for $N = \text{poly}(1/\delta)$. Since $\pi_E \in \Pi$,

$$\min_{t} U_j(\pi^t, f^t) \leq \frac{1}{N}\sum_{t}^{N} U_j(\pi^t, f^t) \leq \delta$$

Since f^t is the best response to π^t:

$$\min_{t}\max_{f\in\mathcal{F}} U_j(\pi^t, f) \leq \delta$$

Integral Probability Metric (IPM):

$$\sup_{f\in\mathcal{F}} \{\mathbb{E}_{x\sim P_1}[f(x)] - \mathbb{E}_{x\sim P_2}[f(x)]\}$$

In our case: (IPM ojective)

$$\sup_{f\in\mathcal{F}} \sum_{t=1}^{T} \{\mathbb{E}_{x\sim\pi}[f(x)] - \mathbb{E}_{x\sim\pi_E}[f(x)]\}$$

Need to connect this IPM to the imitation gap or payoff functions

AdVIL: Adversarial Value-moment Imitation Learning

Algorithm 1 AdVIL

Input: Expert demonstrations $\mathcal{D}_E$, Policy class Π, Discriminator class $\mathcal{F}$, Performance threshold δ, Learning rates $\eta_f > \eta_\pi$

Output: Trained policy π

Set $\pi \in \Pi$, $f \in \mathcal{F}$, $L(\pi, f) = 2\delta$

while $L(\pi, f) > \delta$ **do**

$L(\pi, f) = \mathbb{E}_{(s,a)\sim\mathcal{D}_E}[\mathbb{E}_{x\sim\pi(s)}[f(s,x)] - f(s,a)]$

$f \leftarrow f + \eta_f \nabla_f L(\pi, f)$

$\pi \leftarrow \pi - \eta_\pi \nabla_\pi L(\pi, f)$

end while

Algorithm: Off-Q

Derivation

AdRIL: Adversarial Reward-moment Imitation Learning

Algorithm 2 AdRIL

Input: Expert demonstrations $\mathcal{D}_E$, Policy class Π, Dynamics $\mathcal{T}$, Kernel K, Performance threshold δ
Output: Trained policy π
Set $\pi \in \Pi$, $f = 0$, $\mathcal{D}_\pi = \{\}$, $\mathcal{D}' = \{\}$, $L(\pi, f) = 2\delta$
while $L(\pi, f) > \delta$ **do**
  $f \leftarrow \mathbb{E}_{\tau \sim D_\pi}[\sum_t K(sa, \cdot)] - \mathbb{E}_{\tau \sim D_E}[\sum_t K(sa, \cdot)]$
  $\pi, \mathcal{D}' \leftarrow \texttt{MaxEntRL}(\texttt{T} = \mathcal{T}, \texttt{r} = -f)$
  $\mathcal{D}_\pi \leftarrow \mathcal{D}_\pi \cup \mathcal{D}'$
  $L(\pi, f) = \mathbb{E}_{\tau \sim D'}[\sum_t f(s, a)] - \mathbb{E}_{\tau \sim D_E}[\sum_t f(s, a)]$
end while

DAeQuIL: DAgger-esque Qu-moment Imitation Learning

Algorithm 3 DAeQuIL

Input: Queryable expert π_E, Policy class Π, Discriminator class $\mathcal{F}$, Performance threshold δ, Behavioral cloning loss $\ell_{BC} : \Pi \to \mathbb{R}$, Strongly convex fn $R : \Pi \to \mathbb{R}$
Output: Trained policy π
Optimize: $\pi \leftarrow \arg\min_{\pi' \in \Pi} \ell_{BC}(\pi')$.
Set $L(\pi) = 2\delta$, $\mathcal{D} = []$, $F = []$, $t = 1$
while $L(\pi) > \delta$ **do**
  Rollout π to generate $\mathcal{D}_\pi \leftarrow [(s,a), \ldots]$.
  Relabel $\mathcal{D}_\pi$ to $\mathcal{D}_E \leftarrow [(s,a')|a' \sim \pi_E(s), \forall s \in \mathcal{D}_\pi]$
  $L(f) = \mathbb{E}_{(s,a)\sim\mathcal{D}_\pi}[f(s,a)] - \mathbb{E}_{(s,a)\sim\mathcal{D}_E}[f(s,a)]$
  Append: $F \leftarrow F \cup [\arg\max_{f' \in \mathcal{F}} L(f')]$.
  Append: $\mathcal{D} \leftarrow \mathcal{D} \cup [(s,t) \,|\, \forall s \in \mathcal{D}_\pi]$.
  $L(\pi) = \mathbb{E}_{(s,t)\in\mathcal{D}}[F[t](s,\pi(s))] + \ell_{BC}(\pi) + R(\pi)$
  Optimize: $\pi \leftarrow \arg\min_{\pi' \in \Pi} L(\pi')$.
  $t \leftarrow t + 1$
end while