Reinforcement Learning via Fenchel-Rockafellar Duality

Kei Ishikawa

Journal Club on January 26, 2023

Motivation

- Several influential papers on offline reinforcement learning (RL) in 2019-2020:
  - DualDice [2]
  - AlgaeDice [4]
  - GenDice [5]
  - ValueDice [1]
  - and their summary [3] (today's paper)
- Dice (stationary DIstribution Correction Estimation) leverages
  - The linear structure of RL
  - Fenchel-Rockafellar duality and Lagrange Duality

  Similar ideas are used in "distributionally robust" methods and neural estimation of f-divergence.

Outline

Outline

Fenchel Conjugate

- The Fenchel conjugate f_* of function $f : \Omega \to \mathbb{R}$ is defined as

$$f_*(y) := \max_{x \in \Omega} \langle x, y \rangle - f(x),$$

  where $\langle \cdot, \cdot \rangle$ denotes the inner product defined on Ω.
- For a proper, convex, lower semi-continuous f, one has duality $f_{**} = f$; i.e,

$$f(x) = \max_{y \in \Omega^*} \langle x, y \rangle - f_*(y),$$

  where Ω^* denotes the domain of f_*.
  - f is proper iff $\{x \in \Omega : f(x) < \infty\}$ is non-empty and $f(x) > -\infty$ for all $x \in \Omega$.
  - f is lower semi-continuous iff $\{x \in \Omega : f(x) > \alpha\}$ is an open set for all $\alpha \in \mathbb{R}$.

Fenchel Conjugate

Functions	Conjugates
$\frac{1}{2}x^2$	$\frac{1}{2}y^2$
$\delta_{\{a\}}(x)$[1]	$\langle a, y\rangle$
$\delta_{\mathbb{R}_+}(x)$	$\delta_{\mathbb{R}_-}(y)$
$\langle a, x\rangle + b \cdot f(x)$	$b \cdot f_* \left(\frac{y-a}{b}\right)$
$f(ax)$	$f_*(\frac{y}{a})$
$f(x+b)$	$f_*(y) - \langle b, y\rangle$
$\mathrm{D}_f(x\|\|p)$ (unrestricted x)	$\mathbb{E}_{z\sim p}[f_*(y(z))]$
$\mathrm{D}_{\mathrm{KL}}(x\|\|p)$ where $x \in \Delta(\mathcal{Z})$	$\log \mathbb{E}_{Z\sim p}[\exp y(Z)]$

Table: Common functions and their Fenchel conjugates.

[1] δ_C is an indicator function of that is zero if $x \in C$ and infinity otherwise.

f-divergences

- For a convex function f and distributions p, q over some domain $\mathcal{Z}$, the f-divergence between them is defined as,

$$\mathrm{D}_f(p\|q) = \mathbb{E}_{z\sim q}\left[f\left(\frac{p(z)}{q(z)}\right)\right].$$

  - Non-negativity:

$$\mathrm{D}_f(p\|q) \geq 0, \mathrm{D}_f(p\|q) = 0 \text{ iff } p = q$$

  - Variational representation (not used):

$$\mathrm{D}_f(p\|q) = \sup_{\Omega\to\mathrm{effdom}(f_*)} \mathbb{E}_p[g] - \mathbb{E}_q[f_* \circ g]$$

  - Examples: KL-divergence, total variation, α-divergences

Fenchel-Rockafellar Duality

- Primal problem:

$$\min_{x \in \Omega} J_{\text{primal}}(x) := f(x) + g(Ax),$$

where $f, g : \Omega \to \mathbb{R}$ are convex, lower semi-continuous functions, and A is a linear operator (e.g, a matrix).

- Dual problem:

$$\max_{y \in \Omega^*} J_{\text{dual}} := -f_*(-A_* y) - g_*(y),$$

where we use A_* to denote the adjoint linear operator of A

  - A_* is the linear operator for which $\langle y, Ax \rangle = \langle A_* y, x \rangle$, for all x, y.
  - In the common case of A simply being a real-valued matrix, A_* is the transpose of A.

Fenchel-Rockafellar Duality

- Under mild conditions (constraint qualification), we can derive the above as

$$
\begin{aligned}
\min_{x\in\Omega} J_{\mathrm{primal}}(x) &= \min_{x\in\Omega} f(x) + g(Ax) \\
&= \min_{x\in\Omega} \max_{y\in\Omega^*} f(x) + \langle y, Ax \rangle - g_*(y) \\
&= \max_{y\in\Omega^*} \{ \min_{x\in\Omega} f(x) + \langle y, Ax \rangle \} - g_*(y) \\
&= \max_{y\in\Omega^*} \{ -\max_{x\in\Omega} \langle -A_* y, x \rangle - f(x) \} - g_*(y) \\
&= \max_{y\in\Omega^*} -f_*(-A_* y) - g_*(y) \\
&= \max_{y\in\Omega^*} J_{\mathrm{dual}}(y).
\end{aligned}
$$

- The relationship between primal and dual solutions (when ∇f_* exists)

$$
x^* := \arg\min_{x\in\Omega} J_{\mathrm{primal}}(x) = \nabla f_*(-A_* y^*)
$$

Two different Dual Problems of Linear Constraints)

- Fenchel-Rockafellar dual

$$
\begin{aligned}
&\min_x f(x) \text{ s.t. } Ax = b \\
&= \min_x f(x) + \delta_{\{b\}}(Ax) \\
&= \max_y -f_*(-A_* y) - \langle b, y \rangle
\end{aligned}
$$

- Lagrange dual

$$
\begin{aligned}
&\min_x f(x) \text{ s.t. } Ax = b \\
&= \min_x f(x) + \max_y y^T(Ax - b) \\
&= \min_x \max_y f(x) + y^T(Ax - b)
\end{aligned}
$$

Outline

A Quick Introduction to Reinforcement Learning

- Markov Processes
  - Model: $(\mathcal{S}, T(s'|s), \mu_0(s_0))$
    - $\mathcal{S}$: a set of all (discrete) states
    - $T(s'|s)$: transition probability $\Pr(S_{t+1} = s'|S_t = s)$
    - $\mu_0(s)$: probability of initial state $\Pr(S_0 = s)$
  - Realization: $\{S_t\}_{t=0}^{\infty}$
  - Recursive formulae for $\Pr(S_t = s)$:
    - $\Pr(S_{t+1} = s') = \sum_{s \in \mathcal{S}} T(s'|s) \Pr(S_t = s)$
    - Using transition operator $\mathcal{P}$ and its adjoint $\mathcal{P}_*$ defined as

$$\mathcal{P} : f(s) \mapsto \sum_{s' \in \mathcal{S}} f(s')T(s'|s) = \mathbb{E}[f(S')|S = s],$$

$$\mathcal{P}_* : f(s) \mapsto \sum_{s' \in \mathcal{S}} T(s|s')f(s'),$$

the recursive formula for $p_t(s) := \Pr(S_t = s)$ simplifies to

$$p_{t+1}(s) = \mathcal{P}_* p_t(s)$$

A Quick Introduction to Reinforcement Learning

- Markov Decision Processes (MDPs)
  - Model: $(\mathcal{S}, \mathcal{A}, T(s'|s,a), \mu_0(s_0), R(s,a))$ and policy $\pi(a|s)$
    - $\mathcal{A}$: a set of all (discrete) actions
    - $T(s'|s,a)$: transition probability $\Pr(S_{t+1} = s'|S_t = s, A_t = a)$
    - $R(s,a)$: reward function that gives reward $r_t = R(s_t, a_t)$
    - $\pi(a|s)$: action probability of $\Pr(A_t = a|S_t = s)$
  - Realization: $\{(S_t, A_t)\}_{t=0}^{\infty}$
  - Recursive formulae for $\Pr(S_t = s, A_t = a)$:
    - $\Pr(S_{t+1} = s', A_{t+1} = a') = \pi(a'|s') \sum_{s \in \mathcal{S}, a \in \mathcal{A}} T(s'|s,a) \Pr(S_t = s, A_t = a)$
    - Using transition operator $\mathcal{P}^\pi$ and its adjoint $\mathcal{P}_*^\pi$ defined as

$$\mathcal{P}^\pi : f(s,a) \mapsto \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} f(s',a')\pi(a'|s')T(s'|s,a) = \mathbb{E}[f(S',A')|S' = s,$$

$$\mathcal{P}_*^\pi : f(s,a) \mapsto \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} \pi(a|s)T(s|s',a')f(s',a'),$$

the recursive formula for $p_t(s,a) := \Pr(S_t = s, A_t = a)$ simplifies to

$$p_{t+1}(s,a) = \mathcal{P}_*^\pi p_t(s,a)$$

A Quick Introduction to Reinforcement Learning

- Reinforcement Learning (RL)
  - Given MDP $(\mathcal{S}, \mathcal{A}, T(s'|s,a), \mu_0(s_0), r(s,a))$ and discount rate $0 < \gamma < 1$, find an optimal policy by solving

$$\max_{\pi} \rho(\pi) := (1-\gamma)\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t)]$$

- Online / Offline RL
  - Online RL: You can generate new sample trajectories $\{(S_t, A_t)\}_{t=0}^{\infty}$ from an arbitrary (or sometimes fixed) policy $\pi(a|s)$
  - Offline RL: You can only access the recorded sample trajectories $\{(S_t, A_t)\}_{t=0}^{\infty}$ from some policy $\pi(a|s)$

Outline

Offline Policy Evaluation with Distribution Correction

- Assumptions
  - Observable data: $(S', A, S) \sim \mathcal{D}$ where $(S, A) \sim d^{\mathcal{D}}(s, a)$ and $S'|S = s, A = a \sim T(s'|s, a)$
- Offline Estimation of $\rho(\pi) := (1-\gamma)\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t)]$
  - Let us define $d^{\pi}(s, a) := (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \Pr(S_t = s, A_t = a)$, so that $\rho(\pi) = \mathbb{E}_{(S,A)\sim d^{\pi}}[r(S, A)]$.
  - If we can estimate $\zeta(s, a) := \frac{d^{\pi}(s,a)}{d^{\mathcal{D}}(s,a)}$, we can estimate policy value $\rho(\pi)$ as

$$\hat{\rho}(\pi) = \frac{1}{N}\sum_{n=1}^{N} \hat{\zeta}(S_n, A_n) r(S_n, A_n)$$

    where $(S_n, A_n) \sim d^{\mathcal{D}}$.

Difficulty of Path-wise Distribution Correction

- Assumption: $\mathcal{D} = \{\{(S_t^{(n)}, A_t^{(n)})\}_{t=0}^{\infty} : n = 1, \ldots, N\}$ sampled from the MDP with base policy π_0 so that $d^{\mathcal{D}} = d^{\pi_0}$
- Distribution correction:

$$\begin{aligned}
\rho(\pi) &= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_\pi[r(S_t, A_t)] \\
&= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{\pi_0}\left[\left(\frac{\Pr(\{(S_\tau, A_\tau)\}_{\tau=0}^{t}|\pi)}{\Pr(\{(S_\tau, A_\tau)\}_{\tau=0}^{t}|\pi_0)}\right) r(S_t, A_t)\right] \\
&= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{\pi_0}\left[\prod_{\tau=0}^{t}\left(\frac{\pi(A_\tau|S_\tau)}{\pi_0(A_\tau|S_\tau)}\right) r(S_t, A_t)\right]
\end{aligned}$$

  - The empirical version of the above tends to have high variance, as estimated product term $\prod_{\tau=0}^{t}\left(\frac{\pi(A_\tau|S_\tau)}{\pi_0(A_\tau|S_\tau)}\right)$ is often difficult.

Estimation of Distribution Correction $\zeta = d^\pi / d^\mathcal{D}$

- Sufficient condition of d^π:

$$d^\pi(s,a) = (1-\gamma)\pi(a|s)\mu_0(s) + \gamma \mathcal{P}_*^\pi d^\pi(s,a) \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}$$

  - It follows from $d^\pi(s,a) := (1-\gamma)\sum_{t=0}^{\infty}(\gamma \mathcal{P}_*^\pi)^t(\pi \times \mu_0)(s,a)$
- Solve the equivalent optimization problem:

$$\min_d \mathrm{D}_f(d \| d^\mathcal{D})$$

subject to the above equality constraints.

Estimation of Distribution Correction $\zeta = d^\pi / d^\mathcal{D}$

- Primal:

$$\min_d \mathrm{D}_f(d \| d^\mathcal{D}) \text{ s.t. } (1 - \gamma \mathcal{P}^\pi_*) d(s,a) = (1-\gamma)\pi(a|s)\mu_0(s)$$

- Fenchel-Rockafellar Dual:

$$\begin{aligned} &\max_{Q:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} -\langle (1-\gamma)\pi(a|s)\mu_0(s), Q(S,A)\rangle \\ &\qquad\qquad - \mathbb{E}_{(S,A)\sim d^\mathcal{D}}\left[f_*\left((\gamma \mathcal{P}^\pi - 1)Q(S,A)\right)\right] \\ &= \max_{Q:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} -(1-\gamma)\mathbb{E}_{(S,A)\sim\pi(a|s)\mu_0(s)}[Q(S,A)] \\ &\qquad\qquad - \mathbb{E}_{(S,A)\sim d^\mathcal{D}}\left[f_*\left((\gamma \mathcal{P}^\pi - 1)Q(S,A)\right)\right] \end{aligned}$$

  - The primal solution can be recovered from dual solution $Q^*(s,a)$ as

$$d^\pi(s,a) = d^\mathcal{D}(s,a) \cdot f'_*\left((\gamma\mathcal{P}^\pi - 1)Q(s,a)\right)$$

    - ★ This is because $d^\pi(s,a) = \frac{\mathrm{d}}{\mathrm{d}x(s,a)} \mathbb{E}_{(S,A)\sim d^\mathcal{D}}[f_*(x)]\Big|_{x=(\gamma\mathcal{P}^\pi - 1)Q^*}$

Estimation of Distribution Correction $\zeta = d^\pi / d^\mathcal{D}$

- Lagrange Dual:

$$
\begin{aligned}
&\min_{d:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \max_{Q:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \mathbb{E}_{(S,A)\sim d^\mathcal{D}} \left[f\left(\frac{d(S,A)}{d^\mathcal{D}(S,A)} \right) \right] \\
&\quad + \left\langle Q(s,a), (1-\gamma\mathcal{P}_*^\pi)\left(\frac{d(s,a)}{d^\mathcal{D}(s,a)} \right) d^\mathcal{D}(s,a) - (1-\gamma)\pi(a|s)\mu_0(s) \right\rangle \\
&= \min_{\zeta:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \max_{Q:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \mathbb{E}_{(S,A)\sim d^\mathcal{D}} \left[f(\zeta(S,A)) \right] \\
&\qquad - (1-\gamma)\mathbb{E}_{(S,A)\sim\pi(a|s)\mu_0(s)}[Q(s,a)] \\
&\qquad + \mathbb{E}_{(S,A)\sim d^\mathcal{D}(s,a)}[\zeta(S,A)(1-\gamma\mathcal{P}^\pi)Q(S,A)]\rangle \\
&= \min_{\zeta:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \max_{Q:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \mathbb{E}_{(S,A)\sim d^\mathcal{D}} \left[f(\zeta(S,A)) \right] \\
&\qquad - (1-\gamma)\mathbb{E}_{(S,A)\sim\pi(a|s)\mu_0(s)}[Q(s,a)] \\
&\qquad + \mathbb{E}_{\substack{(S',A',S,A) \\ \sim\pi(a'|s')T(s'|s,a)d^\mathcal{D}(s,a)}} [\zeta(S,A)Q(S,A) - \zeta(S,A)\gamma Q(S',A')]
\end{aligned}
$$

where we used reparametrization $\zeta = d/d^\mathcal{D}$

Outline

Extensions

- Undiscounted RL ($\gamma \nearrow 1$)
- Policy Optimization
- Imitation Learning

Undiscounted RL

- We are interested in estimating

$$\rho(\pi) := \lim_{\gamma \nearrow 1} (1-\gamma)\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t)]$$

- Stationary distribution correction
  - Let us define

$$d^\pi(\pi)(s,a) := \lim_{\gamma \nearrow 1} (1-\gamma)\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t \Pr(S_t = s, A_t = a)]$$

    so that $\rho(\pi) = \mathbb{E}_{(S,A)\sim d^\pi}[r(S,A)]$.
  - The sufficient conditions for d^π are

$$d^\pi = \mathcal{P}_*^\pi d^\pi \text{ and } \sum_{s\in\mathcal{S}, a\in\mathcal{A}} d^\pi(s,a) = 1$$

Undiscounted RL

- Primal:

$$\min_{d} \mathrm{D}_f(d \| d^{\mathcal{D}}) \text{ s.t. } d^\pi = \mathcal{P}_*^\pi d^\pi \text{ and } \sum_{s \in \mathcal{S}, a \in \mathcal{A}} d^\pi(s,a) = 1$$

- Fenchel-Rockafellar Dual:

$$\max_{\lambda \in \mathbb{R}, Q: \mathcal{S} \times \mathcal{A} \to \mathbb{R}} \lambda - \mathbb{E}_{(S,A) \sim d^{\mathcal{D}}} \left[f_* \left(\lambda + (\mathcal{P}^\pi - 1) Q(S,A) \right) \right]$$

  - The primal solution can be recovered from dual solution $\lambda^*, Q^*(s,a)$ as

$$d^\pi(s,a) = d^{\mathcal{D}}(s,a) \cdot f_*' \left(\lambda^* + (\mathcal{P}^\pi - 1) Q(s,a) \right)$$

Undiscounted RL

- Lagrange Dual[2]:

$$
\begin{aligned}
&\min_{d:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \max_{\lambda\in\mathbb{R}, Q:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \mathbb{E}_{(S,A)\sim d^{\mathcal{D}}}\left[f\left(\frac{d(S,A)}{d^{\mathcal{D}}(S,A)}\right)\right] \\
&\quad + \left\langle Q(s,a), (1-\mathcal{P}_*^{\pi})\left(\frac{d(s,a)}{d^{\mathcal{D}}(s,a)}\right) d^{\mathcal{D}}(s,a)\right\rangle \\
&\quad + \lambda\left(1 - \sum_{s\in\mathcal{S}, a\in\mathcal{A}}\left(\frac{d(s,a)}{d^{\mathcal{D}}(s,a)}\right) d^{\mathcal{D}}(s,a)\right) \\
&= \min_{\zeta:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \max_{\lambda\in\mathbb{R}, Q:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \mathbb{E}_{(S,A)\sim d^{\mathcal{D}}}[f(\zeta(S,A))] + \lambda \\
&\quad - \mathbb{E}_{\substack{(S',A',S,A) \\ \sim\pi(a'|s')T(s'|s,a)d^{\mathcal{D}}(s,a)}}\left[\zeta(S,A)\cdot(\lambda + Q(S',A') - Q(S,A))\right]
\end{aligned}
$$

where we used reparametrization $\zeta = d/d^{\mathcal{D}}$

[2] This is called GenDice

Imitation Learning

- Problem Setting:
  - Assumption: $\mathcal{D} = \{\{(S_t^{(n)}, A_t^{(n)})\}_{t=0}^{\infty} : n = 1, \ldots, N\}$ sampled from the MDP with base policy π_0 so that $d^{\mathcal{D}} = d^{\pi_0}$
  - We are interested in imitating π_0 with π^* so that

$$\pi^* = \arg\min_{\pi} \mathrm{D}_{\mathrm{KL}}(d^{\pi} \| d^{\mathcal{D}})$$

Imitation Learning

- Fenchel-Rockafellar Dual (Donsker-Varadhan representation):

$$
\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}(d^{\pi} \| d^{\mathcal{D}}) \\
&= \min_{d \in \Delta(\mathcal{S} \times \mathcal{A})} \mathrm{D}_{\mathrm{KL}}(d \| d^{\mathcal{D}}) \\
&\qquad \text{s.t. } d(s,a) = (1-\gamma)\pi(a|s)\mu_0(s) + \gamma \mathcal{P}_*^{\pi} d(s,a) \\
&= \max_{\nu : \mathcal{S} \times \mathcal{A} \to \mathbb{R}} -\log \mathbb{E}_{(S,A) \sim d^{\mathcal{D}}} \left[\exp\left((1 - \gamma \mathcal{P}^{\pi}) \nu(S,A) \right) \right] \\
&\qquad + (1-\gamma) \mathbb{E}_{(S,A) \sim \pi(a|s)\mu_0(s)} [\nu(S,A)]
\end{aligned}
$$

- Imitation Learning[3]:

$$
\begin{aligned}
\pi^* = \arg\min_{\pi} \max_{\nu : \mathcal{S} \times \mathcal{A} \to \mathbb{R}} & -\log \mathbb{E}_{(S,A) \sim d^{\mathcal{D}}} \left[\exp\left((1 - \gamma \mathcal{P}^{\pi}) \nu(S,A) \right) \right] \\
& + (1-\gamma) \mathbb{E}_{(S,A) \sim \pi(a|s)\mu_0(s)} [\nu(S,A)]
\end{aligned}
$$

[3] This is called ValueDice

Policy Optimization

- Problem Setting:
  - We are interested in finding maximizer

$$\pi^* = \arg\max_{\pi} \rho(\pi) - \mathrm{D}_f(d^{\pi} \| d^{\pi_0})$$

of regularized policy value

Policy Optimization

- Fenchel-Rockafellar Dual:

$$
\begin{aligned}
&\rho(\pi) - \mathrm{D}_f(d^\pi \| d^{\mathcal{D}}) \\
&= \max_{d:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \mathrm{D}_f(d \| d^{\mathcal{D}}) - \mathbb{E}_{(S,A)\sim d}[r(S,A)] \\
&\qquad \text{s.t. } d(s,a) = (1-\gamma)\pi(a|s)\mu_0(s) + \gamma \mathcal{P}_*^\pi d(s,a) \\
&= \min_{Q:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \mathbb{E}_{(S,A)\sim d^{\mathcal{D}}}\left[f_*\left(R(S,A) - (1-\gamma\mathcal{P}^\pi)Q(S,A)\right)\right] \\
&\qquad\qquad + (1-\gamma)\mathbb{E}_{(S,A)\sim\pi(a|s)\mu_0(s)}[Q(S,A)]
\end{aligned}
$$

- Policy Optimization[4]:

$$
\begin{aligned}
\pi^* = \arg\max_{\pi} \min_{Q:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} &\mathbb{E}_{(S,A)\sim d^{\mathcal{D}}}\left[f_*\left(R(S,A) - (1-\gamma\mathcal{P}^\pi)Q(S,A)\right)\right] \\
&+ (1-\gamma)\mathbb{E}_{(S,A)\sim\pi(a|s)\mu_0(s)}[Q(S,A)]
\end{aligned}
$$

[4] This is called AlgaeDice

(My Personal) Take Aways

- Convex optimization is not only for linear/kernelized models
- Reinforcement learning has a useful linear structure
- Neural networks can be used to approximately solve these problems
- Interpreting the conditional expectation as a linear operator and taking its adjoint can be a useful trick

References

Ilya Kostrikov, Ofir Nachum, and Jonathan Tompson.
Imitation learning via off-policy distribution matching.
arXiv preprint arXiv:1912.05032, 2019.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li.
Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections.
Advances in Neural Information Processing Systems, 32, 2019.

Ofir Nachum and Bo Dai.
Reinforcement learning via fenchel-rockafellar duality.
arXiv preprint arXiv:2001.01866, 2020.

Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans.
Algaedice: Policy gradient from arbitrary experience.
arXiv preprint arXiv:1912.02074, 2019.

Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans.
Gendice: Generalized offline estimation of stationary values.
arXiv preprint arXiv:2002.09072, 2020.