DICE: Dual Stationary Distribution Correction Estimation

Boyang Liu

December 29, 2019

Abstract

From Fenchel Conjugate to Off-policy RL
keywords: f-divergence, fenchel conjugate, off-policy RL

1 Introduction

For many scenarios in machine learning, it is important to estimate two distribution difference. One typical metric to evaluate two distribution discrepancy is KL-divergence as following:

$$D_{KL}(p(x)\|q(x)) = \sum_{x\in X} p(x) \ln \frac{p(x)}{q(x)} \tag{1}$$

KL-divergence is very hard to estimate in high dimension since it needs to estimate the density function of both distribution p and q. In order to solve this problem, we first introduce a broad family of distribution divergence, which is called f-divergence. Later we will see how to estimate f-divergence by using convex conjugate. At last, the convex conjugate will bring a nice lower bound estimation of original distribution while the solving the lower bound does not need to explicitly estimate the density function of p and q. Instead, the lower bound of the f-divergence only requires accessing the distribution, which will finally leads to an off-policy RL algorithm.

2 f-divergence

Definition 2.1 (f-divergence)**.** Let P, Q are two distributions, $p(x)$, $q(x)$ are the density functions, then, the given any convex function f, which $f(1) = 0$, the f-divergence is defined as:

$$D_f(P\|Q) = \int_x q(x) f\left(\frac{p(x)}{q(x)}\right) dx \tag{2}$$

Example 2.1. For KL-divergence, the f is $x\log(x)$, which is a valid f for f-divergence:

$$\int_x p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$$

Example 2.2. For reverse KL-divergence, the f is $-\log(x)$, which is a valid f for f-divergence:

$$\int_x q(x) \log\left(\frac{q(x)}{p(x)}\right) dx$$

Example 2.3. For chi-square divergence, the f is $(x-1)^2$, which is a valid f for f-divergence:

$$\int_x \frac{(p(x)-q(x))^2}{q(x)} dx$$

The reason why f needs to be convex and $f(1)$ must equal to 0 is to make sure that the divergence is a positive value, and 0 means that two distributions are equal.

$$\begin{aligned} D_f(P\|Q) &= \int_x q(x) f\left(\frac{p(x)}{q(x)}\right) dx && (3)\\ &\geq f\left(\int_x q(x)\frac{p(x)}{q(x)} dx\right) && (4)\\ &= f(1) && (5)\\ &= 0 && (6) \end{aligned}$$

3 Convex Conjugate

Definition 3.1. The convex (fenchel) conjugate for function f is:

$$f^*(y) = \sup_{x \in \mathrm{dom} f} \left(y^T x - f(x)\right) \tag{7}$$

An important property of fenchel conjugate is that for any closed convex function, we have $f(x) = f^{**}(x)$. In other words, if function f is convex, we have

$$f(x) = \sup_{y \in \mathrm{dom} f} \left(x^T y - f^*(y)\right)$$

Example 3.1. For function $y = \frac{1}{2}x^2$, we have $f^*(u) = \max_x ux - \frac{1}{2}f(x^2)$, thus we have $f^*(u) = \frac{1}{2}u^2$, and $f(x) = f^{**}(x)$.

The key point is that fenchel conjugate can express a function by an optimization of its conjugate. Also, it can express some of the optimization problem to a function.

4 Revisit f-divergence

$$\begin{aligned} D_f(p\|q) &= \mathbb{E}_{a\sim q}\left[f\left(\frac{p(a)}{q(a)}\right)\right] \\ &= \mathbb{E}_{a\sim q}\left[\max_x \frac{p(a)}{q(a)} \cdot x - f_*(x)\right] \\ &\geq \max_{x: A\to\mathbb{R}} \mathbb{E}_{a\sim q}\left[\frac{p(a)}{q(a)} \cdot x(a) - f_*(x(a))\right] \\ &= \max_{x: A\to\mathbb{R}} \mathbb{E}_{a\sim q}\left[\frac{p(a)}{q(a)} \cdot x(a)\right] - \mathbb{E}_{a\sim q}\left[f_*(x(a))\right] \\ &= \max_{x: A\to\mathbb{R}} \mathbb{E}_{a\sim p}[x(a)] - \mathbb{E}_{a\sim q}\left[f_*(x(a))\right] \end{aligned} \tag{8}$$

This is nice since we do not need to estimate the density function. Notice GAN also use this trick to minimize the distribution between fake data and real data. See more details in f-gan [1] paper.

5 Off-policy Policy Evaluation (OPE)

5.1 Problem Setting

We would like to estimate the value function of specific policy π, but we only have access to some fixed data $\mathcal{D} := \left\{ \left(s^{(i)}, a^{(i)}, r^{(i)}, s^{(i)\prime} \right) \right\}_{i=1}^{N}$.

Mathematically, we would like to estimate the following term

$$\rho(\pi) := (1-\gamma) \cdot \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s,a) | s_0 \sim \beta_0, a_t \sim \pi\left(s_t\right), s_{t+1} \sim T\left(s_t, a_t\right) \right], \tag{9}$$

with the limitation that we only have access to $(s,a) \sim d^{\mathcal{D}}, s' \sim T(s,a)$

5.2 Reduction of OPE to Density Ratio Estimation

We can marginalize the above equation by the state action distribution to get rid of the horizon term:

$$\rho(\pi) = \mathbb{E}_{(s,a)\sim d^{\pi}}[r(s,a)], \tag{10}$$

where $d^{\pi}(s,a) := (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \cdot \Pr\left[s_t = s, a_t = a | s_0 \sim \beta_0, \pi\right]$. By employing the importance sampling trick, we have

$$\rho(\pi) = \mathbb{E}_{(s,a)\sim d^{\pi}}[r(s,a)] = \mathbb{E}_{(s,a)\sim d^{D}}\left[\frac{d^{\pi}(s,a)}{d^{\mathcal{D}}(s,a)} \cdot r(s,a) \right] \tag{11}$$

Now, we can see that if we can estimate the density ratio $w_{\pi/\mathcal{D}}(s,a) := \frac{d^{\pi}(s,a)}{d^{\mathcal{D}}(s,a)}$ without interacting with the environment, we solve the OPE problem.

To estimate the ratio is highly non-trival. Again, it is very hard to estimate the density if the dimension is high. More important, we cannot even access d^{π} due to the off-policy limitation. Here is where DICE kicks in.

6 DICE

6.1 Key Step 1: bridging the f-divergence and density ratio

By using the conjugate form of f-divergence, we have

$$-D_f\left(d^{\pi} \| d^{\mathcal{D}}\right) = \min_{x:\mathcal{S}\times A\to\mathbb{R}} \mathbb{E}_{(s,a)\sim d^{D}}\left[f_*(x(s,a))\right] - \mathbb{E}_{(s,a)\sim d^{\pi}}\left[x(s,a)\right] \tag{12}$$

By taking gradient w.r.t. x, we find the optimality conditions for x is:

$$f_*'(x(s,a)) = w_{\pi/\mathcal{D}}(s,a) := \frac{d^{\pi}(s,a)}{d^{\mathcal{D}}(s,a)} \tag{13}$$

Now, the road map is clearer. We optimize the conjugate function of f-divergence, then we use this function to estimate the ratio. However, to optimize the conjugate function form of f-divergence, we still need to access samples from π, which we cannot. We need to get rid of the second expectation $\mathbb{E}_{(s,a)\sim d^\pi}[x(s,a)]$.

6.2 Key step 2: construct a new MDP by new reward

The DICE use change of variables to solve $\mathbb{E}_{(s,a)\sim d^\pi}[x(s,a)]$. Recall we have

$$\rho(\pi) = \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)], \tag{14}$$

where $d^\pi(s,a) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \cdot \Pr[s_t = s, a_t = a | s_0 \sim \beta_0, \pi]$. Now, we see that the term $\mathbb{E}_{(s,a)\sim d^\pi}[x(s,a)]$ is actually estimate the value function of π, but change the reward function to $x(s,a)$ instead of using $r(s,a)$. In other words, we can use Q-function to estimate it by using x as a reward!

Thus, we define the Bellman operator of the new Q-function $\nu(s,a)$

$$\nu(s,a) = x(s,a) + \mathcal{B}_\pi \nu(s,a), \tag{15}$$

where

$$\mathcal{B}_\pi \nu(s,a) := \gamma \mathbb{E}_{s'\sim T(s,a), a'\sim\pi(s')}\left[\nu\left(s', a'\right)\right] \tag{16}$$

. Remember we are going to estimate $\nu(s_0,a)|a\sim\pi(.|s_0)$, we have

$$(1-\gamma)\cdot\mathbb{E}_{s_0\sim\beta_0, a_0\sim\pi(s_0)}\left[\nu\left(s_0,a_0\right)\right] = (1-\gamma)\cdot\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\cdot x\left(s_t,a_t\right)|\pi\right] \tag{17}$$

6.3 Key step 3: change of variables

Currently, we aim to solve:

$$\min_{x:S\times A\to\mathbb{R}} \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[f_*(x(s,a))\right] - \mathbb{E}_{(s,a)\sim d^\pi}[x(s,a)], \tag{18}$$

where we already knew that $f_*'\left(x^*(s,a)\right) = w_{\pi/\mathcal{D}}(s,a)$ and $\mathbb{E}_{(s,a)\sim d^\pi}[x(s,a)] = (1-\gamma)\cdot\mathbb{E}_{s_0\sim\beta_0,a_0\sim\pi(s_0)}\left[\nu\left(s_0,a_0\right)\right]$.

With change of variables $\nu(s,a) = x(s,a) + \mathcal{B}_\pi\nu(s,a)$, we have

$$\min_{\nu:S\times A\to\mathbb{R}} \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[f_*\left(\left(\nu - \mathcal{B}_\pi\nu\right)(s,a)\right)\right] - (1-\gamma)\cdot\mathbb{E}_{s_0\sim\beta_0,a_0\sim\pi(s_0)}\left[\nu\left(s_0,a_0\right)\right]. \tag{19}$$

Now, the optimality condition becomes to

$$f_*'\left(\nu^*(s,a) - \mathcal{B}_\pi\nu^*(s,a)\right) = w_{\pi/\mathcal{D}}(s,a) \tag{20}$$

In DualDICE paper, the f is chosen to be $\frac{1}{2}x^2$, which violates the f-divergence definition. If we ignore this point, by setting $f(x) = \frac{1}{2}x^2$, we have $f^* = f$. Now, the objective becomes to

$$\min_{\nu:S\times A\to\mathbb{R}} \frac{1}{2}\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[\left(\nu(s,a) - \mathcal{B}_\pi\nu(s,a)\right)^2\right] - (1-\gamma)\mathbb{E}_{s_0\sim\beta_0,a_0\sim\pi(s_0)}\left[\nu\left(s_0,a_0\right)\right] \tag{21}$$

6.4 Step 4: another conjugate to reduce the effect of Bellman operator

Directly applying the definition of convex conjugate, we have

$$\min_{\nu:S\times A\to\mathbb{R}}\max_{\zeta:A\to\mathbb{R}} J(\nu,\zeta) := \mathbb{E}_{(s,a,s')\sim d^{\mathcal{D}},a'\sim\pi(s')}\left[\left(\nu(s,a)-\gamma\nu\left(s',a'\right)\right)\zeta(s,a)-f^*(\zeta(s,a))\right] \\ -(1-\gamma)\mathbb{E}_{s_0\sim\beta,a_0\sim\pi(s_0)}\left[\nu\left(s_0,a_0\right)\right], \tag{22}$$

where ζ is the variable of the conjugate function.

Before we have:

$$f'_*\left(\nu^*(s,a)-\mathcal{B}_\pi\nu^*(s,a)\right)=w_{\pi/\mathcal{D}}(s,a) \tag{23}$$

Now we have:

$$\begin{aligned} f'\left(\zeta^*(s,a)\right) &= \nu^*(s,a)-\mathcal{B}_\pi\nu^*(s,a) \\ \Rightarrow \zeta^*(s,a) &= f'_*\left(\nu^*(s,a)-\mathcal{B}_\pi\nu^*(s,a)\right)=w_{\pi/\mathcal{D}}(s,a) \end{aligned} \tag{24}$$

Notice this step is only a trick to improve the estimation. The only difference is that the expectation term is now linear instead of quadratic. In following work of DICE (ValueDICE), this step is ignored.

7 Off-policy RL by DICE

The RL objective is to maximize the long term reward:

$$\max_\pi \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)] \tag{25}$$

In order to use DICE trick to make it off-policy, we add another term:

$$\max_\pi \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)]-\alpha D_f\left(d^\pi\|d^{\mathcal{D}}\right) \tag{26}$$

Again, apply the dual form of f-divergence:

$$\begin{aligned} &\max_\pi \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)]-\alpha D_f\left(d^\pi\|d^{\mathcal{D}}\right) \\ &= \max_\pi \min_x \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)]-\alpha\cdot\mathbb{E}_{(s,a)\sim d^\pi}[x(s,a)]+\alpha\cdot\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[f_*(x(s,a))\right] \\ &= \max_\pi \min_x \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)-\alpha\cdot x(s,a)]+\alpha\cdot\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[f_*(x(s,a))\right] \end{aligned} \tag{27}$$

Again, we could use the change of variable by constructing a new MDP. Let $\nu(s,a) := -\alpha\cdot x(s,a)+\mathcal{B}_\pi\nu(s,a)$, where

$$\mathcal{B}_\pi\nu(s,a) := r(s,a)+\gamma\mathbb{E}_{s'\sim T(s,a),a'\sim\pi(s')}\left[\nu\left(s',a'\right)\right] \tag{28}$$

. Now, objective becomes

$$\max_\pi\min_\nu(1-\gamma)\cdot\mathbb{E}_{s_0\sim\beta_0,a_0\sim\pi(s_0)}\left[\nu\left(s_0,a_0\right)\right]+\alpha\cdot\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[f_*\left(\frac{1}{\alpha}\left(\mathcal{B}_\pi\nu(s,a)-\nu(s,a)\right)\right)\right] \tag{29}$$

Again, this is completely off-policy formulation. You may want to use the same trick (step 4) to reduce the effect of Bellman operator.

Now we could analyze the minimax optimization

$$\max_\pi\min_\nu J(\pi,\nu) := (1-\gamma)\cdot\mathbb{E}_{s_0\sim\beta_0,a_0\sim\pi(s_0)}\left[\nu\left(s_0,a_0\right)\right]+\alpha\cdot\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[f_*\left(\frac{1}{\alpha}\left(\mathcal{B}_\pi\nu(s,a)-\nu(s,a)\right)\right)\right]$$

To have a closer look of it, let us assume we already find the optimal nu function, recall we have $f'_*\left(\frac{1}{\alpha}\left(\mathcal{B}_\pi\nu^*_\pi(s,a)-\nu^*_\pi(s,a)\right)\right)=f'_*\left(x^*_\pi(s,a)\right)=\frac{d^\pi(s,a)}{d^{\mathcal{D}}(s,a)}$, if we take gradient to π, we can see that

$$
\begin{aligned}
&\frac{\partial}{\partial \pi} J\left(\pi, \nu_{\pi}^{*}\right) \\
&=(1-\gamma) \cdot \frac{\partial}{\partial \pi} \mathbb{E}_{s_0 \sim \beta_0, a_0 \sim \pi\left(s_0\right)}\left[\nu_{\pi}^{*}\left(s_0, a_0\right)\right]+\alpha \cdot \mathbb{E}_{(s, a) \sim d^D}\left(\left[\frac{d^{\pi}(s, a)}{d^D(s, a)} \frac{\partial}{\partial \pi}\left(\frac{1}{\alpha}\left(\mathcal{B}_{\pi} \nu_{\pi}^{*}(s, a)-\nu_{\pi}^{*}(s, a)\right)\right)\right]\right. \\
&=(1-\gamma) \cdot \frac{\partial}{\partial \pi} \mathbb{E}_{s_0 \sim \beta_0, a_0 \sim \pi\left(s_0\right)}\left[\nu_{\pi}^{*}\left(s_0, a_0\right)\right]+\gamma \cdot \mathbb{E}_{(s, a) \sim d^{\pi}}\left[\frac{\partial}{\partial \pi} \mathbb{E}_{s^{\prime} \sim T(s, a), a^{\prime} \pi\left(s^{\prime}\right)}\left[\nu_{\pi}^{*}\left(s^{\prime}, a^{\prime}\right)\right]\right] \\
&=\mathbb{E}_{(s, a) \sim d^{\pi}}[\nu(s, a) \nabla \log \pi(s, a)]
\end{aligned}
$$

We can see that this is actually reward shaping and we are doing on-policy optimization on new reward.

8 ValueDice

The ValueDICE use DICE trick to do imitation learning. We knew that adversarial imitation learning tries to estimate $D_{KL}(d^{\pi}||d^{exp})$, in ValueDice paper, they use the Donsker-Varadhan representation:

$$
-D_{\mathrm{KL}}\left(d^{\pi} \| d^{\exp }\right)=\min _{x: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}} \log \mathbb{E}_{(s, a) \sim d^{\exp }}\left[e^{x(s, a)}\right]-\mathbb{E}_{(s, a) \sim d^{\pi}}[x(s, a)] \tag{30}
$$

Similarly, we have

$$
x^{*}(s, a)=\log \frac{d^{\pi}(s, a)}{d^{\exp }(s, a)}+C \tag{31}
$$

By change of variable, we have:

$$
x(s, a)=\nu(s, a)-\mathcal{B}^{\pi} \nu(s, a) \tag{32}
$$

Now, the KL divergence reduces to:

$$
-D_{\mathrm{KL}}\left(d^{\pi} \| d^{\exp }\right)=\min _{\nu: s \times \mathcal{A} \rightarrow \mathbb{R}} \log \mathbb{E}_{(s, a) \sim d^{\exp }}\left[e^{\nu(s, a)-\mathcal{B}^{\pi} \nu(s, a)}\right]-\mathop{\mathbb{E}}_{(s, a) \sim d^{\pi}}\left[\nu(s, a)-\mathcal{B}^{\pi} \nu(s, a)\right] \tag{33}
$$

The second term again can be express by a new 'Q function'

$$
\min _{\nu: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}} J_{\mathrm{DICE}}(\nu):=\log_{(s, a) \sim d^{exp}}\left[e^{\nu(s, a)-\mathcal{B}^{\pi} \nu(s, a)}\right]-(1-\gamma) \cdot \mathop{\mathbb{E}}_{s_0 \sim p_0 a_0 \sim \pi(. | s_0)}\left[\nu\left(s_0, a_0\right)\right] \tag{34}
$$

Everything is the same as the DICE except they use the DV representation of KL.

Now, we can have the objective of ValueDICE.

$$
\max _{\pi} \min _{\nu: S \times \mathcal{A} \rightarrow \mathbb{R}} J_{\mathrm{DICE}}(\pi, \nu):=\log_{(s, a) \sim d^{\exp }}\left[e^{\nu(s, a)-B^{*} \nu(s, a)}\right]-(1-\gamma) \cdot \mathop{\mathbb{E}}_{s_0 \sim p_0(\cdot)}\left[\nu\left(s_0, a_0\right)\right] \tag{35}
$$

References

[1] Nowozin, S., Cseke, B., & Tomioka, R. (2016). f-gan: Training generative neural samplers using variational divergence minimization. In Advances in neural information processing systems (pp. 271-279).