2020 10 19 AlgaeDICE: Policy Gradient from Arbitrary Experience

Chen Gong

19 October 2020

1 Conference Name

33rd Conference on Neural Information Processing Systems (NeurIPS 2019)

2 Thesis statement

2.1 大致思想描述

传统强化学习中，和环境的交互往往受限于较大的消耗和可行性问题。传统 RL 算法中，目标函数中包含求 On-Policy 样本的期望，求解非常的困难。本文中，作者介绍了一种新的表示 max-return 优化的方法，让其可以表达为关于任意分布和 Off-Policy 数据分布的期望形式。实际上，此篇文章的思想和 DualDice 有点类似。

3 研究背景

默认阅读此文章的同学有较好的 RL 基础，简单的问题不再过多的陈述。RL 中，我们希望学习到一个 return-maximizing 的策略：

$$\max_{\pi} J_{\mathrm{P}}(\pi) := (1-\gamma)\mathbb{E}_{s_0\sim\mu_0, a_0\sim\pi(s_0)}\left[Q_\pi\left(s_0, a_0\right)\right] \tag{1}$$

其中，Q 函数的定义为：

$$Q_\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r\left(s_t, a_t\right) \mid s_0 = s, a_0 = a, s_t \sim T\left(s_{t-1}, a_{t-1}\right), a_t \sim \pi\left(s_t\right) \text{ for } t \geq 1\right] \tag{2}$$

而此目标函数可以根据策略的标准化状态访问分布而等价的写为其对偶形式：

$$\max_{\pi} J_D(\pi) := \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)], \tag{3}$$

其中，

$$d^\pi(s,a) = (1-\gamma)\cdot\sum_{t=0}^{\infty}\gamma^t \mathrm{Pr}(s_t = s, a_t = a \mid s_0 \sim \mu_0, \forall t \geq 0, a_t \sim \pi\left(s_t\right), s_{t+1} \sim T\left(s_t, a_t\right)) \tag{4}$$

而优化 π 需要知道策略梯度，根据策略梯度理论，可以得到 $J_P(\pi)$ 的梯度为：

$$\frac{\partial}{\partial \pi} J_{\mathrm{P}}(\pi)=\mathbb{E}_{(s,a)\sim d^{\pi}}\left[Q_{\pi}(s,a)\nabla \log \pi(a \mid s)\right] \tag{5}$$

详细推导可见 CS 598："Notes on Importance Sampling and Policy Gradient"。显然，为了估计这个梯度，我们需要估计 Q 值函数 $Q(s,a)$，需要从 d^{π} 中进行采样。这就要求我们首先需要和环境进行交互，但是和环境交互可能是非常困难的，其次要求我们估算 Q 函数。于是就诞生了一系列的 AC 算法，不断地交替更新 π（the actor）和 Q 近似函数 Q_{θ}（the critic）。critic 的更新方式根据：

$$Q_{\pi}(s,a)=\mathcal{B}_{\pi}Q_{\pi}(s,a):=r(s,a)+\gamma\cdot\mathbb{E}_{s'\sim T(s,a),a'\sim\pi(s')}\left[Q_{\pi}\left(s',a'\right)\right] \tag{6}$$

其中，$\mathcal{B}_{\pi}$ 为关于 π 的期望 Bellman 算子。其中，目标函数为：

$$\min_{Q_{\theta}} J_{\mathrm{critic}}(Q_{\theta}):=\frac{1}{2}\mathbb{E}_{(s,a)\sim\beta}[(\mathcal{B}_{\pi}Q_{\theta}-Q_{\theta})(s,a)^{2}], \tag{7}$$

其中，β 为某些分布。尽管 critic 可以使用任意的 β 来进行 Off-Policy learning，但为了实现较好的性能，AC 算法通常保证 replay buffer 中有较多的新数据，这里实际上就是为了保证采样策略和目标策略之间不要相差的太大。理论工作表明，β 最好的选择就是 d^{π}，这是算法将是 full On-Policy 的。

本文中,将只注重于 Off-Policy 设定。给定数据集 $\mathcal{D}=\{(s_k,a_k,r_k,s'_k)\}_{k=1}^{N}$,其中 $r_k=r(s_k,a_k)$; $s'_k\sim T(s_k.a_k)$，a_k 的样本采集过程未知。其中，令 $d^{\mathcal{D}}$ 表示未知的 state-action 分布，并且初始分布集合为 $\mathcal{U}=\{s_{0,k}\}_{k=1}^{N}$，其中 $s_{0,k}\sim\mu_0$。

4 Methods

文章中首先给出 AlgaeDICE 的非正式推导，作为公式 (3) 中对偶目标函数的正则化。之后，文章将更正式地给出我们算法，是通过最大收益目标线性规划公式的拉格朗日对偶推导出来的。

4.1 AlgaeDICE via Density Regularizaition

4.1.1 A Regularized Off-Policy Max-Return Objective

Max-return 目标函数 (3) 中只含有 On-Policy 分布 d^{π}，为了将 Off-Policy 的分布 $d^{\mathcal{D}}$ 合并到目标函数里，这里合并的正则化项。

$$\max_{\pi} J_{\mathrm{D},f}(\pi):=\mathbb{E}_{(s,a)\sim d^{\pi}}[r(s,a)]-\alpha D_{f}\left(d^{\pi}\|d^{\mathcal{D}}\right) \tag{8}$$

其中，$\alpha>0$ 且 D_f 为凸函数 f 的 f 散度：

$$D_{f}\left(d^{\pi}\|d^{\mathcal{D}}\right)=\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[f(\omega_{\pi/\mathcal{D}}(s,a))] \tag{9}$$

其中，$\omega_{\pi/\mathcal{D}}(s,a):=\frac{d^{\pi}(s,a)}{d^{\mathcal{D}}(s,a)}$。此正则化项是鼓励目标策略靠近行为策略，实际训练偏保守。我读到这儿的时候，思考过为什么要加这个正则化项。作者强调这个正则化项的引入是为了使后续的推导，而不是对最优策略施加强约束。此问题先保留着。实际上，α 和 f 的选择，可以帮助我们控制控制 "探索" 和 "开发"，这个在之后的实验中有体现的。

那么问题来了，$D_f\left(d^\pi \| d^{\mathcal{D}}\right)$ 怎么计算？这里采用了 f 函数的对偶形式，令对偶函数为：$x : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$：

$$\begin{aligned}
&\max_\pi \min_{x:S\times A\to\mathbb{R}} J_{\mathrm{D},f}(\pi, x) \\
&\qquad := \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)] + \alpha \cdot \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[f_*(x(s,a))\right] - \alpha \cdot \mathbb{E}_{(s,a)\sim d^\pi}[x(s,a)] \\
&\qquad = \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a) - \alpha \cdot x(s,a)] + \alpha \cdot \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[f_*(x(s,a))\right],
\end{aligned} \tag{10}$$

这里直接使用的是 f 散度对偶形式。对于目标函数 (10)，我们无法从 d^π 中进行采样，所以想想办法将 d^π 这项消掉。类似于 DualDICE 中的思想，做变量转换，定义 $\nu : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ 为 Bellman 方程的不动点，

$$\nu(s,a) := -\alpha \cdot x(s,a) + \mathcal{B}_\pi \nu(s,a) \tag{11}$$

同样的，$x(s,a) = \frac{1}{\alpha}(\mathcal{B}_\pi \nu - \nu)(s,a)$。并且，在 x 和 r 有界的情况下，ν 通常是存在且有界的。那么有，

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)] - \alpha \mathbb{E}_{(s,a)\sim d^\pi}[x(s,a)] \\
&\qquad = \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)] + \mathbb{E}_{(s,a)\sim d^\pi}\left[\nu(s,a) - r(s,a) - \gamma \mathbb{E}_{s'\sim T(s,a), a'\sim\pi(s')}\left[\nu\left(s', a'\right)\right]\right] \\
&\qquad = \mathbb{E}_{(s,a)\sim d^\pi}\left[\nu(s,a) - \gamma \mathbb{E}_{s'\sim T(s,a), a'\sim\pi(s')}\left[\nu\left(s', a'\right)\right]\right] \\
&\qquad = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{s\sim\beta_t, a\sim\pi(s)}[\nu(s,a)] - (1-\gamma)\sum_{t=0}^{\infty} \gamma^{t+1} \mathbb{E}_{s'\sim\beta_{t+1}, a'\sim\pi(s')}\left[\nu\left(s', a'\right)\right] \\
&\qquad = (1-\gamma)\mathbb{E}_{\mu_0\pi}\left[\nu\left(s_0, a_0\right)\right]
\end{aligned} \tag{12}$$

所以，得到，

$$\begin{aligned}
&\max_\pi \min_{\nu:S\times A\to\mathbb{R}} J_{\mathrm{D},f}(\pi, \nu) \\
&\qquad := (1-\gamma)\mathbb{E}_{a_0\sim\pi(s_0)0\ \sim\ \mu_0}\left[\nu\left(s_0, a_0\right)\right] + \alpha \cdot \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[f_*\left(\left(\mathcal{B}_\pi \nu - \nu\right)(s,a)/\alpha\right)\right]
\end{aligned} \tag{13}$$

那么，显然目标函数是 full Off-Policy 的。于是我们可以得到第一个理论，

Theorem 1 (Primal AlgaeDICE) *Under mild conditions on $d^{\mathcal{D}}, \alpha, f$, the regularized max-return objective may be expressed as a max-min optimization:*

$$\max_\pi\ \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)] - \alpha D_f(d^\pi \| d^{\mathcal{D}}) \equiv \max_\pi \min_{\nu:S\times A\to\mathbb{R}} J_{\mathrm{D},f}(\pi, \nu)\,. \tag{13}$$

Remark (Fenchel AlgaeDICE) f_* 中含有 $\mathcal{B}_\pi$ 项，此项的计算非常的困难，因为其中包含对 T 的期望，当动作空间很大的时候，这一项的求解非常难。**这里小编是这样理解的，首先观察公式 (13) 是先采样得到 (s,a) 再求解 $\mathcal{B}$，相当于求期望的期望，而我们采样得到了一个 (s,a) 并求不出 $\mathcal{B}$，这个问题也被称为 "Double Sample" 问题**。本文中使用对偶方法来表示 f_*，其中，

$$f_*\left(\left(\mathcal{B}_\pi \nu - \nu\right)(s,a)/\alpha\right) = \max_{\zeta\in\mathbb{R}} \frac{1}{\alpha}\left(\mathcal{B}_\pi \nu - \nu\right)(s,a) \cdot \zeta - f(\zeta) \tag{14}$$

那么，最终我们的目标函数为：

$$\begin{aligned}
&\max_\pi \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)] - \alpha D_f\left(d^\pi \| d^{\mathcal{D}}\right) = \max_\pi \min_{\nu:S\times A\to\mathbb{R}} \max_{\zeta:S\times A\to\mathbb{R}} (1-\gamma)\mathbb{E}_{s_0\sim\mu_0, a_0\sim\pi(s_0)}\left[\nu\left(s_0, a_0\right)\right] + \\
&\qquad \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}, s'\sim T(s,a), a'\sim\pi(s')}\left[\left(\gamma\nu\left(s', a'\right) + r(s,a) - \nu(s,a)\right) \cdot \zeta(s,a) - \alpha \cdot f(\zeta(s,a))\right]
\end{aligned} \tag{15}$$

实际上，在 mild 条件下，公式 (15) 中的 min-max 内积是强对偶问题，因此我们可以交换 $\min_\nu$ 和 $\max_\zeta$ 来将其转换为 max-max-min 的形式。

4.1.2 Consistent Policy Gradient using Off-Policy Data

此部分证明的是，公式 (10): $J_{\mathrm{D,f}}(\pi, x)$ 中和 On-Policy 形式的公式 (13): $J_{\mathrm{D,f}}(\pi, \nu)$ 中关于 π 的梯度是一样的。公式 (10) 中对 x 求偏导，并令其等于 0，可得，

$$f_*'\left(x_\pi^*(s,a)\right) = w_{\pi/\mathcal{D}}(s,a) \tag{16}$$

根据公式 (11) 可推导出，

$$f_*'\left(\left(\mathcal{B}_\pi \nu_\pi^* - \nu_\pi^*\right)(s,a)/\alpha\right) = w_{\pi/\mathcal{D}}(s,a) \tag{17}$$

函数 $J_{\mathrm{D,f}}(\pi, \nu_\pi^*)$ 关于 π 的梯度可以表达为：

$$\begin{aligned}
\frac{\partial}{\partial \pi} J_{\mathrm{D},f}\left(\pi, \nu_\pi^*\right) &= (1-\gamma)\frac{\partial}{\partial \pi}\mathbb{E}_{a_0\sim\pi(s_0)}\left[\nu_\pi^*\left(s_0,a_0\right)\right] + \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[\frac{\partial}{\partial \pi} f_*\left(\left(\mathcal{B}_\pi \nu_\pi^* - \nu_\pi^*\right)(s,a)/\alpha\right)\right] \\
&= (1-\gamma)\frac{\partial}{\partial \pi}\mathbb{E}_{a_0\sim\pi(s_0)}\left[\nu_\pi^*\left(s_0,a_0\right)\right] + \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[w_{\pi/\mathcal{D}}(s,a)\frac{\partial}{\partial \pi}\left(\mathcal{B}_\pi \nu_\pi^* - \nu_\pi^*\right)(s,a)\right] \\
&= (1-\gamma)\frac{\partial}{\partial \pi}\mathbb{E}_{a_0\sim\pi(s_0),s_0\sim\mu_0}\left[\nu_\pi^*\left(s_0,a_0\right)\right] + \gamma\cdot\mathbb{E}_{(s,a)\sim d^{\pi}}\left[\frac{\partial}{\partial \pi}\mathbb{E}_{a'\sim\pi(s')}\left[\nu_\pi^*\left(s',a'\right)\right]\right] \\
&= \mathbb{E}_{(s,a)\sim d^{\pi}}\left[\nu_\pi^*(s,a)\nabla\log\pi(a\mid s)\right],
\end{aligned} \tag{18}$$

倒数第二行到最后一行的推导，使用了 Danskin's theorem，所以，可以证明，对偶函数 ν 到达最优时，off-policy 的目标函数 $J_{\mathrm{D,f}}(\nu, \pi)$ 的梯和 On-Policy 策略梯度是一样的。下面的想法则是用 Q 函数来表达 $\nu_\pi^*(s,a)$，根据 Bellman 方程 $Q(s,a) = r(s,a) + \mathcal{B}_\pi Q(s,a)$，将奖励调整为：$\tilde{r}(s,a) := r(s,a) - \alpha\cdot x_\pi^*(s,a)$。根据公式 (16) 中的表达式，且 f 和 f_* 的梯度是互为逆的，于是有

$$\begin{aligned}
\tilde{r}(s,a) :=& r(s,a) - \alpha\cdot x_\pi^*(s,a) \\
=& r(s,a) - \alpha\cdot f_*^{'(-1)}(w_{\pi/\mathcal{D}}(s,a)) \\
=& r(s,a) - \alpha\cdot f'\left(w_{\pi/\mathcal{D}}(s,a)\right)
\end{aligned} \tag{19}$$

于是得到了文章中的 Theorem 2，

Theorem 2 *If the dual function ν is sufficiently optimized, the gradient of the off-policy objective $J_{\mathrm{D},f}(\pi,\nu)$ with respect to π is exactly the (regularized) on-policy policy gradient:*

$$\frac{\partial}{\partial \pi}\min_\nu J_{\mathrm{D},f}(\pi,\nu) = \mathbb{E}_{(s,a)\sim d^{\pi}}\left[\tilde{Q}^{\pi}(s,a)\nabla\log\pi(a|s)\right], \tag{18}$$

where, $\tilde{Q}^{\pi}(s,a)$ is the Q-value function of π with respect to rewards $\tilde{r}(s,a) := r(s,a) - \alpha\cdot f'(w_{\pi/\mathcal{D}}(s,a))$.

并且，即使我们推出的是公式 (13) 中的策略梯度，而其在公式 (15) 中也是可以成立的。因为，ζ_π^* 等于 $\omega_{\pi/\mathcal{D}}(s,a)$，和 π 无关。

4.1.3 Connection to Actor-Critic

考虑 $f(x)=\frac{1}{2}x^2$，且 $f^*(x)=\frac{1}{2}x^2$。目标函数为：

$$\max_{\pi}\min_{\nu:S\times A\to\mathbb{R}} J_{\mathrm{D},f}(\pi,\nu):=(1-\gamma)\mathbb{E}_{s_0\sim\mu_0,a_0\sim\pi(s_0)}\left[\nu\left(s_0,a_0\right)\right]+\frac{1}{2\alpha}\cdot\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[\left(\left(\mathcal{B}_\pi\nu-\nu\right)(s,a)\right)^2\right] \quad (20)$$

其中，第二项就是 Actor-Critic 算法中的 Critic 的目标函数。有强化学习基础的同学应该了解，而实际 AC 算法中，通常不能使用 Off-Policy 目标函数，因为 d^π 和 $d^{\mathcal{D}}$ 之间的不匹配的问题。相比之下，我们的推导通过引入目标函数的第一项，将 Off-Policy 的 AC 算法转换为 On-Policy 的 AC 算法，而无需使用重要权重。此外，标准 AC 算法有两个独立的价值目标和策略目标，但本文提出的目标是一个单一、统一的目标。策略函数和价值函数都是针对算法相同的 off-policy 目标函数进行训练。

4.2 A Lagrangian View of AlgaeDICE

本节中将从 Q 函数的线性动态规划的 Lagrangian 表达式来推导 AlgaeDICE。首先需要一些假设：

Assumption 1 (Bounded rewards) *The rewards of the MDP are bounded by some finite constant* $R_{\max}$*:* $\|r\|_\infty\le R_{\max}$.

For the next assumption, we introduce the *transpose Bellman operator*:

$$\mathcal{B}_\pi^\top\rho\left(s',a'\right):=\gamma\sum_{s,a}\pi\left(a'|s'\right)T\left(s'|s,a\right)\rho\left(s,a\right)+\left(1-\gamma\right)\mu_0\left(s'\right)\pi\left(a'|s'\right). \quad (19)$$

Assumption 2 (MDP regularity) *The transposed Bellman operator* $\mathcal{B}_\pi^\top$ *has a unique fixed point solution.*[2]

Assumption 3 (Bounded ratio) *The target density ratio is bounded by some finite constant* $W_{\max}$*:* $\left\|w_{\pi/\mathcal{D}}\right\|_\infty\le W_{\max}$.

Assumption 4 (Characterization of f) *The function* f *is convex with domain* $\mathbb{R}$ *and continuous derivative* f'*. The convex (Fenchel) conjugate of* f *is* f_**, and* f_* *is closed and strictly convex. The derivative* f_*' *is continuous and its range is a superset of* $[0,W_{\max}]$.

为了方便表示，令 $\nu\in\mathcal{N}:=[-\frac{C}{1-\gamma},\frac{C}{1-\gamma}]$。其中，$C=R_{\max}+|\alpha|\cdot f'(W_{\max})$。推导从有线性规划特征的 Q 函数和其对偶形式开始：

Theorem 3 *Given a policy* π*, the average return of* π *may be expressed in primal and dual forms as*

$$\begin{aligned}\min_{\nu:S\times A\to\mathbb{R}}\ & J_{\mathrm{P}}(\pi,\nu):=(1-\gamma)\,\mathbb{E}_{\mu_0\pi}\left[\nu\left(s_0,a_0\right)\right]\\ \text{s.t.}\ & \nu\left(s,a\right)\ge\mathcal{B}_\pi\nu(s,a), \quad (19)\\ & \forall\left(s,a\right)\in S\times A,\end{aligned}$$

and,

$$\begin{aligned}\max_{\rho:S\times A\to\mathbb{R}_+}\ & J_{\mathrm{D}}(\pi,\rho):=\mathbb{E}_\rho\left[r\left(s,a\right)\right]\\ \text{s.t.}\ & \rho\left(s,a\right)=\mathcal{B}_\pi^T\rho(s,a), \quad (20)\\ & \forall\left(s,a\right)\in S\times A,\end{aligned}$$

respectively, where $\mathcal{B}_\pi^T\rho(s,a):=(1-\gamma)\mu_0(s)\pi(a|s)+\gamma\sum_{\tilde{s},\tilde{a}}\pi(\tilde{a}|\tilde{s})T(s|\tilde{s},\tilde{a})\rho(\tilde{s},\tilde{a})$ *is the* transpose Bellman operator. *The optimal primal* ν_π^* *is* Q_π *and the optimal dual* ρ_π^* *is* d^π.

Proof: 首先，Bellman 算子 $\mathcal{B}_\pi$ 是单调的，$\nu_1 \geq \nu_2 \longrightarrow \mathcal{B}_\pi \nu_1 \geq \mathcal{B}_\pi \nu_2$。那么，对于任意的 ν，有 $\nu \geq (\mathcal{B}_\pi)\nu \geq (\mathcal{B}_\pi)^2 \nu \geq (\mathcal{B}_\pi)^3 \nu \geq \ldots \geq (\mathcal{B}_\pi)^\infty \nu = Q_\pi$。其中，

$$\begin{aligned} &\max_{\rho: S\times A\to\mathbb{R}_+} \mathbb{E}_\rho[R(s,a)], \\ &\quad \text{s.t. } \rho(s',a') - \gamma \sum_{s,a} \pi(a' \mid s') T(s' \mid s,a) \rho(s,a) = (1-\gamma)\mu_0(s')\pi(a' \mid s'), \\ &\quad \forall (s',a') \in S \times A. \end{aligned} \tag{21}$$

将等式约束进行调整可以得到，

$$\rho^* = (1-\gamma)\left(I - \gamma (P_\pi)^\top\right)^{-1} (\mu\pi) \tag{22}$$

其中，$P_\pi(s',a'|s,s) = \pi(a'|s')T(s'|s,a)$。利用等比数列求和公式，$\left(I - \gamma (P_\pi)^\top\right)^{-1} = \sum_{t=0}^{\infty} \gamma^t (P_\pi)^t$，那么可以得到，

$$\rho^* = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t (P_\pi)^t (\mu\pi) = d^\pi$$

□

那么，可以重新表达对偶变量 $\zeta(s,a) = \frac{\rho(s,a)}{d^{\mathcal{D}}(s,a)}$，可得

$$\min_{\nu: S\times A\to\mathbb{R}} \max_{\zeta: S\times A\to\mathbb{R}_+} (1-\gamma)\mathbb{E}_{s_0\sim\mu_0, a_0\sim\pi(s_0)}[\nu(s_0,a_0)] + \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[\zeta(s,a)(\mathcal{B}_\pi\nu - \nu)(s,a)] \tag{23}$$

其中，最优值 ζ_π^* 对应的是 $\omega_{\pi/\mathcal{D}}$。在实践中，(23) 中的线性结构可以引起数值的不稳定性。因此，受增广拉格朗日方法的启发，引入正则化。

$$\begin{aligned} \min_\nu \max_\zeta L(\nu,\zeta;\pi) :=& (1-\gamma)\mathbb{E}_{s_0\sim\mu_0, a_0\sim\pi(s_0)}[\nu(s_0,a_0)] + \\ & \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[\zeta(s,a)(\mathcal{B}_\pi\nu - \nu)(s,a)] - \alpha \cdot \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[f(\zeta(s,a))] \end{aligned} \tag{24}$$

有趣的是，正则化项会影响原问题的最优解 ν_π^*，但是不会影响对偶问题的最优解 ζ_π^*。

Theorem 4 *Under Assumptions 1–4, the solution to (23) is given by,*

$$\begin{aligned} \nu_\pi^*(s,a) &= -\alpha f'\left(w_{\pi/\mathcal{D}}(s,a)\right) + \mathcal{B}_\pi \nu_\pi^*(s,a), \\ \zeta_\pi^*(s,a) &= w_{\pi/\mathcal{D}}(s,a). \end{aligned}$$

The optimal value is $L(\nu_\pi^*, \zeta_\pi^*; \pi) = \mathbb{E}_{d^\pi}[r(s,a)] - \alpha D_f(d^\pi \| d^{\mathcal{D}})$.

Proof: 利用 Fenchel 对偶得，

$$\begin{aligned} &\max_{\zeta: \mathcal{S}\times\mathcal{A}\to\mathbb{R}} \mathbb{E}_{d^{\mathcal{D}}}[\zeta(s,a)(\mathcal{B}_\pi\nu - \nu)(s,a)] - \alpha\mathbb{E}_{d^{\mathcal{D}}}[f(\zeta(s,a))] \\ &\quad = \alpha\mathbb{E}_{d^{\mathcal{D}}}\left[f_*\left(\frac{1}{\alpha}(\mathcal{B}_\pi\nu - \nu)(s,a)\right)\right] \end{aligned} \tag{25}$$

融入到公式 (24) 中可得，

$$L(\nu,\zeta_\pi^*;\pi) = \min_{\nu: \mathcal{S}\times\mathcal{A}\to\mathcal{N}} (1-\gamma)\mathbb{E}_{\mu_0\pi}[\nu(s_0,a_0)] + \alpha\mathbb{E}_{d^{\mathcal{D}}}\left[f_*\left(\frac{1}{\alpha}(\mathcal{B}_\pi\nu - \nu)(s,a)\right)\right] \tag{26}$$

前面详细的推导过这个变量替换，这里不做过多的解释，有

$$\begin{aligned} L\left(\nu_{\pi}^{*}, \zeta_{\pi}^{*}; \pi\right) &= \min_{x \in \mathcal{C}} \mathbb{E}_{(s,a)\sim d^{\pi}}[r(s,a)] - \alpha \mathbb{E}_{(s,a)\sim d^{\pi}}[x(s,a)] + \alpha \mathbb{E}_{d^{\mathcal{D}}}\left[f_{*}(x(s,a))\right] \\ &= \mathbb{E}_{(s,a)\sim d^{\pi}}[r(s,a)] - \alpha\left(\max_{x \in \mathcal{C}} \mathbb{E}_{d^{\pi}}[x(s,a)] - \mathbb{E}_{d^{\mathcal{D}}}\left[f_{*}(x(s,a))\right]\right) \end{aligned} \tag{27}$$

根据公式 (10) 中类似的推导，可以得到，

$$L\left(\nu_{\pi}^{*}, \zeta_{\pi}^{*}; \pi\right) = \mathbb{E}_{(s,a)\sim d^{\pi}}[r(s,a)] - \alpha D_{f}\left(d^{\pi} \| d^{\mathcal{D}}\right) \tag{28}$$

为了表达 ν_{π}^{*}，

$$x^{*}(s,a) = f'\left(w_{\pi/\mathcal{D}}(s,a)\right) \Rightarrow \nu_{\pi}^{*}(s,a) = \mathcal{B}_{\pi}\nu_{\pi}^{*}(s,a) - \alpha f'\left(w_{\pi/\mathcal{D}}(s,a)\right) \tag{29}$$

然后，根据 Fenchel 共轭的定义有，

$$\zeta_{\pi}^{*}(s,a) = \operatorname{argmax}_{\zeta}\, \zeta \cdot x_{\pi}^{*}(s,a) - f(\zeta) = f_{*}'\left(x_{\pi}^{*}(s,a)\right) = w_{\pi/\mathcal{D}}(s,a) \tag{30}$$

其中，第二个等号用到 $f'\left(\zeta_{\pi}^{*}(s,a)\right) = x_{\pi}^{*}(s,a) \Rightarrow \zeta_{\pi}^{*}(s,a) = f_{*}'\left(x_{\pi}^{*}(s,a)\right)$。 □

那么，这样可以得到和公式 (15) 中一样的结果了。从 LP 的角度我们得到了相同的推导，可以用来探索强对偶性。在假设中，$\omega_{\pi/\mathcal{D}}$ 和 r 都是有界的，并且 $(\nu_{\pi}^{*}, \zeta_{\pi}^{*}) \in \mathcal{H} \times \mathcal{F}$。所以，此对偶是有强对偶，有，

$$\min_{\nu \in \mathcal{H}} \max_{\zeta \in \mathcal{F}} L(\nu, \zeta; \pi) = \max_{\zeta \in \mathcal{F}} \min_{\nu \in \mathcal{H}} L(\nu, \zeta; \pi). \tag{31}$$

为了提高计算效率，优化策略为，

$$\begin{aligned} \max_{\pi} \ell(\pi) := \max_{\zeta \in \mathcal{F}} \min_{\nu \in \mathcal{H}} (1-\gamma) \mathbb{E}_{(s_0,a_0)\sim \mu_0 \pi}\left[\nu\left(s_0, a_0\right)\right] + \\ \mathbb{E}_{d^{\mathcal{D}}}\left[\zeta(s,a)\left(\mathcal{B}_{\pi}\nu - \nu\right)(s,a)\right] - \alpha \cdot \mathbb{E}_{d^{\mathcal{D}}}[f(\zeta(s,a))]. \end{aligned} \tag{32}$$

4.3 Extensions to $\gamma = 1$ or $\alpha = 0$

尽管 AlgaeDICE 是基于 $\gamma \in [0, 1)$ 和 $\alpha > 0$ 推导出来的。从 LP 中关于 Q_{π} 的表达式，用 Lagrangian 的观点也可以推广到 $\gamma = 1$ 和 $\alpha = 0$ 的情况。$\alpha = 0$ 的情况比较简单，当 $\gamma = 1$ 时，此问题将转化为关于无折扣 Q_{π} 函数的 LP 的 Lagrangian 对偶。详细的推导需要看文章，“AlgaeDICE: Policy Gradient from Arbitrary Experience”。

4.4 Off-Policy Evaluation (OPE)

公式 (23) 可以直接看成是 OPE 问题，通常在 behavior-agnostic 的 setting 下，OPE 问题通常转换为估计 $\omega_{\pi/\mathcal{D}}$，可以通过对拉格朗日函量引入不同的正则化来重新转换为特殊情况。正如文章中所说的，拉格朗日对偶方程的解同时得到了 Q 值和 $\omega_{\pi/\mathcal{D}}$ 的原始变量和对偶变量。

5 My Reflections

本文提出了一种，behavior-agnostic 的 Off-Policy 的策略梯度计算方法。基于具有线性规划特征的 Q 函数，通过推导 Lagrangian 鞍点表达式的方法来得出此策略梯度估计方法。由此得到的算法，AlgaeDICE，自动修正收集到的非策略数据的分布位移，并使用该 Off-Policy 数据实现了 On-Policy 的策略梯度估计。