Off-Policy Evaluation via the Regularized Lagrangian

Chen Gong

02 June 2021

目录

1 Core idea

DAI Bo 老师提出的 DICE(distribution correction estimation) family 在 OPE(Off-policy evaluation) 问题中，在 behavior-agnostic 的数据上取得了 SOTA 的效果。本文将这些 evaluation 方法统一为同一线性规划的正则 Lagrangian 估计。这种统一将为改进 DICE 提供新帮助，将 DICE 扩展到一个更大的 space，并实现更好的性能。更重要的是，通过数学和实验分析 estimators 的扩展空间，我们发现对偶解在优化稳定性和估计偏差之间的权衡提供了更大的灵活性，在实践中通常提供更好的估计。

文章中关注的问题叫做 behavior-agnostic OPE，也就是用当前环境中其他策略获得的数据来估计当前策略的好坏，而且通常对于收集数据的策略，我们不知道其显式信息。**所谓 DICE 就是对目标策略访问特定的 state-action 对的几率，与它们出现在数据集中的可能性之间的比率进行建模**。DICE 实际上算是一种分布矫正器，可以直接用于估计目标策略的值。

目前有很多种版本的 DICE，看着似乎不一样。DualDICE 是通过 change-of-varialbes 推导出来的。尽管这些方法有这些明显的差异，但这些算法都涉及一个相似形式的 max-min 优化，这表明在不同的推导过程中存在一个共同的联系。

实际上本文分析，之前的 DICE 都可以用带 Lagrangian 正则化项的线性规划问题推导而来。该 LP 与 OPE 问题有着密切的关系，具有 Q-LP 的原始形式和 d-LP 的对偶形式。作者认为 d-LP 的对偶形式比 Q-LP 的原始形式更加直观（我感觉差不多，作者估计是想在这里吹自己的方法）。在考虑 d-LP 形式的时候，需要考虑一些关键的选择，将其转化为一个稳定的 max-min 优化问题，比如，是否包含冗余约束，是否正则化原变量或对偶变量，还要考虑如何得到 OPE 的无偏估计。**作者提出了一种统一用带 Lagrangian 正则化项的 LP 问题来考虑现有的 OPE 方法，提出 DICE family 只是众多方程中的一个小部分，通过统一的框架，或许可以得到非常多种潜在的 OPE 方法**。

2 背景

2.1 Policy Evaluation

这里在 DualDICE，ALgaeDICE 中都分析过了，这里做简要描述。策略评估被定义为：

$$\rho(\pi) := (1-\gamma)\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t R\left(s_t,a_t\right) \mid s_0 \sim \mu_0, \forall t, a_t \sim \pi\left(s_t\right), s_{t+1} \sim T\left(s_t,a_t\right)\right] \tag{1}$$

在策略评估 setting 中，正在评估的策略称为目标策略。策略的价值可以用两种等价的方式表示：

$$\rho(\pi) = (1-\gamma)\cdot\mathbb{E}_{a_0\sim\pi(s_0)}\left[Q^\pi\left(s_0,a_0\right)\right] = \mathbb{E}_{(s,a)\sim d^\pi}[R(s,a)] \tag{2}$$

其中，Q^π 和 d^π 分别是状态价值函数和 π 的访问概率，具体解读请看 [DualDICE 论文解读]，并且它们满足：

$$\begin{aligned} Q^\pi(s,a) &= R(s,a) + \gamma\cdot\mathcal{P}^\pi Q^\pi(s,a), \text{ where } \mathcal{P}^\pi Q(s,a) := \mathbb{E}_{s'\sim T(s,a),a'\sim\pi(s')}\left[Q\left(s',a'\right)\right] \\ d^\pi(s,a) &= (1-\gamma)\mu_0(s)\pi(a\mid s) + \gamma\cdot\mathcal{P}_*^\pi d^\pi(s,a), \text{ where } \mathcal{P}_*^\pi d(s,a) := \pi(a\mid s)\sum_{\tilde{s},\tilde{a}} T(s\mid\tilde{s},\tilde{a})d(\tilde{s},\tilde{a}) \end{aligned} \tag{3}$$

其中，$\mathcal{P}_*^\pi$ 和 $\mathcal{P}^\pi$ 互为转置的，因为线性算子的伴随算子是它的转置，这里不多解释了。

Assumption 1 (MDP ergodicity). 假设公式 (3) 有唯一的最优解。

2.2 Undiscounted MDP

当 $\gamma = 1$ 时，策略价值将被定义为平均回报。

$$\rho(\pi) := \lim_{t_{\text{stop}} \to \infty} \mathbb{E}\left[\frac{1}{t_{\text{stop}}} \sum_{t=0}^{t_{\text{stop}}} R\left(s_t, a_t\right) \mid s_0 \sim \mu_0, \forall t, a_t \sim \pi\left(s_t\right), s_{t+1} \sim T\left(s_t, a_t\right)\right]$$

而其原始问题 prime LP 和对偶问题 $d-$LP 分别被定义为：

$$\max_{Q: S \times A \to \mathbb{R}} \lambda, \quad \text{s.t.}, \quad Q(s,a) = R(s,a) + \mathcal{P}^\pi Q(s,a) - \lambda$$

$$\max_{d: S \times A \to \mathbb{R}} \mathbb{E}_d[R(s,a)], \quad \text{s.t.}, \quad d(s,a) = \mathcal{P}_*^\pi d(s,a) \text{ and } \sum_{s,a} d(s,a) = 1$$

这里简单推导一下，prime LP：

$$\begin{aligned}
Q(s,a) =& \frac{1}{n+1} \sum_{i=0}^{n} R(s_i, a_i) \\
=& \frac{1}{n+1} R(s,a) + \sum_{i=1}^{n} R(s_i, a_i) \\
=& \frac{1}{n+1} R(s,a) + \frac{n}{n+1} R(s,a) - \frac{n}{n+1} R(s,a) + \sum_{i=1}^{n} R(s_i, a_i) \\
=& \frac{1}{n+1} R(s,a) + \frac{n}{n+1} R(s,a) + \sum_{i=1}^{n} \left(R(s_i, a_i) - R(s,a)\right) \\
=& R(s,a) + \mathcal{P}^\pi Q(s,a) - \lambda
\end{aligned}$$

由于有额外约束 $\sum_{s,a} d(s,a) = 1$，且马尔可夫链具有动态性，有唯一的 $d^* = d^\pi$。但是不同于带折扣的情况，对于 Q^* 是一个常数，其最优解 Q^* 将独立于 Q。

2.3 Off-policy Evaluation via the DICE Family

OPE 目的是用固定的经验数据集来估计 $\rho(\pi)$。假设确定的数据集为 $\mathcal{D} = \left\{\left(s_0^{(i)}, s^{(i)}, a^{(i)}, r^{(i)}, s'^{(i)}\right)\right\}_{i=1}^{N}$，其中，$s_0^{(i)} \sim \mu_0$，$(s^{(i)}, a^{(i)}) \sim d^{\mathcal{D}}$，其中 $d^{\mathcal{D}}$ 是未知的分布，$r^{(i)} = R(s^{(i)}, a^{(i)})$，$s'^{(i)} \sim T(s^{(i)}, a^{(i)})$。有时使用，$(s,a,r,s') \sim d^{\mathcal{D}}$ or $(s,a,r) \sim d^{\mathcal{D}}$ as a shorthand for $(s,a) \sim d^{\mathcal{D}}, r = R(s,a), s' \sim T(s,a)$。DICE 方法利用了以下 OPE 的表达式:

$$\rho(\pi) = \mathbb{E}_{(s,a) \sim d^{\mathcal{D}}}[\hat{\zeta}^*(s,a) \cdot R(s,a)] \tag{4}$$

其中，优化问题的最优结果 $\zeta^*(s,a) := \frac{d^\pi(s,a)}{d^{\mathcal{D}}(s,a)}$ 是分布修正率。DICE 方法就是在不知道 d^π 和 $d^{\mathcal{D}}$ 的情况下估计 $\zeta(s,a)$ 然后用公式 (4) 解 OPE 问题。

Assumption 2 (Boundedness). 分布修正率是有界的，$\|\zeta^*\|_\infty \le C < \infty$。

当 $\gamma < 1$ 时，DualDICE 选择一个最优解对应于该比值的凸目标，通过变量变换将对 d^π 的依赖转化为对 μ_0 的依赖。GenDICE，最小化连续的 On -policy 状态操作分布之间的分歧，并引入一个规范化约束，以确保 off-policy 数据集的估计比率平均为 1。并且 DualDICE 和 GenDICE 利用 Fenchel 对偶性将难处理的凸目标简化为 min-max 目标，使随机或连续行动空间中的抽样和优化成为可能。GradientDICE 采用线性参数化方法对 GenDICE 进行了扩展，使 min-max 优化算法具有凸凹收敛性保证。

3 DICE 估计的统一化观点

本节讨论的很多有，1. 讨论使用正则化项或者冗余约束可以提高优化的稳定性；虽然一般来说，这可能会在最终值估计中引入偏差，但在许多有效的 setting 下，OPE 仍然是无偏的。2. 我们表明，现有的 DICE 算法只是目前已有的统一框架的几种选择，而还有相当多算法仍未探索。

3.1 Linear Programming Representation for the d^π-distribution

Theorem 1. *Given a policy π, under Assumption 1, its value $\rho(\pi)$ defined in (1) can be expressed by the following d-LP:*

$$\max_{d:S\times A\to\mathbb{R}} \mathbb{E}_d\left[R\left(s,a\right)\right], \quad \text{s.t.,} \quad d(s,a) = \underbrace{(1-\gamma)\mu_0(s)\pi(a|s) + \gamma\cdot\mathcal{P}_*^\pi d(s,a)}_{\mathcal{B}_*^\pi\cdot d}. \tag{6}$$

*We refer to the d-LP above as the **dual** problem. Its corresponding **primal** LP is*

$$\min_{Q:S\times A\to\mathbb{R}} (1-\gamma)\,\mathbb{E}_{\mu_0\pi}\left[Q\left(s,a\right)\right], \quad \text{s.t.,} \quad Q(s,a) = \underbrace{R(s,a) + \gamma\cdot\mathcal{P}^\pi Q(s,a)}_{\mathcal{B}^\pi\cdot Q}. \tag{7}$$

作者这里通过简单的说事推理得到了这是个强对偶问题，实际大家想想也知道当 $d^* = d^\pi$ 时，必然有 $Q^* = Q^\pi$。尽管，$d-$LP 提供了一种有效的 OPE 方法，但是由于大量的约束，直接求解非常的困难，甚至 s, a 的数量几乎是不可数的。并且在 offline 的设定中，这个问题将加剧，因为只能随机的从 (s_0, s, a, r, s') 中进行采样。为了解决此问题，引入了 Lagrangian 方法，将带约束的优化问题，转换为无约束优化问题，

$$\max_d \min_Q L(d,Q) := (1-\gamma)\cdot\mathbb{E}_{a_0\sim\pi(s_0),s_0\sim\mu_0}\left[Q\left(s_0,a_0\right)\right] + \sum_{s,a} d(s,a)\cdot\left(R(s,a) + \gamma\mathcal{P}^\pi Q(s,a) - Q(s,a)\right) \tag{5}$$

同样上述方程可以被等价的写为：

$$\begin{aligned}\max_\zeta \min_Q L_D(\zeta,Q) :=& (1-\gamma)\cdot\mathbb{E}_{a_0\sim\pi(s_0),s_0\sim\mu_0}\left[Q\left(s_0,a_0\right)\right] \\ &+\mathbb{E}_{(s,a,r,s')\sim d^{\mathcal{D}},a'\sim\pi(s')}\left[\zeta(s,a)\cdot\left(r+\gamma Q\left(s',a'\right) - Q(s,a)\right)\right].\end{aligned} \tag{6}$$

其中，$\zeta(s,a) = \frac{d(s,a)}{d^{\mathcal{D}}(s,a)}$。此 Lagrangian 原问题和对偶问题的最优解为：$Q^* = Q^\pi$，$\zeta^* = \frac{d^\pi}{d^{\mathcal{D}}}$。看似上述优化问题非常的完美，实际上，在实践中，容易遇到许多优化难题。即使在 tabular 情况下，由于缺乏曲率，拉格朗日量不是强凸强凹，因此在随机梯度下降-上升 (SGDA) 下不能保证最终解的收敛性。在实际应用中，这些优化问题在神经网络参数化的连续情况下会变得更加严重。为了减轻这个问题，作者提出了一些方法来使优化更加稳定，并讨论了这些机制如何权衡最终估计的偏差。而且，从本文引入的方法可以推导出现有的 DICE 家族成员，一个更大的领域仍未探索。

3.2 Regularizations and Redundant Constraints

增广拉格朗日方法 (ALM) 正是为了避免优化不稳定性而提出的，该方法通过在不改变最优解的情况下添加额外的正则化来引入强凸性，这里我参考了一些资料其主要意思是利用 ALM 可以理论

上保证收敛，建议参考资料[增广拉格朗日法]。如果直接使用 ALM，$h_p(Q) := \|\mathcal{B}^\pi \cdot Q - Q\|_{d^\mathcal{D}}^2$ 或者 $h_d(d) := D_f(d\|\mathcal{B}_*^\pi \cdot d)$，也不好求解。因为在非线性运算符 $h_p(Q)$ 和 $h_d(d)$ 中含有条件期望算子 $\mathcal{B}^\pi$ 和 $\mathcal{B}_*^\pi$，这通常被认为是双采样问题。比如在对 $\nabla_Q h_p(Q)$ 的梯度计算中，

$$\nabla_Q h_p(Q) := 2\left\|\mathcal{B}^\pi \cdot Q - Q\right\|_{d^\mathcal{D}} \nabla_Q \left\|\mathcal{B}^\pi \cdot Q - Q\right\|_{d^\mathcal{D}} \tag{7}$$

只有对于每个 (s,a) 独立采样两次，得到两个 (r,s') 并且用于上述两项中才能得到无偏的梯度估计。而这将导致较大的偏差（个人猜想是独立采样做不到，直接应该是有关联的）。所以，导致传统的随机梯度下降不再适用。

收到 ALM 的启发，此部分将探索其他的正则化项来对公式 (6) 中的原拉格朗日问题引入强对偶性。除了正则化，作者还使用了冗余约束，这有助于在不影响最优解的情况下为优化添加更多的结构。我们稍后将分析这些对原始问题的修改将导致对 $\rho(\pi)$ 的有偏估计。

首先将提出一个全面的统一目标，其中包括所有正规化和冗余约束的选择:

$$\begin{aligned}\max_{\zeta\geq 0}\min_{Q,\lambda} L_D(\zeta,Q,\lambda) :=& (1-\gamma)\cdot\mathbb{E}_{a_0\sim\pi(s_0)}\left[Q\left(s_0,a_0\right)\right]+\lambda \\ &+\mathbb{E}_{(s,a,r,s')\sim d^\mathcal{D}\sim\mu_0}\left[\zeta(s,a)\cdot\left(\alpha_R\cdot R(s,a)+\gamma Q\left(s',a'\right)-Q(s,a)-\lambda\right)\right] \\ &+\alpha_Q\cdot\mathbb{E}_{(s,a)\sim d^\mathcal{D}}\left[f_1(Q(s,a))\right]-\alpha_\zeta\cdot\mathbb{E}_{(s,a)\sim d^\mathcal{D}}\left[f_2(\zeta(s,a))\right]\end{aligned} \tag{8}$$

下面将依次解释每一项的作用，

1. Primal和Dual正则化项：为了在拉格朗日函数中引入更好的曲率，我们引入了原始正则化和对偶正则化，$\alpha_Q\cdot\mathbb{E}_{(s,a)\sim d^\mathcal{D}}\left[f_1(Q(s,a))\right]$ 和 $\alpha_\zeta\cdot\mathbb{E}_{(s,a)\sim d^\mathcal{D}}\left[f_2(\zeta(s,a))\right]$。其中，$f_1$ 和 f_2 代表下半连续的凸函数。

2. Reward: 奖赏的缩放可以看作是对偶正则化的延伸，$\alpha_R\in\{0,1\}$。

3. Positivity: 公式 (6) 中拉格朗日问题的最优解是 $\zeta^*(s,a)=\frac{d^\pi(s,a)}{d^\mathcal{D}(s,a)}\geq 0$。因此，我们考虑给对偶变量增加一个正性约束。这可以解释为修改原始的 d-LP，在其目标上增加一个条件 $d\geq 0$。

4. Normalization: 规范化约束也来自于最优解的性质，$\mathbb{E}_{d^\mathcal{D}}[\zeta(s,a)]=1$，相当于增加约束 $\sum_{s,a}d(s,a)=1$，应用拉格朗日对偶法可得 $\lambda-\mathbb{E}_{d^\mathcal{D}}[\lambda\zeta(s,a)]$。

后两个选项来自最优对偶解的性质，这表明包含它们不会影响最优对偶解。而前两个约束 (Prime and Dual Regularization 和 reward scaling) 通常会影响优化的解决方案。解的偏差是否影响最终估计结果取决于所使用的估计方法。

Remark (Robust optimization justification):

除了考虑公式 (8) 中的正则化项 $\alpha_Q\cdot\mathbb{E}_{(s,a)\sim d^\mathcal{D}}\left[f_1(Q(s,a))\right]$ 和 $\alpha_\zeta\cdot\mathbb{E}_{(s,a)\sim d^\mathcal{D}}\left[f_2(\zeta(s,a))\right]$ 来源于 ALM 思想，将引入强对偶性之外。也可以解释为引入对 Bellman 残差的一些扰动的鲁棒性。本文中以对偶正则化为例，利用 Fenchel-Rockafellar 对偶为例，

$$\alpha_\zeta\cdot\mathbb{E}_{(s,a)\sim d^\mathcal{D}}\left[f_2(\zeta(s,a))\right]=\alpha_\zeta\left\{\max_{\delta(s,a)\in\Omega}\langle\zeta,\delta\rangle-\mathbb{E}_{(s,a)\sim d^\mathcal{D}}\left[f_2^*(\delta(s,a))\right]\right\}$$

其中，令 Ω 为函数 f_2^* 的定义域，$f_2^*(\cdot)=(\cdot)^2$，融入到公式 (8) 中有，

$$\begin{aligned}\max_{\zeta\geq 0}\min_{Q,\lambda,\delta\in\Omega} L_D(\zeta,Q,\lambda) :=&(1-\gamma)\cdot\mathbb{E}_{a_0\sim\pi(s_0)}\left[Q\left(s_0,a_0\right)\right]+\lambda\\ &+\mathbb{E}_{(s,a,r,s')\sim d^{\mathcal{D}}\sim\mu_0}\left[\zeta(s,a)\cdot\left(\alpha_R\cdot R(s,a)+\gamma Q\left(s',a'\right)-Q(s,a)-\lambda-\alpha_\zeta\delta(s,a)\right)\right]\\ &+\alpha_Q\cdot\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[f_1(Q(s,a))\right]+\alpha_\zeta\cdot\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[\delta^2(s,a)\right]\end{aligned} \tag{9}$$

可以看成是松弛变量，或者是贝尔曼残差：$\alpha_R R(s,a)+\gamma Q(s',a')-Q(s,a)$ 在 L_2 球中的扰动。对于不同的正则化，扰动将出现在不同的对偶空间中。从这个角度来看，除了稳定性的考虑外，Dual 正则化还将减少由于近似 Bellman 差分时的采样效应而产生的统计误差，以及由 Q 近似引起的近似误差。

Prime 正则化可以解释为在平稳 $d(s,a)$ 分布下引入松弛变量，$\alpha_Q\cdot\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[f_1(Q(s,a))\right]=\alpha_Q\cdot\left\{\max_{\delta(s,a)}\langle Q,\delta\rangle-\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[\delta^2(s,a)\right]\right\}$。代入公式 (8) 可得，

$$\begin{aligned}\max_{\zeta\geq 0,\delta}\min_{Q,\lambda} L_D(\zeta,Q,\lambda,\delta) :=&(1-\gamma)\cdot\mathbb{E}_{a_0\sim\pi(s_0)}\left[Q\left(s_0,a_0\right)\right]+\alpha_Q\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[\delta(s,a)\cdot Q(s,a)]+\lambda\\ &+\mathbb{E}_{(s,a,r,s')\sim d^{\mathcal{D}}\sim\mu_0 a'\sim\pi\left(s'\right)}\left[\zeta(s,a)\cdot\left(\alpha_R\cdot R(s,a)+\gamma Q\left(s',a'\right)-Q(s,a)-\lambda\right)\right]\\ &-\alpha_Q\cdot\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[\delta^2(s,a)\right]-\alpha_\zeta\cdot\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[f_2(\zeta(s,a))\right]\end{aligned} \tag{10}$$

可以理解为的拉格朗日量，

$$\begin{aligned}\max_{\zeta\geq 0,\delta}\quad & \alpha_R\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[\zeta(s,a)\cdot R(s,a)]-\alpha_Q\cdot\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[\delta^2(s,a)\right]-\alpha_\zeta\cdot\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[f_2(\zeta(s,a))\right]\\ \text{s.t.}\quad & (1-\gamma)\mu_0\pi+\alpha_Q d^{\mathcal{D}}\cdot\delta+\gamma\cdot\mathcal{P}_*^\pi\cdot\left(d^{\mathcal{D}}\cdot\zeta\right)=\left(d^{\mathcal{D}}\cdot\zeta\right)\\ & \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[\zeta]=1\end{aligned} \tag{11}$$

这样就可以理解了，Primal 正则化项可以作为 $d(s,a)$ 的对偶项。事实上，Primal 正则化项可看成是对 d 的扰动，和前面的描述差不多。

所以，在给定估计量 $\hat{Q},\hat{\lambda},\hat{\zeta}$ 的情况下，有三种潜在的估计 $\rho(\pi)$ 的方法。

1. **Primal estimator:** $\hat{\rho}_Q(\pi):=(1-\gamma)\cdot\mathbb{E}_{a_0\sim\pi(s_0),s_0\sim\mu_0}\left[\hat{Q}\left(s_0,a_0\right)\right]+\hat{\lambda}$
2. **Dual estimator:** $\hat{\rho}_\zeta(\pi):=\mathbb{E}_{(s,a,r)\sim d^{\mathcal{D}}}[\hat{\zeta}(s,a)\cdot r]$
3. **Lagrangian:** $\hat{\rho}_{Q,\zeta}(\pi):=\hat{\rho}_Q(\pi)+\hat{\rho}_\zeta(\pi)+\mathbb{E}_{(s,a,r,s')\sim d^{\mathcal{D}},a'\sim\pi(s')}\left[\hat{\zeta}(s,a)\left(\gamma\hat{Q}\left(s',a'\right)-\hat{Q}(s,a)-\hat{\lambda}\right)\right]$

Theorem 2 (Regularization profiling).

对于 $\alpha_Q,\alpha_\zeta,\alpha_R,\lambda,\zeta\geq 0$ 是否存在，一共有 32 种选法。假设，Q^*,ζ^* 是不同配置下的最优解。很显然，如果是无偏估计的话，那么，对于原问题估计 $\hat{\rho}^Q$ 的解为 $Q^*=Q^\pi$，而对于对偶问题的估计 $\hat{\rho}^\zeta$ 的解为 $\zeta^*=\frac{d^\pi}{d^{\mathcal{D}}}$。而对于拉格朗日估计器 $\hat{\rho}_{Q,\zeta}$ 有两种表达方式，

$$\begin{aligned}\hat{\rho}_{Q,\zeta}(\pi)&=\hat{\rho}_Q(\pi)+\sum_{s,a}d^{\mathcal{D}}(s,a)\zeta(s,a)\left(R(s,a)+\gamma\mathcal{P}^\pi Q(s,a)-Q(s,a)\right)\\ &=\hat{\rho}_\zeta(\pi)+\sum_{s,a}Q(s,a)\left((1-\gamma)\mu_0(s)\pi(a\mid s)+\gamma\mathcal{P}_*^\pi d^{\mathcal{D}}\times\zeta(s,a)-d^{\mathcal{D}}\times\zeta(s,a)\right)\end{aligned} \tag{12}$$

可以，清晰的发现，当 $Q** = Q$ 时，第一行的第二项为 0，那么有 $\hat{\rho}_{Q,\zeta}(\pi) = \rho(\pi)$。同样，当 $\zeta^* = \frac{d^\pi}{d^\mathcal{D}}$ 时，有 $\hat{\rho}_{Q,\zeta}(\pi) = \rho(\pi)$。所以，大家现在应该可以之前老是提到的双鲁棒性了，只要 Q 和 ζ 中有一个是无偏估计，拉格朗日估计就是无偏的。

$$L(Q, \zeta^*) = L(Q^*, \zeta) = L(Q^*, \zeta^*) = \rho(\pi) \tag{13}$$

由于前面提到，ζ 和 λ 是从解的结构性质中引出的，他们不会影响解的性质。所以，我们只要关注 $\alpha_Q, \alpha_\zeta, \alpha_R$ 的配置是否会影响 Q^*, ζ^* 即可。

其中 $\alpha_Q = 0, \alpha_\zeta = 0$ 的情况，肯定是无偏的，这就是 AlgaeDICE，[AlgaeDICE 论文解读]。如果，$\alpha_Q > 0, \alpha_\zeta > 0$ 也通常是有偏的。所以，考虑 $\alpha_Q = 0, \alpha_\zeta > 0$ 或 $\alpha_Q > 0, \alpha_\zeta = 0$ 的情况，如下表所示。并给出阴影部分的证明，而非阴影部分，请参考 [AlgaeDICE 论文解读]，[DualDICE 论文解读]。

Regularizer (w./w.o. λ)		Case		$Q^*(s,a)$	$\zeta^*(s,a)$	$L(Q^*, \zeta^*)$
$\alpha_\zeta = 0$	$\alpha_R = 1$	ζ free	i	Q^π	$\frac{d^\pi}{d^\mathcal{D}} + \alpha_Q \frac{(\mathcal{I} - \gamma \mathcal{P}_*^\pi)^{-1}\left(d^\mathcal{D} \cdot f_1'(Q^\pi)\right)}{d^\mathcal{D}}$	$\alpha_R(1-\gamma)\cdot \mathbb{E}_{\mu_0}[Q^\pi]$ $+\alpha_Q \mathbb{E}_{(s,a)\sim d^\mathcal{D}}\left[f_1\left(Q^\pi\right)\right]$
		$\zeta \geq 0$	ii	$f_1^{*\prime}\left(\frac{1}{\alpha_Q}\left((\alpha_Q f_1'\left(Q^\pi\right) + \frac{(1-\gamma)\mu_0\pi}{d^\mathcal{D}})_+ - \frac{(1-\gamma)\mu_0\pi}{d^\mathcal{D}}\right)\right)$	$\frac{1}{d^\mathcal{D}}\left(\mathcal{I} - \gamma \cdot \mathcal{P}^\pi\right)^{-1} \cdot$ $d^\mathcal{D}\left(\alpha_Q f_1'\left(Q^\pi\right) + \frac{(1-\gamma)\mu_0\pi}{d^\mathcal{D}}\right)_+$	$(1-\gamma)\cdot \mathbb{E}_{\mu_0}[Q^*]$ $+\mathbb{E}_{d^\mathcal{D}}[\zeta^*(s,a)\cdot(\alpha_R \cdot r$ $+\gamma Q^*(s',a') - Q^*(s,a))]$ $+\alpha_Q \cdot \mathbb{E}_{d^\mathcal{D}}[f_1(Q^*(s,a))]$
$\alpha_Q > 0$	$\alpha_R = 0$	ζ free	iii	$f_1^{*\prime}(0)$		$-\alpha_Q f_1^*(0)$
		$\zeta \geq 0$	iv			
$\alpha_\zeta > 0$	$\alpha_R = 1$	ζ free	v	$-\alpha_\zeta\left(\mathcal{I} - \mathcal{P}^\pi\right)^{-1} f_2'(\frac{d^\pi}{d^\mathcal{D}})$ $+\alpha_R Q^\pi$ (Nachum et al., 2019a,b)	$\frac{d^\pi}{d^\mathcal{D}}$ (Nachum et al., 2019a,b)	$\alpha_R \cdot \mathbb{E}_{(s,a,r,s')\sim d^\mathcal{D}}[r]$ $-\alpha_\zeta \cdot D_f(d^\pi \| d^\mathcal{D})$ (Nachum et al., 2019a,b)
		$\zeta \geq 0$	vi			
$\alpha_Q = 0$	$\alpha_R = 0$	ζ free	vii			
		$\zeta \geq 0$	viii			

在 Assumption 1 和 Assumption 2 下，公式 (8) 满足强对偶性。

- $iii) - iv)$，此时 $\alpha_Q > 0, \alpha_\zeta = 0, \alpha_R = 0$。公式 (8) 被写为，

$$\begin{aligned}\max_{\zeta \geq 0} \min_{Q,\lambda} L_D(\zeta, Q, \lambda) :=& (1-\gamma)\cdot \mathbb{E}_{a_0 \sim \pi(s_0)}\left[Q\left(s_0, a_0\right)\right] + \alpha_Q \cdot \mathbb{E}_{(s,a)\sim d^\mathcal{D}}\left[f_1(Q(s,a))\right] + \lambda \\ &+ \mathbb{E}_{(s,a,r,s')\sim d^\mathcal{D}, a'\sim\pi(s')}[\zeta(s,a)\cdot(\gamma Q(s',a') - Q(s,a) - \lambda)],\end{aligned} \tag{14}$$

等价于，

$$\begin{aligned}&\max_{\zeta \geq 0} \min_{Q} L_D(\zeta, Q) = \left\langle (1-\gamma)\mu_0\pi + \gamma \cdot \mathcal{P}_*^\pi \cdot \left(d^\mathcal{D} \cdot \zeta\right) - d^\mathcal{D} \cdot \zeta, Q \right\rangle + \alpha_Q \mathbb{E}_{d^\mathcal{D}}\left[f_1(Q)\right] \\ &\text{s.t.} \quad \mathbb{E}_{d^\mathcal{D}}[\zeta] = 1\end{aligned} \tag{15}$$

根据 Fenchel 对偶公式，$f(x) = \max_\zeta x \cdot \zeta - f^*(\zeta)$，w.r.t. Q，可得，

$$\begin{aligned}&\max_\zeta \quad L_D\left(\zeta, Q^*\right) = -\alpha_Q \mathbb{E}_{d^\mathcal{D}}\left[f_1^*\left(\frac{(1-\gamma)\mu_0\pi + \gamma\cdot\mathcal{P}_*^\pi\cdot\left(d^\mathcal{D}\cdot\zeta\right) - d^\mathcal{D}\cdot\zeta}{\alpha_Q d^\mathcal{D}}\right)\right] \\ &\text{s.t.} \quad \mathbb{E}_{d\mathcal{D}}[\zeta] = 1\end{aligned} \tag{16}$$

然后，$\max_\zeta$ 显然就是 $f_1^*(0)$，那么有，

$$d^\mathcal{D} \cdot \zeta^* = (1-\gamma)\mu_0\pi + \gamma \cdot \mathcal{P}_*^\pi \cdot \left(d^\mathcal{D} \cdot \zeta^*\right) \Rightarrow d^\mathcal{D} \cdot \zeta^* = d^\pi \tag{17}$$

所以有，

$$L\left(\zeta^*, Q^*\right) = -\alpha_Q f_1^*(0)$$

而，

$$\begin{aligned} Q^* &= \underset{Q}{\operatorname{argmax}} \left\langle (1-\gamma)\mu_0\pi + \gamma \cdot \mathcal{P}_*^\pi \cdot \left(d^{\mathcal{D}} \cdot \zeta^*\right) - d^{\mathcal{D}} \cdot \zeta^*, Q \right\rangle + \alpha_Q \mathbb{E}_{d^{\mathcal{D}}}\left[f_1(Q)\right] \\ &= f_1^{*\prime}(0) \end{aligned} \tag{18}$$

这里用到一个推导，$f^{*'}(\cdot) = f^{(-1)'}(\cdot)$。

- $i) - ii)$，和前面的推导相比，多一项 $\alpha_R \mathbb{E}_{d^{\mathcal{D}}}[\zeta \cdot R]$，

$$\begin{aligned} \max_{\zeta} \min_{Q} L_D(\zeta, Q) :=& (1-\gamma) \cdot \mathbb{E}_{a_0 \sim \pi(s_0)}\left[Q\left(s_0, a_0\right)\right] + \alpha_Q \cdot \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[f_1(Q(s,a))\right] \\ &+ \mathbb{E}_{\substack{(s,a,r,s')\sim d^{\mathcal{D}} \\ a' \sim \pi(s')}}\left[\zeta(s,a) \cdot \left(\alpha_R \cdot R(s,a) + \gamma Q\left(s', a'\right) - Q(s,a)\right)\right]. \end{aligned} \tag{19}$$

其中，考虑不考虑 $\zeta > 0$ 的情况，和前面的推导一样，首先 w.r.t. Q 使用 Fenchel 对偶，可得

$$\max_{\zeta} L_D\left(\zeta, Q^*\right) = \alpha_R \left\langle d^{\mathcal{D}} \cdot \zeta, R \right\rangle - \alpha_Q \mathbb{E}_{d^{\mathcal{D}}}\left[f_1^*\left(\frac{d^{\mathcal{D}} \cdot \zeta - (1-\gamma)\mu_0\pi - \gamma \cdot \mathcal{P}_*^\pi \cdot \left(d^{\mathcal{D}} \cdot \zeta\right)}{\alpha_Q d^{\mathcal{D}}}\right)\right] \tag{20}$$

然后，令

$$\nu = \frac{d^{\mathcal{D}} \cdot \zeta - (1-\gamma)\mu_0\pi - \gamma \cdot \mathcal{P}_*^\pi \cdot \left(d^{\mathcal{D}} \cdot \zeta\right)}{d^{\mathcal{D}}} \Rightarrow d^{\mathcal{D}} \cdot \zeta = \left(\mathcal{I} - \gamma \cdot \mathcal{P}_*^\pi\right)^{-1}\left((1-\gamma)\mu_0\pi + d^{\mathcal{D}} \cdot \nu\right)$$

$$\begin{aligned} L_D\left(\zeta^*, Q^*\right) &= \max_{\nu} \left\langle \left(\mathcal{I} - \gamma \cdot \mathcal{P}_*^\pi\right)^{-1}\left((1-\gamma)\mu_0\pi + d^{\mathcal{D}} \cdot \nu\right), \alpha_R R \right\rangle - \alpha_Q \mathbb{E}_{d^{\mathcal{D}}}\left[f_1^*\left(\frac{\nu}{\alpha_Q}\right)\right] \\ &= \alpha_R(1-\gamma)\mathbb{E}_{a_0\sim\pi(s_0), s_0\sim\mu_0}\left[Q^\pi\left(s_0, a_0\right)\right] + \max_{\nu}\left\{\mathbb{E}_{d^{\mathcal{D}}}\left[\nu \cdot \left(Q^\pi\right)\right] - \alpha_Q \mathbb{E}_{d^{\mathcal{D}}}\left[f_1^*\left(\frac{\nu}{\alpha_Q}\right)\right]\right\} \\ &= \alpha_R(1-\gamma)\mathbb{E}_{a_0\sim\pi(s_0), s_0\sim\mu_0}\left[Q^\pi\left(s_0, a_0\right)\right] + \alpha_Q \max_{\nu}\left\{\mathbb{E}_{d^{\mathcal{D}}}\left[\frac{\nu}{\alpha_Q} \cdot \left(Q^\pi\right)\right] - \mathbb{E}_{d^{\mathcal{D}}}\left[f_1^*\left(\frac{\nu}{\alpha_Q}\right)\right]\right\} \\ &= \alpha_R(1-\gamma)\mathbb{E}_{a_0\sim\pi(s_0), s_0\sim\mu_0}\left[Q^\pi\left(s_0, a_0\right)\right] + \alpha_Q \mathbb{E}_{d^{\mathcal{D}}}\left[f_1\left(Q^\pi\right)\right] \end{aligned}$$

其中，根据 KKT 条件，$\frac{\nu^*}{\alpha_Q} = f_1'(Q^\pi)$，$Q^\pi - (\mathcal{I} - \gamma \cdot \mathcal{P}^\pi)^{-1}$。

$$\begin{aligned} &\frac{\partial \frac{\nu}{\alpha_Q} \cdot \left(Q^\pi\right) - f_1^*\left(\frac{\nu}{\alpha_Q}\right)}{\partial \nu} = 0 \\ &\frac{\left(Q^\pi\right)}{\alpha_Q} - \frac{1}{\alpha_Q} \cdot f_1^{*'}\left(\frac{\nu}{\alpha_Q}\right) = 0 \\ &Q^\pi = f_1^{*'}\left(\frac{\nu}{\alpha_Q}\right) \\ &\nu^* = \alpha_Q f_1'(Q^\pi) \end{aligned}$$

根据变量替换结果有，

$$\begin{aligned} \nu^* &= \frac{d^{\mathcal{D}} \cdot \zeta^* - (1-\gamma)\mu_0\pi - \gamma \cdot \mathcal{P}_*^\pi \cdot \left(d^{\mathcal{D}} \cdot \zeta^*\right)}{d^{\mathcal{D}}} = \alpha_Q f_1'(Q^\pi) \\ \zeta^* &= \frac{\left(\mathcal{I} - \gamma \cdot \mathcal{P}_*^\pi\right)^{-1}\left((1-\gamma)\mu_0\pi\right)}{d^{\mathcal{D}}} + \alpha_Q \frac{\left(\mathcal{I} - \gamma \cdot \mathcal{P}_*^\pi\right)^{-1}\left(d^{\mathcal{D}} \cdot f_1'\left(Q^\pi\right)\right)}{d^{\mathcal{D}}} \\ &= \frac{d^\pi}{d^{\mathcal{D}}} + \alpha_Q \frac{\left(\mathcal{I} - \gamma \cdot \mathcal{P}_*^\pi\right)^{-1}\left(d^{\mathcal{D}} \cdot f_1'\left(Q^\pi\right)\right)}{d^{\mathcal{D}}}, \end{aligned} \tag{21}$$

同理对公式 (20) 使用 KKT 条件，可得

$$f_1'(Q) = \frac{d^{\mathcal{D}} \cdot \zeta - (1-\gamma)\mu_0\pi - \gamma \cdot \mathcal{P}_*^{\pi} \cdot \left(d^{\mathcal{D}} \cdot \zeta\right)}{\alpha_Q d^{\mathcal{D}}} \tag{22}$$

$$\begin{aligned} Q^* =& (f_1')^{-1}\left(\frac{d^{\mathcal{D}} \cdot \zeta^* - (1-\gamma)\mu_0\pi - \gamma \cdot \mathcal{P}_*^{\pi} \cdot \left(d^{\mathcal{D}} \cdot \zeta^*\right)}{\alpha_Q d^{\mathcal{D}}}\right) \\ =& (f_1')^{-1}\left(\frac{\nu^*}{\alpha_Q}\right) \\ =& Q^{\pi} \end{aligned} \tag{23}$$

如果，有约束令 $\zeta > 0$，令

$$\exp(\nu) = \frac{\left(\mathcal{I} - \gamma \cdot \mathcal{P}_*^{\pi}\right)\left(d^{\mathcal{D}} \cdot \zeta\right)}{d^{\mathcal{D}}} \Rightarrow d^{\mathcal{D}} \cdot \zeta = \left(\mathcal{I} - \gamma \cdot \mathcal{P}_*^{\pi}\right)^{-1} d^{\mathcal{D}} \cdot \exp(\nu)$$

同样代入公式（20）中可得，

$$L_D\left(\zeta^*, Q^*\right) = \max_{\nu} \mathbb{E}_{d^{\mathcal{D}}}\left[\exp(\nu) \cdot Q^{\pi}\right] - \alpha_Q \mathbb{E}_{d^{\mathcal{D}}}\left[f_1^*\left(\frac{1}{\alpha_Q}\left(\exp(\nu) - \frac{(1-\gamma)\mu_0\pi}{d^{\mathcal{D}}}\right)\right)\right] \tag{24}$$

和前面的过程是一样的，使用 KKT 条件，可得，

$$\begin{aligned} & \exp\left(\nu^*\right)\left(Q^{\pi} - f_1^{*\prime}\left(\frac{1}{\alpha_Q}\left(\exp(\nu^*) - \frac{(1-\gamma)\mu_0\pi}{d^{\mathcal{D}}}\right)\right)\right) = 0 \\ =& \exp\left(\nu^*\right) = \left(\alpha_Q f_1'\left(Q^{\pi}\right) + \frac{(1-\gamma)\mu_0\pi}{d^{\mathcal{D}}}\right)_+ \\ \Rightarrow & d^{\mathcal{D}} \cdot \zeta^* = \left(\mathcal{I} - \gamma \cdot \mathcal{P}^{\pi}\right)^{-1} \cdot d^{\mathcal{D}}\left(\alpha_Q f_1'\left(Q^{\pi}\right) + \frac{(1-\gamma)\mu_0\pi}{d^{\mathcal{D}}}\right)_+ \\ \Rightarrow & \zeta^* = \frac{1}{d^{\mathcal{D}}}\left(\mathcal{I} - \gamma \cdot \mathcal{P}^{\pi}\right)^{-1} \cdot d^{\mathcal{D}}\left(\alpha_Q f_1'\left(Q^{\pi}\right) + \frac{(1-\gamma)\mu_0\pi}{d^{\mathcal{D}}}\right)_+ \end{aligned} \tag{25}$$

其中，第一行到第二行的推导，

$$\begin{aligned} & \exp\left(\nu^*\right)\left(Q^{\pi} - f_1^{*\prime}\left(\frac{1}{\alpha_Q}\left(\exp(\nu^*) - \frac{(1-\gamma)\mu_0\pi}{d^{\mathcal{D}}}\right)\right)\right) = 0 \\ & Q^{\pi} - f_1^{*\prime}\left(\frac{1}{\alpha_Q}\left(\exp(\nu^*) - \frac{(1-\gamma)\mu_0\pi}{d^{\mathcal{D}}}\right)\right) = 0 \\ & Q^{\pi} = f_1^{*\prime}\left(\frac{1}{\alpha_Q}\left(\exp(\nu^*) - \frac{(1-\gamma)\mu_0\pi}{d^{\mathcal{D}}}\right)\right) \\ & f_1'(Q^{\pi}) = \frac{1}{\alpha_Q}\left(\exp(\nu^*) - \frac{(1-\gamma)\mu_0\pi}{d^{\mathcal{D}}}\right) \\ & \exp\left(\nu^*\right) = \left(\alpha_Q f_1'\left(Q^{\pi}\right) + \frac{(1-\gamma)\mu_0\pi}{d^{\mathcal{D}}}\right)_+ \end{aligned}$$

至于 Q^* 也是一样的处理方法，

$$\begin{aligned} Q^* &= f_1^{*\prime}\left(\frac{1}{\alpha_Q}\left(\exp\left(\nu^*\right) - \frac{(1-\gamma)\mu_0\pi}{d^{\mathcal{D}}}\right)\right) \\ &= f_1^{*\prime}\left(\frac{1}{\alpha_Q}\left(\left(\alpha_Q f_1'\left(Q^{\pi}\right) + \frac{(1-\gamma)\mu_0\pi}{d^{\mathcal{D}}}\right)_+ - \frac{(1-\gamma)\mu_0\pi}{d^{\mathcal{D}}}\right)\right) \end{aligned} \tag{26}$$

很显然，关于 ζ^*, Q^* 都是有偏估计。

- $v) - viii)$

  为了简单，我们忽略 $\zeta > 0$ 和 λ，结论不受影响，因为最优解 ζ 自动满足这些约束。根据公式 (8) 和选择的配置，可得，

$$\begin{aligned}\min_{Q}\max_{\zeta} L_D(\zeta, Q) :=&(1-\gamma)\cdot\mathbb{E}_{a_0\sim\pi(s_0)}\left[Q\left(s_0,a_0\right)\right]-\alpha_\zeta\cdot\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[f_2(\zeta(s,a))\right]\\&+\mathbb{E}_{\substack{(s,a,r,s')\sim d^{\mathcal{D}},\\ a'\sim\pi(s')}}\left[\zeta(s,a)\cdot\left(\alpha_R\cdot R(s,a)+\gamma Q\left(s',a'\right)-Q(s,a)\right)\right]\end{aligned} \tag{27}$$

  w.r.t. ζ 使用 Fenchel 对偶，可以得到，

$$\min_{Q} L_D\left(\zeta^*, Q\right) := (1-\gamma)\cdot\mathbb{E}_{a_0\sim\pi(s_0)}\left[Q\left(s_0,a_0\right)\right]+\alpha_\zeta\mathbb{E}_{d^{\mathcal{D}}}\left[f_2^*\left(\frac{1}{\alpha_\zeta}\left(\mathcal{B}^\pi\cdot Q(s,a)-Q(s,a)\right)\right)\right] \tag{28}$$

  其中，$\mathcal{B}^\pi\cdot Q(s,a) := \alpha_R\cdot R(s,a)+\gamma\mathcal{P}^\pi Q(s,a)$，令 $\nu(s,a) = \mathcal{B}\cdot Q(s,a) - Q(s,a)$，那么可以推导出，

$$Q(s,a) = \left(\mathcal{I}-\gamma\cdot\mathcal{P}^\pi\right)^{-1}\left(\alpha_R\cdot R-\nu\right) \tag{29}$$

  推导也很简单，$\mathcal{B}^\pi\cdot Q(s,a) - Q(s,a) := \alpha_R\cdot R(s,a)+\gamma\mathcal{P}^\pi Q(s,a) - Q(s,a)$，融合入公式 (28)，可以得到，

$$\begin{aligned}L_D\left(\zeta^*, Q^*\right) =&\min_{\nu}(1-\gamma)\cdot\mathbb{E}_{\substack{a_0\sim\pi(s_0),\\ s_0\sim\mu_0}}\left[\left(\left(\mathcal{I}-\gamma\cdot\mathcal{P}^\pi\right)^{-1}\left(\alpha_R\cdot R-\nu\right)\right)\left(s_0,a_0\right)\right]\\&+\alpha_\zeta\mathbb{E}_{d^{\mathcal{D}}}\left[f_2^*\left(\frac{1}{\alpha_\zeta}\nu(s,a)\right)\right]\\=&\alpha_R\mathbb{E}_{d^\pi}[R(s,a)]-\alpha_\zeta\max_{\nu}\left(\mathbb{E}_{d^\pi}\left[\frac{\nu\left(s_0,a_0\right)}{\alpha_\zeta}\right]+\mathbb{E}_{d^{\mathcal{D}}}\left[f_2^*\left(\frac{1}{\alpha_\zeta}\nu(s,a)\right)\right]\right)\\=&\alpha_R\mathbb{E}_{d^\pi}[R(s,a)]-\alpha_\zeta D_f\left(d^\pi\|d^{\mathcal{D}}\right)\end{aligned} \tag{30}$$

  其中，$d^\pi = (\mathcal{I}-\gamma\cdot\mathcal{P}_*^\pi)^{-1}(1-\gamma)(\mu_0\pi)$。同理，根据 KKT 条件，可得，$\nu^* = \alpha_\zeta f_2'(\frac{d^\pi}{d^{\mathcal{D}}})$。结合公式 (29) 和公式 (30) 可以推导出，

$$\begin{aligned}Q^* &= -\left(\mathcal{I}-\gamma\cdot\mathcal{P}^\pi\right)^{-1}\nu^* + \left(\mathcal{I}-\gamma\cdot\mathcal{P}^\pi\right)^{-1}\left(\alpha_R\cdot R\right)\\&= -\alpha_\zeta\left(\mathcal{I}-\gamma\cdot\mathcal{P}^\pi\right)^{-1} f_2'\left(\frac{d^\pi}{d^{\mathcal{D}}}\right)+\alpha_R Q^\pi\end{aligned} \tag{31}$$

$$\begin{aligned}\zeta^*(s,a) &= \operatorname*{argmax}_{\zeta}\zeta\cdot\nu^*(s,a)-\alpha_\zeta f_2(\zeta(s,a))\\&= f_2^{*\prime}\left(\frac{1}{\alpha_\zeta}\nu^*(s,a)\right) = \frac{d^\pi(s,a)}{d^D(s,a)}\end{aligned} \tag{32}$$

□

根据上述的推导可以得到下表，

Regularization (with or without λ)			$\hat{\rho}_Q$	$\hat{\rho}_\zeta$	$\hat{\rho}_{Q,\zeta}$
$\alpha_\zeta = 0$ $\alpha_Q > 0$	$\alpha_R = 1$	ζ free	**Unbiased**	Biased	**Unbiased**
		$\zeta \geq 0$	Biased		Biased
	$\alpha_R = 0$	ζ free		**Unbiased**	**Unbiased**
		$\zeta \geq 0$			
$\alpha_\zeta > 0$ $\alpha_Q = 0$	$\alpha_R = 1$	ζ free			
		$\zeta \geq 0$			
	$\alpha_R = 0$	ζ free			
		$\zeta \geq 0$			

由于只要 $\hat{Q}, \hat{\lambda}, \hat{\zeta}$ 中有一个是无偏的，那么可以得到拉格朗日估计是无偏的。所以，似乎拉格朗日估计更加适合于 behavior-agnostic 的 Off-Policy 估计。然而，接下来的实验分析中看到，情况并非如此。相反，近似对偶解通常比近似原始解更精确。由于两者都不准确，所以拉格朗日估计在两者中都存在误差，而对偶估计则存在误差 $\hat{\rho}_\zeta$ 将表现出更健壮的表现，因为它仅仅依赖于近似的 $\hat{\zeta}$。

3.3 Recovering Existing OPE Estimators

通过对正则化项系数的设置，可以还原出目前的 DICE family。

- **DualDICE**: ($\alpha_Q = 0, \alpha_\zeta = 1, \alpha_R = 0$, without $\zeta \geq 0$ and λ)。DualDICE 还导出了一个无约束的原始形式，并且只在原始变量上进行优化。这种形式产生有偏估计，但避免了极大极小优化中的困难，是一个优化稳定性和解无偏之间的权衡。

  首先，拉格朗日问题可以化简为只和 Q 相关的函数，

$$\min_Q (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0), \\ s_0 \sim \mu_0}} \left[Q\left(s_0, a_0\right) \right] + \alpha_Q \cdot \mathbb{E}_{(s,a) \sim d^D} \left[f_1(Q(s,a)) \right] + \alpha_\zeta \cdot \mathbb{E}_{(s,a) \sim d^D} \left[f_2^* \left(\frac{1}{\alpha_\zeta} \left(\alpha_R \cdot R(s,a) + \gamma \mathcal{P}^\pi Q(s,a) - Q(s,a) \right) \right) \right] \tag{33}$$

  我们称其为无约束原始形式，因为优化现在只在原始变量上进行。这里用到了 w.r.t. ζ 的 Fenchel 共轭，且

$$\zeta^*(s,a) = f_2^{*\prime} \left(\left(\alpha_R \cdot R(s,a) + \gamma \mathcal{P}^\pi Q^*(s,a) - Q^*(s,a) \right) / \alpha_\zeta \right)$$

  虽然 uncontrained primal form 更简单，但在实践中它有一个缺点，不知道 $\mathcal{P}^\pi$。也就是说，在实践中，我们必须通过此形式，

$$\min_Q (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0), \\ s_0 \sim \mu_0}} \left[Q\left(s_0, a_0\right) \right] + \alpha_Q \cdot \mathbb{E}_{(s,a) \sim d^D} \left[f_1(Q(s,a)) \right] + \alpha_\zeta \cdot \mathbb{E}_{\substack{(s,a,r,s') \sim d^D \\ a' \sim \pi(s')}} \left[f_2^* \left(\frac{1}{\alpha_\zeta} \left(\alpha_R \cdot R(s,a) + \gamma Q(s',a') - Q(s,a) \right) \right) \right] \tag{34}$$

  这是真实目标的有偏估计，因此导致有偏解，因为对下一步样本的期望是在一个平方函数内进行的 (我们选择 f2 作为平方函数)。尽管如此，在某些情况下 (例如，在简单和离散的环境中)，这种偏差可能是为了换取更简单的优化而需要的。

- **GenDICE 和 GradientDICE**: 当 $\alpha_Q = 0, \alpha_\zeta = 0, \alpha_R = 1$ with λ)。而 GenDICE 中要求 $\zeta \geq 0$, 而 GradientDICE 中不需要。
- **DR-MWQL 和 MEL**: $\alpha_Q = 0, \alpha_\zeta = 0, \alpha_R = 1$ 和 $\alpha_Q = 1, \alpha_\zeta = 0, \alpha_R = 0$ both without $\zeta \geq 0$ and λ)。
- **LSTDQ**: 对于线性参数化 $d(s,a) = \alpha^T \phi(s,a)$ 和 $Q(s,a) = \beta^T \phi(s,a)$，任何 3.2 表中的无偏估计，都可以用拉格朗日对偶的形式推出来，至于为什么要是无偏估计呢? 因为不是如果不是无偏估计，算出来的 $\hat{Q} \neq Q^\pi$，所以导致估计出来的参数计算出的不是 Q。

  – 当 $\alpha_Q = 1, \alpha_\zeta = 0, \alpha_R = 1$。将线性形式代入公式 (8) 可得，

$$\begin{aligned}\max_v \min_w L_D(v,w) :=& (1-\gamma) \cdot w^\top \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}} \left[\phi\left(s_0, a_0\right)\right] + \alpha_Q \cdot \mathbb{E}_{(s,a) \sim d^{\mathcal{D}}} \left[f_1\left(w^\top \phi(s,a)\right)\right] \\ &+ v^\top \mathbb{E}_{\substack{(s,a,r,s') \\ a' \sim \pi(s')}} \left[\phi(s,a) \cdot \left(\alpha_R \cdot R(s,a) + \gamma w^\top \phi\left(s', a'\right) - w^\top \phi(s,a)\right)\right].\end{aligned} \tag{35}$$

  因为这里是关于 $\rho_Q(\pi)$ 的无偏估计，使用 KKT 条件可得，

$$\begin{aligned}&\mathbb{E}_{\substack{(s,a,r,s') \sim d^{\mathcal{D}}, \\ a' \sim \pi(s')}} \left[\phi(s,a) \cdot \left(\alpha_R \cdot R(s,a) + \gamma w^\top \phi\left(s', a'\right) - w^\top \phi(s,a)\right)\right] = 0 \\ \Rightarrow \quad & w = \underbrace{\mathbb{E}_{\substack{(s,a,r,s') \sim d^{\mathcal{D}} \\ a' \sim \pi(s')}} \left[\phi(s,a) \cdot \left(\phi(s,a) - \gamma \phi\left(s', a'\right)\right)\right]^{-1}}_{\Xi} \mathbb{E}_{(s,a) \sim d^{\mathcal{D}}} \left[\alpha_R \cdot R(s,a) \phi(s,a)\right] \\ \Rightarrow\ & Q^*(s,a) = w^T \phi(s,a)\end{aligned} \tag{36}$$

  所以，可以完成对 $\hat{\rho}_Q(\pi)$ 的估计，

$$\begin{aligned}\hat{\rho}_Q(\pi) &= (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}} \left[\hat{Q}\left(s_0, a_0\right)\right] \\ &= (1-\gamma) \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}} [\phi(s,a)]^\top \Xi^{-1} \mathbb{E}_{(s,a) \sim d^{\mathcal{D}}} [R \mid (s,a) \phi(s,a)]\end{aligned} \tag{37}$$

  – 当 $\alpha_Q = 0, \alpha_\zeta = 1, \alpha_R = \{0,1\}$。

$$\begin{aligned}\max_v \min_w L_D(v,w) :=& (1-\gamma) \cdot w^\top \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}} \left[\phi\left(s_0, a_0\right)\right] - \alpha_\zeta \cdot \mathbb{E}_{(s,a) \sim d^{\mathcal{D}}} \left[f_2\left(v^\top \phi(s,a)\right)\right] \\ &+ v^\top \mathbb{E}_{\substack{(s,a,r,s') \sim \mathcal{D} \\ a' \sim \pi(s')}} \left[\phi\left[\phi(s,a) \cdot \left(\alpha_R \cdot R(s,a) + \gamma w^\top \phi\left(s', a'\right) - w^\top \phi(s,a)\right)\right]\right].\end{aligned} \tag{38}$$

  同样，使用 KKT 条件，可得

$$v^\top \mathbb{E}_{\substack{(s,a,r,s') \sim d^{\mathcal{D}} \\ a' \sim \pi(s')}} \left[\phi(s,a) \cdot \left(\gamma \phi\left(s', a'\right) - \phi(s,a)\right)\right] + (1-\gamma) \cdot \mathbb{E}_{a_0 \sim \pi(s_0)} \left[\phi\left(s_0, a_0\right)\right] = 0 \tag{39}$$

  从而推导出，

$$v = (1-\gamma) \cdot \Xi^{-1} \mathbb{E}_{a_0 \sim \pi(s_0)} \left[\phi\left(s_0 \sim \mu_0, a_0\right)\right] \tag{40}$$

  于是，可以得到关于 $\hat{\rho}_\zeta(\pi)$，

$$\begin{aligned}\hat{\rho}_\zeta(\pi) &= \mathbb{E}_{(s,a,r) \sim d^{\mathcal{D}}} [R \cdot \phi(s,a)]^\top v \\ &= (1-\gamma) \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}} [\phi(s,a)]^\top \Xi^{-1} \mathbb{E}_{(s,a) \sim d^{\mathcal{D}}} [R(s,a) \phi(s,a)]\end{aligned} \tag{41}$$

- **Algae Q-LP**: ($\alpha_Q = 0, \alpha_\zeta = 1, \alpha_R = 0/1$, without $\zeta \geq 0$ and λ)，[AlgaeDICE 论文解读]。
- **BestDICE**: ($\alpha_Q = 0, \alpha_\zeta = 1, \alpha_R = 0/1$, with $\zeta \geq 0$ and λ)，文章中发现了能够获得最佳性能的 DICE 变体，这是基于文章提出的统一的框架发现的。

4 Experiments

作者大概总结了四条实验性的结论。

- 对偶估计 $\hat{\rho}_\zeta$ 在大多数的配置下是无偏的，并且在所有估计器中产生最佳的性能，并且进一步对 MDP 奖励的缩放和偏移表现出很强的鲁棒性。**所以最好的方法是用拉格朗日正则化方程求解得到 $\hat{\zeta}$，然后利用 Dual 估计器来对策略进行评估**。
- 对偶正则化项（$\alpha_\zeta > 0$）比原正则化项更好；而且当 $\alpha_R = 1$ 时效果更好一点。
- 冗余约束 (λ 和 $\zeta \geq 0$) 的加入提高了估计的稳定性和估计性能。
- 正如预期的那样，使用无约束原始形式的优化比使用极大极小正则化拉格朗日优化更稳定，但也更有偏差。

所以，基于以上的发现，作者提出的 DICE 估计器设置为：使用拉格朗日正则化表达式，对对偶估计 $\hat{\rho}_\zeta$ 进行优化，其中正则化对偶变量 $\alpha_\zeta > 0, \alpha_R = 1$，冗余约束 $(\lambda, \zeta \geq 0)$。大部分实验都非常好理解，这里只挑两个我感觉有意思的分析一下，其他的大家读论文就 OK 了，非常的容易理解。

4.1 Choice of Estimator ($\hat{\rho}_Q, \hat{\rho}_\zeta, \hat{\rho}_{Q,\zeta}$)

我们发现对偶估计器在不同的任务和行为策略中都一致地产生最好的估计。相比之下，原始估计要差得多。虽然拉格朗日估计比原估计有改进，但它通常比对偶估计具有更高的方差。据推测，拉格朗日量并不能从双重鲁棒性中获得理想的效果，因为这两个解决方案在这个实际设置中都有偏差。

为了更完善地评估双重估计器，我们研究了当奖励函数按一个常数缩放、移动一个常数或取幂时的表现。为控制优化的困难，我们首先将原变量和对偶变量参数化为线性函数，并利用随机梯度下降法求解 (8) 中 $\alpha_Q = 0, \alpha_\zeta = 1, \alpha_R = 1$ 的凸凹极大极小目标。

当使用原始奖励时，Prime、Dual 和拉格朗日估计最终收敛到大致相同的值 (尽管原始估计收敛得更慢)。当奖励的比例是 10 或 100 倍，或移动了 5 或 10 个单位时，Prime 估计结果便会受到严重影响，并且在相同次数的梯度更新中也不会收敛。当使用神经网络参数化进行同样的评估时，Dual 估计继续表现出对奖励转换的敏感性，而 Dual 估计在转换回原始奖励后保持大致相同。我们进一步实现了训练原变量稳定性的目标网络，得到了相同的结论。

4.2 Choice of Optimization (Lagrangian or Unconstrained Primal Form)

到目前为止，实验使用极小极大优化通过拉格朗日学习原始和对偶变量。我们现在考虑解决 d-LP 的无约束原始形式，理论上其可以令优化更加稳定。而实验结果表示确实是这样的，无约束原始优化减少了 Grid 上的方差，并在 Cartpole 上产生了更好的估计。另一方面，Reacher 有一个连续的行动空间，这在对下一步样本进行期望时造成了困难，导致了无约束原始形式的偏差。基于实验结果，我

们通常主张拉格朗日，除非任务是离散的动作，且动力学的随机性是非常的低（这在显示中几乎不可能）。

5 总结

本文通过 d-LP 的正则化拉格朗日方程提出了 OPE 的统一框架。在这种框架下，通过框架中正则化器的特定 (次优) 选择、(冗余) 约束的选择和将优化解转换为策略值的方法选择，可以推导出各种不同的 DICE 算法。通过系统地研究这些不同配置选择下的框架优化的数学性质和实验效果，我们发现，与原始估计 (即估计 Q 值) 进行比较，对偶估计 (用状态动作分布 $d(s,a)$ 来估计 policy value) 在保证估计的无偏性方面更加的优秀，并且在优化框架中更加的灵活，也更加的鲁棒，在不同的配置下都取得了较好的效果。本文的研究还揭示了在文献中没有发现的，但表现优越性能的 OPE 评估器。这些发现为 offline 环境下的 OPE 研究提出了新方向。