Reinforcement Learning via Fenchel-Rockafellar Duality

Chen Gong

20 April 2021

目录

1 Core idea

本篇文章是 DAI Bo 老师和 Ofir Nachum 合作的文章。文章中总结了这种对偶性如何应用于各种强化学习 (RL) 设置，包括策略评估或优化，在线或离线学习，以及折扣或无折扣奖励。在强化学习中使用对偶方法，产生了很多非常有意思的 idea。包括使用与行为无关的离线数据来进行策略评估和求解 on-policy 策略梯度，通过最大似然优化来求解策略。文章中，对这些方法进行汇总，提供自己的观点，希望将使研究人员能够更好地使用和应用凸共轭工具，以在 RL 中取得进一步的进展。

作者认为强化学习中有两个主要的困难：1. 决策问题是一个序列，任何早期的决定都会或多或少的影响后续的状态。2. 我们不知道环境的更多信息，只能从环境中进行采样。

本文中，总结概括了线性规划（LP），描述了一些关于 RL 的凸问题，比如将 RL 问题描述为一个求解具有线性约束的目标凸函数。也许这种推广最有用的性质是，当原始问题涉及一个严格的凸目标时，Fenchel-Rockafellar 对偶的应用，可以引出了一个无约束的对偶问题。

对偶的理论非常的漂亮，对偶有什么用呢?

2 凸共轭背景

对偶性是优化和机器学习的一个基本和强大的工具，特别是在凸分析领域，使得我们容易使用更容易处理的替代方法来重新规划优化问题。在本节中，我们将简要概述几个关键的凸对偶结果，这些结果将在后面的 RL 算法中发挥重要作用。

2.1 Fenchel Conjugate

对于函数 $f : \Omega \to \mathbb{R}$ 的共轭函数 f_*：

$$f_*(y) := \max_{x\in\Omega} \langle x, y\rangle - f(x) \tag{1}$$

Definition 1：如果函数 $f : \{x \in \Omega : f(x) < \infty\}$ 是非空的，并且，对于 $\forall x \in \Omega : f(x) > \infty$，我们称函数 f 是适定的。

Definition 2：下半连续：对于适定，凸，下半连续的函数 f，其共轭函数 f_* 也是适定的，凸，下半连续的函数。并且具有对偶性 $f_{**} = f$；

$$f(x) = \max_{y\in\Omega^*} \langle x, y\rangle - f_*(y) \tag{2}$$

后文中，没有特殊说明，f 即为凸函数。而且作者假设，文章中所有提到的凸函数都是适定和下半连续的。下表中描述了一系列普通函数和其凸函数。

2.1.1 指示函数

指示函数 $\delta_C(x)$ 定义为：

$$\delta_C(x) := \begin{cases} 0 & x \in C \\ \infty & \text{otherwise} \end{cases}$$

如果 C 是闭凸集，$\delta_C(x)$ 是适定的，凸，下半连续的函数。指示函数非常强大，**可以将有约束的优化问题，转化为无约束优化问题**。比如，优化问题 $\min_{Ax=0} f(x)$ 可以等价的表达为 $\min_x f(x) + \delta_{\{0\}}(Ax)$。且 $\delta_{\{a\}}(x)$ 的共轭函数是线性函数，反之亦然。

Function	Conjugate	Notes
$\frac{1}{2}x^2$	$\frac{1}{2}y^2$	
$\frac{1}{p}\|x\|^p$	$\frac{1}{q}\|y\|^q$	For $p, q > 0$ and $\frac{1}{p}+\frac{1}{q}=1$.
$\delta_{\{a\}}(x)$	$\langle a, y\rangle$	$\delta_C(x)$ is 0 if $x \in C$ and ∞ otherwise.
$\delta_{\mathbb{R}_+}(x)$	$\delta_{\mathbb{R}_-}(y)$	$\mathbb{R}_\pm := \{x \in \mathbb{R} \mid \pm x \geq 0\}$.
$\langle a, x\rangle + b \cdot f(x)$	$b \cdot f_*\left(\frac{y-a}{b}\right)$	
$D_f(x\\|p)$	$\mathbb{E}_{z\sim p}[f_*(y(z))]$	For $x : \mathcal{Z} \to \mathbb{R}$ and p a distribution over $\mathcal{Z}$.
$D_{\mathrm{KL}}(x\\|p)$	$\log \mathbb{E}_{z\sim p}[\exp y(z)]$	For $x \in \Delta(\mathcal{Z})$, *i.e.*, a normalized distribution over $\mathcal{Z}$.

图 1

2.1.2 f 散度

对于凸函数 f 和定义域 $\mathcal{Z}$ 上的分布 p，f 散度被定义为:

$$D_f(x\|p) = \mathbb{E}_{z\sim p}\left[f\left(\frac{x(z)}{p(z)}\right)\right].$$

如果，定义域是无限实值函数的集合，则共轭 $D_f(x\|p)$ at $y : \mathcal{Z} \to \mathbb{R}$，通过 interchangeability principle 可以写成:

$$\begin{aligned} g(y) &= \max_{x:\mathcal{Z}\to\mathbb{R}} \sum_z x(z)\cdot y(z) - \mathbb{E}_{z\sim p}[f(x(z)/p(z))] \\ &= \mathbb{E}_{z\sim p}\left[\max_{x(z)\in\mathbb{R}} x(z)\cdot y(z)/p(z) - f(x(z)/p(z))\right] \\ &= \mathbb{E}_{z\sim p}\left[f_*(y(z))\right] \end{aligned}$$

Lemma 1: Interchangeability principle ξ 为 Ξ 中的随机变量，有 $\xi \in \Xi$。函数 $g(\cdot, \xi) : \mathbb{R} \to (-\infty, +\infty)$ 是适定和上半连续的凹函数（也可以是下半连续的凸函数）。这时有，

$$\mathbb{E}_\xi\left[\max_{u\in\mathbb{R}} g(u, \xi)\right] = \max_{u(\cdot)\in\mathcal{G}(\Xi)} \mathbb{E}_\xi[g(u(\xi), \xi)] \tag{3}$$

其中，$\mathcal{G}(\Xi) = \{u(\cdot) : \Xi \to \mathbb{R}\}$。

所以，可以看到其共轭函数的具体性质和 f 的选择有很大的关系。KL 散度是 f 散度一种特殊形式。

2.2 Fenchel-Rockafellar Duality

对于原问题，

$$\min_{x\in\Omega} J_{\mathrm{P}}(x) := f(x) + g(Ax) \tag{4}$$

其中，$f, g \to \Omega, \mathbb{R}$ 是凸且下半连续的函数。A 是线性算子（比如矩阵运算），其对偶问题为:

$$\max_{y\in\Omega^*} J_{\mathrm{D}} := -f_*\left(-A_* y\right) - g_*(y) \tag{5}$$

其中，A_* 表示 A 矩阵的伴随线性算子，满足 $\langle y, Ax\rangle = \langle A_* y, x\rangle$。通常情况下，当 A 简单的表示为实值矩阵，A_* 为 A 的转置。

而此处的对偶问题是这样推导来的。

Lemma 2: Fenchel Inequation：$f : \mathbb{R}^n \to \overline{\mathbb{R}}$ 是适定凸函数，那么

$$\langle x, x^*\rangle \leq f(x) + f^*(x^*) \tag{6}$$

这是根据共轭函数的定义来的。然后，利用 Fenchel 不等式可得，

$$\begin{gathered} f(x)+f^*\left(-A^* x^*\right)+g(A x)+g^*\left(x^*\right) \geqslant\left\langle x,-A^* x^*\right\rangle+\left\langle A x, x^*\right\rangle=0 \\ f(x)+f^*\left(-A^* x^*\right) \geqslant-g(A x)-g^*\left(x^*\right) \end{gathered}$$

根据对偶问题的推导是，

$$\begin{aligned} \min_{x\in\Omega} f(x)+g(Ax) &= \min_{x\in\Omega}\max_{y\in\Omega^*} f(x)+\langle y, Ax\rangle - g_*(y) \\ &= \max_{y\in\Omega^*}\left\{\min_{x\in\Omega} f(x)+\langle y, Ax\rangle\right\} - g_*(y) \\ &= \max_{y\in\Omega^*}\left\{-\max_{x\in\Omega}\langle -A_* y, x\rangle - f(x)\right\} - g_*(y) \\ &= \max_{y\in\Omega^*} -f_*\left(-A_* y\right) - g_*(y) \end{aligned} \tag{7}$$

所以，我们得到了原问题的对偶形式：

$$\min_{x\in\Omega} J_{\mathrm{P}}(x) = \min_{y\in\Omega^*} J_{\mathrm{D}}(x) \tag{8}$$

对偶问题的解为：$y^* := \arg\max_y J_{\mathrm{D}}(y)$。可以用来寻找原问题的解。如果 f_*' 是唯一的，$x^* = f_*'(-A_* y^*)$ 是原问题的解。更加一般的说可以用 $x^* \in \partial f_*\left(-A_* y^*\right) \cap A^{-1}\partial g_*\left(y^*\right)$ 来表示所有原问题的解的集合。

2.2.1 The Largrangian

Fenchel-Rockafellar 对偶可以用来推导 Lagrangian 对偶。考虑约束优化问题，

$$\min_x f(x) \quad \text{s.t.} \quad Ax \geq b. \tag{9}$$

原问题可以用指示函数来改写为：$\min_x f(x) + g(Ax)$，其中 $g(Ax) = \delta_{\mathbb{R}\_}(-Ax+b)$，将其转换为无约束优化问题。其 Fenchel-Rockafellar 对偶形式为：

$$\max_y \langle y, b\rangle - f_*(A_* y) \quad s.t. \quad y \geq 0 \tag{10}$$

这里的 $\langle y, b\rangle$ 这里用到了平移定理。将 $g^*(y) = \langle y, b\rangle + g*$，其中 $g : -y + b \leq 0$，其对偶形式为 $y \geq 0$。

定理 3.3.8

$f \in \Gamma(\mathbb{R}^n)$。记$\tau_a : x \to x - a$为平移算子，那么

$$(f \circ \tau_a)^* = \langle \cdot, a \rangle + f^*, \tag{3.6}$$

$$(f + \langle \cdot, a \rangle)^* = f^* \circ \tau_a. \tag{3.7}$$

证明 直接计算，对任意x^*有

$$\begin{aligned}(f \circ \tau_a)^*(x^*) &= \sup_x \{\langle x, x^* \rangle - f(x-a)\} \\ &= \sup_x \{\langle a, x^* \rangle + \langle x - a, x^* \rangle - f(x-a)\} \\ &= \langle x^*, a \rangle + \sup_x \{\langle x, x^* \rangle + f(x)\} \\ &= \langle x^*, a \rangle + f^*(x^*).\end{aligned}$$

图 2: 平移定理及其证明

其中，f_* 是 Fenchel conjugate 的后一项，结合公式 (9) 和公式 (10)，研究的问题可以写为：

$$\min_x \max_{y \geq 0} \langle y, b \rangle - \langle x, A_* y \rangle + f(x) \tag{11}$$

其中，$\langle y, Ax \rangle = \langle x, A_* y \rangle$，那么公式 (11) 可以被表达为：

$$\min_x \max_{y \geq 0} \underbrace{\langle y, b - Ax \rangle + f(x)}_{L(x,y)}. \tag{12}$$

其中，$L(x, y)$ 就是公式 (9) 的 Lagrangian 原问题，以此，我们可以进一步推导出大名鼎鼎的 Lagrange 对偶：

$$\max_{y \geq 0} \min_x L(x, y) = \min_x \max_{y \geq 0} L(x, y). \tag{13}$$

2.2.2 LP Duality

Frenchel-Rockafellar 对偶也可以推广到著名的 Linear Programming (LP) 对偶中。如果考虑函数 $f(x) = \langle c, x \rangle + \delta_{\mathbb{R}_+}(x)$ 和 $g(x) = \delta_{\{b\}}(x)$。其原问题和对偶问题分别为：

$$\begin{aligned}&\min\nolimits_{x \geq 0} \langle c, x \rangle \quad && \text{s.t.} \quad && Ax = b, \\ &\max\nolimits_y -\langle b, y \rangle \quad && \text{s.t.} \quad && A_* y \geq -c,\end{aligned} \tag{14}$$

将 $y \to -y$，对偶式子可以等价的表达为：

$$\max_y \langle b, y \rangle \quad \text{s.t.} \quad A_* y \leq c, \tag{15}$$

因此，Fenchel-Rockafellar 对偶为我们提供了较强的 LP 对偶性质。即如果原问题是可解的，则其结果与对偶问题的结果相同。

3 强化学习背景

这里默认阅读此文章的有一定的强化学习基础，就不多说了。整篇文章的逻辑可以用下图表示。接下来将以这幅图为主要框架介绍本文。

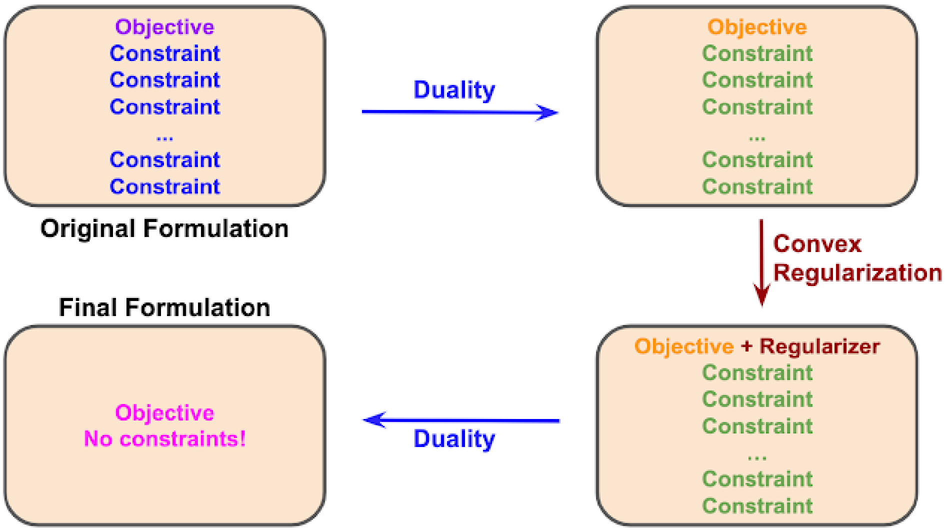

图 3: 全文逻辑

3.1 Policy Evaluation and Optimization

策略评估公式为：

$$\rho(\pi) := (1-\gamma) \cdot \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 \sim \beta, \forall t, a_t \sim \pi\left(s_t\right), r_t \sim R\left(s_t, a_t\right), s_{t+1} \sim T\left(s_t, a_t\right)\right] \tag{16}$$

策略评估是评价策略 π 怎么样，而策略优化问题则是寻找是 $\rho(\pi)$ 最大的 π^*。如果，奖励函数 R 和策略 π 之间是独立的，那么可以用确定性的策略来表示 π^*。如果使用熵正则化的形式来表示奖励，则 π^* 只能是随机策略，其中 $\tilde{R}(s,a) = R(s,a) - \log \pi(a|s)$。

3.2 Online vs Offline RL

在策略评估和策略优化中，一个重要的难点就是不知道环境的显示信息，比如不知道函数 R, T, μ_0 是什么。相反，对环境的访问是以经验的形式给予的 $s_t, a_t, r_t, s_{t+1}, a_{t+1}, r_{t+1} \cdots$，轨迹是通过与环境的互动收集。通过收集经验的形式，可以将 RL 算法分为 online 和 offline 学习。

online 学习的设定中，环境中的经验可以通过蒙特卡罗算法收集。有了这种对环境的访问，政策评估和优化问题就可以很容易地解决。但是，蒙特卡罗采样的效率太低了，需要随机的从环境中采样。Online 学习经常困于采样效率低的问题，所以，如何尽可能少的采样而得到近似的策略评估和策略优化的近似解，非常重要。

实际应用中，在训练时和环境交互来收集数据是比较理想的。根据一般的情况是，交互的是 offline 环境。即为交互的环境被限制在一个静态的数据集中：$\mathcal{D} = \{(s^{(i)}, a^{(i)}, r^{(i)}, s^{(i)\prime})\}_{i=1}^{N}$，其中 $(s^{(i)}, a^{(i)}) \sim$

$d^{\mathcal{D}}$,$\mathcal{D}$ 是未知的动作状态对分布,$r^{(i)} = R(s^{(i)}, a^{(i)})$,$s^{(i)'} \sim T(s^{(i)}, a^{(i)})$,初始状态为$\mathcal{U} = \{s_0^{(i)}\}_{i=1}^{M} \sim u_0$。**通常情况下，我们假设轨迹长度是无限的，这样可以用 $d^{\mathcal{D}}, T, \mu_0$ 来表示期望**。在 offline RL 中最大的挑战是，采样策略和 target 策略之间差距较大而造成的分布偏移。同时注意，offline 学习一定是 Off-Policy 学习，而 Off-Policy 学习不一定是 offline。

3.3 Q-values and State-Action Visitations

对于策略 π，Q 值被定义为从 (s,a) 开始，使用策略 π 将累加获得的折扣奖励：

$$Q^{\pi}(s,a) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{t} R\left(s_{t}, a_{t}\right) \mid s_{0}=s, a_{0}=a, \forall t>0, a_{t} \sim \pi\left(s_{t}\right), s_{t} \sim T\left(s_{t-1}, a_{t-1}\right)\right] \tag{17}$$

Q 值满足单步的贝尔曼迭代：

$$Q^{\pi}(s,a) = R(s,a) + \gamma \cdot \mathcal{P}^{\pi} Q^{\pi}(s,a), \tag{18}$$

其中，$\mathcal{P}^{\pi}$ 表示为策略转移算子：

$$\mathcal{P}^{\pi} Q(s,a) = \mathbb{E}_{s' \sim T(s,a), a' \sim \pi(s')}[Q(s', a')] \tag{19}$$

state-action visitations: d^{π}（有时也被称为 occupancies 或者 density，这里翻译成中文，感觉都很奇怪，就不翻译了。）

$$d^{\pi}(s,a) := (1-\gamma) \sum_{t=0}^{\infty} \gamma^{t} \operatorname{Pr}\left(s_{t}=s, a_{t}=a \mid s_{0} \sim \mu_{0}, \forall t, a_{t} \sim \pi\left(s_{t}\right), s_{t+1} \sim T\left(s_{t}, a_{t}\right)\right) \tag{20}$$

其中，$d^{\pi}(s,a)$ 表示 π 和马尔可夫决策过程 $\mathcal{M}$ 交互过程中，遇到 (s,a) 的概率。和 Q-values 类似，visitations 也满足单步的 Bellman 转移：

$$d^{\pi}(s,a) = (1-\gamma)\mu_0(s)\pi(a|s) + \gamma \cdot \mathcal{P}_{*}^{\pi} d^{\pi}(s,a), \tag{21}$$

其中，

$$\mathcal{P}_{*}^{\pi} d^{\pi}(s,a) := \pi(a|s) \sum_{\tilde{s},\tilde{a}} T(s|\tilde{s},\tilde{a}) d(\tilde{s},\tilde{a}) \tag{22}$$

关于此部分的详细解答,已在 [DualDICE 论文解读] 的 2.1.1 和 2.1.2 小节中做了详细的描述。Q-values 和 Visitations 在强化学习中都发挥着较大的作用，比如，可以用来表示策略评估：

$$\rho(\pi) = (1-\gamma) \cdot \mathbb{E}_{a_0 \sim \pi(s_0)}\left[Q^{\pi}\left(s_0, a_0\right)\right] = \mathbb{E}_{(s,a) \sim d^{\pi}}[R(s,a)] \tag{23}$$

同样策略梯度也可以用 Q-values 和 Visitation 表示：

$$\frac{\partial}{\partial \pi} \rho(\pi) = \mathbb{E}_{(s,a) \sim d^{\pi}}\left[Q^{\pi}(s,a) \nabla \log \pi(a \mid s)\right] \tag{24}$$

接下来主要问题即为如何估计 $Q(s,a)$ 和 d^{π}。

4 Policy Evaluation

接下来则是描述将 Fenchel-Rockafellar 对偶应用与强化学习的策略估计中。

4.1 The Linear Programming From of Q

观察公式 (23) 可得，$\rho_(\pi)$ 在 Q^π 或 d^π 上的等价表达式暗示了对偶性，这可以通过下面的具有 LP 性质的 $\rho(\pi)$ 刻画得到，称为 Q-LP:

$$\begin{aligned} \rho(\pi) = & \min_Q (1-\gamma) \cdot \mathbb{E}_{a_0 \sim \pi(s_0)} \left[Q\left(s_0, a_0\right) \right] \\ & \text{s.t. } Q(s,a) \geq R(s,a) + \gamma \cdot \mathcal{P}^\pi Q(s,a), \quad \forall (s,a) \in S \times A \end{aligned} \tag{25}$$

其中，LP 的最优结果满足 $Q^*(s,a) = Q^\pi(s,a)$。

$$\begin{aligned} \rho(\pi) = & \max_{d \geq 0} \sum_{s,a} d(s,a) \cdot R(s,a) \\ & \text{s.t. } d(s,a) = (1-\gamma)\mu_0(s)\pi(a \mid s) + \gamma \cdot \mathcal{P}_*^\pi d(s,a), \forall s \in S, a \in A \end{aligned} \tag{26}$$

而且，公式 (26) 中的 $|\mathcal{S}| \times |\mathcal{A}|$ 个等式约束独立决定 d 的取值，并不需要考虑公式 (26) 中的目标函数。详细的推导要参考 AlgaeDICE 那篇文章了，https://arxiv.org/abs/1912.02074。这实际是第一步，即为策略评估的部分。

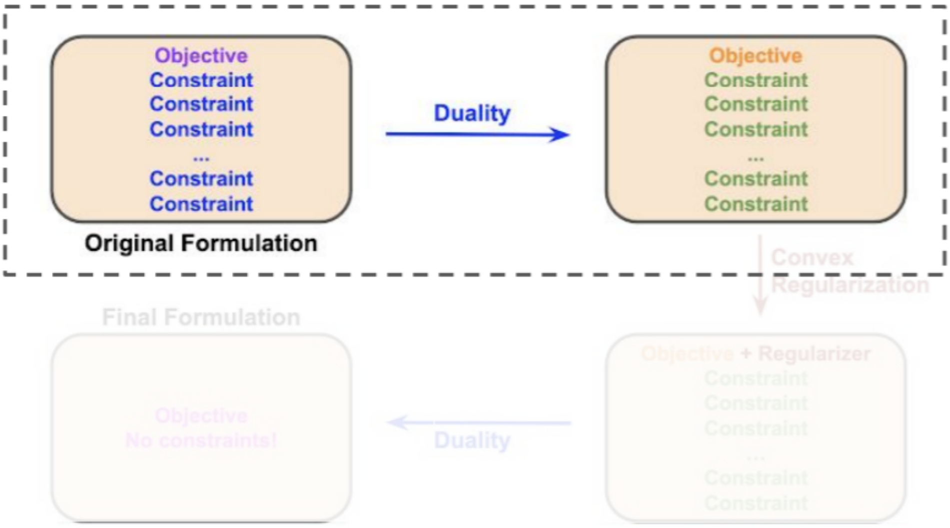

图 4: 策略评估

4.2 Policy Evaluation via the Lagrangian

在 Q-LP 的原始或对偶形式中潜在的大量约束，这给估计 $\rho(\pi)$ 带来了挑战。文章中从无约束优化的角度介绍了一种计算更加简单的方法，此方法使用 Q-LP 的 Lagrangian 来解决策略评估问题(**肯定有小伙伴对这个公式怎么来的一头雾水 1 此处需要参考公式 (12)，大家就知道为什么这里的 Lagrangian 系数是** $d(s,a)$ **了**。):

$$\rho(\pi) = \min_Q \max_{d \geq 0} (1-\gamma) \cdot \mathbb{E}_{a_0 \sim \pi(s_0)} \left[Q\left(s_0, a_0\right) \right] + \sum_{s,a} d(s,a) \cdot \left(R(s,a) + \gamma \cdot \mathcal{P}^\pi Q(s,a) - Q(s,a) \right) \tag{27}$$

实际中，$|\mathcal{S}| \times |\mathcal{A}|$ 可能是无限的。所以，公式 (27) 中涉及到对全空间的求和，实际非常困难。而且，在 offline 的 setting 中，我们只能访问分布 $d^\mathcal{D}$，所以需要用重要性采样来做变量替换，$\zeta(s,a) = \frac{d(s,a)}{d^\mathcal{D}(s,a)}$。

那么可以将公式 (27) 重写为：

$$
\begin{aligned}
& \min_{Q} \max_{\zeta \geq 0} L(Q, \zeta) \\
:= & (1-\gamma) \cdot \mathbb{E}_{a_0 \sim \pi(s_0)}\left[Q\left(s_0, a_0\right)\right]+\mathbb{E}_{(s, a) \sim d^{\mathcal{D}}}\left[\zeta(s, a) \cdot\left(R(s, a)+\gamma \cdot \mathcal{P}^\pi Q(s, a)-Q(s, a)\right)\right] \\
= & (1-\gamma) \cdot \mathbb{E}_{a_0 \sim \pi(s_0)}\left[Q\left(s_0, a_0\right)\right]+\mathbb{E}_{\substack{(s, a, s^{\prime}) \sim d^{\mathcal{D}} \\ a^{\prime} \sim \pi(s^{\prime})}}\left[\zeta(s, a) \cdot\left(R(s, a)+\gamma Q\left(s^{\prime}, a^{\prime}\right)-Q(s, a)\right)\right] .
\end{aligned} \tag{28}
$$

并且，最优化的结果满足：$\zeta^*=\frac{d^\pi(s,a)}{d^{\mathcal{D}}(s,a)}$。那么，如果需要估计 $\rho(\pi)$，计算得出 $L(\hat{Q}^*,\hat{\zeta}^*)$ 即可。并且，其有一个 doubly robust 性质，

$$
L(Q,\zeta^*)=L(Q^*,\zeta)=L(Q^*,\zeta^*)=\rho(\pi) \tag{29}
$$

因此，这样有一个好处，至少对 Q,ζ 中的一个变量是鲁棒的，比如 Q 取最佳的时候，ζ 的性能差一点没有关系。但是，由于目标函数和约束条件都是线性的，会造成优化过程中收敛不稳定的问题，下面将来介绍解决方法。

4.3 Changing the Problem Before Applying Duality

之前提到了，公式 (26) 中有 $|\mathcal{S}|\times|\mathcal{A}|$ 个等式约束，根据等式约束就可以直接求解得到 $d(s,a)$ 了，而不需要改目标函数的形式，这里存在 over-constrain 问题。那么，可以用新的目标函数来代替原目标函数，$\max_d -h(d)$，且 h 函数不会改变最优结果 $d^*=d^\pi$。**并且，因为不能和环境直接进行交互，不能直接得到公式中的** $R(s,a)$。因此，最近许多工作的主要思想是选择一个适当的 h 函数，从而使这个问题的拉格朗日或 Fenchel-Rockafellar 对偶更容易接近，并可能避免与原始 LP 相关的不稳定性问题。

尽管，优化问题发生了改变，但是这并不影响问题的求解。而且，如果目标函数重写为：$\zeta(s,a)=\frac{d(s,a)}{d^{\mathcal{D}}(s,a)}$。在问题求解的过程中，通过近似解 $\hat{\zeta}^*$，可以对 π 进行策略评估。

$$
\rho(\pi)=\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[\hat{\zeta}^*(s,a)\cdot R(s,a)] \tag{30}
$$

4.3.1 Constant Function $h(d):=0$

如果令 $h(d)=0$，公式 (28) 中的优化问题，转变为：

$$
\min_{Q} \max_{\zeta} L(Q, \zeta)=(1-\gamma) \cdot \mathbb{E}_{a_0 \sim \pi(s_0)}\left[Q\left(s_0, a_0\right)\right]+\mathbb{E}_{\substack{(s, a, s^{\prime}) \sim d^{\mathcal{D}} \\ a^{\prime} \sim \pi(s^{\prime})}}\left[\zeta(s, a) \cdot\left(\gamma Q\left(s^{\prime}, a^{\prime}\right)-Q(s, a)\right)\right] \tag{31}
$$

优化问题的最优结果 $\zeta^*=\frac{d^\pi}{d^{\mathcal{D}}}$，并且得到了 ζ^* 的近似解，就可以拿来求解公式 (30) 中的 $\rho(\pi)$ 了。不同于之前公式 (28) 中的等式，优化公式中不包含有关 Q 函数的奖励，在实践中会更好优化。并且，此处关于 Q 和 ζ 的 Lagrangian 表达式都是线性的。这可以选择 h 的严格凸形式来修正，例如，使用 f-散度。

4.3.2 $f-$Divergence $h(d):=D_f(d\|d^{\mathcal{D}})$

目标函数使用 f 散度，可以引出一系列 Off-Policy 评估方法，并且在最近的文章，比如 DualDICE 中都有概述。具体地说，DualDICE 提出的各种估计方法，分别对应于将拉格朗日或 Fenchel-Rockafellar 对偶性应用于优化问题，

$$
\begin{aligned}
& \max_d -D_f\left(d \| d^{\mathcal{D}}\right) \\
& \text{s.t. } d(s, a)=(1-\gamma) \mu_0(s) \pi(a \mid s)+\gamma \cdot \mathcal{P}_*^\pi d(s, a) \quad \forall s \in S, a \in A
\end{aligned} \tag{32}
$$

此处是对目标函数做修改，

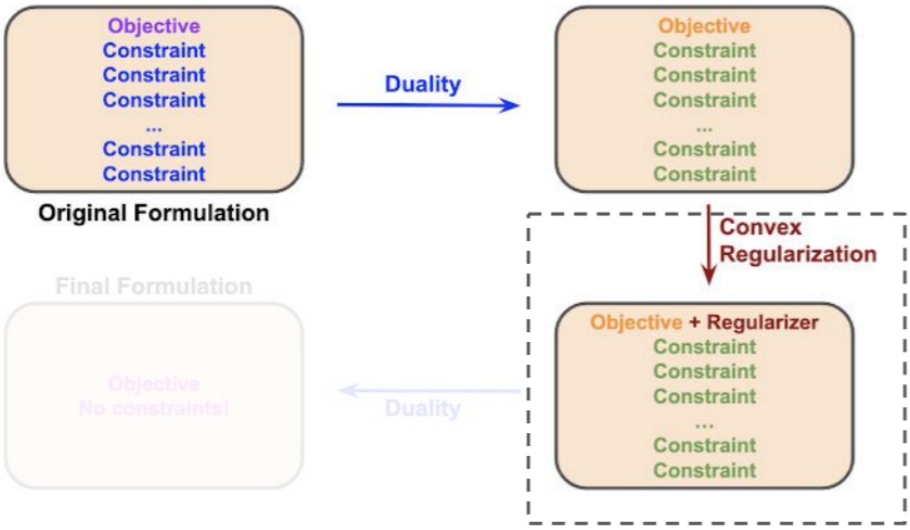

图 5: 修改目标函数

Lagrangian Duality: 对上述问题采用 Lagrangian 对偶可得，

$$\begin{aligned}\max_d \min_Q L(Q,d) &:= -D_f\left(d \| d^{\mathcal{D}}\right)+\sum_{s,a} Q(s,a)\cdot\left((1-\gamma)\mu_0(s)\pi(a\mid s)+\gamma\cdot\mathcal{P}_*^\pi d(s,a)-d(s,a)\right)\\ &=(1-\gamma)\cdot\mathbb{E}_{a_0\sim\pi(s_0)}\left[Q\left(s_0,a_0\right)\right]-D_f\left(d\|d^{\mathcal{D}}\right)+\sum_{s,a} Q(s,a)\cdot\left(\gamma\cdot\mathcal{P}_*^\pi d(s,a)-d(s,a)\right)\end{aligned} \tag{33}$$

由于 $\mathcal{P}_*^\pi$ 实际上就是 $\mathcal{P}^\pi$ 的转置矩阵，有 $\langle y, Ax\rangle=\langle x, A_* y\rangle$，可得

$$L(Q,d)=(1-\gamma)\cdot\mathbb{E}_{a_0\sim\pi(s_0)}\left[Q\left(s_0,a_0\right)\right]-D_f\left(d\|d^{\mathcal{D}}\right)+\sum_{s,a} d(s,a)\cdot\left(\gamma\cdot\mathcal{P}^\pi Q(s,a)-Q(s,a)\right) \tag{34}$$

在此等式中进行变量替换 $\zeta(s,a)=\frac{d(s,a)}{d^{\mathcal{D}}(s,a)}$，可得

$$\begin{aligned}&\max_\zeta \min_Q L(Q,\zeta)\\ :=&(1-\gamma)\cdot\mathbb{E}_{a_0\sim\pi(s_0)}\left[Q\left(s_0,a_0\right)\right]-\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[f(\zeta(s,a))]+\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[\zeta(s,a)\cdot\left(\gamma\cdot\mathcal{P}^\pi Q(s,a)-Q(s,a)\right)\right]\\ =&(1-\gamma)\cdot\mathbb{E}_{a_0\sim\pi(s_0)}\left[Q\left(s_0,a_0\right)\right]+\mathbb{E}_{(s,a,s')\sim d^{\mathcal{D}}}\left[\zeta(s,a)\cdot\left(\gamma\cdot\mathcal{P}^\pi Q(s,a)-Q(s,a)\right)-f(\zeta(s,a))\right].\end{aligned} \tag{35}$$

这样，就从 Lagrangian 的角度推出了 DualDICE 的鞍点形式。

Fenchel-Rockafellar Duality: 同样可以从 Fenchel-Rockafellar 对偶的角度来思考，公式 (32) 可以用 Fenchel-Rockafellar 对偶的形式重写。前面详细分析过了，任何的有约束优化问题，可以用冲击函数，将其转换为无约束优化问题。这里采用同样的技巧，将公式 (32) 进行重写，

$$\max_d -g(-Ad)-h(d) \tag{36}$$

其中，

$$g:=\delta_{\{(1-\gamma)\mu_0\times\pi\}} \text{ and } A:=\gamma\cdot\mathcal{P}_*^\pi-I \tag{37}$$

如果使用 Fenchel-Rockafellar 对偶，线性算子 A 的伴随算子为 $A_*:=\gamma\cdot\mathcal{P}^\pi-I$。其中，$h$ 的共轭函数为 $h_*(\cdot)=\mathbb{E}_{d^{\mathcal{D}}}[f_*(\cdot)]$。同时 g 函数的共轭函数为：$g_*(\cdot)=(1-\gamma)\mathbb{E}_{\mu_0\times\pi}[\cdot]$。所以，原问题的对偶问题，被写为，

$$\min_Q g_*(Q)+h_*\left(A_*Q\right)=\min_Q(1-\gamma)\cdot\mathbb{E}_{a_0\sim\pi(s_0)}\left[Q\left(s_0,a_0\right)\right]+\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[f_*\left(\gamma\cdot\mathcal{P}^\pi Q(s,a)-Q(s,a)\right)\right]. \tag{38}$$

看到此表达式，非常的干净。首先没有和 d^π 相关的变量，可以非常好的适用于 offline-setting。然后，没有约束等式，可以采用基于梯度下降的方法找到 Q^*。并且对于最优解 Q^*，可以利用 Fenchel-Rockafellar 对偶，复原出对偶变量的最优解，

$$\begin{aligned} d^{\mathcal{D}}(s,a)\cdot f_*'\left(\gamma\cdot\mathcal{P}^\pi Q^*(s,a)-Q^*(s,a)\right)&=d^*(s,a),\\ f_*'\left(\gamma\cdot\mathcal{P}^\pi Q^*(s,a)-Q^*(s,a)\right)&=\frac{d^\pi(s,a)}{d^{\mathcal{D}}(s,a)}\end{aligned} \tag{39}$$

为什么可以得出这个，我相信很多小伙伴都看得一脸懵逼，实际上这里省略了一步，$(1-\gamma)\cdot\mathbb{E}_{a_0\sim\pi(s_0)}\left[Q\left(s_0,a_0\right)\right]=\mathbb{E}_{(s,a)\sim d^\pi(s,a)}[\gamma\cdot\mathcal{P}^\pi Q(s,a)-Q(s,a)]$，详细内容请看 [DualDICE 论文解读]。如果令 $f(x)=\frac{1}{2}x^2$，可以推出，

$$\begin{aligned} Q^*=\arg\min_Q(1-\gamma)\cdot\mathbb{E}_{a_0\sim\pi(s_0)}\left[Q\left(s_0,a_0\right)\right]+\frac{1}{2}\mathbb{E}_{(s,a)\sim d^{D}}\left[\left(\gamma\cdot\mathcal{P}^\pi Q(s,a)-Q(s,a)\right)^2\right]\\ \Rightarrow\gamma\cdot\mathcal{P}^\pi Q^*(s,a)-Q^*(s,a)=\frac{d^\pi(s,a)}{d^D(s,a)},\quad\forall s\in S,a\in A.\end{aligned} \tag{40}$$

也就是说，如果通过最小化 Bellman 残差的平方来得到 Q^*，同时最小化初始状态 Q 值，那么最优 Bellman 残差就是 on-policy 和 offline 的 state-action 分布之间的密度比。有趣的是，在 DualDICE 的原文中没有显式的使用 Lagrangian 或者 Fenchel Rockafellar 来进行推导，而是使用了一种变量替换的方法，所以，有时也被称为 DualDICE trick。但是其本质上也是利用 $\mathcal{P}^\pi$ 和 $\mathcal{P}_*^\pi$ 之间的关系，也是 Fenchel-Rockafellar 对偶的另一种应用形式。

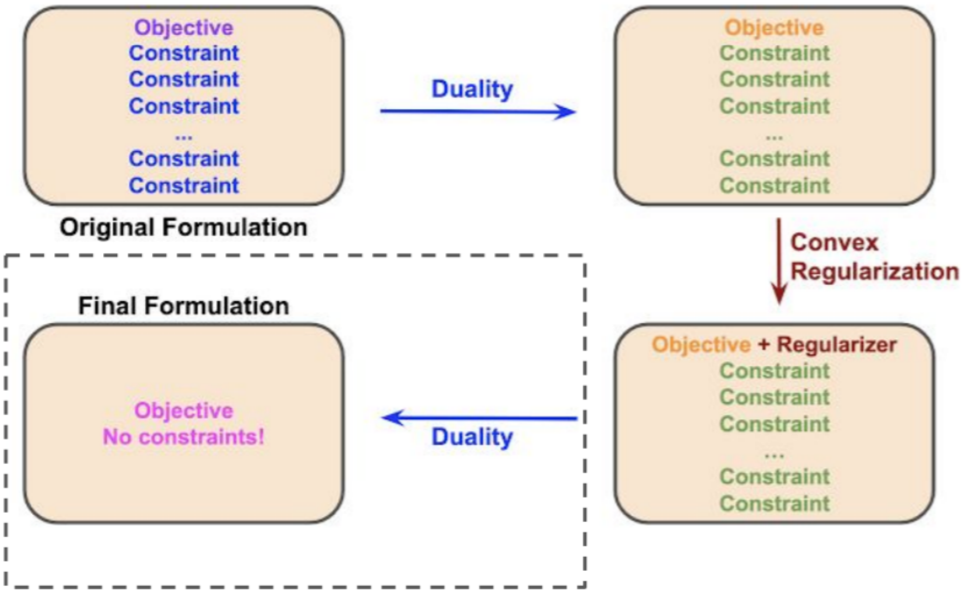

图 6: 再一次使用对偶将此问题转换为无约束优化问题

4.4 Policy Evaluation 小结

首先，策略评估问题可以表达为一个线性规划问题 ($Q-$LP)，线性规划问题的最优解即为 Q^π。此线性规划问题的对偶问题的最优解是 d^π。使用 $Q-$LP 的 Lagrangian 形式可以得到对 $\rho(\pi)$ 的双鲁棒估计。并且，发现改变目标函数的形式，并不会影响 $d^*=d^\pi$。将目标函数变为一个合适的形式非常的强大，可以使用 Fenchell - Rockafellar 对偶，将其转换为无约束的优化问题，因此更易于使用随机梯度下降法求解，也更适应于 offline setting。

5 Policy Gradient

5.1 Policy Gradient Theorem

第四小节中,描述了如何估计 $\rho(\pi)$,本小节将聚焦于如何优化 π,也可以写为寻找最优解,$\arg\max_\pi \rho(\pi)$。考虑公式 (27) 中给出的 $\rho(\pi)$ 的 Lagrangian 形式，可以类似的推导出 policy gradient 理论。将公式 (27) 表达为 $L(Q,d;\pi)$，使用 Danskin’s 理论可以推导出，

$$\frac{\partial}{\partial \pi}\rho(\pi)=\frac{\partial}{\partial \pi}\min_Q \max_{d\geq 0} L(Q,d;\pi)=\frac{\partial}{\partial \pi}L\left(Q^*,d^*;\pi\right) \tag{41}$$

其中,Q^* 和 d^* 是问题 $\min_Q \max_{d\geq 0} L(Q,d;\pi)=\max_{d\geq 0}\min_Q L(Q,d;\pi)$ 的解。下面将求解 $L\left(Q^*,d^*;\pi\right)$ 关于 π 的梯度。

对于公式 (27) 的第一项有,

$$\frac{\partial}{\partial \pi}(1-\gamma)\cdot \mathbb{E}_{a_0\sim\pi(s_0),s_0\sim\mu_0}\left[Q^*\left(s_0,a_0\right)\right]=(1-\gamma)\cdot \mathbb{E}_{a_0\sim\pi(s_0),s_0\sim\mu_0}\left[Q^*\left(s_0,a_0\right)\nabla \log \pi\left(a_0\mid s_0\right)\right] \tag{42}$$

其中，用到等式 $\frac{\partial}{\partial p}\mathbb{E}_{z\sim p}[h(z)]=\mathbb{E}_{z\sim p}[h(z)\nabla \log p(z)]$。

对于公式 (27) 的第二项有，

$$
\begin{aligned}
\frac{\partial}{\partial \pi} \mathbb{E}_{(s,a)\sim d^*}[R(s,a) + \gamma \cdot \mathcal{P}^\pi Q^*(s,a) - Q^*(s,a)] &= \mathbb{E}_{(s,a)\sim d^*}\left[\gamma \cdot \frac{\partial}{\partial \pi} \mathbb{E}_{s'\sim T(s,a), a'\sim\pi(s')}\left[Q^*\left(s', a'\right)\right]\right] \\
&= \gamma \cdot \mathbb{E}_{(s,a)\sim d^*, s'\sim T(s,a), a'\sim\pi(s')}\left[Q^*\left(s', a'\right) \nabla \log \pi\left(a' \mid s'\right)\right].
\end{aligned} \tag{43}
$$

根据之前的内容，不难得到，当 $\forall s, a, \quad Q^*(s,a) = Q^\pi(s,a)$，有 $d^* = d^\pi$，且 $d(s,a) > 0$。并且有，

$$
d^\pi(s,a) = (1-\gamma)\mu_0(s)\pi(a \mid s) + \gamma\pi(a \mid s)\sum_{\tilde{s},\tilde{a}} T\left(s' \mid \tilde{s}, \tilde{a}\right) d^\pi(\tilde{s}, \tilde{a}) \tag{44}
$$

合并公式 (42) 和公式 (43)，可得

$$
\frac{\partial}{\partial \pi} L\left(Q^*, d^*; \pi\right) = \mathbb{E}_{(s,a)\sim d^\pi}\left[Q^\pi(s,a) \nabla \log \pi(a \mid s)\right] \tag{45}
$$

根据此等式，我们可以得到使用 offline 数据计算出来的 policy gradient 和 On-Policy setting 下的 policy gradient 是一样的。详细的推导可以参考文章，[AlgaeDICE 论文解读]。而且，作者这里省略了一些东西，公式中并不简单就是 $Q^\pi(s,a)$，或者可以说是将 reward 进行调整后，获得了一种新的 Q 函数的表达形式。

5.2 Offline Policy Gradient via the Lagrangian

Sutton 最开始提出的 Policy Gradient 理论是 On-Policy setting，因为用之前的数据来更新当前的策略没有意义。而实际中我们只能使用 offline 数据来优化策略。使用公式 (27)，可以将策略优化问题改写为：

$$
\begin{aligned}
&\max_\pi \min_Q \max_{\zeta \geq 0} L(Q, \zeta, \pi) \\
:=& (1-\gamma) \cdot \mathbb{E}_{a_0\sim\pi(s_0), s_0\sim\mu_0}\left[Q\left(s_0, a_0\right)\right] + \mathbb{E}_{(s,a,s')\sim d^{\mathcal{D}}, a'\sim\pi(s')}\left[\zeta(s,a) \cdot \left(R(s,a) + \gamma Q\left(s', a'\right) - Q(s,a)\right)\right].
\end{aligned} \tag{46}
$$

根据 5.1 中的推导，只要知道了 Q^*, ζ^*，就可以用来计算 $\frac{\partial}{\partial} L(Q, \zeta, \pi)$，于是可以准确的得到 On-Policy gradient 下的梯度，$\mathbb{E}_{(s,a)\sim d^\pi}\left[Q^\pi(s,a) \nabla \log \pi(a \mid s)\right]$。这样就是用 offline data 计算 On-Policy 策略梯度。

5.3 Fenchel-Rockafellar Duality for Regularized Optimization

前面提到的实际应用中，由于 Lagrangian 的线性特征和 min-max 形式，可能会导致数值解的不稳定性。在之前的描述中，我们只关心可以改变目标函数的形式，使其更容易求解，且 $d^* = d^\pi$。然而，在我们当前的策略优化 setting 中，改变目标函数将改变最优策略。并且使用正则化项修正后的目标函数，可以作为最大奖励策略目标函数来使用，在许多应用中找到最优的正则化策略仍然是可取的。

这样，考虑将 $f-$divergence 作为策略优化原问题的正则化项，

$$
\begin{aligned}
\rho(\pi) - D_f\left(d^\pi \| d^{\mathcal{D}}\right) = &\max_d - D_f\left(d \| d^{\mathcal{D}}\right) + \sum_{s,a} d(s,a) \cdot R(s,a) \\
&\text{s.t. } d(s,a) = (1-\gamma)\mu_0(s)\pi(a \mid s) + \gamma \cdot \mathcal{P}_*^\pi d(s,a), \forall s \in S, a \in A
\end{aligned} \tag{47}
$$

而根据 Fenchel-Rockafellar 对偶，可以得到如下的对偶表达式：

$$\begin{aligned}\rho(\pi)-D_f\left(d^\pi \| d^{\mathcal{D}}\right)= & \min_Q(1-\gamma) \cdot \mathbb{E}_{a_0 \sim \pi\left(s_0\right), s_0 \sim \mu_0}\left[Q\left(s_0, a_0\right)\right]+ \\ & \mathbb{E}_{(s, a) \sim d^\pi}\left[f_*\left(R(s, a)+\gamma \cdot \mathcal{P}^\pi Q(s, a)-Q(s, a)\right)\right]\end{aligned} \tag{48}$$

优化 π 相当于求解如下表达式，

$$\max_\pi \min_Q(1-\gamma) \cdot \mathbb{E}_{a_0 \sim \pi\left(s_0\right), s_0 \sim \mu_0}\left[Q\left(s_0, a_0\right)\right]+\mathbb{E}_{(s, a) \sim d^{\mathcal{D}}}\left[f_*\left(R(s, a)+\gamma \cdot \mathcal{P}^\pi Q(s, a)-Q(s, a)\right)\right] \tag{49}$$

其中，$\pi^*=\rho(\pi)-D_f\left(d^\pi \| d^{\mathcal{D}}\right)$。在 AlgaeDICE 的文章中的推导，当 $f(x)=\frac{1}{2}x^2$ 的时候，实际上就是使用 offline 数据的 AC 算法。尽管目标函数 (49) 的最优化结果 π^* 会因为正则化项 $D_f(d\|d^{\mathcal{D}})$ 而发生改变，但是 $d^*=d^\pi$ 并不会受影响。观察到公式 (47) 仍然是 over-constrain 的，通过等式约束实际可以解出 $d(s,a)$。因此，可以表明公式 (49) 中的目标函数，仍然可以计算 offline setting 下的 on-policy 策略梯度。

5.3.1 Regularization with the KL-Divergence

如果考虑正则化项用的是 KL 散度，此约束优化问题为，

$$\begin{aligned}\rho(\pi)-D_{\mathrm{KL}}\left(d^\pi \| d^{\mathcal{D}}\right)= & \max_{d \in \Delta(|S| \times|A|)}-D_{\mathrm{KL}}\left(d \| d^{\mathcal{D}}\right)+\sum_{s, a} d(s, a) \cdot R(s, a) \\ & \text { s.t. } d(s, a)=(1-\gamma) \mu_0(s) \pi(a \mid s)+\gamma \cdot \mathcal{P}_*^\pi d(s, a), \forall s \in S, a \in A .\end{aligned} \tag{50}$$

同理，使用 fenchel-Rockafellar 对偶，可以得到 offline 的策略优化目标函数，

$$\max_\pi \min_Q(1-\gamma) \cdot \mathbb{E}_{a_0 \sim \pi\left(s_0\right), s_0 \sim \mu_0}\left[Q\left(s_0, a_0\right)\right]+\log \mathbb{E}_{(s, a) \sim d^{\mathcal{D}}}\left[\exp \left\{R(s, a)+\gamma \cdot \mathcal{P}^\pi Q(s, a)-Q(s, a)\right\}\right] \tag{51}$$

这一目标函数的美妙之处就显现出来了，使用基于梯度的方法对 π 求导，可以得到，

$$\begin{aligned}&(1-\gamma) \cdot \mathbb{E}_{a_0 \sim \pi\left(s_0\right), s_0 \sim \mu_0}\left[Q\left(s_0, a_0\right) \nabla \log \pi\left(a_0 \mid s_0\right)\right] \\ &\quad+\gamma \cdot \mathbb{E}_{\left(s, a, s^{\prime}\right) \sim d^{\mathcal{D}}, a^{\prime} \sim \pi\left(s^{\prime}\right)}\left[\operatorname{softmax}_{d^{\mathcal{D}}}\left(R+\gamma \cdot \mathcal{P}^\pi Q-Q\right)(s, a) \cdot Q\left(s^{\prime}, a^{\prime}\right) \nabla \log \pi\left(a^{\prime} \mid s^{\prime}\right)\right]\end{aligned} \tag{52}$$

此推导非常简单，利用链式求导法则一步一步来就行，

$$\operatorname{softmax}_p(h)(z):=\frac{\exp \{h(z)\}}{\mathbb{E}_{\tilde{z} \sim p}[\exp \{h(\tilde{z}\}]} \tag{53}$$

因此，我们看到 KL-divergence，引出了一个与极大似然策略学习相似的对偶表达式，而极大似然策略学习也是一些近期研究的热点。同样，最大似然策略学习与 Q 值函数的 log-average-exp 目标的使用也与 REPS 算法非常相似。不过，这里的策略目标函数只是类似于最大似然学习，而不是完全等价的。

5.4 Imitation Learning

公式 (47) 中的关于 Q-LP 的使用 $f-$divergence 的对偶形式，同样可以应用在模仿学习中。其中将 offline 数据集 $\mathcal{D}$ 视为来自专家策略的样例，目标是在新的任务中使用专家策略。如果，将公式 (47) 中的奖励设置为 0，优化即为于找到一个策略 π，该策略 π 使 $D_f(d^\pi\|d^{\mathcal{D}})$ 最小化。很多的模仿学习的目的也是精确优化这个 f-divergence 目标。但是目前的方法都需要和环境交互。于是，考虑使用同样的离线策略评估和离线策略优化技术，可以衍生出离线模仿学习算法。类似于没有奖励 $R(s,a)$ 的 AlgaeDICE。

5.5 本节小结

下面简要总结了 section 5 的主要内容。

1. 很多策略评估中的方法可以用到策略优化中，简单来看，策略评估问题为求：$\rho(\pi)$，策略优化问题为：$\max_\pi \rho(\pi)$。
2. 由于内部优化的最优解通常是 d^π 或 $d^\pi/d^\mathcal{D}$，因此可以利用 Danskin 定理来证明使用 offline 数据计算的关于 π 的梯度是 on-policy 的策略梯度。
3. 在使用对偶之前适当的修改目标函数是非常有效的。适当的正则化项可以推导出一个无约束的 Fenchel Rockafella 对偶问题。
4. 同样的技术也可以用于 offline 的模仿学习，即为在 RL 的基础上，忽略奖励即可。
5. 对于文中所提出的关于 $D_f(d\|d^\mathcal{D})$ 作为正则化项，Fenchell-Rockafellar 对偶可以通过平方 Bellman 误差最小化和最大似然策略学习，来实现与 AC 算法类似的目标函数。**那么通过选择其他不同的正则化项，是否可以实现其他有趣的目标函数呢？**

6 RL with the Linear Programming From of V

max-min 问题，对于这种问题，理论上很难激励随机优化。有没有更好的办法？实际上为什么 max-min 问题和 bilinear 问题难以优化，我也没有想得很明白。最直接的方式是将策略优化问题改写成一个凸问题。我们首先从 $d(s,a)$ 的角度，介绍 V-LP 的对偶特性，

$$\begin{aligned}&\max_{d\geq 0}\sum_{s,a} d(s,a)\cdot R(s,a)\\ &\text{s.t.}\sum_a d(s,a)=(1-\gamma)\mu_0(s)+\gamma\cdot\mathcal{T}_* d(s),\quad \forall s\in\mathcal{S}\end{aligned}\tag{54}$$

其中，$\mathcal{T}_*$ 表示转置转移算子，

$$\mathcal{T}_* d(s):=\sum_{\tilde{s},\tilde{a}} T(s|\tilde{s},\tilde{a})\cdot d(\tilde{s},\tilde{a})\tag{55}$$

和公式 (26) 中表达的类似，但是这里建立的是关于 $d(s)$ 的平衡方程。并且重要的是，公式 (54) 并不是 over-constrained 的。此问题的解为 $\rho(\pi^*)$，其获得的 max-reward 的最优策略为 π^*，其解为 $d^*=d^{\pi^*}$。公式 (54) 的对偶形式，也就是我们通常看到的形式，

$$\begin{aligned}&\min_V(1-\gamma)\cdot\mathbb{E}_{s_0\sim\mu_0}\left[V\left(s_0\right)\right]\\ &\text{s.t. } V(s)\geq R(s,a)+\gamma\cdot\mathcal{T}V(s,a),\quad\forall s\in S,a\in A\end{aligned}\tag{56}$$

其中，$\mathcal{T}$ 是转移算子，

$$\mathcal{T}V(s,a):=\mathbb{E}_{s'\sim T(s,a)}[V(s')]\tag{57}$$

此问题的最优解 V^* 是关于 π^* 的价值函数 V^{π^*}，其中

$$V^\pi(s):=\mathbb{E}_{a\sim\pi(s)}[Q^\pi(s,a)]\tag{58}$$

实际上使用公式 (54) 和公式 (56) 中的原问题和对偶问题表示形式，可以得到很多已有的基于表格的和 on-policy 算法。但是，和 Q-LP 不同的是，由于使用 $|\mathcal{S}|$ 个来对 $d(s)$ 进行约束，所以不能忽略公式 (54) 第一行中的 $|\mathcal{S}|\times|\mathcal{A}|$ 关于 $d\geq 0$ 的约束。即为，想和前面的方法一样，采用凸函数 $-D_f(d\|d^{\mathcal{D}})$ 来代替原函数，必须考虑 $d\geq 0$ 为额外的线性约束。利用两个函数 $V:\mathcal{S}\rightarrow\mathbb{R};K:\mathcal{S}\times\mathcal{A}\rightarrow\mathbb{R}_+$ 可以得到对偶表达式，

$$\min_{K\geq 0,V}(1-\gamma)\cdot\mathbb{E}_{s_0\sim\mu_0}\left[V\left(s_0\right)\right]+\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[f_*(K(s,a)+R(s,a)+\gamma\cdot\mathcal{T}V(s,a)-V(s))\right] \tag{59}$$

此目标函数，显然比之前的使用 Q-LP 推导的目标函数更好，可以看成是对之前问题的改进。这里的目标函数只涉及到 V,K 的最小化，而不需要求解关于 π 和 Q 函数的 max-min 问题。通过求解此目标函数可以得到最优策略的价值函数 V^*，但是我们的原始目标是得到最优策略本身。而要想得到最优策略 $\pi^*(a|s)$，首先要从 V^*,K^* 中得到 d^*，

$$d^*(s,a)=d^{\mathcal{D}}(s,a)\cdot f_*'\left(K^*(s,a)+R(s,a)+\gamma\cdot\mathcal{T}V^*(s,a)-V^*(s)\right) \tag{60}$$

使用 Bayes rule 可以得到最优策略，

$$\pi^*(a\mid s)=\frac{d^*(s,a)}{\sum_{\tilde{a}}d^*(s,\tilde{a})}=\frac{d^{\mathcal{D}}(s,a)\cdot f_*'\left(K^*(s,a)+R(s,a)+\gamma\cdot\mathcal{T}V^*(s,a)-V^*(s)\right)}{\sum_{\tilde{a}}d^{\mathcal{D}}(s,\tilde{a})\cdot f_*'\left(K^*(s,a)+R(s,\tilde{a})+\gamma\cdot\mathcal{T}V^*(s,\tilde{a})-V^*(s)\right)} \tag{61}$$

有关此部分证明可以仔细阅读 [Dual AC]。其大致证明思路为说明，$d^*(s,a)$ 对于每个 s 对应的是 a^*，而公式 (61) 可以看成是归一化的形式。**通过这样的方式，可以成功的利用 V^*,K^* 推导出** π^*。可以看到，使用 Fenchel-Rockafellar 对偶替代 Q-LP。避免了求解 Q-LP 产生的 max-min 问题，但是现在的问题是并没有直接求解出提供 π^*。必须用额外的步骤来推导出 π。在实际实现中 (随机) 中，从 V,K 中推导出 π^* 可能非常困难。

6.1 Max-Likelihood Policy Learning

使用 KL 散度作为正则化项有两大好处，1. 有效的避免数值解的不稳定性，2. 保证 d 的值是正的。在通用的 f 散度中，必须考虑 $d\geq 0$，然而由于 KL 散度的对偶形式是 log-expected-exponentde，所以一定是非负的。所以，将公式 (64) 中的目标函数，改写为 $\mathbb{D}_{\mathrm{KL}}(d\|d^{\mathcal{D}})$，再使用 Fenchel-Rockafellar 对偶可以得到更简单的目标函数，

$$\min_{V}(1-\gamma)\cdot\mathbb{E}_{s_0\sim\mu_0}\left[V\left(s_0\right)\right]+\log\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[\exp\{R(s,a)+\gamma\cdot\mathcal{T}V(s,a)-V(s)\}] \tag{62}$$

同样可以得到最优策略的 visitation 为：

$$d^{\pi^*}(s,a)=d^{\mathcal{D}}(s,a)\cdot\operatorname{softmax}_{d^{\mathcal{D}}}\left(R+\gamma\cdot\mathcal{T}V^*-V^*\right)(s,a). \tag{63}$$

那么，按照同样的方法，可以推导出最优策略的形式，

$$\pi^*(a\mid s)=d^{\mathcal{D}}(a\mid s)\cdot\operatorname{softmax}_{d^{\mathcal{D}}(\cdot|s)}\left(R(s,\cdot)+\gamma\cdot\mathcal{T}V^*(s,\cdot)-V^*(s)\right)(a) \tag{64}$$

6.2 Policy Evaluation with the $V-$LP

实际上也可以用 $V-$LP 来做策略评估。由于策略评估过程中，$\pi(a|s)$ 是固定的，那么可以做此分解 $d(s,a)=\mu(s)\pi(a|s)$。公式 (54) 可以做如下改写，

$$\begin{aligned}&\max_\mu\textstyle\sum_{s,a}\mu(s)\pi(a\mid s)\cdot R(s,a)\\&\text{s.t. }\ \mu(s)=(1-\gamma)\mu_0(s)+\gamma\cdot\mathcal{T}_*(\mu\times\pi)(s)\quad\forall s\in S.\end{aligned} \tag{65}$$

这里用到了推导：$\sum_a d(s,a)=\sum_a \mu(s)\pi(a|s)=\mu(s)\sum_a \pi(a|s)=\mu(s)$。由于这里固定了 $\pi(a|s)$，此 LP 问题变成了 over-constrained 的。和之前的思路类似，这里同样在使用 Lagrangian 或者 Fenchel-Rockafellar 对偶前，对目标函数进行替换。需要注意的是，使用 $V-$LP 进行 offline 策略评估会导致需要数据分布策略 $d^{\mathcal{D}}(a|s)$ 的先验知识的目标，以便将 offline 样本纠正为 on-policy 的样本，正如 $\mu\times\pi$ 在公式 (65) 中所需要的那样。(涉及到 $\mathcal{T}_*=\sum_{\tilde{s},\tilde{a}}\mu(\tilde{s})\pi(\tilde{s}|\tilde{a})T(s|\tilde{s},\tilde{s}))$ 这与 Q-LP 的 behavior-agnostic 目标不一致。

7 Undiscounted Settings

在之前的 setting 中，考虑的都是带折扣的奖励情况，$\gamma\in(0,1)$。实际上当 $\gamma=1$ 的时候，是 RL 领域的一大难点。因为此时 Q 值的概念和 Bellman 算子的收敛是很难掌握的。另一方面，通过微小的改动，可以将 Fenchel-Rockafellar 对偶应用到 undiscounted setting 中。在这一节中，我们将先前的推导推广到 $\gamma=1$ 的情况，从而得到几种实用的算法。

7.1 Policy Evaluation

当 $\gamma=1$ 的情况下，策略评估问题为：

$$\rho(\pi):=\lim_{t_{\text{stop}}\to\infty}\mathbb{E}\left[\frac{1}{t_{\text{stop}}}\sum_{t=0}^{t_{\text{stop}}}R\left(s_t,a_t\right)\middle| s_0\sim\mu_0,\forall t,a_t\sim\pi\left(s_t\right),s_{t+1}\sim T\left(s_t,a_t\right)\right] \tag{66}$$

在确定性环境下，比如状态空间和动作空间都是有限的，$\rho(\pi)$ 可以改写为：

$$\rho(\pi)=\mathbb{E}_{(s,a)\sim d^\pi}[R(s,a)] \tag{67}$$

其中 undiscounted on-policy 分布 d^π 被定义为归一化分布，满足，(**这实际上是平衡方程**)

$$d^\pi(s,a)=\mathcal{P}_*^\pi d^\pi(s,a) \tag{68}$$

那么，可以将公式 $\rho(\pi)$ 表达为，

$$\begin{aligned}\rho(\pi)=\max_{d\geq 0}&\sum_{s,a}d(s,a)\cdot R(s,a)\\ \text{s.t.}&\sum_{s,a}d(s,a)=1\text{ and}\\ &d(s,a)=\mathcal{P}_*^\pi d(s,a),\forall s\in S,a\in A\end{aligned} \tag{69}$$

实际上对比原公式，就增加了一个 $\sum_{s,a}d(s,a)=1$ 来确保其被归一化。实际上，此处可以采用和 Section 4 类似的方法来求解。首先将目标函数改写为 $D_f(d\|d^{\mathcal{D}})$ 采用 Fenchel-Rockafellar 对偶法可以得到，这里写得实在是简单，具体的推导请移步 [AlgaeDICE] 中的附录 A.1。

$$\min_{Q,\lambda}-\lambda+\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[f_*\left(\lambda+\mathcal{P}^\pi Q(s,a)-Q(s,a)\right)\right] \tag{70}$$

对于给定的最优解 Q^*,λ^*，有

$$f_*'\left(\lambda^*+\mathcal{P}^\pi Q^*(s,a)-Q^*(s,a)\right)=\frac{d^\pi(s,a)}{d^{\mathcal{D}}(s,a)} \tag{71}$$

7.2 Regularized Lagrangian

在公式 (70) 前面加上 $\max_\pi$ 就得到了策略优化的公式，

$$\max_{\pi}\min_{Q,\lambda} -\lambda + \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[f_*\left(\lambda + R(s,a) + \mathcal{P}^\pi Q(s,a) - Q(s,a)\right)\right] \tag{72}$$

与 Section 5 相似，可以使用 Danskin 定理来证明此目标函数的对 π 的梯度，是在 On-Policy setting 上的策略梯度 (尽管是正则化的 Q 值)。

和公式 (54) 和 (56) 中的表达形式类似，策略优化问题可以写为：

$$\begin{aligned}
&\max_{d\geq 0}\sum_{s,a} d(s,a)\cdot R(s,a)\\
&\text{s.t. } \sum_{s,a} d(s,a) = 1 \text{ and}\\
&\qquad \sum_{a} d(s,a) = \mathcal{T}_* d(s), \quad \forall s\in S.
\end{aligned} \tag{73}$$

可以通过它的拉格朗日或通过添加一个适当的正则化器并应用 Fenchell-Rockafellar 对偶性来解决这个问题。可以对公式 (73) 增加一个正则化项 $-D_{\mathrm{KL}}(d\|d^{\mathcal{D}})$，其 Fenchel-Rockafellar 对偶形式为，

$$\min_V \log \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[\exp\{R(s,a) + \mathcal{T}V(s,a) - V(s)\}] \tag{74}$$

然后，通过最大似然优化，从 V^* 中得到最优策略 π^*，

$$\pi^* = \arg\max_\pi \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[\frac{1}{Z(s)}\exp\left\{R(s,a) + \mathcal{T}V^*(s,a) - V^*(s)\right\}\log\pi(a\mid s)\right] \tag{75}$$

8 Conclusion

总体评价这篇文章省去了很多的细节，只是作为一个大致了解 DICE Family 的文章，而且很多的符号都采用了简写，需要配合对应的 DICE 发表的原文阅读才能读得清晰。

此篇文章展示了很多 Fenchel-Rockafellar 对偶在 RL 问题中的应用。文章中使用的技术可以简单概括为：

1. 当提出一个似乎很难解决的问题时,可以将问题写成约束凸优化形式,并求解其 Fenchell-Rockafellar 对偶，或其拉格朗日形式;
2. 如果对偶仍然很难解决 (例如，当原始目标是线性的，产生一个带有约束的对偶)，考虑修改原始目标，例如，通过使用适当的凸正则化器。

看似非常简单，但在我们的推导中一直反复出现，导致了几种算法来解决策略评估、策略优化和模仿学习问题，而且不管 online 或 offline 问题，以及 discounted 或 un-discounted 问题都可以求解。

作者希望可以将 RL 和优化紧密的联合起来，可以将此方法扩展到其他的 RL settings 中，比如 multi-agent RL，safe RL，exploration for RL，或者其他问题。

与此同时，这些新的表达公式与凸优化算法 (特别是在随机梯度下降和函数逼近设置中) 的相互作用如何，以及这些基于对偶性的公式是否比基于 dp 的方法更有效，这些问题给优化研究带来了新的挑战和问题。