DualDice: Behavior-Agnostic Estimation of Discounted Stationary Distribution Corrections

Chen Gong

09 March 2021

目录

1 Core idea

本篇文章是 DAI Bo 老师在 Google Brain 做出来的作品，发表在 NIPS 2019 上，其理论非常的漂亮。本文的主要工作是给出了对 discounted stationary distribution radios (DSDR) 的估计方法：DualDICE。DSTR 是一个修正项，它量化了新策略将经历某一确定的 (s,a) 的概率，此 (s,a) 通过其在一个数据集中出现的概率进行归一化。DualDICE，与其他算法相比的优势在于，不需要知道生成数据的行为策略的知识，并且它避免了直接使用重要性权重的情况，可以有效的避免重要性权重连乘带来的方差过大的问题。并且，作者提供了算法收敛性的理论保证。

而为什么要设计这样的算法呢?因为强化学习很多时候从一些现实的环境中学习到较好的策略，而现实环境中交互成本可能非常高。只能通过一些从多个行为策略，甚至是不知道什么行为策略中采集到的数据来进行学习。

现在很多做法是考虑，估计 discounted stationary distribution radios (DSDR) 或者 corrections。这些变量衡量当前目标策略经历某一 (s,a) 的可能性，该 (s,a) 在 off-policy 中出现的概率是标准化的。合适的估计可以提高策略估计的准确性和策略学习的稳定性，但是其并不容易被计算。因为其计算不仅依赖于目标策略在相关状态下采取预期动作的概率，而且依赖于目标策略与环境动态的相互作用而导致其进入相关状态的概率。

2 背景

马尔可夫决策过程：$\langle S, A, R, T, \beta\rangle$。注意，任何有限的环境可以被考虑为无限视界，其方法为增加一个终止状态 s_T，且 $P(s_T|s_T,a)=1, R(s_T,a)=0$。

2.1 Off-Policy 策略评估

对于目标策略，其策略估计被定义为：

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

off-policy 策略评估问题是使用一个固定的数据集 $\mathcal{D}$：$\{(s,a,r,s')\}$。$\mathcal{D}$ 的来源多种多样，来自各种不同的行为策略。在一种特殊情况下，$\mathcal{D}$ 中所有的轨迹都来自于已知的行为策略 μ。那么，可以采用重要性采样 (IS) 来估计目标函数，

$$\rho(\pi) = (1-\gamma)\left(\prod_{t=0}^{H-1}\frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}\right)\left(\sum_{t=0}^{H-1}\gamma^t r_t\right)$$

尽管有很多的方法来改进 IS，但是依然无法避免指数级的高方差。这个问题在我 offline RL 那篇文章中详细解释过了，这里不过多阐述了。**为了避免每个动作的重要性权重出错，而连乘带来方差的指数级累加。所以采用了直接对状态长距离的发生的重要性权重比例进行估算**。而要想较好的理解这里，需要看看 Qiang Liu 的 Breaking the Curse of Horizon: Infinite-Horizon Off-Policy Estimation。

令 $p_\pi(\cdot)$ 表示策略 π 下采样的轨迹的概率分布：$\tau\sim\{s_t,a_t,r_t\}_{t=0}^{\infty}$。而 π 的期望回报为：

$$\rho(\pi) := \lim_{T\to\infty}\mathbb{E}_{\boldsymbol{\tau}\sim p_\pi}\left[R^T(\boldsymbol{\tau})\right], \quad R^T(\boldsymbol{\tau}) := \left(\sum_{t=0}^{T}\gamma^t r_t\right) / \left(\sum_{t=0}^{T}\gamma^t\right), \tag{2}$$

而令 $R_\pi^T(\tau)$ 表示 τ 取 $0:T$ 时刻。其中，当 $\gamma=1$ 和 $\gamma\in(0,1)$ 时，

$$\text{Average:}\quad R(\tau):=\lim_{T\to\infty}\frac{1}{T+1}\sum_{t=0}^{T}r_t,\quad \text{Discounted:}\quad R(\tau):=(1-\gamma)\sum_{t=0}^{\infty}\gamma^t r_t \tag{3}$$

其中，$(1-\gamma)=1/\sum_{t=0}^{\infty}\gamma^t$ 为归一化因子。Off-Policy 价值估计的问题为估计 R_π 的期望奖励，只能使用通过 π_0 产生的一系列轨迹 $\tau^i=\{s_t^i,a_t^i,r_t^i\}_{t=0}^T$。

2.1.1 Discounted Stationary Distribution 相关概念介绍

定义 $d_{\pi,t}(\cdot)$ 为从初始状态 s_0 执行策略 π,t 时刻状态 s_t 的分布，其平均访问分布（即为在**平稳分布**的假设条件下，从 0 时刻开始，状态 s 在一个轨迹中，被访问的平均概率。**从直观上理解，对比公式 (1) 我感觉这是将对一段轨迹中的所有可能出现的状态的概率分布的平均值计算出来，从而将一段很长的轨迹空间，考虑为单独对状态空间进行建模，从而消除了轨迹长度的影响。利用每个状态 s_t 出现的平均概率乘上此状态下的奖励 r_t，同样能计算出 $\rho(\pi)$，所以在某些文献中也被称为边缘性概率分布**）被定义为：

$$d^\pi(s)=\lim_{T\to\infty}\left(\sum_{t=0}^{T}\gamma^t d^{\pi,t}(s)\right)/\left(\sum_{t=0}^{T}\gamma^t\right) \tag{4}$$

$$d^\pi(s)=(1-\gamma)\cdot\sum_{t=0}^{\infty}\gamma^t\Pr(s_t=s\mid s_0\sim\beta,\forall t,a_t\sim\pi(s_t),r_t\sim R(s_t,a_t),s_{t+1}\sim T(s_t,a_t)) \tag{5}$$

通常假设 $T\to\infty$，任何有限的情况可以被考虑为无限视界。当 $\gamma\in(0,1]$，$d^\pi(s)=(1-\gamma)\sum_{t=0}^{\infty}\gamma^t d^{\pi,t}(s)$；而当 $\gamma=1$ 时，$d^\pi(s)$ 为 s_t 的平稳分布。

$$d^\pi(s)=\lim_{T\to\infty}\frac{1}{T+1}\sum_{t=0}^{T}d^{\pi,t}(s)=\lim_{t\to\infty}d^{\pi,t}(s)$$

于是公式 (1) 中的策略价值函数可以被重写为：

$$\rho(\pi)=\mathbb{E}_{(s,a)\sim d^\pi,r\sim R(s,a)}[r] \tag{6}$$

为了简化起见，之后用 $(s,a)\sim d^\pi$ 表示从 $d^\pi(s,a)$ 中采样得到的 (s,a) 对，其中 $d^\pi(s,a):=d^\pi(s)\pi(a|s)$。如果 $\mathcal{D}$ 是由行为策略 μ 中收集而来的，策略的估计值可以写为：

$$\rho(\pi)=\mathbb{E}_{(s,a)\sim d^\mu,r\sim R(s,a)}\left[w_{\pi/\mu}(s,a)\cdot r\right] \tag{7}$$

其中，$w_{\pi/\mu}(s,a)=\frac{d^\pi(s,a)}{d^\mu(s,a)}$，即为 Discounted Stationary Distribution Correction，而 $w_{\pi/\mu}(s,a)$ 的计算非常困难，也就是 Bo Dai 此篇文章主要解决的问题。所以，对于 $\{s_t^i,a_t^i,r_t^i\}_{i=1}^m\sim\mu$，

$$\rho(\pi)=\frac{1}{m}\sum_{i=1}^{m}\sum_{t=1}^{T}w_t^i r_t^i,\qquad w_t^i:=\gamma^t\frac{w_{\pi/\mu}(s_t^i,a_t^i)}{\sum_{t',i'}w_{\pi/\mu}(s_{t'}^{i'},a_{t'}^{i'})}$$

同样，这里对每个状态动作对 (s,a) 在一段轨迹中出现的平均概率进行建模，而不是直接对 $\tau=\{(s_t,a_t)\}_{t=0}^T$ 进行建模。很显然，这里没有连乘的操作了，这样或许可以减小方差。而在 Qiang Liu 的 “Breaking the Curse of Horizon: Infinite-Horizon Off-Policy Estimation” 中，举了一个例子来验证。

2.1.2 Learning Stationary Distribution Corrections

基于以下平稳马尔科夫过程的稳态性质有：

$$
\begin{aligned}
d^{\pi}\left(s'\right) &= (1-\gamma)\sum_{t=0}^{\infty}\gamma^{t}d^{\pi,t}\left(s'\right) \\
&= (1-\gamma)\beta\left(s'\right) \mid +(1-\gamma)\sum_{t=1}^{\infty}\gamma^{t}d^{\pi,t}\left(s'\right) \\
&= (1-\gamma)\beta\left(s'\right)+(1-\gamma)\gamma\sum_{t=0}^{\infty}\gamma^{t}d^{\pi,t+1}\left(s'\right) \\
&= (1-\gamma)\beta\left(s'\right)+(1-\gamma)\gamma\sum_{t=0}^{\infty}\gamma^{t}\sum_{s}\boldsymbol{T}_{\pi}\left(s'\mid s\right)d^{\pi,t}(s) \quad //d^{\pi,t+1}(s')=\sum_{s,a}T_{\pi}\left(s'\mid s\right)d^{\pi,t}(s) \\
&= (1-\gamma)\beta\left(s'\right)+\gamma\sum_{s}\boldsymbol{T}_{\pi}\left(s'\mid s\right)\left((1-\gamma)\sum_{t=0}^{\infty}\gamma^{t}d^{\pi,t}(s)\right) \\
&= (1-\gamma)\beta\left(s'\right)+\gamma\sum_{s}\boldsymbol{T}_{\pi}\left(s'\mid s\right)d^{\pi}(s) \\
&= (1-\gamma)\beta\left(s'\right)+\gamma\sum_{s,a}\boldsymbol{T}\left(s'\mid s,a\right)\pi(a\mid s)d^{\pi}(s)
\end{aligned}
\tag{8}
$$

这个恒等式简单地反映了平稳分布的流动守恒：学过随机过程的同学都熟悉平稳分布状态方程：$\pi P=\pi$（π 为平稳状态概率，P 为转移矩阵）。对于行为策略 μ，公式 (3) 可以用平稳分布修正来等价地重写，$s\in S'$，

$$
\begin{aligned}
\sum_{s'}d^{\pi}\left(s'\right) &= \sum_{s'}(1-\gamma)\beta\left(s'\right)+\gamma\sum_{s,a,s'}\boldsymbol{T}\left(s'\mid s,a\right)\pi(a\mid s)d^{\pi}(s) \\
\mathbb{E}_{s'\sim d^{\mu}}\left[w_{\pi/\mu(s')}\right] &= (1-\gamma)\beta(s')+\gamma\sum_{s,a,s'}\frac{\boldsymbol{T}\left(s'\mid s,a\right)\pi(a\mid s)d^{\pi}(s)}{\boldsymbol{T}\left(s'\mid s,a\right)\mu(a\mid s)d^{\mu}(s)}d(s',a,s) \\
\mathbb{E}_{s'\sim d^{\mu}}\left[w_{\pi/\mu}(s')\right] &= (1-\gamma)\beta(s')+\gamma\mathbb{E}_{(s,a,s')\sim d^{\mu}}\left[\frac{\pi(a\mid s)}{\mu(a\mid s)}w_{\pi/\mu}(s')\right]
\end{aligned}
\tag{9}
$$

所以，可以推导出，

$$
\mathbb{E}_{(s_t,a_t,s_{t+1})\sim d^{\mu}}\left[\mathrm{TD}\left(s_t,a_t,s_{t+1}\mid w_{\pi/\mu}\right)\mid s_{t+1}=s'\right]=0 \tag{10}
$$

其中，

$$
\mathrm{TD}\left(s,a,s'\mid w_{\pi/\mu}\right):=-w_{\pi/\mu}\left(s'\right)+(1-\gamma)\beta\left(s'\right)+\gamma w_{\pi/\mu}(s)\cdot\frac{\pi(a\mid s)}{\mu(a\mid s)} \tag{11}
$$

而其中 TD 可以看成 $w_{\pi/\mu}$ 的时间差分。因此，之前的工作通过使用 d^{μ} 中的样本来优化损失函数，使 TD 误差最小化。文中强调，虽然 $w_{\pi/\mu}$ 与时间差异有关，但它不满足通常意义上的 Bellman 递归。**请注意公式 (8) 是“向后”写成的：平稳分布下状态 s' 的平均访问概率写成之前状态 s 的函数，反之也可以。这将是我们的算法和前面这些方法之间的一个关键区别**。

2.1.3 Off-Policy Estimation with Multiple Unknown Behavior Policies

此文中探究的从任意行为策略中采样得到的数据混合到一起，得到 $\mathcal{D}$。所以，$\mathcal{D}$ 不再是一个可知的，单一的策略。目标策略值被等价写为：

$$
\rho(\pi)=\mathbb{E}_{(s,a,r)\sim\mathcal{D}}\left[w_{\pi/\mathcal{D}}(s,a)\cdot r\right] \tag{12}
$$

Assumption 1 对于任何 $d^\pi(s,a)>0$，有 $d^\mu(s,a)>0$。而且，修正项是由一个有限常数 C 限定的：$\|w_{\pi/\mathcal{D}}\|_\infty \le C$。

3 DualDICE

本节开始将引出本文的方法。用 DualDICE 来估计 Discounted Stationary Distribution Corrections: $w_{\pi/\mathcal{D}}(s,a)=\frac{d^\pi(s,a)}{d^{\mathcal{D}}(s,a)}$。在本文的设定中，我们不知道有关分布 $\mathcal{D}$ 的显示信息，只能从 $\mathcal{D}=\{(s,a,r,s')\}\sim d^{\mathcal{D}}$ 中进行采样。我们也假设访问样品从初始状态分布 。我们首先介绍一个关键结果，稍后我们将推导这个结果并将其用作我们算法的关键。文章中首先介绍的是核心思想，然后介绍推导过程和理论支撑。

3.1 The Key Idea

其主要思想是考虑优化一个有界的函数，$\nu: S\times A\to\mathbb{R}$，优化问题的解即为我们要求的目标。

$$\min_{\nu:S\times A\to\mathbb{R}} J(\nu):=\frac{1}{2}\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[\left(\nu-\mathcal{B}^\pi\nu\right)(s,a)^2\right]-(1-\gamma)\mathbb{E}_{s_0\sim\beta,a_0\sim\pi(s_0)}\left[\nu\left(s_0,a_0\right)\right] \tag{13}$$

其中，使用 $\mathcal{B}^\pi$ 来表示关于策略 π，奖励为 0 的贝尔曼算子：$\mathcal{B}^\pi\nu(s,a)=\gamma\mathbb{E}_{s'\sim T(s,a),a'\sim\pi(s')}\left[\nu\left(s',a'\right)\right]$。公式 (13) 的第一项是平方奖励函数，求 min 操作会使得 $\nu^*\equiv 0$。而第二项的增加，会使得 $\nu^*>0$。这两项合在一起会得到一个最优结果 ν^*。

更巧的是，ν^* 的贝尔曼残差就是我们想求的分布修正：

$$\left(\nu^*-\mathcal{B}^\pi\nu^*\right)(s,a)=w_{\pi/\mathcal{D}}(s,a) \tag{14}$$

那么，可以看到只需要 $d^{\mathcal{D}},\beta,\pi$ 就可以估计出分布修正，而这些都是已知的。所以，我们并不需要知道 $d^{\mathcal{D}}$ 的具体形式了。

3.2 Derivation

对于任意的标量 $m\in\mathbb{R}_{>0},n\in\mathbb{R}_{\ge 0}$，对于凸函数 $\min_x J(x):=\frac{1}{2}mx^2-nx$，其最优解为 $x^*=\frac{n}{m}$。所以，根据这个观察，令 $\mathcal{C}$ 是 $\mathbb{R}$ 的某个有界子集，其中包含 $[0,C]$，下面这个优化问题，

$$\begin{aligned}\min_{x:S\times A\to\mathcal{C}} J_1(x):=&\frac{1}{2}\sum_{s,a}d^{\mathcal{D}}(s,a)\left[x(s,a)^2\right]-\sum_{s,a}d^\pi(s,a)[x(s,a)]\\ =&\frac{1}{2}\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[x(s,a)^2\right]-\mathbb{E}_{(s,a)\sim d^\pi}[x(s,a)]\end{aligned} \tag{15}$$

显然有，$\forall(s,a)\in\mathcal{S}\times\mathcal{A}$ 有 $x^*(s,a)=w_{\pi/\mathcal{D}}(s,a)=\frac{d^\pi(s,a)}{d^{\mathcal{D}}(s,a)}$。但是第二项要求 $(s,a)\sim d^\pi(s,a)$，我们无法获得。下一步则是想办法将其转换为我们有的条件。

3.2.1 Change of Variables

令 $\nu: S\times A\to\mathbb{R}$，为任意的状态价值函数，满足：

$$\nu(s,a):=x(s,a)+\gamma\mathbb{E}_{s'\sim T(s,a),a'\sim\pi(s')}\left[\nu\left(s',a'\right)\right],\forall(s,a)\in S\times A \tag{16}$$

因为 $x(s,a) \in \mathcal{C}$，并且 $\gamma \in [0,1]$，变量 $\nu(s,a)$ 定义明确且有界。定义采用策略 π 在 t 时刻状态的访问概率为，

$$\beta_t(s) := \Pr\left(s = s_t \mid s_0 \sim \beta, a_k \sim \pi\left(s_k\right), s_{k+1} \sim T\left(s_k, a_k\right) \text{ for } 0 \leq k < t\right) , \tag{17}$$

接下来可以推导，

$$\begin{aligned}
\mathbb{E}_{(s,a)\sim d^\pi}[x(s,a)] =& \mathbb{E}_{(s,a)\sim d^\pi}\left[\nu(s,a) - \gamma \mathbb{E}_{s'\sim T(s,a), a'\sim\pi(s')}\left[\nu\left(s', a'\right)\right]\right] \\
=& (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{s\sim\beta_t, a\sim\pi(s)}\left[\nu(s,a) - \gamma \mathbb{E}_{s'\sim T(s,a), a'\sim\pi(s')}\left[\nu\left(s', a'\right)\right]\right] \\
=& (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{s\sim\beta_t, a\sim\pi(s)}[\nu(s,a)] - (1-\gamma)\sum_{t=0}^{\infty}\gamma^{t+1}\mathbb{E}_{s\sim\beta_{t+1}, a\sim\pi(s)}[\nu(s,a)] \\
=& (1-\gamma)\mathbb{E}_{s\sim\beta, a\sim\pi(s)}[\nu(s,a)]
\end{aligned} \tag{18}$$

其中，第一行到第二行的推导使用了 $d^\pi(s,a)$ 的定义。而且，

$$\left(\nu - \mathcal{B}^\pi \nu\right)(s,a) = x(s,a) \rightarrow \left(\nu^* - \mathcal{B}^\pi \nu^*\right)(s,a) = x^*(s,a) = w_{\pi/\mathcal{D}}(s,a). \tag{19}$$

所以，公式 (15) 可以转换为公式 (13)，所有的变量我们都可以轻易的获得。**那么下一步则是专注于解** ν^*。然而，在实践中，求解目前形式的目标函数有两个困难：

1. $\left(\nu - \mathcal{B}^\pi \nu\right)(s,a)^2$ 中包含条件期望的平方。通常环境的动态变化是随机的，而且动作空间可能很大或者是连续的，使用随机优化技术可能有很大的误差。（例如，当环境是随机的时候，其蒙特卡罗估计的梯度一般是有偏的。）

2. 即使计算得到了 ν^*，而 $\left(\nu^* - \mathcal{B}^\pi \nu^*\right)(s,a)$，和第一点列出的原因相同，其值同样很难计算。特别环境是随机的，动作空间是连续的时候。

3.2.2 Fenchel Duality

分析可得，上述的两个问题，都是由于 $\mathcal{B}^\pi$ 中存在求期望。而 Fenchel 对偶可以较好的解决这个问题。对于任意的凸函数 $f(x)$，可以写成 $f(x) = \max_\zeta x \cdot \zeta - f^*(\xi)$，其中 $f^*(x)$ 就是 f 函数的 Fenchel 共轭函数。当 $f(x) = \frac{1}{2}x^2$，其 Fenchel 共轭函数为 $f^*(\zeta) = \frac{1}{2}\zeta^2$。所以，目标函数可以重新表达为：

$$\min_{\nu: S\times A\rightarrow\mathbb{R}} J(\nu) := \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[\max_\zeta \left(\nu - \mathcal{B}^\pi \nu\right)(s,a)\cdot\zeta - \frac{1}{2}\zeta^2\right] - (1-\gamma)\mathbb{E}_{s_0\sim\beta, a_0\sim\pi(s_0)}\left[\nu\left(s_0, a_0\right)\right] \tag{20}$$

根据替换准则可将最大化标量 ζ 的操作，变成一个 min-max 的鞍点优化问题。这里是因为关于 ζ 的最大化操作是与 (s,a) 无关的，而且关于 ζ 的函数是单峰，且连续的。Interchangeability principle 为：

Example 2 Consider the setting of case (ℵ3) and let $\mathcal{R}(Z) := \max_{\omega\in\Omega} Z(\omega)$. This functional is monotone and continuous, but is not strictly monotone. The equality (2.2) takes here the form

$$\max_{\omega\in\Omega} \underbrace{\inf_{x\in X} f(x,\omega)}_{F(\omega)} = \inf_{\chi\in\mathcal{X}} \underbrace{\max_{\omega\in\Omega} f(\chi(\omega),\omega)}_{\mathcal{R}(f_\chi)} \tag{2.8}$$

and the implication (2.3) becomes

$$\bar{\chi}(\cdot) \in \arg\min_{x\in X} f(x,\cdot) \Rightarrow \bar{\chi} \in \arg\min_{\chi\in\mathcal{X}} \left\{ \max_{\omega\in\Omega} f(\chi(\omega),\omega) \right\}. \tag{2.9}$$

图 1: interchangeability principle: http://www.optimization-online.org/DB_FILE/2017/04/5983.pdf

$$\begin{aligned} \min_{\nu:S\times A\to\mathbb{R}} \max_{\zeta:S\times A\to\mathbb{R}} J(\nu,\zeta) := \mathbb{E}_{(s,a,s')\sim d^{\mathcal{D}},a'\sim\pi(s')}\big[& \left[\nu(s,a) - \gamma\nu\left(s',a'\right)\right)\zeta(s,a) - \zeta(s,a)^2/2\big] \\ & - (1-\gamma)\mathbb{E}_{s_0\sim\beta,a_0\sim\pi(s_0)}\left[\nu\left(s_0,a_0\right)\right]. \end{aligned} \tag{21}$$

对内部优化问题使用 KKT 条件（ζ 是凸的且为二次），对于任意的 ν，最优值 $\zeta^*_\nu = \nu - \mathcal{B}^\pi\nu$。所以，平稳分布修正实际上就是公式 (21) 的 minimax 问题的鞍点 (ν^*,ζ^*)：

$$\zeta^*(s,a) = \left(\nu^* - \mathcal{B}^\pi\nu^*\right)(s,a) = w_{\pi/\mathcal{D}}(s,a) \tag{22}$$

这样就得到了一个实际计算中较方便的目标函数。通过上述的推导，这样计算的好处有三点：

1. 对于所有的目标都是无偏估计，梯度容易使用从 $d^{\mathcal{D}},\pi,\beta$ 中进行随机采样而得到。

2. 公式 (21) 中描述的 min-max 目标函数中，ν 是线性的，ζ 是凹函数。所以，可以保证优化算法的收敛性，之后会给出详细的证明。

3. 公式 (11) 的优化目标 $\zeta^*(s,a)$ 即为我们要求的平稳分布修正：$\zeta^*(s,a)$，不需要额外的计算。

而共轭函数有什么用呢?

1. 原函数约束很多，不一定是凸函数，也就是说原函数是一个也许有很多极小值的多维空间函数，它是不容易求最小值的。用这种方式来拟合（训练学习），容易陷入局部最小值，得到的结果不够泛化。举例：一个训练好的分类器，对一些东西分类很准（拟合误差达到局部极小值），但是对另一些东西分类很烂（拟合误差不是全局最小）。通过求共轭函数，我们把它原函数映射到另一个多维空间（自变量都变了），变成一个新函数，这个函数是凸的，而且它的最大值小于等于原函数的最小值。这样求原函数最小值问题，变成一个无约束凸函数的求最大值问题。那就很简单了，只要求新函数的唯一鞍点（梯度为零）。这样原本难以进行全局最优拟合的问题，变成可以拟合最优了，人工智能的精度立马登上一个新高峰。https://www.zhihu.com/question/268862097/answer/371323504。

2. 平时所说的共轭函数就是基于相同的线性超平面而构建的对偶关系。好处是一个函数即便不是凸函数，但通过共轭法获得一个凸函数。更妙的是，再次通过这种共轭又能得到与原函数不错的近似函数。

3. 也就是说，通过两次共轭就可以得到原函数的凸包函数。如果使用这个凸包函数来代理原函数，在优化求解上则会获得许多优良的性质，但又能在特性上损失不大。这也许就是使用共轭函数的最大意义和用途。https://www.zhihu.com/question/263754316/answer/1290371489

3.2.3 Extension to General Convex Functions

除了使用二次函数惩罚函数：$(\nu - \mathcal{B}^\pi \nu)(s,a)^2$，可以将上述推导推广到更一般的凸惩罚函数。考虑一个通用的凸惩罚函数类 $f : \mathbb{R} \to \mathbb{R}$。回想一下，$\mathcal{C}$ 是 $\mathbb{R}$ 的一个有界子集，其中包含平稳分布修正的区间 $[0, C]$。如果，$\nabla f \in \mathcal{C}$，凸优化问题为：

$$\min_x J(x) := m \cdot f(x) - nx \tag{23}$$

f 需要满足 KKT 条件：$\nabla_x f(x) = 0 -> f'(x^*) = \frac{n}{m}$。类似与公式 (15)，优化问题可以写作，

$$\min_{x: S \times A \to \mathcal{C}} J_1(x) := \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[f(x(s,a))] - \mathbb{E}_{(s,a)\sim d^{\pi}}[x(s,a)] \tag{24}$$

并且，需要满足 $f'(x^*(s,a)) = w_{\pi/\mathcal{D}}(s,a)$。

同样，使用变量替换：$\nu := x + \mathcal{B}^\pi \nu$，公式 (24) 被重写为：

$$\min_{\nu: S \times A \to \mathbb{R}} J(\nu) := \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[f\left((\nu - \mathcal{B}^\pi \nu)(s,a)\right)\right] - (1-\gamma)\mathbb{E}_{s_0 \sim \beta, a_0 \sim \pi(s_0)}\left[\nu\left(s_0, a_0\right)\right] \tag{25}$$

在此目标下，对 f 应用 Fenchel 对偶性将进一步引出以下求鞍点问题:

$$\begin{aligned}\min_{\nu: S \times A \to \mathbb{R}} \max_{\zeta: S \times A \to \mathbb{R}} J(\nu, \zeta) := \mathbb{E}_{(s,a,s')\sim d^{\mathcal{D}}, a' \sim \pi(s')}\left[\left(\nu(s,a) - \gamma \nu\left(s', a'\right)\right)\zeta(s,a) - f^*(\zeta(s,a))\right] \\ - (1-\gamma)\mathbb{E}_{s_0 \sim \beta, a_0 \sim \pi(s_0)}\left[\nu\left(s_0, a_0\right)\right]\end{aligned} \tag{26}$$

通过内部优化问题 KKT 条件，对于 $\forall \nu$，最优值 ζ_ν^* 满足，

$$f^{*\prime}\left(\zeta_\nu^*(s,a)\right) = (\nu - \mathcal{B}^\pi \nu)(s,a). \tag{27}$$

凸函数的导数 f' 是其 Fenchel 共轭函数导数 $f^{*'}$ 的反函数。通过求鞍点 (ζ^*, ν^*)，可以计算出平稳分布修正 $w_{\pi/\mathcal{D}}(s,a)$，

$$\zeta^*(s,a) = f'\left((\nu^* - \mathcal{B}^\pi \nu^*)(s,a)\right) = f'\left(x^*(s,a)\right) = w_{\pi/\mathcal{D}}(s,a) \tag{28}$$

令人惊讶的是，公式 (26) 中的优化问题保留了使该方法在学习 $w_{\pi/\mathcal{D}}(s,a)$ 时非常实用的所有计算优点：所有数量及其梯度都可以从随机样本中无偏估计；目标 ν 是线性的，ζ 是凹的，因而表现良好；这个问题的优化器立即通过 $\zeta^*(s,a)$ 的值提供所需的平稳分布修正，而不需要任何额外的计算。

3.3 DualDICE 伪代码

A Pseudocode

Algorithm 1 DualDICE

Inputs: Convex function f and its Fenchel conjugate f^*, off-policy data $\hat{\mathcal{D}} = \{(s^{(i)}, a^{(i)}, r^{(i)}, s'^{(i)})\}_{i=1}^N$, sampled initial states $\hat{\beta} = \{s_0^{(i)}\}_{i=1}^M$, target policy π, networks $\nu_{\theta_1}(\cdot,\cdot)$, $\zeta_{\theta_2}(\cdot,\cdot)$, learning rates η_ν, η_ζ, number of iterations T, batch size B.
for $t = 1, \ldots, T$ **do**
  Sample batch $\{(s^{(i)}, a^{(i)}, r^{(i)}, s'^{(i)})\}_{i=1}^B$ from $\hat{\mathcal{D}}$.
  Sample batch $\{s_0^{(i)}\}_{i=1}^B$ from $\hat{\beta}$.
  Sample actions $a'^{(i)} \sim \pi(s'^{(i)})$, for $i = 1, \ldots, B$.
  Sample actions $a_0^{(i)} \sim \pi(s_0^{(i)})$, for $i = 1, \ldots, B$.
  Compute empirical loss $\hat{J} = \frac{1}{B}\sum_{i=1}^B (\nu_{\theta_1}(s^{(i)}, a^{(i)}) - \nu_{\theta_1}(s'^{(i)}, a'^{(i)}))\zeta_{\theta_2}(s^{(i)}, a^{(i)}) - f^*(\zeta_{\theta_2}(s^{(i)}, a^{(i)})) - (1-\gamma)\nu_{\theta_1}(s_0^{(i)}, a_0^{(i)})$.
  Update $\theta_1 \leftarrow \theta_1 - \eta_\nu \nabla_{\theta_1} \hat{J}$.
  Update $\theta_2 \leftarrow \theta_2 + \eta_\zeta \nabla_{\theta_2} \hat{J}$.
end for
Return $\zeta_{\theta_2}(\cdot,\cdot)$.

图 2: DualDICE 伪代码

实际上看伪代码此算法并不难理解。使用网络来优化 ν, ζ，来求鞍点。理论写的挺复杂，实现还挺简单的。。。。。。。

3.4 理论保证

$\{s_i, a_i, r_i, s_i'\}_{i=1}^N \sim d^{\mathcal{D}}, \{s_0^i\}_{i=1}^N \sim \beta$，目标分布为：$a_i' \sim \pi(s_i')$，$a_0^i \sim \pi(s_0^i)$ for $i = 1, \cdots, N.$。$\hat{\mathbb{E}}_{d^{\mathcal{D}}}$ 表示采样出的期望，令 $f(x) = \frac{1}{2}x^2$。我们考虑使用随机梯度下降法从参数族 $\mathcal{F}, \mathcal{H}$ 中寻找公式 (21) 中的最优值 ν, ζ，用 $\hat{\nu}, \hat{\zeta}$ 表示 OPT 的输出。对 off-policy policy estimate (OPE) 问题，估计出的 $\hat{\nu}, \hat{\zeta}$ 的质量有以下保证。

在一些假设下有，

$$\mathbb{E}\left[\left(\hat{\mathbb{E}}_{d^{\mathcal{D}}}[\hat{\zeta}(s,a)\cdot r] - \rho(\pi)\right)^2\right] = \widetilde{\mathcal{O}}\left(\epsilon_{\text{approx}}\left(\mathcal{F}, \mathcal{H}\right) + \epsilon_{opt} + \frac{1}{\sqrt{N}}\right) \tag{29}$$

ϵ_{opt} 表示最优误差，$\epsilon_{\text{approx}}\left(\mathcal{F}, \mathcal{H}\right)$ 表示由于 $\mathcal{F}, \mathcal{H}$ 产生的近似误差。根据此结论可以等出，随着采样数 N 的增加，$\frac{1}{\sqrt{N}} \to 0$。同时，$\epsilon_{\text{approx}}\left(\mathcal{F}, \mathcal{H}\right)$ 和 ϵ_{opt} 直接有一种隐式的平衡关系。这部分的详细分析在文章的附录部分。

4 Proof

本部分是对公式 (29) 的证明，首先对误差进行分解，然后我们分析了统计误差和优化误差，最后讨论了整体误差。

证明之前需要几个假设条件，在假设 1 中假设了，$\|\nu\|_\infty \leq C$，可以推出 $\|(\nu - \mathcal{B}^\pi \nu)\|_\infty \leq \frac{1+\gamma}{1-\gamma} C$。同样可以得到 $\|w\|_\infty \leq C$。假设观察到的奖励是有界限的，$\|\hat{r}(s,a)\|_\infty \leq C_r$，奖励的均值方差都在此区间内 $[-C_r, C_r]$。

重复一下 dualdice 的目标函数，

$$\begin{aligned} J(\nu,\zeta) =& \mathbb{E}_{(s,a,s')\sim d^{\mathcal{D}}, a'\sim\pi(s')} \left[\left(\nu(s,a) - \gamma\nu\left(s',a'\right)\right)\zeta(s,a) - \zeta(s,a)^2/2 \right] \\ & - (1-\gamma)\mathbb{E}_{s_0\sim\beta, a_0\sim\pi(s_0)} \left[\nu\left(s_0,a_0\right)\right] \end{aligned} \tag{30}$$

并且引入 ζ 之前的目标函数形式为，

$$J(\nu) = \frac{1}{2}\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}} \left[\left(\nu - \mathcal{B}^\pi \nu\right)(s,a)^2 \right] - (1-\gamma)\mathbb{E}_{s_0\sim\beta, a_0\sim\pi(s_0)} \left[\nu\left(s_0,a_0\right)\right] \tag{31}$$

此处的定义值得好好看看。

1. 令 $\hat{J}(\nu,\zeta)$ 表示对 $J(\nu,\zeta)$ 的估计，其中最优解为 $(\hat{\nu}^*, \hat{\zeta}^*)$。

2. $\nu^*_{\mathcal{F}} = \arg\min_{\nu\in\mathcal{F}} J(\nu)$ and $\nu^* = \arg\min_{\nu\in S\times A\to\mathbb{R}} J(\nu)$ 前者相比于后者，多了一个限制条件 $\nu \in \mathcal{F}$。

3. 令 $L(\nu) = \max_{\zeta\in\mathcal{H}} J(\nu,\zeta)$ 和 $\hat{L}(\nu) = \max_{\zeta\in\mathcal{H}} \hat{J}(\nu,\zeta)$ 表示原问题。$\ell(\zeta) = \min_{\nu\in\mathcal{F}} J(\nu,\zeta), \hat{\ell}(\zeta) = \min_{\nu\in\mathcal{F}} \hat{J}(\nu,\zeta)$。

4. 我们使用一些优化算法来优化 $\hat{J}(\nu,\zeta)$，其中样本 $\{s_i, a_i, r_i, s_i'\}_{i=1}^N \sim d^{\mathcal{D}}, \{s_0^i\}_{i=1}^N \sim \beta$，目标分布动作为 $a_i' \sim \pi\left(s_i'\right), a_0^i \sim \pi(s_0^i)$。用 $(\hat{\nu}, \hat{\zeta})$ 表示优化算法的输出。

5. 令观测到的数据为：

$$\begin{aligned} \bar{R}(s,a) =& \mathbb{E}_{\cdot|s,a}[r] \\ \rho(\pi) =& \mathbb{E}_{d^{\mathcal{D}}}\left[w_{\pi/\mathcal{D}}(s,a)\cdot\bar{R}(s,a)\right]. \end{aligned} \tag{32}$$

我们首先考虑使用 $\left(\hat{\nu} - \hat{\mathcal{B}}^\pi\hat{\nu}\right)(s,a)$ 作为 $w_{\pi/\mathcal{D}}(s,a)$ 的估计引起的估计误差。其中，$\hat{\mathcal{B}}$ 表示从 $d^{\mathcal{D}}, \pi$ 中采样计算的 Bellman backup。我们随后会将其与 DualDICE 的真实做法相结合，其使用 $\hat{\zeta}(s,a)$ 作为 $w_{\pi/\mathcal{D}}(s,a)$ 的估计。

4.1 误差分解

$\left(\hat{\nu} - \hat{\mathcal{B}}^\pi\hat{\nu}\right)(s,a)$ 和 $w_{\pi/\mathcal{D}}(s,a)$ 之间的均方误差可以被分解为：

$$\begin{aligned} & \left(\hat{\mathbb{E}}_{d^{\mathcal{D}}}\left[\left(\hat{\nu} - \hat{\mathcal{B}}^\pi\hat{\nu}\right)(s,a)\cdot r\right] - \mathbb{E}_{d^{\mathcal{D}}}\left[w_{\pi/\mathcal{D}}(s,a)\cdot\bar{R}(s,a)\right]\right)^2 \\ =& \left(\hat{\mathbb{E}}_{d^{\mathcal{D}}}\left[\left(\hat{\nu} - \hat{\mathcal{B}}^\pi\hat{\nu}\right)(s,a)\cdot r\right] - \hat{\mathbb{E}}_{d^{\mathcal{D}}}\left[\left(\hat{\nu} - \hat{\mathcal{B}}^\pi\hat{\nu}\right)(s,a)\cdot\bar{R}(s,a)\right]\right. \\ & + \hat{\mathbb{E}}_{d^{\mathcal{D}}}\left[\left(\hat{\nu} - \hat{\mathcal{B}}^\pi\hat{\nu}\right)(s,a)\cdot\bar{R}(s,a)\right] - \hat{\mathbb{E}}_{d^{\mathcal{D}}}\left[\left(\hat{\nu}^* - \hat{\mathcal{B}}^\pi\hat{\nu}^*\right)(s,a)\cdot\bar{R}(s,a)\right] \\ & \left. + \hat{\mathbb{E}}_{d^{\mathcal{D}}}\left[\left(\hat{\nu}^* - \hat{\mathcal{B}}^\pi\hat{\nu}^*\right)(s,a)\cdot\bar{R}(s,a)\right] - \mathbb{E}_{d^{\mathcal{D}}}\left[w_{\pi/\mathcal{D}}(s,a)\cdot\bar{R}(s,a)\right]\right)^2 \\ \leq& 4\underbrace{\left(\hat{\mathbb{E}}_{d^{\mathcal{D}}}\left[\left(\hat{\nu} - \hat{\mathcal{B}}^\pi\hat{\nu}\right)(s,a)\cdot r\right] - \hat{\mathbb{E}}_{d^{\mathcal{D}}}\left[\left(\hat{\nu} - \hat{\mathcal{B}}^\pi\hat{\nu}\right)(s,a)\cdot\bar{R}(s,a)\right]\right)^2}_{\epsilon_r} \\ & + 4\underbrace{\left(\hat{\mathbb{E}}_{d^{\mathcal{D}}}\left[\left(\hat{\nu} - \hat{\mathcal{B}}^\pi\hat{\nu}\right)(s,a)\cdot\bar{R}(s,a)\right] - \hat{\mathbb{E}}_{d^{\mathcal{D}}}\left[\left(\hat{\nu}^* - \hat{\mathcal{B}}^\pi\hat{\nu}^*\right)(s,a)\cdot\bar{R}(s,a)\right]\right)^2}_{\epsilon_1} \\ & + 4\underbrace{\left(\hat{\mathbb{E}}_{d^{\mathcal{D}}}\left[\left(\hat{\nu}^* - \hat{\mathcal{B}}^\pi\hat{\nu}^*\right)(s,a)\cdot\bar{R}(s,a)\right] - \mathbb{E}_{d^{\mathcal{D}}}[w_{\pi/\mathcal{D}}(s,a)\cdot\bar{R}(s,a)]\right)^2}_{\epsilon_2}. \end{aligned} \tag{33}$$

这一步推导比较简单就是加一项，减一项，然后拆开。

$$\begin{aligned}(a-b)^2 =&(a-c+c-d+d-b)^2\\ =&(a-c)^2+(c-d)^2+(d-b)^2+2(a-c)(c-d)+2(a-c)(c-d)+2(c-d)(d-b)\\ \leq&3(a-c)^2+3(c-d)^2+3(d-b)^2\\ \leq&4(a-c)^2+4(c-d)^2+4(d-b)^2\end{aligned} \tag{34}$$

第一项 ϵ_r 是由观察到的奖励的随机性引起的，我们有，

$$\epsilon_r=\left(\hat{\mathbb{E}}_{d^{\mathcal{D}}}\left[\left(\hat{\nu}-\hat{\mathcal{B}}^{\pi}\hat{\nu}\right)(s,a)\cdot(r(s,a)-\overline{R}(s,a))\right]\right)^2\leq\left(\frac{1+\gamma}{1-\gamma}\right)^2C^2\left(\hat{\mathbb{E}}_{d^{\mathcal{D}}}[r(s,a)]-\hat{\mathbb{E}}_{d^{\mathcal{D}}}[\overline{R}(s,a)]\right)^2 \tag{35}$$

第二项是优化过程中要引起的误差，估计值和计算出的最优解之间的误差：

$$\epsilon_1\leq C_r^2\left\|\left(\hat{\nu}-\hat{\mathcal{B}}^{\pi}\hat{\nu}\right)-\left(\hat{\nu}^*-\hat{\mathcal{B}}^{\pi}\hat{\nu}^*\right)\right\|_{\hat{\mathcal{D}}}^2\leq C_r^2(\underbrace{\left\|\hat{\zeta}-\hat{\zeta}^*\right\|_{\hat{\mathcal{D}}}^2+\left\|\left(\hat{\nu}^*-\hat{\mathcal{B}}^{\pi}\hat{\nu}^*\right)-\left(\hat{\nu}-\hat{\mathcal{B}}^{\pi}\hat{\nu}\right)\right\|_{\hat{\mathcal{O}}}^2}_{\hat{\epsilon}_{ppt}}) \tag{36}$$

对于最后一项有：

$$\begin{aligned}\epsilon_2\leq&2\underbrace{\left(\hat{\mathbb{E}}_{d^{\mathcal{D}}}\left[\left(\hat{\nu}^*-\hat{\mathcal{B}}^{\pi}\hat{\nu}^*\right)(s,a)\cdot r(s,a)\right]-\mathbb{E}_{d^{\mathcal{D}}}\left[\left(\hat{\nu}^*-\mathcal{B}^{\pi}\hat{\nu}^*\right)(s,a)\cdot r(s,a)\right]\right)^2}_{\epsilon_{\text{stat}}}\\ &+2\left(\mathbb{E}_{d^{\mathcal{D}}}\left[\left(\hat{\nu}^*-\mathcal{B}^{\pi}\hat{\nu}^*\right)(s,a)\cdot r(s,a)\right]-\mathbb{E}_{d^{\mathcal{D}}}\left[w_{\pi/\mathcal{D}}(s,a)\cdot r(s,a)\right]\right)^2\\ \leq&2\epsilon_{\text{stat}}+2\left(\mathbb{E}_{d^{\mathcal{D}}}\left[\left(\hat{\nu}^*-\mathcal{B}^{\pi}\hat{\nu}^*\right)(s,a)\cdot r(s,a)\right]-\mathbb{E}_{d^{\mathcal{D}}}\left[\left(\nu^*-\mathcal{B}^{\pi}\nu^*\right)(s,a)\cdot r(s,a)\right]\right)^2\end{aligned} \tag{37}$$

此公式最后一行的推导由公式 (14) 得到。对于公式 (37) 中的第一项，是由于确定性采样得到的，之后会进行详细分析。对于第二项有：

$$\begin{aligned}&\left(\mathbb{E}_{d\mathcal{D}}\left[\left(\hat{\nu}^*-\mathcal{B}^{\pi}\hat{\nu}^*\right)(s,a)\cdot r(s,a)\right]-\mathbb{E}_{d^{\mathcal{D}}}\left[\left(\nu^*-\mathcal{B}^{\pi}\nu^*\right)(s,a)\cdot r(s,a)\right]\right)^2\\ \leq&\mathbb{E}_{d^{\mathcal{D}}}\left[r(s,a)^2\cdot\left(\left(\hat{\nu}^*-\mathcal{B}^{\pi}\hat{\nu}^*\right)(s,a)-\left(\nu^*-\mathcal{B}^{\pi}\nu^*\right)(s,a)\right)^2\right]\\ \leq&C_r^2\left\|\left(\hat{\nu}^*-\mathcal{B}^{\pi}\hat{\nu}^*\right)-\left(\nu^*-\mathcal{B}^{\pi}\nu^*\right)\right\|_{\mathcal{D}}^2\\ \leq&\frac{2C_r^2}{\eta}\left(J\left(\hat{\nu}^*\right)-J\left(\nu^*\right)\right)\end{aligned} \tag{38}$$

最后一行不等式来自 f 的 1-strongly convexity。由于 ν^* 是最优解，所以其一阶导数等于 0 根据下面的不等式推出。

如果梯度是L-Lipschitz的，就有了一个**二次函数的上界**：

$$f(x)\leqslant f(x_0)+<\nabla f(x_0),x-x_0>+\frac{L}{2}\|x-x_0\|^2\quad\forall x$$

如果是 μ -strongly convex的，就有了一个**二次函数的下界**：

$$f(x_0)+<\nabla f(x_0),x-x_0>+\frac{\mu}{2}\|x-x_0\|^2\leqslant f(x)\quad\forall x$$

当我们考虑 $J(\hat{\nu}^*)$ 和 $J(\nu^*)$ 之间的误差，可以被分解为：

$$\begin{aligned} J\left(\hat{\nu}^*\right)-J\left(\nu^*\right) &= J\left(\hat{\nu}^*\right)-J\left(\nu_{\mathcal{F}}^*\right)+J\left(\nu_{\mathcal{F}}^*\right)-J\left(\nu^*\right) \\ &= J\left(\hat{\nu}^*\right)-L\left(\hat{\nu}^*\right)+L\left(\hat{\nu}^*\right)-L\left(\nu_{\mathcal{F}}^*\right)+L\left(\nu_{\mathcal{F}}^*\right)-J\left(\nu_{\mathcal{F}}^*\right)+J\left(\nu_{\mathcal{F}}^*\right)-J\left(\nu^*\right) \end{aligned} \tag{39}$$

通过分解可以得到，

$$\begin{aligned} J\left(\nu_{\mathcal{F}}^*\right)-J\left(\nu^*\right) &= \mathbb{E}_{\mathcal{D}}\left[f\left(\nu_{\mathcal{F}}^*-\mathcal{B}^\pi\nu_{\mathcal{F}}^*\right)-f\left(\nu^*-\mathcal{B}^\pi\nu^*\right)\right]-\mathbb{E}_{\beta\pi}\left[\nu_{\mathcal{F}}^*-\nu^*\right] \\ &\leq \kappa\left\|\nu_{\mathcal{F}}^*-\nu^*\right\|_{\mathcal{D},1}+\kappa\left\|\mathcal{B}^\pi\left(\nu_{\mathcal{F}}^*-\nu^*\right)\right\|_{\mathcal{D},1}+\left\|\nu_{\mathcal{F}}^*-\nu^*\right\|_{\beta\pi,1} \\ &\leq \max\left(\kappa+\kappa\left\|\mathcal{B}^\pi\right\|_{\mathcal{D},1},1\right)\left(\left\|\nu_{\mathcal{F}}^*-\nu^*\right\|_{\mathcal{D},1}+\left\|\nu_{\mathcal{F}}^*-\nu^*\right\|_{\beta\pi,1}\right) \\ &\leq \max\left(\kappa+\kappa\left\|\mathcal{B}^\pi\right\|_{\mathcal{D},1},1\right)\cdot\epsilon_{\text{approx}}\left(\mathcal{F}\right) \end{aligned} \tag{40}$$

其中，$\epsilon_{\text{approx}}\left(\mathcal{F}\right):=\sup_{\nu\in S\times A\to\mathbb{R}}\inf_{\nu\in\mathcal{F}}\left(\left\|\nu_{\mathcal{F}}-\nu\right\|_{\mathcal{D},1}+\left\|\nu_{\mathcal{F}}-\nu\right\|_{\beta\pi,1}\right)$。

对于 $L\left(\nu_{\mathcal{F}}^*\right)-J\left(\nu_{\mathcal{F}}^*\right)$ 有

$$L\left(\nu_{\mathcal{F}}^*\right)-J\left(\nu_{\mathcal{F}}^*\right)=\max_{\zeta\in\mathcal{H}}J\left(\nu_{\mathcal{F}}^*,\zeta\right)-\max_{\zeta\in S\times A\to\mathbb{R}}J\left(\nu_{\mathcal{F}}^*,\zeta\right)\leq 0 \tag{41}$$

对于 $L\left(\hat{\nu}^*\right)-L\left(\nu_{\mathcal{F}}^*\right)$，

$$\begin{aligned} L\left(\hat{\nu}^*\right)-L\left(\nu_{\mathcal{F}}^*\right) &= L\left(\hat{\nu}^*\right)-\hat{L}\left(\hat{\nu}^*\right)+\hat{L}\left(\hat{\nu}^*\right)-\hat{L}\left(\nu_{\mathcal{F}}^*\right)+\hat{L}\left(\nu_{\mathcal{F}}^*\right)-L\left(\nu_{\mathcal{F}}^*\right) \\ &\leq L\left(\hat{\nu}^*\right)-\hat{L}\left(\hat{\nu}^*\right)+\hat{L}\left(\nu_{\mathcal{F}}^*\right)-L\left(\nu_{\mathcal{F}}^*\right) \\ &\leq 2\sup_{\nu\in\mathcal{F}}\left|L(\nu)-\hat{L}(\nu)\right| \\ &= 2\sup_{\nu\in\mathcal{F}}\left|\max_{\zeta\in\mathcal{H}}J(\nu,\zeta)-\max_{\zeta\in\mathcal{H}}\hat{J}(\nu,\zeta)\right| \\ &\leq 2\sup_{\nu\in\mathcal{F},\zeta\in\mathcal{H}}\left|\hat{J}(\nu,\zeta)-J(\nu,\zeta)\right| \\ &= 2\cdot\epsilon_{est}(\mathcal{F}) \end{aligned} \tag{42}$$

对于第一项有：$\hat{L}\left(\hat{\nu}^*\right)-\hat{L}\left(\nu_{\mathcal{F}}^*\right)\leq 0$。

对于 $J\left(\hat{\nu}^*\right)-L\left(\hat{\nu}^*\right)$ 有，

$$\begin{aligned} J\left(\hat{\nu}^*\right)-L\left(\hat{\nu}^*\right) &= \max_{\zeta\in S\times A\to\mathbb{R}}J\left(\hat{\nu}^*,\zeta\right)-\max_{\zeta\in\mathcal{H}}J\left(\hat{\nu}^*,\zeta\right) \\ &\leq \left(L+\frac{1+\gamma}{1-\gamma}C\right)\underbrace{\left\|\zeta_{\mathcal{H}}^*-\zeta^*\right\|_{\mathcal{D},1}}_{\leq\epsilon_{\text{ approx }}\left(\mathcal{H}\right)}, \end{aligned} \tag{43}$$

其中，$\epsilon_{\text{approx}}\left(\mathcal{H}\right):=\sup_{\zeta\in S\times A\to\mathbb{R}}\inf_{\zeta\in\mathcal{H}}\left(\left\|\zeta_{\mathcal{H}}-\zeta\right\|_{\mathcal{D},1}+\left\|\zeta_{\mathcal{H}}-\zeta\right\|_{\beta\pi,1}\right)$ 合并误差可得：

$$\begin{aligned} &\left(\hat{\mathbb{E}}_{d^{\mathcal{D}}}\left[\left(\hat{\nu}-\hat{\mathcal{B}}^\pi\hat{\nu}\right)(s,a)\cdot\hat{r}(s,a)\right]-\rho(\pi)\right)^2 \\ &\qquad\leq \frac{16C_r^2}{\eta}\left(\max\left(\kappa+\kappa\left\|\mathcal{B}^\pi\right\|_{\mathcal{D},1},1\right)\epsilon_{approx}(\mathcal{F})+\left(L+\frac{1+\gamma}{1-\gamma}C\right)\epsilon_{approx}(\mathcal{H})\right) \\ &\qquad\qquad +4\epsilon_r+8\epsilon_{\text{stat}}+\frac{32C_r^2}{\eta}\epsilon_{\text{est}}\left(\mathcal{F}\right)+4\hat{\epsilon}_{\text{opt}}\,. \end{aligned} \tag{44}$$

对偶 OPE 估计：在实现 DualDICE 时，使用估计，

$$\hat{\mathbb{E}}_{d^{\mathcal{D}}}[\hat{\zeta}(s,a)\cdot r]$$

误差可以被分解为：

$$\begin{aligned}
&\left(\hat{\mathbb{E}}_{d^{\mathcal{D}}}[\hat{\zeta}(s,a)\cdot r]-\mathbb{E}_{d^{\mathcal{D}}}\left[w_{\pi/\mathcal{D}}(s,a)\cdot\bar{R}(s,a)\right]\right)^2\\
\leq&2\left(\hat{\mathbb{E}}_{d^{\mathcal{D}}}[\hat{\zeta}(s,a)\cdot r]-\hat{\mathbb{E}}_{d^{\mathcal{D}}}\left[\left(\hat{\nu}-\hat{\mathcal{B}}^{\pi}\hat{\nu}\right)(s,a)\cdot r\right]\right)^2\\
&+2\left(\hat{\mathbb{E}}_{d^{\mathcal{D}}}\left[\left(\hat{\nu}-\hat{\mathcal{B}}^{\pi}\hat{\nu}\right)(s,a)\cdot r\right]-\mathbb{E}_{d^{\mathcal{D}}}\left[w_{\pi/\mathcal{D}}(s,a)\cdot\bar{R}(s,a)\right]\right)^2
\end{aligned}\tag{45}$$

其中，第二项已经求解过了，即为公式 (44) 中的结果。而第一项为：

$$\left(\hat{\mathbb{E}}_{d^{\mathcal{D}}}[\hat{\zeta}(s,a)\cdot r]-\hat{\mathbb{E}}_{d^{\mathcal{D}}}\left[\left(\hat{\nu}-\hat{\mathcal{B}}^{\pi}\hat{\nu}\right)(s,a)\cdot r\right]\right)^2\leq C_r^2\left\|\hat{\zeta}-\left(\hat{\nu}-\hat{\mathcal{B}}^{\pi}\hat{\nu}\right)\right\|_{\hat{\mathcal{D}}}^2\tag{46}$$

而，

$$\begin{aligned}
&\left\|\hat{\zeta}-\left(\hat{\nu}-\hat{\mathcal{B}}^{\pi}\hat{\nu}\right)\right\|_{\hat{\mathcal{D}}}^2\\
=&\left\|\hat{\zeta}-\hat{\zeta}^*+\hat{\zeta}^*-\left(\hat{\nu}^*-\hat{\mathcal{B}}^{\pi}\hat{\nu}^*\right)+\left(\hat{\nu}^*-\hat{\mathcal{B}}^{\pi}\hat{\nu}^*\right)-\left(\hat{\nu}-\hat{\mathcal{B}}^{\pi}\hat{\nu}\right)\right\|_{\hat{\mathcal{D}}}^2\\
\leq&4\left\|\hat{\zeta}-\hat{\zeta}^*\right\|_{\hat{\mathcal{D}}}^2+4\left\|\left(\hat{\nu}^*-\hat{\mathcal{B}}^{\pi}\hat{\nu}^*\right)-\left(\hat{\nu}-\hat{\mathcal{B}}^{\pi}\hat{\nu}\right)\right\|_{\hat{\mathcal{D}}}^2+4\left\|\hat{\zeta}^*-\left(\hat{\nu}^*-\hat{\mathcal{B}}^{\pi}\hat{\nu}^*\right)\right\|_{\hat{\mathcal{D}}}^2
\end{aligned}\tag{47}$$

前两项加起来是 ϵ_{opt}，后一项等于 0。其中，$\frac{16C_r^2}{\eta}\left(\max\left(\kappa+\kappa\left\|\mathcal{B}^{\pi}\right\|_{\mathcal{D},1},1\right)\epsilon_{approx}(\mathcal{F})+\left(L+\frac{1+\gamma}{1-\gamma}C\right)\epsilon_{approx}(\mathcal{H})\right)$ 是 $\epsilon_{approx}(\mathcal{F},\mathcal{H})$。那么接下来主要分析的是，$\epsilon_r,\epsilon_{\text{stat}},\epsilon_{\text{est}}(\mathcal{F})$。

4.2 统计学误差

4.2.1 Bounding ϵ_r

$$\begin{aligned}
\mathbb{E}[\epsilon_r]&\leq\left(\frac{1+\gamma}{1-\gamma}\right)^2C^2\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}r_i-\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}r_i\right]\right)^2\right]\\
&=\left(\frac{1+\gamma}{1-\gamma}\right)^2C^2\mathbb{V}\left(\frac{1}{N}\sum_{i=1}^{N}r_i\right)\\
&\leq\frac{1}{N}\left(\frac{1+\gamma}{1-\gamma}\right)^2C^2\sup_{s,a}\mathbb{V}(r\mid s,a)=\mathcal{O}\left(\frac{1}{N}\right)
\end{aligned}\tag{48}$$

4.2.2 Bounding $\epsilon_{\text{est}}(\mathcal{F})$

$$\epsilon_{est}(\mathcal{F})=\sup_{\nu\in\mathcal{F},\zeta\in\mathcal{H}}|\hat{J}(\nu,\zeta)-J(\nu,\zeta)|$$

此推导需要用到三个引理，

Lemma 4. *[35] Let $\mathcal{G}$ be a permissible class of $\mathcal{Z} \to [-M, M]$ functions and $\{Z_i\}_{i=1}^N$ are* i.i.d. *samples from some distribution. Then, for any given $\epsilon > 0$,*

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{N}\sum_{i=1}^{N} g(Z_i) - \mathbb{E}\left[g(Z)\right]\right| > \epsilon\right) \leq 8\mathbb{E}\left[\mathcal{N}_1\left(\frac{\epsilon}{8}, \mathcal{G}, \{Z_i\}_{i=1}^N\right)\right] \exp\left(\frac{-N\epsilon^2}{512M^2}\right).$$

The covering number can then be bounded in terms of the function class's pseudo-dimension:

Lemma 5. *[Corollary 3, [20]] For any set $\mathcal{X}$, any points $x^{1:N} \in \mathcal{X}^N$, any class $\mathcal{F}$ of functions on $\mathcal{X}$ taking values in $[0, M]$ with pseudo-dimension $D_{\mathcal{F}} < \infty$, and any $\epsilon > 0$,*

$$\mathcal{N}_1\left(\epsilon, \mathcal{F}, x^{1:N}\right) \leq e\left(D_{\mathcal{F}} + 1\right)\left(\frac{2eM}{\epsilon}\right)^{D_{\mathcal{F}}}.$$

With the above technical lemmas, we are ready to bound $\epsilon_{est}\left(\mathcal{F}\right)$.

Lemma 6 (Statistical error $\epsilon_{est}\left(\mathcal{F}\right)$)**.** *Under Assumption 1, if f^* is L-Lipschitz continuous, with at least probability $1-\delta$,*

$$\epsilon_{est}\left(\mathcal{F}\right) = \mathcal{O}\left(\sqrt{\frac{\log N + \log\frac{1}{\delta}}{N}}\right).$$

证明:

$$h_{\nu,\zeta}\left(s, a, s', a', s_0, a_0\right) = \left(\nu(s,a) - \gamma\nu\left(s', a'\right)\right)\zeta(s,a) - f^*(\zeta(s,a)) - (1-\gamma)\nu(s_0, a_0) \tag{49}$$

令 $\mathcal{Z} = \underbrace{S\times A\times S\times A}_{d^D\pi} \times \underbrace{S\times A}_{\beta\pi}, Z_i = (s_i, a_i, s_i', a_i', s_0^i, a_0^i)$, $\mathcal{G} = h_{\mathcal{F}\times\mathcal{H}}$。我们首先要证明 $h_{\nu,\zeta}$ 有界。首先 $\nu \in \mathcal{F}$, $\zeta \in \mathcal{H}$ 被 $\frac{1}{1-\gamma}C$ 和 C 约束。并且 $f^*(\cdot)$ 是 L-Lipzchitz 连续。

$$\begin{aligned}
\|h_{\nu,\zeta}\|_\infty &\leq (1+\gamma)\|\nu\|_\infty\|\zeta\|_\infty + (1-\gamma)\|\nu\|_\infty + \|f^*(\zeta)\|_\infty \\
&\leq \frac{1+\gamma}{1-\gamma}C^2 + C + \|f^*(\zeta) - f^*(0)\|_\infty + |f^*(0)| \\
&\leq \frac{1+\gamma}{1-\gamma}C^2 + C + L\|\zeta\|_\infty + |f^*(0)| \\
&\leq \frac{1+\gamma}{1-\gamma}C^2 + C + LC + |f^*(0)|
\end{aligned} \tag{50}$$

这样有,

$$\begin{aligned}
\mathbb{P}\left(\sup_{\nu\in\mathcal{F},\zeta\in\mathcal{H}}|\hat{J}(\nu,\zeta) - J(\nu,\zeta)| \geq \epsilon\right) &= \mathbb{P}\left(\sup_{\nu\in\mathcal{F},\zeta\in\mathcal{H}}|\frac{1}{N}\sum_{i=1}^{N} h_{\nu,\zeta}\left(Z_i\right) - \mathbb{E}\left[h_{\nu,\zeta}\right]| \geq \epsilon\right) \\
&\leq 8\mathbb{E}\left[\mathcal{N}_1\left(\frac{\epsilon}{8}, \mathcal{G}, \{Z_i\}_{i=1}^N\right)\right]\exp\left(\frac{-N\epsilon^2}{512M_1^2}\right).
\end{aligned} \tag{51}$$

然后，求 $\mathcal{G}$ 中的距离边界，

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\left|h_{\nu_1,\zeta_1}\left(Z_i\right)-h_{\nu_2,\zeta_2}\left(Z_i\right)\right| \\
\leq&\frac{\left(L+\frac{1+\gamma}{1-\gamma}C\right)}{N}\sum_{i=1}^{N}\left|\zeta_1\left(s_i,a_i\right)-\zeta_2\left(s_i,a_i\right)\right|+\frac{C}{N}\sum_{i=1}^{N}\left|\nu_1\left(s_i,a_i\right)-\nu_2\left(s_i,a_i\right)\right| \\
&+\frac{\gamma C}{N}\sum_{i=1}^{N}\left|\nu_1\left(s_i',a_i'\right)-\nu_2\left(s_i',a_i'\right)\right|+\frac{(1-\gamma)}{N}\sum_{i=1}^{N}\left|\nu_1\left(s_0^i,a_i^0\right)-\nu_2\left(s_0^i,a_i^0\right)\right|
\end{aligned} \tag{52}$$

可以推出

$$\begin{aligned}
&\mathcal{N}_1\left(\left(L+\frac{2+\gamma-\gamma^2}{1-\gamma}C+(1-\gamma)\right)\epsilon',\mathcal{G},\{Z_i\}_{i=1}^N\right) \\
&\leq\mathcal{N}_1\left(\epsilon',\mathcal{H},\{s_i,a_i\}_{i=1}^N\right)\mathcal{N}_1\left(\epsilon',\mathcal{F},\{s_i,a_i\}_{i=1}^N\right)\mathcal{N}_1\left(\epsilon',\mathcal{F},\{s_i',a_i'\}_{i=1}^N\right)\mathcal{N}_1\left(\epsilon',\mathcal{F},\left\{s_0^i,a_0^i\right\}_{i=1}^N\right)
\end{aligned} \tag{53}$$

其中，使用引理 5 可得，

$$\mathcal{N}_1\left(\left(L+\frac{2+\gamma-\gamma^2}{1-\gamma}C+(1-\gamma)\right)\epsilon',\mathcal{G},\{Z_i\}_{i=1}^N\right)\leq e^4\left(D_{\mathcal{F}}+1\right)^3\left(D_{\mathcal{H}}+1\right)\left(\frac{4eM_1}{\epsilon'}\right)^{3D_{\mathcal{F}}+D_{\mathcal{H}}} \tag{54}$$

做一个变量替换，

$$\begin{aligned}
&\mathcal{N}_1\left(\tfrac{\epsilon}{8},\mathcal{G},\{Z_i\}_{i=1}^N\right) \\
&\leq e^4\left(D_{\mathcal{F}}+1\right)^3\left(D_{\mathcal{H}}+1\right)\left(\frac{32\left(L+\frac{2+\gamma-\gamma^2}{1-\gamma}C+(1-\gamma)\right)eM_1}{\epsilon}\right)^{3D_{\mathcal{F}}+D_{\mathcal{H}}} \qquad :=C_1\left(\tfrac{1}{\epsilon}\right)^{D_1}
\end{aligned} \tag{55}$$

其中，$C_1=e^4\left(D_{\mathcal{F}}+1\right)^3\left(D_{\mathcal{H}}+1\right)\left(32\left(L+\frac{2+\gamma-\gamma^2}{1-\gamma}C+(1-\gamma)\right)eM_1\right)^{D_1}$ 和 $D_1=3D_{\mathcal{F}}+D_{\mathcal{H}}$。那么结合公式 (51) 有，

$$\mathbb{P}\left(\sup_{\nu\in\mathcal{F},\zeta\in\mathcal{H}}|\hat{J}(\nu,\zeta)-J(\nu,\zeta)|\geq\epsilon\right)\leq 8C_1\left(\frac{1}{\epsilon}\right)^{D_1}\exp\left(\frac{-N\epsilon^2}{512M_1^2}\right) \tag{56}$$

可以解出，当 $\epsilon=\sqrt{\frac{C_2\left(\log N+\log\frac{1}{\delta}\right)}{N}}$ with $C_2=\max\left(\left(8C_1\right)^{\frac{2}{D_1}},512M_1D_1,512M_1,1\right)$，有

$$8C_1\left(\frac{1}{\epsilon}\right)^{D_1}\exp\left(\frac{-N\epsilon^2}{512M_1^2}\right)\leq\delta \tag{57}$$

从而推导出引理 6。

4.2.3 Bounding ϵ_{stat}

$$\epsilon_{stat}=\mathcal{O}\left(\frac{\log N+\log\frac{1}{\delta}}{N}\right) \tag{58}$$

4.3 误差合并

按照公式 (44) 合并就可以得出结论：

$$\mathbb{E}\left[\left(\hat{\mathbb{E}}_d\mathcal{D}[\hat{\zeta}(s,a)\cdot\hat{r}(s,a)]-\mathbb{E}_{d^{\mathcal{D}}}\left[w_{\pi/\mathcal{D}}(s,a)\cdot r(s,a)\right]\right)^2\right]=\widetilde{\mathcal{O}}\left(\epsilon_{approx}(\mathcal{F},\mathcal{H})+\epsilon_{opt}+\frac{1}{\sqrt{N}}\right) \tag{59}$$