arXiv:2301.08215v1 [cs.LG] 19 Jan 2023

利用决策估计系数为交互式决策提供紧确保证

DylanJ. Foster NoahGolowich*
YanjunHandylanfoster@microsoft.com nzg@mit.edu yjhan@mit.edu

2023年1月19日

摘要

强化学习和交互式决策中的一个基础问题是，理解哪些建模假设能够带来样本高效的学习保证，以及哪些算法设计原则能够实现最优样本复杂度。近期，Foster 等人(2021年) 提出了决策估计系数（DEC），这是一种统计复杂度度量，可为涵盖多臂老虎机以及采用函数逼近的强化学习在内的一般问题类，推导出最优样本复杂度的上界和下界。本文提出 DEC 的一种新变体——受约束决策-估计系数，并利用它推导出新的下界，在以下三个方面改进了先前工作：

- 这些结果在期望意义下成立，且对所考虑的算法类没有任何限制。
- 这些结果在全局范围内成立，不依赖局部化这一概念，而该概念由Foster 等人(2021年) 使用。
- 最有趣的是，它们允许定义 DEC 所依据的参考模型是非适当的，这确立了不适当参考模型发挥着根本性作用。

我们给出了与同一量成比例的遗憾上界，从而弥合了 Foster 等人 (2021年中除一个以外的所有上界与下界之间的间隙。我们的结果适用于遗憾框架和 PAC 框架，并运用了若干新的分析和算法设计技术，我们预期这些技术将在更广泛的领域发挥作用。

目录

1 引言 **2**
1.1 交互式决策.2 1.2 背景：决策估计系数.4 1.3 受约束决策估计系数与结果概述.5 1.4 预备知识.6 1.5 组织结构.7

2 下界 **7**
2.1 主要结果.7 2.2 PAC下界的证明（定理 2.1）.10

3 上界 **13**
3.1 在线估计.14 3.2 PAC 的算法与上界.14 3.3 PAC 上界的证明（定理 3.1）.16 3.4 算法与遗憾上界.18

4 决策估计系数：结构性质 **22**

*由 Fannie & John Hertz 基金会奖学金和美国国家科学基金会研究生奖学金资助。

4.1 受约束 DEC 与偏移 DEC 之间的关系.22 4.2 局部化.24 4.3 参考模型：次优性的作用.25 4.4 参考模型：凸性与随机化的作用.26

5 相较先前工作的改进 275
.1 其他 onal Related Work 29
.6 其他示例 29 A 预备知识 33A.1 极小极大定理.33**B 第 2 节中省略的证明 33**B.1 遗憾下界的证明（定理 2.2）.33**C 第 3 节中省略的证明 39**C.1 遗憾上界的证明（定理 3.2）.39**D 第4节的证明和补充结果 45**D.1 技术引理.45 D.2 决策估计系数的其他性质.46 D.3 省略的证明.49

E 第 5 节中省略的证明 **58**

1 引言

强化学习和决策中的样本效率是一项根本性挑战。最先进的算法（Lillicrap 等人，2015年；Mnih 等人，2015年；Silver 等人，2016年）通常需要数百万轮交互才能达到人类水平的性能，这在实践中难以承受，也构成了可靠部署的障碍。为了在具有挑战性的现实世界领域中持续取得进展——在这些领域中，智能体必须在高维状态空间和观测空间中导航——设计能够利用用户领域知识（通过建模和函数逼近）以实现泛化并提高样本效率的算法至关重要。因此，一个基础性问题是：哪些建模假设能够带来样本高效的学习保证，以及哪些算法能够实现最优样本复杂度。

强化学习的非渐近理论中，已有许多关于样本高效学习何以可能的充分条件（Dean 等人，2020年；Yang 和 Wang，2019年；Jin 等人，2020年；Modi 等人，2020年；Ayoub 等人，2020年；Krishnamurthy 等人，2016年；Du 等人，2019年；Li，2009年；Dong 等人，2019年；Zhou 等人，2021年），也提出了旨在统一这些条件的结构性质（Russo 和 Van Roy，2013年；Jiang 等人，2017年；Sun 等人，2019年；Wang 等人，2020年；Du 等人，2021年；Jin 等人，2021年），但必要条件通常仍难以确定。不过，最近 Foster 等人（2021）引入了决策估计系数（DEC），这是一种统一的统计复杂度概念，可在一般决策框架下给出最优样本复杂度的上界和下界。Foster 等人（2021）的结果表明，对于交互式决策，DEC 所起的作用类似于 VC 维及其相关概念在统计学习中的作用，但其定量保证仍有进一步收紧的空间。本文提出了 DEC 的一种新变体——受约束决策-估计系数，并利用它弥合 Foster 等人（2021）中上界和下界之间的若干间隙。

1.1 交互式决策

我们考虑*Foster* 等人提出的结构化观测决策（DMSO）框架Foster 等人（2021年），这是交互式决策的一般设定，涵盖多臂老虎机问题（包括结构化和上下文多臂老虎机）以及带函数逼近的强化学习。

该协议由 T轮次组成。对于每个轮次 $t = 1, \ldots\ldots, T$:

1. 学习者选择一个决策 $\pi^t \in \Pi$，其中 Π 是决策空间。

2. 学习者接收奖励 $r^t \in \mathcal{R} \subseteq \mathbb{R}$ 和观测 $o^t \in \mathcal{O}$，它们是通过 $(r^t, o^t) \sim M^\star(\pi^t)$ 采样得到的，其中$M^\star: \Pi \to \Delta(\mathcal{R} \times \mathcal{O})$ 是底层模型。

我们将 $\mathcal{R}$ 称为奖励空间，将 $\mathcal{O}$ 称为观测空间。模型 $M^\star$被形式化为一个条件分布，起到底层环境的作用，且学习者并不知道该模型。不过，假设学习者可以访问一个包含 $M^\star$的模型类 $\mathcal{M} \subset (\Pi \to \Delta(\mathcal{R} \times \mathcal{O}))$。

假设 1.1（可实现性）。学习者可以访问模型类 $\mathcal{M}$，该模型类包含真实模型 $M^\star$。

模型类 $\mathcal{M}$ 代表学习者对底层环境的先验知识。对于结构化老虎机问题，模型对应奖励分布，模型类编码了奖励景观中的结构（光滑性、线性性、凸性等）；对于强化学习问题，模型对应马尔可夫决策过程（MDP），模型类通常编码转移概率或价值函数中的结构。具体来说：

- 多臂老虎机。在多臂老虎机问题中，$\mathcal{M}$ 是奖励分布，π^t 称为动作或臂，Π称为动作空间；除了奖励之外，没有其他观测（$\mathcal{O} = \{\varnothing\}$）。通过选择 $\mathcal{M}$，使平均奖励函数 f^M 具有适当的结构，可以描述具有连续或无限动作空间以及结构化奖励的多臂老虎机问题，包括线性老虎机（Dani 等人，2007年；Abernethy 等人，2008年；Bubeck 等人，2012年）、老虎机凸优化（Kleinberg，2004年；Flaxman 等人，2005年；Bubeck 等人，2017年；Lattimore，2020年）以及非参数老虎机（Kleinberg，2004年；Bubeck 等人，2011年；Magureanu 等人，2014年）。

- **强化学习**。在幕式强化学习中，每个模型 $M \in \mathcal{M}$ 都指定一个非平稳、时域为H的马尔可夫决策过程 $M = \left\{\{\mathcal{S}_h\}_{h=1}^H, \mathcal{A}, \{P_h^M\}_{h=1}^H, \{R_h^M\}_{h=1}^H, d_1\right\}$，其中 $\mathcal{S}_h$ 是第 h层的状态空间，$\mathcal{A}$ 是动作空间，P_h^M: $\mathcal{S}_h \times \mathcal{A} \to \Delta(\mathcal{S}_{h+1})$是第 h层的概率转移核，R_h^M: $\mathcal{S}_h \times \mathcal{A} \to \Delta(\mathbb{R})$是第 h层的奖励分布，而 $d_1 \in \Delta(\mathcal{S}_1)$ 是初始状态分布。决策 $\pi = \{\pi_h : \mathcal{S}_h \to \Delta(\mathcal{A})\}_{h=1}^H$ 是策略（从状态到动作的映射）。给定策略 π后，一个回合按如下方式进行（从 $s_1 \sim d_1$开始）。对于 $h = 1$、……、 H:

  – $a_h \sim \pi_h(s_h)$。– $r_h \sim R_h^M(s_h, a_h)$ 和 $s_{h+1} \sim P_h^M(\cdot \mid s_h, a_h)$。

  这一过程会产生 $(r, o) \sim M(\pi)$，其中 $r = \sum_{h=1}^H r_h$ 是该回合的累积奖励，而观测 $o = (s_1, a_1, r_1) \ldots (s_H, a_H, r_H)$ 是该回合的轨迹（观测状态、动作和奖励的序列）。通过适当选择 $\mathcal{M}$，可以涵盖标准的马尔可夫决策过程类别（例如表格型马尔可夫决策过程或线性系统）（Dean 等（2020）；Yang 和 Wang（2019）；Jin 等（2020）；Modi 等（2020）；Ayoub 等（2020）；Krishnamurthy 等（2016）；Du 等（2019）；Li（2009）；Dong 等（2019）），以及更一般的结构条件（Jiang 等人，2017；Sun 等人，2019；Wang 等人，2020；Du 等人，2021；Jin 等人，2021）。

对于模型 $M \in \mathcal{M}$, $\mathbb{E}^{M,\ \pi}[\cdot]$表示过程 $(r,\ o) \sim M(\pi)$，$f^M(\pi) := \mathbb{E}^{M,\ \pi}[r]$下的期望，$\pi_M := \arg\max_{\pi \in \Pi} f^M(\pi)$ 表示最优决策。

我们考虑交互式决策的两类保证：遗憾保证和 PAC（概率近似正确）保证。对于遗憾保证，我们关注由下式给出的累积次优性

$$\mathbf{Reg}_{\mathsf{DM}}(T) := \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)\right], \tag{1}$$

其中，p^t 是学习者在第 t轮的随机化分布（关于 π^t的条件分布）。对于 PAC 保证，我们只关注最终次优性。在第 $t = 1$、……、 T轮完成后，学习者可以利用收集到的所有数据来选择最终决策 $\widehat{\pi}$（该决策可以按照分布 $p \in \Delta(\Pi)$进行随机化），我们通过以下方式衡量性能：

$$\mathbf{Risk}_{\mathsf{DM}}(T) := \mathbb{E}_{\widehat{\pi} \sim p}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\widehat{\pi})\right]. \tag{2}$$

有关更多背景和测度论细节，以及其他示例和讨论，请参阅Foster 等人（2021年、2022b）。

1.2 背景：决策估计系数

为了引出我们的结果，我们先回顾一下 Foster 等人 (2021年提出的决策估计系数；在本讨论中，我们仅关注遗憾保证。对于具有共同支配测度 $\mathbb{P}$ 和 $\mathbb{Q}$ 的概率测度 ν，平方海灵格距离定义为

$$D^2_{\mathsf{H}}(\mathbb{P},\mathbb{Q}) = \int\left(\sqrt{\frac{d\mathbb{P}}{d\nu}} - \sqrt{\frac{d\mathbb{Q}}{d\nu}}\right)^2 d\nu. \tag{3}$$

对于模型类 $\mathcal{M}$、参考模型 $M:\ \Pi \to \overline{\Delta}(\mathcal{R}\times\mathcal{O})$ 和尺度参数 $\gamma>0$，决策估计系数定义为

$$\textsf{r-dec}^{\textsf{o}}_{\gamma}(\mathcal{M},\overline{M}) = \inf_{p\in\Delta(\Pi)}\sup_{M\in\mathcal{M}}\mathbb{E}_{\pi\sim p}\left[f^M(\pi_M) - f^M(\pi) - \gamma\cdot D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right]. \tag{4}$$

这里，我们对 Foster 等人 (2021年)的记法稍作调整，添加前缀 r- （表示“遗憾”）和上标“o”（表示“偏移”），以便与稍后将介绍的其他变体区分开来。

Foster 等人 (2021年）（另见Foster 等人 (2022b）） 使用 DEC 为交互式决策的最优遗憾给出上界和下界：

- On 在上界方面，存在一种算法（从估计到决策），能够获得 s

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \lesssim \inf_{\gamma}\max\left\{\sup_{\overline{M}\in\mathrm{co}(\mathcal{M})}\textsf{r-dec}^{\textsf{o}}_{\gamma}(\mathcal{M},\overline{M})\cdot T + \gamma\cdot\mathbf{Est}_{\mathsf{H}}(T)\right\}. \tag{5}$$

  这里，术语 $\mathbf{Est}_{\mathsf{H}}(T)$ 表示使用该类别进行统计估计所需的样本复杂度；对于有限类这一特殊情况，它等于 $\mathbf{Est}_{\mathsf{H}}(T) \le \log|\mathcal{M}|$。

- 在下界方面，任何算法都必须满足

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \gtrsim \inf_{\gamma}\max\left\{\sup_{\overline{M}\in\mathcal{M}}\textsf{r-dec}^{\textsf{o}}_{\gamma}(\mathcal{M}',\overline{M})\cdot T + \gamma\right\}, \tag{6}$$

  其中 $\mathcal{M}'$ 是一个 特定的“局部化”子类 $\mathcal{M}$（粗略地说，是满足 $\|f^M - f^{\overline{M}}\|_\infty \lesssim \frac{\gamma}{T}$ 的模型） – – .

尽管 Foster 等人（2021年）的结果表明，对于各种模型类别（例如估计复杂度较低的凸类），DEC 刻画了其可学习性，但其定量速率仍有改进空间。本文旨在解决以下不足：

- 为了给出在期望意义下（而非以低概率）成立的下界，下界 (6)将范围限制在 $\mathcal{M}'\subseteq\mathcal{M}$的一个局部化子类上，该子类的复杂度可能低于原始类别。这对许多特殊情况给出了合理的结果，但一般而言可能会任意宽松。

- 下界 (6) 仅限于参考模型 $M\in\mathcal{M}$，而上界 (5) 则在参考模型 $M\in\mathrm{co}(\mathcal{M})$上取最大值，这可能导致更大的遗憾。例如，Foster 等人 (2021年的命题 3.1 给出了一个模型类 $\mathcal{M}$，在该模型类中，这一区别会导致上界和下界之间的间隙任意大。

第三个间隙是估计复杂度 $\mathbf{Est}_{\mathsf{H}}(T)$ 在 式（5）中的存在问题，我们在此不予讨论；下界中并不存在这一项。关于这一问题的进一步讨论，请参见 Foster 等人(2021年，2022a。

1.3 约束 训练的决策估计系数及结果概述

为了解决前文提到的问题，我们引入一种新的复杂度度量，即受约束决策估计系数。对于参考模型 M： $\Pi \to \Delta(\mathcal{R} \times \mathcal{O})$，定义，其定义方式如下

$$\textsf{r-dec}^{\textsf{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) = \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi)] \mid \mathbb{E}_{\pi \sim p}\left[D^2_{\textsf{H}}\big(M(\pi), \overline{M}(\pi)\big) \right] \le \varepsilon^2 \right\}, \tag{7}$$

按照约定，当集合

$$\mathcal{H}_{p,\varepsilon}(\overline{M}) := \left\{ M \in \mathcal{M} \mid \mathbb{E}_{\pi \sim p}\left[D^2_{\textsf{H}}\big(M(\pi), \overline{M}(\pi)\big) \right] \le \varepsilon^2 \right\} \tag{8}$$

为空时，上面的值为零；上标“c”表示“受约束”，用于与偏移版本区分。我们关注的主要量是

$$\textsf{r-dec}^{\textsf{c}}_{\varepsilon}(\mathcal{M}) = \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \textsf{r-dec}^{\textsf{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}), \tag{9}$$

其中 $\mathrm{co}(\mathcal{M})$ 表示类别 $\mathcal{M}$的凸包。

受约束与偏移 DEC 的区别仅在于如何纳入表示信息增益的量 $\mathbb{E}_{\pi \sim p}\left[D^2_{\textsf{H}}\big(M(\pi)\ M(\pi)\big) \right]$。受约束 DEC 对信息增益施加硬约束，而偏移 DEC 将信息增益视为针对最大化玩家的惩罚，相当于施加软约束。乍看之下，人们可能会预期这两个量可通过拉格朗日对偶等价起来。事实上，偏移 DEC 是受约束 DEC 的上界：

$$\textsf{r-dec}^{\textsf{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \le \inf_{\gamma} \{ \textsf{r-dec}^{\textsf{o}}_{\gamma}(\mathcal{M}, \overline{M}) \vee 0 + \gamma \varepsilon^2 \}, \quad \forall \overline{M},$$

但强对偶不成立，而且这些复杂度度量通常并不等价；详细讨论见 第 4 节。

遗憾框架的主要结果。 受约束 DEC 具有若干偏移 DEC 不具备的有用性质，其中包括 隐式地 以自适应方式强制实现某种局部化（参见第 4 节）。利用这些性质，我们给出了改进的下界和上界，弥合了 Foster 等人 (2021年）中各界之间除一个以外的所有间隙。我们的主要结果如下。

定理（非正式表述）。 对于任意模型类 $\mathcal{M}$:

- 下界：对于 $\mathcal{M}$ 中的最坏情形模型，任何算法都必须满足

$$\mathbb{E}[\mathbf{Reg}_{\textsf{DM}}(T)] \ge \widetilde{\Omega}(1) \cdot \textsf{r-dec}^{\textsf{c}}_{\underline{\varepsilon}(T)}(\mathcal{M}) \cdot T$$

  对于 $\underline{\varepsilon}(T) = \widetilde{\Theta}\left(\sqrt{1/T}\right)$,

- 上界：存在一种算法（从估计到决策$^{+}$），其性能达到

$$\mathbb{E}[\mathbf{Reg}_{\textsf{DM}}(T)] \le \widetilde{O}(1) \cdot \textsf{r-dec}^{\textsf{c}}_{\bar{\varepsilon}(T)}(\mathcal{M}) \cdot T$$

  $for\ \bar{\varepsilon}(T) = \widetilde{\Theta}\left(\sqrt{\mathbf{Est}_{\textsf{H}}(T)/T}\right) \le \widetilde{\Theta}\left(\sqrt{\log|\mathcal{M}|/T}\right)$.

这一上下界始终比先前工作中的相应界更紧，尤其是下界在以下两个方面优于Foster 等人（2021年）：

- 它在全局范围内成立，并且不再需要局部化这一Foster 等人（2021年）所采用的概念。此外，它在期望意义下成立，且对所考虑的算法类没有任何限制。
- 更有意思的是，它随 $\textsf{r-dec}_{\varepsilon}(\mathcal{M}) = \sup_{M \in \mathrm{co}(\mathcal{M})} \textsf{r-dec}_{\varepsilon}(\mathcal{M} \cup \{M\}, M)$ 缩放，而不是随 $\sup_{M \in \mathcal{M}} \textsf{r-dec}_{\varepsilon}(\mathcal{M}, M)$ 缩放；这表明（结合上界来看），非适当参考模型$M \in \mathrm{co}(\mathcal{M})$ 发挥着根本作用，并非只是Foster 等人（2021年）上界中的一种人为产物。

总之，我们的上界和下界为构建一种基于决策估计系数、简单、易用且统一的交互式决策理论迈出了重要一步。我们的结果唯一尚未解决的问题是估计误差$\mathbf{Est}_{\mathsf{H}}(T)$的作用；这是一个深刻的问题，有待未来研究。

在技术层面，我们的下界运用了若干新的分析技术和结构性结果，这些方法与以往的思路截然不同（Foster 等人，2021年，2022b）。我们的上界建立在 Foster 等人（2021年）提出的从估计到决策范式之上，但其设计与分析要复杂得多，因为受约束 DEC 的某些性质虽有助于证明下界，却给推导上界带来了非平凡的挑战。我们预计这些技术将有更广泛的用途；为此，我们在第 4 节提供了便于使用的受约束 DEC 工具。

注 1.1。此时，读者或许会想知道，定义 (9) 为何纳入集合 $\mathcal{M}\cup\{M\}$，其中 M 可能不属于类别 $\mathcal{M}$。我们将在 第 4 节 说明，尽管偏移 DEC 并不需要这一修改，但它对于受约束 DEC 却起着核心作用。

PAC 框架的主要结果。从遗憾框架转向 PAC 时，我们采用以下偏移与受约束 DEC 的 PAC 对应形式：

$$\textsf{p-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\overline{M})=\inf_{p,q\in\Delta(\Pi)}\sup_{M\in\mathcal{M}}\left\{\mathbb{E}_{\pi\sim p}[f^{M}(\pi_{M})-f^{M}(\pi)]-\gamma\cdot\mathbb{E}_{\pi\sim q}\left[D^{2}_{\mathsf{H}}\left(M(\pi),\overline{M}(\pi)\right)\right]\right\} \tag{10}$$

and

$$\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M},\overline{M})=\inf_{p,q\in\Delta(\Pi)}\sup_{M\in\mathcal{M}}\left\{\mathbb{E}_{\pi\sim p}[f^{M}(\pi_{M})-f^{M}(\pi)]\mid\mathbb{E}_{\pi\sim q}\left[D^{2}_{\mathsf{H}}\left(M(\pi),\overline{M}(\pi)\right)\right]\leq\varepsilon^{2}\right\}, \tag{11}$$

约定当 $\mathcal{H}_{q,\varepsilon}(M)=\varnothing$ 时，公式（11）中的值为零。这些定义与(4)和(7)类似，但允许最小化方选择一个单独的探索分布 q 来评估信息增益（海林格距离），并选择一个利用分布 p 来评估遗憾。由于始终可以选择 $p=q$，因此显然$\textsf{p-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},M)\leq\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},M)$，同样$\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M},M)\leq\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M},M)$。定义[1]

$$\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M})=\sup_{\overline{M}\in\mathrm{co}(\mathcal{M})}\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M},\overline{M}), \tag{12}$$

我们证明，PAC DEC 给出了以下关于 PAC 样本复杂度的下界和上界。

定理（非正式）。对于任意模型类 $\mathcal{M}$：

- 下界：对于 $\mathcal{M}$ 中的最坏情况模型，任何 *PAC* 算法只要经过 T 轮交互，就必须 *hav* e
$$\mathbb{E}[\mathbf{Risk}_{\mathsf{DM}}(T)]\geq\widetilde{\Omega}(1)\cdot\textsf{p-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M})$$
适用于 $\bar{\varepsilon}(T)=\widetilde{\Theta}\left(\sqrt{1/T}\right)$。
- 上界：存在一种算法（从估计到决策$^{+}$），能够达到
$$\mathbb{E}[\mathbf{Risk}_{\mathsf{DM}}(T)]\leq\widetilde{O}(1)\cdot\textsf{p-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M})$$
for $\bar{\varepsilon}(T)=\widetilde{\Theta}\left(\sqrt{\mathbf{Est}_{\mathsf{H}}(T)/T}\right)\leq\widetilde{\Theta}\left(\sqrt{\log|\mathcal{M}|/T}\right)$.

1.4 预备知识

交互式决策。本文中，除非另有说明，我们始终假设 $\mathcal{R}=[0,1]$。我们令 $\mathcal{M}^{+}=\{M\colon\Pi\to\Delta([0,1]\times\mathcal{O})\}$ 表示所有奖励取值于 $[0,1]$ 的可能模型构成的空间。我们采用简写 $g^{M}(\pi)=f^{M}(\pi_{M})-f^{M}(\pi)$。

[1]与遗憾的受约束 DEC 定义相比，定义 $\textsf{p-dec}_{c\varepsilon}(\mathcal{M})=\sup^{M\in\mathrm{co}(\mathcal{M})}\textsf{p-dec}_{c\varepsilon}(\mathcal{M},M)$ 并未将 DEC 应用于类别 $\mathcal{M}\cup\{\overline{M}\}$。详见 第 4.3 节，了解详细说明。

我们采用与 Foster 等人（2021，2022b）相同的概率空间形式化方式。决策与可测空间 $(\Pi, \mathscr{P})$ 相关联，奖励与空间 $(\mathcal{R}, \mathscr{R})$ 相关联，观测与空间 $(\mathcal{O}, \mathscr{O})$ 相关联。直到时间 t 的历史记作 $\mathfrak{H}^t = (\pi^1, r^1, o^1) \ldots, (\pi^t, r^t, o^t)$。我们定义

$$\Omega^t = \prod_{i=1}^{t} (\Pi \times \mathcal{R} \times \mathcal{O}), \quad \text{and} \quad \mathscr{F}^t = \bigotimes_{i=1}^{t} (\mathscr{P} \otimes \mathscr{R} \otimes \mathscr{O})$$

使得 $\mathfrak{H}^t$ 与空间 $(\Omega^t, \mathscr{F}^t)$ 相关联。对于类别 $\mathcal{M}$，我们定义

$$V(\mathcal{M}) := \sup_{M, M' \in \mathcal{M}} \sup_{\pi \in \Pi} \sup_{A \in \mathscr{R} \otimes \mathscr{O}} \left\{ \frac{M(A|\pi)}{M'(A|\pi)} \right\} \vee e;$$

该量有限并非我们任何结果成立的必要条件，但能使我们的下界（定理 2.1 和 2.2）改善一个 $\log(T)$ 因子。

散度。 对于可测空间 $(\Omega, \mathscr{F})$ 上具有共同支配测度的概率分布 $\mathbb{P}$ 和 $\mathbb{Q}$，我们将全变差距离定义为

$$D_{\mathsf{TV}}(\mathbb{P}, \mathbb{Q}) = \sup_{A \in \mathscr{F}} |\mathbb{P}(A) - \mathbb{Q}(A)| = \frac{1}{2} \int_{\Omega} |d\mathbb{P} - d\mathbb{Q}|.$$

1.5 结构安排

第 2 节介绍遗憾和 PAC 框架的主要下界，而 第 3 节 介绍相应的上界。 第 4 节 阐述关于受约束 DEC 和偏移 DEC 的结构性结果，而 第 5 节 详细比较先前工作中的界。最后，我们通过其他示例作结（第 6 节）。除非另有说明，证明均推迟至附录。

其他记号。 对于整数 $n \in \mathbb{N}$，我们用 $[n]$ 表示集合 $\{1, \ldots, n\}$。对于集合 $\mathcal{Z}$，我们用 $\Delta(\mathcal{Z})$ 表示 $\mathcal{Z}$上的所有概率分布构成的集合，并用 $\mathcal{Z}^c$ 表示其补集。我们采用标准的 $\widetilde{O}$ 大 O 记号，并用 $f = O(g)$ 表示 $f = O(g \cdot \max\{1, \mathrm{polylog}(g)\})$。我们仅在非正式陈述中使用 $\lesssim$，以强调不等式中最值得注意的部分。

2 下界

本节基于受约束决策估计系数，分别给出 PAC 和遗憾框架下的下界，随后证明 PAC 的下界，并着重阐述这一结果背后最重要的思路。

2.1 主要结果

我们给出交互式决策的极小极大下界，表明对于任意模型类 $\mathcal{M}$ 和时域 $T \in \mathbb{N}$，任何算法的最坏情况遗憾（或 PAC 样本复杂度）都以下界的形式受限于受约束 DEC，其中半径参数 $\varepsilon > 0$取适当值。PAC 和遗憾的下界形式几乎相同，唯一的区别在于使用 $\mathsf{p\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M})$ 还是 $\mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M})$。为陈述这些结果，定义 $C(T) := \log(T \wedge V(\mathcal{M}))$。

定理 2.1 (主要下界：PAC)。 设 $\varepsilon(T) := c \cdot \sqrt{\frac{1}{TC(T)}}$，其中 $c > 0$ 是一个充分小的数值常数，且 $C := 48\sqrt{2}$。对于所有满足条件的 $T \in \mathbb{N}$

$$\mathsf{p\text{-}dec}^{\mathrm{c}}_{\varepsilon(T)}(\mathcal{M}) \geq C \cdot \varepsilon(T) \tag{13}$$

得到满足时，对于任意 *PAC* 算法，都存在一个模型，使得在 $\mathcal{M}$ 中 t

$$\mathbb{E}\left[\mathbf{Risk}_{\mathsf{DM}}(T)\right] \geq \Omega(1) \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{p\text{-}dec}^{\mathrm{c}}_{\varepsilon(T)}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \geq \Omega(1) \cdot \mathsf{p\text{-}dec}^{\mathrm{c}}_{\varepsilon(T)}(\mathcal{M}). \tag{14}$$

定理 2.2(主要下界：遗憾)。 存在通用常数 C、$C' > 0$ 和 c、$c' > 0$，使得以下结论成立。令 $\varepsilon(T)$：$= c \cdot \sqrt{TC(T)}$。对于所有满足条件的 $T \in \mathbb{N}$

$$\text{r-dec}^{\text{c}}_{\underline{\varepsilon}(T)}(\mathcal{M}) \geq C \cdot \underline{\varepsilon}(T) \tag{15}$$

满足上述条件时，对于任意遗憾最小化算法，都存在一个属于 $\mathcal{M}$ 的模型，使得

$$\begin{aligned}\mathbb{E}[\mathbf{Reg}_{\text{DM}}(T)] &\geq c' \cdot \sup_{\overline{M} \in \mathcal{M}^+} \text{r-dec}^{\text{c}}_{\underline{\varepsilon}(T)}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \cdot \frac{T}{\log T} - C' \cdot \frac{\sqrt{T}}{\log(T)} \\ &\geq c' \cdot \text{r-dec}^{\text{c}}_{\underline{\varepsilon}(T)}(\mathcal{M}) \cdot \frac{T}{\log T} - C' \cdot \frac{\sqrt{T}}{\log(T)}.\end{aligned} \tag{16}$$

定理 2.1 和 2.2 表明，受约束 DEC 是交互式决策的一个基本限制。重要的是，这些下界消除了先前工作所要求的局部化概念，并表明，即使允许使用非适当参考模型 $M \in \text{co}(\mathcal{M})$，DEC 仍然是一个下界；因此，它们总是比 Foster 等人 (2021年,2022b) 中的下界更紧。稍后我们将证明（第 3 节），通过算法可以达到我们的下界，但半径存在差异；这一差异取决于 $\mathcal{M}(\sqrt{\log|\mathcal{M}|/T}$ 相较于 $1/\sqrt{T}$ 在有限类情况下的表现）的估计容量。我们将先前工作的详细比较留到第 5 节，现在先建立对下界行为的直观理解，并概述我们的证明技巧。

注 2.1. 当 $V(\mathcal{M}) = O(1)$ 时，定理 2.1 和 2.2 中有 $\underline{\varepsilon}(T) \propto 1/\sqrt{T}$。在 $V(\mathcal{M})$ 无界的一般情况下，我们有 $\varepsilon(T) \propto 1/\sqrt{T\log(T)}$，下界会多出一个对数因子。对于大多数标准模型类，$V(\mathcal{M}) = \infty$，但存在一个子类 $\mathcal{M}' \subseteq \mathcal{M}$ with$V(\mathcal{M}') = O(1)$，且 $\text{p-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}') \gtrsim \text{p-dec}^{\text{c}}_{\varepsilon}(\mathcal{M})$（相应地，$\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}') \gtrsim \text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M})$）。在这种情况下，将定理 2.1（相应地，定理 2.2）应用于 $\mathcal{M}'$，即可推出一个满足 $\varepsilon(T) \propto 1/\sqrt{T}$ 的更紧下界。示例见 Foster 等人（2021）。

注 2.2. 定理 2.1（分别地，定理 2.2）随量 $\sup_{M \in \mathcal{M}^+} \text{p-dec}^{\text{c}}_{\varepsilon}(\mathcal{M} \cup \{M\}, M) \geq \text{p-dec}^{\text{c}}_{\varepsilon}(\mathcal{M})$（分别地，$\sup_{M \in \mathcal{M}^+} \text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M} \cup \{M\}, M) \geq \text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M})$）缩放，这允许使用任意参考模型$M \notin \text{co}(\mathcal{M})$。我们在第 4.4 节证明，对参考模型 $M \in \mathcal{M}^+$ 取最大值不会使 DEC 超过 $M \in \text{co}(\mathcal{M})$ 所达到的值，因此这一结果并不与我们的上界矛盾。我们以这种形式陈述下界，是因为：1）我们的证明直接使用 $M \in \mathcal{M}^+$，不依赖 $\text{co}(\mathcal{M})$ 的结构；2）允许使用 $M \in \mathcal{M}^+$ 通常能简化计算。

理解下界。下面我们简要说明标准模型类的下界表现如何。我们聚焦于遗憾，因为这是先前工作主要关注的对象。

- $\sqrt{T}$-阶。 对于多臂老虎机和强化学习文献中广泛研究的类别，我们有

  $$\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) \propto \varepsilon \cdot \sqrt{C_{\text{prob}}},$$

  其中 $C_{\text{prob}} > 0$ 是一个依赖于问题的常数，反映了某种内在复杂度。在这种情况下，只要 (15)满足条件 C_{prob} 大于某个数值常数，定理 2.2 便给出[2]

  $$\mathbb{E}[\mathbf{Reg}_{\text{DM}}(T)] \geq \widetilde{\Omega}\big(\sqrt{C_{\text{prob}} \cdot T}\big).$$

  例子（参见第 6 节）包括具有 A 个动作的多臂老虎机，其中 $C_{\text{prob}} \geq A$（从而得到 $\mathbb{E}[\text{Reg}_{\text{DM}}(T)] \geq \widetilde{\Omega}(\sqrt{AT})$）；维度为 d 的线性老虎机，其中 $C_{\text{prob}} \geq d$（从而得到 $\mathbb{E}[\text{Reg}_{\text{DM}}(T)] \geq \widetilde{\Omega}(\sqrt{dT})$）；以及具有 S 个状态、A 个动作和时域 H 的表格型强化学习，其中 $C_{\text{prob}} \geq HSA$（从而得到 $\mathbb{E}[\text{Reg}_{\text{DM}}(T)] \geq \widetilde{\Omega}(\sqrt{HSAT})$）。

[2]这要求 $V(\mathcal{M}) = O(1)$ 使得 $\varepsilon(T) \propto 1/\sqrt{T}$；参见 注 2.1。

- 非参数速率。 对于非参数模型类，其最优遗憾的阶高于$\sqrt{T}$，通常有

$$\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) \propto \varepsilon^{1-\rho}$$

对于某个 $\rho \in (0,1)$。在这种情况下，只要 T 是一个足够大的常数，公式（15）中的条件就成立，而 定理 2.2 给出

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \geq \widetilde{\Omega}(T^{\frac{1+\rho}{2}}).$$

一个标准示例是定义在 $[0,1]^d$ 上的 Lipschitz 多臂老虎机，其中我们有 $\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) \propto \varepsilon^{1-\frac{d}{d+2}}$，从而得到$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \geq \widetilde{\Omega}(T^{\frac{d+1}{d+2}})$。

- 快速速率。对于低噪声问题（例如无噪声多臂老虎机），DEC 通常表现出阈值行为，其中

$$\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) \propto \mathbb{I}\{\varepsilon \geq 1/\sqrt{C_{\text{prob}}}\},$$

其中 C_{prob} 是一个问题相关参数。例如，如果 $\mathcal{M}$ 由满足以下条件的多臂老虎机实例构成：$\Pi = \{1, \ldots, A\}$ 且奖励为无噪声二元奖励，则可以取 $C_{\text{prob}} \propto A$。对于此类设置，只要满足条件 （15），就有 $T = O(C_{\text{prob}})$，并且 定理 2.2 给出

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \geq \widetilde{\Omega}\big(\min\{C_{\text{prob}}, T\}\big).$$

对于 PAC 保证，情况与遗憾类似，但由于归一化，下界会乘以一个 $1/T$ 因子。简言之：

- 每当 $\text{p-dec}_{\varepsilon}(\mathcal{M}) \propto \varepsilon \cdot \sqrt{C_{\text{prob}}}$ 时，定理 2.1 给出 $\mathbb{E}[\textbf{风险}_{\mathsf{DM}}(T)] \geq \widetilde{\Omega}\left(\sqrt{\frac{C_{\text{prob}}}{T}}\right)$，这意味着需要 $\widetilde{\Omega}\left(\frac{C_{\text{prob}}}{\varepsilon^2}\right)$ 个样本才能学习一个 ε-最优策略。
- 每当 $\text{p-dec}_{\varepsilon}(\mathcal{M}) \propto \varepsilon^{1-\rho}$ 对 $\rho \in (0, 1)$ 成立时，定理 2.1 给出 $\mathbb{E}[\textbf{风险}_{\mathsf{DM}}(T)] \geq \widetilde{\Omega}(T^{-(1-\rho)/2})$，这意味着需要 $\widetilde{\Omega}(\varepsilon^{-\frac{2}{1-\rho}})$ 个样本才能学习一个 ε-最优策略。
- 每当 $\text{p-dec}_{\varepsilon}(\mathcal{M}) \propto \mathbb{I}\{\varepsilon \geq 1/\sqrt{C_{\text{prob}}}\}$ 时，其中 C_{prob} 是一个问题相关参数，定理 2.1 给出 $\mathbb{E}[\textbf{风险}_{\mathsf{DM}}(T)] \geq \Omega(1)$，直到 $T = \Omega(C_{\text{prob}})$；也就是说，至少需要 $\Omega(C_{\text{prob}})$ 个样本才能达到超越常数次优性的水平。

更多示例和细节请参见第 6 节。

证明技巧：PAC。现在，我们重点介绍定理 2.1证明背后的一些关键思想。证明的大致结构如下。对于任意算法，都可以构造一对“困难”的模型M_1，$M_2 \in \mathcal{M}$ 使得：

1. 算法在 M_1 和 M_2 下产生的决策 π^t 与观测 (r^t, o^t) 的联合分布在全变差意义下相近（即 $D_{\mathsf{TV}}(\mathbb{P}^{M_1}, \mathbb{P}^{M_2}) \leq 1/4$，其中 $\mathbb{P}^M$ 表示 M 下 $\mathfrak{H}^T$ 的概率分布）。2. 任何风险远低于 DEC 的算法，都必须根据底层模型是 M_1 还是 M_2，在 Π 中查询截然不同的决策。

由于 $\mathbb{P}^{M_1}$ 在全变差意义下足够接近，算法以恒定概率无法区分这些模型；又由于这些模型的最优决策（近似）互斥，因此下界得证。

我们根据所考虑的算法，以对抗式的方式选择模型对（M_1，M_2），从而推广Foster 等人 （2021年、2022b）所采用的方法。我们固定一个任意模型$M_1 \in \mathcal{M}$，然后令 $q_{M_1} := \mathbb{E}^{M_1}\left[\frac{1}{T}\sum_{t=1}^{T} q^t(\cdot \mid \mathfrak{H}^{t-1})\right]$ 和 $p_{M_1} := \mathbb{E}^{M_1}[p(\cdot \mid \mathfrak{H}^T)]$ 分别表示学习者在该模型下的平均策略，选择使代入（p_{M_1}，q_{M_1}）后的公式（11） 取得最大值的模型 M_2 。这种方法足以证明下界随 $\sup_{M \in \mathcal{M}} \text{p-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}, M)$ 缩放，但不足以纳入不适当参考模型并证明下界随

$\text{p-dec}^{c}_{\varepsilon}(\mathcal{M}) = \sup_{M \in \text{co}(\mathcal{M})} \text{p-dec}^{c}_{\varepsilon}(\mathcal{M} \cup \{M\}, M)$。事实上，除非类别 $\mathcal{M}$ 是凸的，否则没有理由认为，对于类别 $\mathcal{M}$ 中的模型风险较低的算法，在不适当的混合 $M \in \text{co}(\mathcal{M})$下也会有较低的风险。因此，直接选择 $M_1 = M \in \text{co}(\mathcal{M})$存在问题，因为我们无法将算法在 M_1 下的风险与其在该类别模型下的风险联系起来。

为规避这一问题，我们迭代上述过程：给定任意一个可能非适当的参考模型 $M \in \mathcal{M}^+$，我们首先通过找出在 公式（11） 中取到最大值的模型，得到 M_1，其中代入的是 $(\cdot, q_M)$。有了这个模型后，我们以类似方式得到 M_2，但需以学习者针对 M_1 的行为近似最优这一事件为条件。也就是说，我们找出在公式（11） 中取到最大值的模型，其中代入的是分布 $(p_M(\cdot \mid \mathcal{E}_1)\ q_M)$，这里 $\mathcal{E}_1$ 是 M_1 的近似最优决策集合。由此可得，下界随 $\sup_{M \in \mathcal{M}^+} \text{p-dec}^{c}_{\varepsilon}(\mathcal{M}, M) \geq \text{p-dec}^{c}_{\varepsilon}(\mathcal{M})$缩放，这是因为参考模型 M 仅作为 M_1 与 M_2 之间的中点（其存在使我们能够控制模型之间的全变差），因此无需属于 $\mathcal{M}$。

证明技巧：遗憾。我们关于遗憾下界的证明（定理 2.2）采用了与定理 2.1类似的方法。然而，由于不再有分别用于探索（q_M）和利用（p_M）的独立分布，应用前文讨论的迭代条件化方案时会遇到实质性困难，导致遗憾与信息的分析相互耦合。为解决这一问题，我们采用一种略有不同的两部分方案。

1. 首先，我们证明一个类似于定理 2.2的下界，但在定性上较弱，因为在式（16）中，量 $\text{r-dec}^{c}_{\varepsilon(T)}(\mathcal{M}) = \sup_{M \in \mathcal{M}^+} \text{r-dec}^{c}_{\varepsilon(T)}(\mathcal{M}, M)$ 被 $\sup_{M \in \mathcal{M}} \text{r-dec}^{c}_{\varepsilon(T)}(\mathcal{M}, M)$ 取代；也就是说，该下界仅针对适当的参考模型。其证明采用了与我们针对 PAC 的下界证明类似的方法。

2. 然后，我们利用以下算法结果，将这一较弱的结果提升为定理 2.2的完整结论：对于任意模型类 $\mathcal{M}$ 以及任意 $M \in \mathcal{M}^+$（不一定属于 $\mathcal{M}$），若其满足温和的技术条件，且存在一个算法，其相对于类别 $\mathcal{M}$的期望遗憾至多为 R，那么就存在一个算法，其相对于扩展后的类别 $\mathcal{M} \cup \{M\}$的遗憾至多为 $O(R \cdot \log T)$。之后，通过适当选择 M，并将第一部分中的适当下界应用于类别 $\mathcal{M} \cup \{M\}$的一个略作修改的版本，我们便可证明定理 2.2。

参见附录B 中的证明。

受约束 DEC 的优势。与偏移 DEC 相比，采用受约束 DEC 的主要技术优势在于：通过对所考虑模型之间的海林格距离施加硬约束，我们可以采用比先前工作所考虑的更强的测度变换论证；这对于推导期望意义下（而非低概率）的下界至关重要。具体而言，半径 $\varepsilon(T) \approx 1/\sqrt{T}$ 是最大的可行选择，使得对于任意算法，都能找到一对最坏情况模型，使其全变差距离 $D_{\text{TV}}(\mathbb{P}^{M_1}, \mathbb{P}^{M_2})$ 为一个较小的常数（例如 $1/4$）。只要诱导分布之间的全变差距离为常数，算法就必然以常数概率无法区分 M_1和 M_2；如果 M_1 和 M_2 的最优决策差异显著，这就会导致较大的遗憾。

Foster 等人(2021年) 强调，决策估计系数可以被视为统计估计中连续性模量的交互式对应概念（Donoho 和 Liu，1987年，1991a，b）。我们认为，受约束 DEC 比偏移 DEC 更直接地对应连续性模量：出于类似的技术原因，连续性模量对海林格距离施加硬约束，而基于连续性模量的下界也使用相同的 $1/\sqrt{T}$ 半径。

2.2 PAC 下界的证明（定理 2.1）

接下来，我们证明 PAC下界（定理 2.1）。关于遗憾的证明（定理 2.2与前者类似，但在一些重要的技术细节上有所不同，因此推迟到附录B。

预备知识。 形式化地，对于 $T \in \mathbb{N}$，PAC 框架中的算法 是由一系列映射组成的集合（p，q） $= \big(\{q^t(\cdot \mid \cdot)\}_{t=1}^{T}$，$p(\cdot \mid \cdot)\big)$，它会（自适应地）抽取决策 $\pi^t \sim q^t(\cdot \mid \mathfrak{H}^{t-1})$（对于 $t \in [T]$），然后根据历史 $\mathfrak{H}^T$输出最终决策 $\widehat{\pi} \sim p(\cdot \mid \mathfrak{H}^T)$。我们将 $\mathbb{P}^{M,\ (p,\ q)}$ 定义为底层模型为 M 且算法为（p，q）时 $\mathfrak{H}^T$的概率分布。在整个证明中，我们将使用基本性质 $D_{\mathsf{TV}}(\mathbb{P},\ \mathbb{Q}) \le D_{\mathsf{H}}(\mathbb{P},\ \mathbb{Q})$。

定理 2.1 的证明。 为便于后文引用，我们定义常数

$$c_0 = 1/16,$$

并定义 $C(T) = 2^8 \cdot \log(T \wedge V(\mathcal{M}))$。固定 $T \in \mathbb{N}$和一个算法（p，q） $= \{q^t(\cdot \mid \cdot)\ p(\cdot \mid \cdot)\}_{t=1}^{T}$。对于每个模型 $M \in \mathcal{M}^+$，我们使用缩写 $\mathbb{P}^M \equiv \mathbb{P}^{M,\ (p,\ q)}$，并将相应的期望记为 $\mathbb{E}^M$。此外，我们定义

$$p_M = \mathbb{E}^M\left[p(\cdot \mid \mathfrak{H}^T)\right], \quad \text{and} \quad q_M = \mathbb{E}^M\left[\frac{1}{T}\sum_{t=1}^{T} q^t(\cdot \mid \mathfrak{H}^{t-1})\right].$$

选择一对困难模型。 固定任意参考模型 $M \in \mathcal{M}^+$ 并定义 $\varepsilon := \frac{1}{10\sqrt{C(T)\cdot T}}$以及 $\varepsilon(T) := \varepsilon/\sqrt{2}$。~~我们将~~根据 $\text{p-dec}^{\mathrm{c}}_{\varepsilon(T)}(\mathcal{M}, \overline{M})$证明一个下界。我们将 $\delta := \text{p-dec}^{\mathrm{c}}_{\varepsilon(T)}(\mathcal{M},\ \overline{M})$简写为 (13)，因此假设 $\delta \ge 48\varepsilon$。

首先，任取满足以下条件的模型 $M_1 \in \mathcal{M}$：

$$M_1 \in \left\{M \in \mathcal{M} \ :\ \mathbb{E}_{\pi\sim q_{\overline{M}}}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \le \varepsilon^2 \ \wedge\ \mathbb{E}_{\pi\sim p_{\overline{M}}}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \le \varepsilon^2\right\}. \tag{17}$$

我们将利用这样一个事实：根据 引理 D.3，对于所有 $p \in \Delta(\Pi)$，

$$\begin{aligned}&\mathsf{dec}_{\varepsilon/\sqrt{2}}(\mathcal{M}, \overline{M})\\ &\le \sup_{M\in\mathcal{M}}\left\{\mathbb{E}_{\pi\sim p}\left[g^M(\pi)\right] \mid \mathbb{E}_{\pi\sim q_{\overline{M}}}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \le \varepsilon^2 \ \wedge\ \mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \le \varepsilon^2\right\},\end{aligned} \tag{18}$$

这意味着（结合 公式 (13)），公式 (17) 中的集合非空。对于任意模型 $M \in \mathcal{M}^+$，定义

$$\mathcal{E}^M := \left\{\pi \in \Pi : g^M(\pi) \ge c_0 \cdot \delta\right\},$$

并定义 $\mathcal{A}_1 := \mathcal{E}^{M_1}$。令 $p' := p_M(\cdot \mid \mathcal{A}_1^{\mathrm{c}})$，并设

$$M_2 := \arg\max_{M\in\mathcal{M}}\left\{\mathbb{E}_{\pi\sim p'}\left[g^M(\pi)\right] \mid \mathbb{E}_{\pi\sim q_{\overline{M}}}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \le \varepsilon^2 \ \wedge\ \mathbb{E}_{\pi\sim p'}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \le \varepsilon^2\right\}; \tag{19}$$

as 对于 M_1，引理 D.3 意味着该集合非空。最后，定义 $\mathcal{A}_2 := \mathcal{E}^{M_2} \cap \mathcal{A}_1^{\mathrm{c}}$.

算法风险的下界。 现在我们回顾 Foster 等人 (2021年) 的引理 A.13，该引理指出，对于所有模型 M，

$$D^2_{\mathsf{H}}\big(\mathbb{P}^M, \mathbb{P}^{\overline{M}}\big) \le C(T)\cdot T \cdot \mathbb{E}_{\pi\sim q_{\overline{M}}}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big]. \tag{20}$$

根据数据处理不等式，进一步可得

$$D^2_{\mathsf{TV}}(p_M, p_{\overline{M}}) \le C(T)\cdot T \cdot \mathbb{E}_{\pi\sim q_{\overline{M}}}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big]. \tag{21}$$

由于 $\mathbb{E}_{\pi\sim q_M}\big[D^2_{\mathsf{H}}\big(M_i(\pi) M(\pi)\big)\big] \le \varepsilon^2$对 $i \in \{1,2\}$ 成立，我们的选择 $\varepsilon \le \frac{1}{10\cdot\sqrt{T\cdot C(T)}}$ 意味着

$$D_{\mathsf{TV}}\big(p_{M_i}, p_{\overline{M}}\big) \le \frac{1}{10}, \quad \text{for} \quad i \in \{1,2\}.$$

因此，对于每个 $i \in \{1,2\}$，我们有

$$\begin{aligned}\mathbb{E}_{\pi\sim p_{M_i}}\left[g^{M_i}(\pi)\right] &\geq c_0\delta \cdot p_{M_i}(\pi \in \mathcal{E}^{M_i}) &(22)\\ &\geq c_0\delta \cdot \left(p_{\overline{M}}(\pi \in \mathcal{E}^{M_i}) - D_{\mathsf{TV}}\left(p_{\overline{M}}, p_{M_i}\right)\right)\\ &\geq c_0\delta \cdot \left(p_{\overline{M}}(\pi \in \mathcal{E}^{M_i}) - 1/10\right). &(23)\end{aligned}$$

因此，要证明该定理，只需证明 $p_M(\pi \in \mathcal{E}^{M_i})$ 至少为 $1/4$，且这对至少一个 $i \in \{1,2\}$成立；这将表明 公式（23）中的量至少为 $\frac{3c_0\delta}{20}$ （事实上，任何大于 $1/10$的常数下界都足够）。

下文假设 $p_M(\mathcal{A}_1^{\mathrm{c}}) \geq 1/2$；否则，我们有 $p_M(\mathcal{E}^{M_1}) = p_M(\mathcal{A}_1) \geq 1/2$，此时结论可由 公式（23）立即推出。在继续之前，我们注意到，由于 $\mathbb{E}_{\pi\sim p_M}\left[D^2_{\mathsf{H}}(M_1(\pi), M(\pi))\right] \leq \frac{\varepsilon^2}{()()}$并且

$$\mathbb{E}_{\pi\sim p_{\overline{M}}}\left[\mathbb{1}\left\{\pi \in \mathcal{A}_1^{\mathrm{c}}\right\} \cdot D^2_{\mathsf{H}}\left(M_2(\pi), \overline{M}(\pi)\right)\right] \leq \mathbb{E}_{\pi\sim p'}\left[D^2_{\mathsf{H}}\left(M_2(\pi), \overline{M}(\pi)\right)\right] \leq \varepsilon^2,$$

海林格距离的三角不等式意味着 $\mathbb{E}_{\pi\sim p_M}\left[\mathbb{1}\left\{\pi \in \mathcal{A}_1^{\mathrm{c}}\right\} \cdot D^2_{\mathsf{H}}(M_1(\pi), M_2(\pi))\right] \leq 4\varepsilon^2$.因此，根据公式（D.1）和詹森不等式，我们得到

$$\mathbb{E}_{\pi\sim p_{\overline{M}}}\left[\mathbb{1}\left\{\pi \in \mathcal{A}_1^{\mathrm{c}}\right\} \cdot |f^{M_1}(\pi) - f^{M_2}(\pi)|\right] \leq \sqrt{\mathbb{E}_{\pi\sim p_{\overline{M}}}\left[\mathbb{1}\left\{\pi \in \mathcal{A}_1^{\mathrm{c}}\right\} \cdot D^2_{\mathsf{H}}(M_1(\pi), M_2(\pi))\right]} \leq 2\varepsilon. \tag{24}$$

间隙下界。 接下来，我们首先证明

$$f^{M_2}(\pi_{M_2}) \geq f^{M_1}(\pi_{M_1}) + \frac{\delta}{2} \cdot (1 - 4c_0) - 2\varepsilon. \tag{25}$$

为此，我们注意到，根据 M_2的定义，

$$\mathbb{E}_{\pi\sim p_{\overline{M}}}\left[\mathbb{1}\left\{\pi \in \mathcal{A}_1^{\mathrm{c}}\right\} \cdot g^{M_2}(\pi)\right] \geq p_{\overline{M}}(\mathcal{A}_1^{\mathrm{c}}) \cdot \mathbb{E}_{\pi\sim p'}\left[g^{M_2}(\pi)\right] \geq \frac{1}{2} \cdot \mathsf{p\text{-}dec}^{\mathrm{c}}_{\varepsilon/\sqrt{2}}(\mathcal{M}, \overline{M}) = \frac{\delta}{2},$$

这里我们使用了以下假设： $p_M(\mathcal{A}_1^{\mathrm{c}}) \geq \frac{1}{2}$。因此，

$$\begin{aligned}\mathbb{E}_{\pi\sim p_{\overline{M}}}\left[\mathbb{1}\left\{\pi \in \mathcal{A}_2\right\} \cdot g^{M_2}(\pi)\right] &= \mathbb{E}_{\pi\sim p_{\overline{M}}}\left[\mathbb{1}\left\{\pi \in \mathcal{A}_1^{\mathrm{c}}\right\} \cdot g^{M_2}(\pi)\right] - \mathbb{E}_{\pi\sim p_{\overline{M}}}\left[\mathbb{1}\left\{\pi \in \mathcal{A}_1^{\mathrm{c}} \cap (\mathcal{E}^{M_2})^{\mathrm{c}}\right\} \cdot g^{M_2}(\pi)\right]\\ &\geq \mathbb{E}_{\pi\sim p_{\overline{M}}}\left[\mathbb{1}\left\{\pi \in \mathcal{A}_1^{\mathrm{c}}\right\} \cdot g^{M_2}(\pi)\right] - c_0\delta \geq \frac{\delta}{2} \cdot (1 - 2c_0)\,, &(26)\end{aligned}$$

这里第一个不等式成立，是因为对于 $\pi \in (\mathcal{E}^{M_2})^{\mathrm{c}}$， $g^{M_2}(\pi) < c_0\delta$。

接下来，我们计算

$$\begin{aligned}c_0\delta &\geq \mathbb{E}_{\pi\sim p_{\overline{M}}}\left[\mathbb{1}\left\{\pi \in \mathcal{A}_2\right\} \cdot (f^{M_1}(\pi_{M_1}) - f^{M_1}(\pi))\right]\\ &= \mathbb{E}_{\pi\sim p_{\overline{M}}}\left[\mathbb{1}\left\{\pi \in \mathcal{A}_2\right\} \cdot (f^{M_2}(\pi_{M_2}) - f^{M_1}(\pi))\right] + p_{\overline{M}}(\mathcal{A}_2) \cdot (f^{M_1}(\pi_{M_1}) - f^{M_2}(\pi_{M_2}))\\ &\geq \mathbb{E}_{\pi\sim p_{\overline{M}}}\left[\mathbb{1}\left\{\pi \in \mathcal{A}_2\right\} \cdot (f^{M_2}(\pi_{M_2}) - f^{M_2}(\pi))\right] - 2\varepsilon + p_{\overline{M}}(\mathcal{A}_2) \cdot (f^{M_1}(\pi_{M_1}) - f^{M_2}(\pi_{M_2}))\\ &\geq \frac{\delta}{2} \cdot (1 - 2c_0) - 2\varepsilon + p_{\overline{M}}(\mathcal{A}_2) \cdot (f^{M_1}(\pi_{M_1}) - f^{M_2}(\pi_{M_2}))\,,\end{aligned}$$

其中，第一个不等式成立是因为 $g^{M_1}(\pi) < c_0\delta$ 当 $\pi \in \mathcal{A}_2 \subset \mathcal{A}_1^{\mathrm{c}} = (\mathcal{E}^{M_1})^{\mathrm{c}}$；倒数第二个不等式由 公式（24）以及 $\mathcal{A}_2 \subset \mathcal{A}_1^{\mathrm{c}}$这一事实得出，最后一个不等式则由 公式（26）得出。重新排列并利用 $p_M(\mathcal{A}_2) \in [0,1]$，可得

$$\begin{aligned}f^{M_2}(\pi_{M_2}) - f^{M_1}(\pi_{M_1}) &\geq p_{\overline{M}}(\mathcal{A}_2) \cdot (f^{M_2}(\pi_{M_2}) - f^{M_1}(\pi_{M_1}))\\ &\geq \frac{\delta}{2} \cdot (1 - 4c_0) - 2\varepsilon.\end{aligned}$$

给出失败概率的下界并得出结论。为完成证明，我们利用不等式(25) 对概率 $p_M((\mathcal{E}^{M_2})^c \cap \mathcal{A}_1^c)$ 进行如下界定：

$$\begin{aligned} p_{\overline{M}}((\mathcal{E}^{M_2})^c \cap \mathcal{A}_1^c) \cdot \left(\frac{\delta}{2} \cdot (1 - 4c_0) - 2\varepsilon\right) &\leq \mathbb{E}_{\pi \sim p_{\overline{M}}}\left[\mathbb{1}\left\{(\mathcal{E}^{M_2})^c \cap \mathcal{A}_1^c\right\} \cdot (f^{M_2}(\pi_{M_2}) - f^{M_1}(\pi_{M_1}))\right] \\ &\leq \mathbb{E}_{\pi \sim p_{\overline{M}}}\left[\mathbb{1}\left\{(\mathcal{E}^{M_2})^c \cap \mathcal{A}_1^c\right\} \cdot ((f^{M_2}(\pi_{M_2}) - f^{M_2}(\pi)) - (f^{M_1}(\pi_{M_1}) - f^{M_1}(\pi)))\right] + 2\epsilon \\ &\leq \mathbb{E}_{\pi \sim p_{\overline{M}}}\left[\mathbb{1}\left\{(\mathcal{E}^{M_2})^c \cap \mathcal{A}_1^c\right\} \cdot g^{M_2}(\pi)\right] + 2\varepsilon \\ &\leq c_0\delta + 2\varepsilon, \end{aligned}$$

其中，第一个不等式使用了 公式（25），第二个不等式使用了 公式（24）以及 $(\mathcal{E}^{M_2})^c \cap \mathcal{A}_1^c \subset \mathcal{A}_1^c$这一事实，第三个不等式利用了 $g^{M_1} \geq 0$，最后一个不等式则使用了 $\mathcal{E}^{M_2}$的定义。由于我们假设 $\varepsilon \leq \frac{\delta}{48}$ 且 $c_0 \leq 1/16$，因此可得 $p_M((\mathcal{E}^{M_2})^c \cap \mathcal{A}_1^c) \leq \frac{\delta/8}{\delta/4} = 1/2$。由此，我们证明了

$$p_{\overline{M}}(\mathcal{A}_1) + p_{\overline{M}}(\mathcal{E}^{M_2}) \geq p_{\overline{M}}(\mathcal{A}_1 \cup \mathcal{E}^{M_2}) \geq 1/2,$$

这意味着 $p_M(\mathcal{E}^{M_1}) = p_M(\mathcal{A}_1) \geq 1/4$或 $p_M(\mathcal{E}^{M_2}) \geq 1/4$。因此，根据 公式（23），我们可知，对于某个 $i \in \{1, 2\}$，$\mathbb{E}_{\pi \sim p_{M_i}}[g^{M_i}(\pi)] \geq \frac{3c_0\delta}{20}$，从而得到了所需的下界。

最后需要指出，我们所证明的下界对任意参考模型$M \in \mathcal{M}^+$都成立，因此可以自由选择 $M \in \mathcal{M}^+$，使 $\text{p-dec}^c_{\varepsilon/\sqrt{2}}(\mathcal{M}, M)$ 最大化。

□

注释 2.3。 定理 2.1 的证明结构 与统计学和信息论中的经典两点法（Donoho 和 Liu，1987年，1991a，b；Yu，1997年；Tsybakov，2008年）在表面上有些相似之处，但存在若干根本差异。

1. 首先，在我们的下界中，模型对（M_1，M_2）是根据所考虑的算法以对抗性的方式选取的，而经典方法则不加区别地选取一对模型。仅考虑两个模型时，以对抗方式选择模型对于刻画那些需要在许多不同决策之间进行区分的类别的复杂度至关重要。例如，即使在多臂老虎机这一简单的特殊情形中，这也是使动作数量 A 出现在下界中的必要条件。

2. 其次，也许更重要的是，经典的两点论证无法直接应用，因为函数 $g^M(\pi)$ 不具备该方法所要求的度量结构。特别是，经典的“分离条件”在应用于我们的设置时形式为 $g^{M_1}(\pi) + g^{M_2}(\pi) \geq \delta \ \forall \pi$，但并不成立。相反，证明的关键在于表明，由于我们对M_1 和 M_2的选择，我们有

$$\mathbb{E}_{\pi \sim p_{M_1}}[g^{M_1}(\pi)] + \mathbb{E}_{\pi \sim p_{M_2}}[g^{M_2}(\pi)] \gtrsim \delta.$$

为证明这一不等式，我们利用以下事实：1）根据假设，$\text{p-dec}^c_{\varepsilon(T)}(\mathcal{M}) \gtrsim \varepsilon(T)$；2）奖励 r可被观测，且取值范围为 $[0, 1]$；后者意味着，对于所有 $M, \bar{M} \in \mathcal{M}^+$，$|f^M(\pi) - f^{\bar{M}}(\pi)| \leq D_{\mathsf{H}}\left(M(\pi) \| \bar{M}(\pi)\right)$。

3 上界

现在，我们给出一些算法，其样本复杂度上界与第 2 节中的下界相辅相成。首先，在第3.2节中，我们给出 PAC 框架的一个算法及其上界。随后，我们在第3.3节中证明这一结果，并重点阐述关键的算法设计原则和分析思路。接着，在第3.4节中，我们给出一个用于遗憾的算法及其上界。我们的两个算法都采用从估计到决策范式，该范式由Foster 等人（2021年）提出。我们的遗憾算法建立在 PAC 算法的基础上，但技术上更为复杂。

3.1 在线估计

我们的算法和遗憾界使用在线估计预言机这一基本工具，记作 $\mathbf{Alg}_{\mathsf{Est}}$。它是一种算法，已知某个包含真实模型 $M^\star$ 的类别 $\mathcal{M}$ 后，能够根据数据 $M^\star$ 在线估计底层模型。在每一轮次 t，给定目前为止观测到的数据 $\mathfrak{H}^{t-1} = (\pi^1, r^1, o^1) \ldots, (\pi^{t-1}, r^{t-1}, o^{t-1})$，估计预言机构造一个估计

$$\widehat{M}^t = \mathbf{Alg}_{\mathsf{Est}}\left(\{(\pi^i, r^i, o^i)\}_{i=1}^{t-1}\right)$$

对于真实模型 $M^\star$。我们根据累积海灵格误差衡量预言机的估计性能，并假设该误差满足以下上界。

假设 3.1 (针对 $\mathcal{M}$ 的估计预言机)。 在每个时间点 $t \in [T]$，在线估计预言机 $\mathbf{Alg}_{\mathsf{Est}}$ 针对$\mathcal{M}$，在给定

$$\mathfrak{H}^{t-1} = (\pi^1, r^1, o^1), \ldots, (\pi^{t-1}, r^{t-1}, o^{t-1})$$

且 $(r \quad \frown^i, o^i) \sim M^\star(\pi^i)$ 以及 $\pi^i \sim p^i$ 时，返回一个估计器 $M^t \in \mathrm{co}(\mathcal{M})$，使得只要 $M^\star \in \mathcal{M}$，

$$\mathbf{Est}_{\mathsf{H}}(T) := \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left[D^2_{\mathsf{H}}\left(M^\star(\pi^t), \widehat{M}^t(\pi^t)\right)\right] \le \mathbf{Est}_{\mathsf{H}}(T, \delta), \tag{27}$$

以至少 $1-\delta$ 的概率成立，其中 $\mathbf{Est}_{\mathsf{H}}(T, \delta)$ 是一个已知上界。

满足假设 3.1 的算法可以通过在线条件密度估计（即采用对数损失的在线学习）获得。通常，最优的估计速率 $\mathsf{Est}_{\mathsf{H}}(T, \delta)$ 会反映类别 $\mathcal{M}$ 的统计容量。标准示例包括有限类，其中指数权重算法（也称为 Vovk 聚合算法）可达到 $\mathsf{Est}_{\mathsf{H}}(T, \delta) \le O(\log(|\mathcal{M}|/\delta))$；以及 $\mathbb{R}^d$ 中的参数类，其中可达到 $\mathsf{Est}_{\mathsf{H}}(T, \delta) \le \widetilde{O}(d)$。更多背景信息请参阅 Foster 等人（2021）的第 4 节。

3.2 PAC 的算法与上界

算法 1展示了我们针对 PAC 框架的主要算法，$\mathsf{E2D}^+$。该算法基于Foster 等人(2021年) 提出的从估计到决策范式，每一轮 t均采用以下方案:

- 从在线估计预言机 $\mathbf{Alg}_{\widehat{\mathsf{Est}}}$ 中获得针对 $M^\star$ 的估计器 $M^t \in \mathsf{co}(\mathcal{M})$。
- 从通过求解定义 DEC 的极小极大优化问题而得到的决策分布中 π^t 采样，并将 M^t 代入作为参考模型。

$\mathsf{E2D}^+$ 遵循这一模板，但做了非平凡的调整，以适应 i) 受约束的（而非偏移的）DEC，以及 ii) PAC 保证（而非遗憾）。简而言之，该算法由两个阶段组成：探索阶段 和利用阶段，下面将对此进行概述。

$\widetilde{\ }$探索阶段。在探索阶段（包含轮次 $t = 1$、.....、 J，其中 $J = \Omega(T)$），$\widehat{\ }$算法 1 会反复查询估计预言机 $\mathsf{Alg}_{\mathsf{Est}}$，利用当前数据集 $\mathfrak{H}^{t-1} = (\pi^1, r^1, o^1)$ ..... $(\pi^{t-1}, r^{t-1}, o^{t-1})$ 获取一个估计器 M^t （第 5 行），然后计算一对分布 (p^t, q^t)：将 M^t 代入作为参考模型，求解定义 PAC DEC 的极小极大问题（公式（11））（第 6 行）：

$$(p^t, q^t) := \arg\min_{p,q \in \Delta(\Pi)} \sup_{M \in \mathcal{H}_{q, \bar{\varepsilon}(T)}(\widehat{M}^t)} \mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi)]. \tag{28}$$

$\widehat{\ }$根据定义，上述值始终有界于 $\mathsf{p\text{-}dec}^{\mathsf{c}}_{\bar{\varepsilon}(T)}(\mathcal{M}, M^t)$。回想一下， p^t可以被视为一种利用分布，而 q^t可以被视为一种探索分布。有了这些分布，算法 1 从探索分布 $\pi^t \sim q^t$ 中抽取样本 q^t。分布 p^t在此阶段不会被使用，但会保留供后续的利用阶段使用。

算法 1 从估计到决策（E2D$^+$），用于 PAC

参数：
轮数 $T \in \mathbb{N}$。
失败概率 $\delta > 0$。
在线估计预言机$\mathbf{Alg}_{\mathsf{Est}}$。
定义 $L := \lceil \log 2/\delta \rceil$, $J := \frac{T}{L+1}$, 以及 $\mathbf{Est}_{\mathsf{H}} := \mathbf{Est}_{\mathsf{H}}\left(\lceil 2T \log 2/\delta \rceil, 4\lceil \frac{\delta}{\log 2/\delta} \rceil\right)$。
集合 $\bar{\varepsilon}(T) := 8\sqrt{\frac{\lceil \log 2/\delta \rceil}{T}}$ ~~· 估计$_{\mathsf{H}}$.~~

/* 探索阶段 */

对于 $t = 1, 2, \ldots, J$，**执行**
获取估计 $\widehat{M}^t = \mathbf{Alg}_{\mathsf{Est}}\left(\{(\pi^i, r^i, o^i)\}_{i=1}^{t-1}\right)$。
计算

$$(p^t, q^t) := \arg\min_{p,q \in \Delta(\Pi)} \sup_{M \in \mathcal{H}_{q,\bar{\varepsilon}(T)}(\widehat{M}^t)} \mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi)],$$

按照约定，当 $\mathcal{H}_{q,\varepsilon(T)}(M^t) = \varnothing$ 时，值为零。
采样决策 $\pi^t \sim q^t$ 并使用 (π^t, r^t, o^t) 更新估计预言机 $\mathbf{Alg}_{\mathsf{Est}}$。

/* 利用阶段 */

从均匀分布 $t_L \sim \mathrm{Unif}([J])$ 独立抽取样本索引 L、t_1、……、t_L。对于每个 $\ell \in [L]$，抽取 J 个独立样本 π_ℓ^1、……、$\pi_\ell^J \sim q^{t_\ell}$，并对每个 $j \in [J]$ 观测 $(\pi_\ell^j, r_\ell^j, o_\ell^j)$。对于每个 $\ell \in [L]$ 和 $j \in [J]$，计算

$$\widetilde{M}_\ell^j := \mathbf{Alg}_{\mathsf{Est}}\left(\{(\pi_\ell^i, r_\ell^i, o_\ell^i)\}_{i=1}^{j-1}\right),$$

并令 $\widetilde{M}_\ell := \frac{1}{J}\sum_{j=1}^J \widetilde{M}_\ell^j$。 // $\widetilde{M}_\ell$ 是在 q^{t_ℓ} 条件下对 $M^\star$ 的高质量估计。
输出： 设 $\widehat{p} := p^{t_{\widehat{\ell}}}$，并输出 $\widehat{\pi} \sim \widehat{p}$，其中 $\widehat{\ell} := \arg\min_{\ell \in [L]} \mathbb{E}_{\pi \sim q^{t_\ell}}\left[D^2_{\mathsf{H}}\left(\widehat{M}^{t_\ell}(\pi), \widetilde{M}_\ell(\pi)\right)\right]$。

利用阶段。在利用阶段（或称后处理阶段）中，我们旨在从探索阶段计算得到的集合中识别出一个利用分布 $p^t \in \{p^1, \ldots, p^J\}$，其“良好”之处在于：在 $M^\star$ 下具有足够低的次优性。该阶段的动机如下。根据表达式（28）以及受约束 DEC 的定义，我们可以保证 p^t 具有

$$\mathbb{E}_{\pi \sim p^t}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)\right] \le \mathsf{p\text{-}dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M}, \widehat{M}^t)$$

对于任意轮次 $t \in [J]$，其中 $M^\star \in \mathcal{H}_{q^t,\varepsilon}(T)(M^t)$。这里的挑战在于，估计预言机仅能确保估计器 M^1、……、M^J 的累积估计误差较低。无法保证逐轮估计误差 $\mathbb{E}_{\pi^t \sim q^t}\left[D^2_{\mathsf{H}}\left(M^\star(\pi^t), \widehat{M}^t(\pi^t)\right)\right]$ 会随时间减少；对于任意给定的目标轮次 t，该量都可能大得微不足道。这可能导致一个我们称为错误排除的问题，即 $M^\star \notin \mathcal{H}_{q^t,\varepsilon}(T)(M^t)$。错误排除之所以成问题，是因为对于发生错误排除的轮次 $t \in [J]$，我们无法控制在 $M^\star$ 下的次优性。好消息是，尽管在线估计预言机可能导致某些轮次发生错误排除，但马尔可夫不等式意味着（对于我们所选择的 $\varepsilon(T)$，其取决于 $\mathsf{Est}_{\mathsf{H}}(T, \delta)$），真实模型 $M^\star$ 将至少在一半的轮次 $t \in [J]$中被包含在 $\mathcal{H}_{q^t,\varepsilon}(T)(M^t)$ 中。因此，利用阶段会对少量轮次 t_1、……、$t_L \in [J]$进行采样（对数级数量即可确保至少有一个轮次以高概率是良好的），并执行测试，以便在集合 $\{p^{t_1}, \ldots, p^{t_L}\}$中识别出一个良好的分布 $p^{t_{\widehat{\ell}}}$，随后由算法返回该分布。

更具体地说，利用阶段（第 8 行 至 第 11 行）通过收集大量（即 ，

$\Theta(T/\log T)$) 个来自 q^{t_ℓ}的样本，用于每个 L 轮次 $\{t_\ell\}_{\ell\in[L]}$；随后，对于每个 $\ell \in [L]$，使用$\widetilde{\ }$估计预言机 $\mathrm{Alg}_{\mathrm{Est}}$，基于这些样本生成一个估计模型 $M_\ell \in \mathcal{M}$。由于使用了大量$\widetilde{\ }$样本来生成估计 M_ℓ，因此可以保证它在 q^{t_ℓ}下与真实模型 $M^\star$ 足够接近。$\widehat{\ }$ $\widetilde{\ }$这意味着，选择一个使 $M^{t_{\widehat{\ell}}}$ 与 $M_{\widehat{\ell}}$ 之间的海林格距离最小的轮次 $t_{\widehat{\ell}}$（第 11 行），$\widehat{\ }$我们便可以高概率地得到 $M^\star \in \mathcal{H}_{q^{t_{\widehat{\ell}}},\ \varepsilon}(M^{t_{\widehat{\ell}}})$，从而解决错误排除问题，并确保利用分布 $p^{t_{\widehat{\ell}}}$ 具有较低的风险。

主要结果。我们证明$\mathsf{E2D}^+$ 满足以下 PAC 保证。

定理 *3.1*（主要上界：*PAC*）。固定 $\delta \in \left(0,\ \frac{1}{10}\right)$ 和 $T \in \mathbb{N}$。假设假设 *1.1* 和 *3.1* 成立，并令 $\mathit{Est}_{\mathsf{H}} := \mathit{Est}_{\mathsf{H}}\left(\lceil {}^{2T}\mathit{log}_{\ 2/\delta}\rceil,\ \lceil {}^{\delta}\mathit{log}_{\ 2/\delta}\rceil\right)$。取 $\varepsilon(T) := 8\sqrt{\lceil \mathit{log}_{T}^{2/\delta}\rceil \cdot \mathit{Est}_{\mathsf{H}}}$，算法 *1* 保证以至少 $1-\delta$ 的概率，

$$\mathbf{Risk}_{\mathsf{DM}}(T) \le \text{p-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M}).$$

因此，算法 *1*所达到的期望风险有如下上界

$$\mathbb{E}[\mathbf{Risk}_{\mathsf{DM}}(T)] \le \text{p-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M}) + \delta. \tag{29}$$

这一结果与定理 2.1中的下界相匹配，唯一的差别在于受约束 DEC 的半径选择 $\varepsilon > 0$：下界（定理 2.1）为 $\varepsilon(T) \propto \sqrt{\frac{1}{T}}$，而上界（定理 3.1）为$\varepsilon(T) \propto \sqrt{\log(2/\delta)\cdot \mathbf{Est}_{T}}$ $^{\mathsf{H}}$。正如Foster 等人 (2021年所讨论的，何时可以去除或削弱估计复杂度$\mathbf{Est}_{\mathsf{H}}$ 是一个微妙的问题，因为有些类别 $\mathcal{M}$ 中这一项是必要的，而在另一些类别中则是多余的。这是我们的研究尚未解决的主要问题。

我们来针对一些标准示例具体应用定理 3.1，重点考虑 $\mathcal{M}$ 为有限值且 $\mathbf{Est}_{\mathsf{H}}(T,\delta) \le \log(|\mathcal{M}|/\delta)$ 这一特殊情形，以简化表述。

- 只要 $\text{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \propto \varepsilon \cdot \sqrt{C_{\mathrm{prob}}}$，定理 3.1 就给出 $\mathbb{E}[\text{风险}_{\mathsf{DM}}(T)] \le \widetilde{O}\left(\sqrt{\frac{C_{\mathrm{prob}}\log|\mathcal{M}|}{T}}\right)$，这意味着需要 $\widetilde{O}\left(C_{\mathrm{prob}}\log_{\varepsilon^2}|\mathcal{M}|\right)$ 个样本来学习一个 ε-最优策略。
- 只要 $\text{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \propto \varepsilon^{1-\rho}$对于 $\rho \in (0,\ 1)$，定理 3.1 就给出 $\mathbb{E}[\text{风险}_{\mathsf{DM}}(T)] \le \widetilde{O}((\log|\mathcal{M}|/T)^{(1-\rho)^2})$，这意味着需要 $\widetilde{O}(\log|\mathcal{M}|\cdot\varepsilon^{-\frac{2}{1-\rho}})$ 个样本来学习一个 ε-最优策略。
- 只要 $\text{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \propto \mathbb{I}\{\varepsilon \ge 1/\sqrt{C_{\mathrm{prob}}}\}$，其中 C_{prob} 是一个问题相关参数，定理 3.1$\widetilde{\ }$就以高概率给出 $\mathbf{Risk}_{\mathsf{DM}}(T) = 0$，条件是 $T \ge \Omega(C_{\mathrm{prob}}\cdot\log|\mathcal{M}|)$。

3.3 PAC 上界的证明（定理 3.1）

现在我们来证明定理 3.1。在开始证明之前，我们先简要介绍在线到批量转换这一概念；利用阶段及其分析中会用到它。

背景：在线到批量转换。如前文所述，我们假设可以访问在线预言机$\mathbf{Alg}_{\mathrm{Est}}$。在探索阶段，我们以一种非平凡的方式利用该预言机提供的在线保证——即对于自适应选择的决策序列，它能确保累积估计误差有界；此外，我们的分析还利用了在线预言机能够为离线估计提供保证这一事实。

对于离线估计，我们考虑这样一种设定：存在某个 $p \in \Delta(\Pi)$，使得对所有 t，$p^t = p$ 均成立；算法必须输出单个模型估计 $\widehat{M}$，使得 $\mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\left(\widehat{M}(\pi)\ M^\star(\pi)\right)\right] \le \varepsilon^2$。我们可以通过在线估计预言机，使用以下在线到批量转换过程获得这样的保证：

- 对于每个 $t = 1, \ldots\ldots, T$，通过对样本 $\{(\pi^i, r^i, o^i)\}_{i=1}^{t-1}$运行 **算法**估计，获得 $\widehat{M}^t$，其中 $\pi^t \sim p$，且 $(r^t, o^t) \sim M^\star(\pi^t)$。
- 令 $\widehat{M} := \frac{1}{T}\sum_{t=1}^{T} \widehat{M}^t \in \mathrm{co}(\mathcal{M})$。

显然，根据 假设 3.1 以及平方海灵格距离的凸性，以至少 $1-\delta$ 的概率，上文构造的估计器 M 满足

$$\begin{aligned}\mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\Big(M^\star(\pi), \widehat{M}(\pi)\Big)\right] &= \mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\left(M^\star(\pi), \frac{1}{T}\sum_{t=1}^{T}\widehat{M}^t(\pi)\right)\right] \\ &\le \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\Big(M^\star(\pi), \widehat{M}^t(\pi)\Big)\right] \le \frac{\mathbf{Est}_{\mathsf{H}}(T,\delta)}{T}. \end{aligned} \tag{30}$$

这就是 算法 1 的 第 10 行所采用的策略。标准的离线估计算法（例如最大似然估计）也可以得出类似的保证，但我们采用在线到批量的方法以简化记号，因为 算法 1 在探索阶段已经需要在线估计算法。

定理 定理 3.1的证明。我们首先分析利用阶段。e 回顾一下，我们令 $J := T_{\lceil \log 2/\delta\rceil+1} \ge \frac{T}{2L}$。由 假设 3.1，我们有，以至少 $1-\frac{\delta}{4L}$的概率，

$$\sum_{t=1}^{J}\mathbb{E}_{\pi^t\sim q^t}\left[D^2_{\mathsf{H}}\Big(M^\star(\pi^t), \widehat{M}^t(\pi^t)\Big)\right] \le \mathbf{Est}_{\mathsf{H}}\left(J, \frac{\delta}{4L}\right) \le \overline{\mathbf{Est}}_{\mathsf{H}}.$$

我们将此事件记为 $\mathscr{E}_0$，并在后续推导中以此为条件。由于根据定义有 $\varepsilon(T)^2 \ge \frac{32}{J}\cdot\mathbf{Est}_{\mathsf{H}}$，由马尔可夫不等式可知，若 $s \in [J]$ 是均匀随机选取的，则其概率至少为 $1/2$，

$$\mathbb{E}_{\pi^s\sim q^s}\left[D^2_{\mathsf{H}}\Big(M^\star(\pi^s), \widehat{M}^s(\pi^s)\Big)\right] \le \frac{\bar{\varepsilon}(T)^2}{16}. \tag{31}$$

接下来，我们的目标是证明利用阶段能够识别出这样的索引 $s \in [J]$。事实上，对于任何满足不等式 (31)的 $s \in [J]$，都有 $M^\star \in \mathcal{H}_{q^s,\,\varepsilon(T)}(M^s)$，因此

$$\mathbb{E}_{\pi\sim p^s}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)\right] \le \mathsf{p\text{-}dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M}, \widehat{M}^s) \le \mathsf{p\text{-}dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M})$$

由表达式(28) 以及受约束 DEC 的定义可得。

继续证明之前，先注意，对于均匀采样得到的索引 $t_1, \ldots, t_L \in [J]$，标准的置信度提升论证表明，以至少 $1-2^{-L} \ge 1-\frac{\delta}{2}$ 的概率，存在某个 $\ell \in [L]$，使得 式 (31) 在 $s = t_\ell$ 的情况下成立。我们将此事件记为 $\mathscr{F}$。

接下来，回顾 $\widetilde{M}_\ell = \frac{1}{J}\sum_{j=1}^{J}\widetilde{M}^j_\ell$ 在式 (10)中的算法 1的定义。结合假设 3.1 以及式 (30)，可知对每个 $\ell \in [L]$，存在一个发生概率至少为 $1-\frac{\delta}{4L}$的事件，记作 $\mathscr{E}_\ell$，在 $\mathscr{E}_\ell$ 条件下，我们有

$$\mathbb{E}_{\pi\sim q^{t_\ell}}\left[D^2_{\mathsf{H}}\Big(M^\star(\pi), \widetilde{M}_\ell(\pi)\Big)\right] \le \frac{\mathbf{Est}_{\mathsf{H}}(J,\delta/4L)}{J} \le \frac{\overline{\mathbf{Est}}_{\mathsf{H}}}{J} \le \frac{\bar{\varepsilon}(T)^2}{32},$$

其中，第二个不等式用到了我们对 $\varepsilon(T)$ 的选择。定义 $\mathscr{E} := \bigcap_{\ell=0}^{L}\mathscr{E}_\ell$，使得 $\mathscr{E}$发生的概率至少为 $1-\frac{(L+1)\delta}{4L} \ge 1-\frac{\delta}{2}$。我们定义 $\mathscr{E} = \mathscr{F}\cap\bigcap_{\ell=0}^{L}\mathscr{E}_\ell$，使得 $\mathscr{E}$发生的概率至少为$1-\frac{(L+1)\delta}{4L} - \frac{\delta}{2} \ge 1-\delta$。

我们现在证明，只要事件 $\mathscr{E}$ 成立，利用阶段就会成功。根据海林格距离的三角不等式，令 $\ell \in [L]$ 为满足 式（31）且 $s = t_\ell$的任意索引，我们有

$$\mathbb{E}_{\pi\sim q^{t_\ell}}\left[D^2_{\mathsf{H}}\Big(\widetilde{M}_\ell(\pi), \widehat{M}^{t_\ell}(\pi)\Big)\right] \le 2\cdot\left(\mathbb{E}_{\pi\sim q^{t_\ell}}\left[D^2_{\mathsf{H}}\Big(M^\star(\pi), \widetilde{M}_\ell(\pi)\Big) + D^2_{\mathsf{H}}\Big(M^\star(\pi), \widehat{M}^{t_\ell}(\pi)\Big)\right]\right) \le \frac{\bar{\varepsilon}(T)^2}{4}.$$

根据 第 11 行中的定义，索引 $\ell \in [L]$ 满足

$$\mathbb{E}_{\pi\sim q^{t,\hat{\ell}}}\left[D^2_{\mathsf{H}}\left(\widetilde{M}_{\hat{\ell}}(\pi),\widehat{M}^{t,\hat{\ell}}(\pi)\right)\right] \le \mathbb{E}_{\pi\sim q^{t,\ell}}\left[D^2_{\mathsf{H}}\left(\widetilde{M}_{\ell}(\pi),\widehat{M}^{t,\ell}(\pi)\right)\right] \le \frac{\bar{\varepsilon}(T)^2}{4}.$$

再次利用海林格距离的三角不等式，我们得到在事件 $\mathscr{E}$ 下，

$$\mathbb{E}_{\pi\sim q^{t,\hat{\ell}}}\left[D^2_{\mathsf{H}}\left(\widehat{M}^{t,\hat{\ell}}(\pi),M^\star(\pi)\right)\right] \le 2\cdot\left(\frac{\bar{\varepsilon}(T)^2}{4}+\frac{\bar{\varepsilon}(T)^2}{32}\right) < \bar{\varepsilon}(T)^2,$$

这意味着 $M^\star \in \mathcal{H}_{q^{t,\hat{\ell}},\ \varepsilon_(T)}(M^{t,\hat{\ell}})$。因此，只要 $\mathscr{E}$成立，

$$\begin{aligned}
\mathbf{Risk}_{\mathsf{DM}}(T) = \mathbb{E}_{\pi\sim p^{t,\hat{\ell}}}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)\right] &\le \sup_{M\in\mathcal{H}_{q^{t,\hat{\ell}},\bar{\varepsilon}(T)}(\widehat{M}^{t,\hat{\ell}})} \mathbb{E}_{\pi\sim p^{t,\hat{\ell}}}[f^M(\pi_M) - f^M(\pi)] \\
&= \inf_{p,q\in\Delta(\Pi)} \sup_{M\in\mathcal{H}_{q,\bar{\varepsilon}(T)}(\widehat{M}^{t,\hat{\ell}})} \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] \\
&= \textsf{p-dec}^{\mathrm{c}}_{\bar{\varepsilon}}(\mathcal{M},\widehat{M}^{t,\hat{\ell}}) \\
&\le \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \textsf{p-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M},\overline{M}) = \textsf{p-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M}),
\end{aligned}$$

其中，第一个等号由 $p^{t,\hat{\ell}}$，$q^{t,\hat{\ell}}$ 在第 6 行中的选取方式得到。定理陈述中公式（29）的结论可由以下观察得出：**风险**$_{\mathsf{DM}}$（T）的上界为 1，因为我们假设奖励取值于 $[0,1]$。 □

3.4 遗憾的算法与上界

现在我们介绍面向遗憾框架的E2D$^+$ 的一种变体，算法 2。对于适当的半径 $\varepsilon(T) > 0$，该算法可达到随 $\textsf{r-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M})\cdot T$变化的遗憾界。

在线估计。算法 2与算法 1类似，使用一个在线估计预言机，但我们要求该预言机稍强一些，能够纳入通过子集 $\mathcal{M}' \subseteq \mathcal{M}$指定的模型类上的约束。

假设 3.2 (针对 $\mathcal{M}$ 的受约束估计预言机)。 针对 $\mathcal{M}$ 的受约束估计预言机以约束集 $\mathcal{M}' \subseteq \mathcal{M}$ 作为输入。在每个时间点 $t \in [T]$，在线估计预言机 **算法**$_{\mathsf{Est}}$ 在给定

$$\mathfrak{H}^{t-1} = (\pi^1, r^1, o^1), \ldots, (\pi^{t-1}, r^{t-1}, o^{t-1})$$

以及 $(r^i, o^i) \sim M^\star(\pi^i)$和 $\pi^i \sim p^i$ 后，返回一个估计器 $M^t \in \mathrm{co}(\mathcal{M}')$，使得只要 $M^\star \in \mathcal{M}'$，

$$\mathbf{Est}_{\mathsf{H}}(T) := \sum_{t=1}^{T} \mathbb{E}_{\pi^t\sim p^t}\left[D^2_{\mathsf{H}}\left(M^\star(\pi^t),\widehat{M}^t(\pi^t)\right)\right] \le \mathbf{Est}_{\mathsf{H}}(T,\delta), \tag{32}$$

其成立的概率至少为 $1-\delta$，其中 $\mathbf{Est}_{\mathsf{H}}(T,\delta)$ 是一个已知上界。

这个假设与 假设 3.1 相同，不同之处在于：i) 预言机在学习过程开始前接收一个 约束集 $\mathcal{M}' \subseteq \mathcal{M}$ 作为输入；ii) 预言机必须生成 $M^t \in \mathrm{co}(\mathcal{M}')$；也就是说，估计器必须位于约束集 $\mathcal{M}'$ 的凸包中。据我们所知，所有估计算法只需稍作修改或无需修改即可达到这一保证，包括满足 假设 3.2 的$\mathbf{Est}_{\mathsf{H}}(T,\ \delta) \le O(\log(|\mathcal{M}|/\delta))$。

[3]算法 2 将继续运行；如果执行到此 if 语句，我们可以修改它，使其跳出关于 $s \in \mathcal{S}_i$ 的循环，这在某种程度上更自然。我们采用此处给出的版本，因为它能稍微简化证明。

算法 2 从估计到决策 (E2D$^+$) 的遗憾

参数:

轮数 $T \in \mathbb{N}$ 和错误概率 $\delta \in (0,1)$。

失败概率 $\delta > 0$。

在线估计预言机 $\mathbf{Alg}_{\mathsf{Est}}$。

常数 $C_1 > 1$。 // 详见 附录 C。

定义 $N := \lceil \log T \rceil$ 和 $L := \lceil \log 1/\delta \rceil$。

集合 $\varepsilon_N := \sqrt{\frac{C_1 \mathbf{dim}_{\mathsf{H}}(\mathcal{M}, T^{-1}) L(\delta)}{T}}$ 对于每个 $\varepsilon_i := \sqrt{2^{N-i}} \cdot \varepsilon_N$，执行 $i \in [N]$。

将 $[T]$ 拆分为 $2N$ 个连续区块 $\mathcal{E}_1 \cup \mathcal{R}_1 \cup \cdots \cup \mathcal{E}_N \cup \mathcal{R}_N$，其中 $|\mathcal{R}_i|, |\mathcal{E}_i| \in \left[\frac{2^i}{4}, \frac{2^i}{2}\right]$ 对所有 $i \in [N]$ 均成立。

设置 $\mathcal{M}_1 := \mathcal{M}$。

对于 $i \in [N]$，**执行**

/* 第 i 轮的探索*/

初始化一个 $\mathbf{Alg}_{\mathsf{Est}}$ 实例，时间范围为 $|\mathcal{E}_i|$，失败概率为 δ，模型类为 $\mathcal{M}_i$。 **对于** $t \in \mathcal{E}_i$ **执行**

获得估计 $M^t := \mathbf{Alg}_{\mathsf{Est}}(\{(\pi^s, r^s, o^s)\}_{s \in \mathcal{E}_i,\ s<t})$，其中 $M^t \in \mathsf{co}(\mathcal{M}_i)$。 计算

$$p^t := \arg\min_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{H}_{p,\varepsilon_i}(\widehat{M}^t) \cup \{\widehat{M}^t\}} \mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi)].$$

采样决策 $\pi^t \sim p^t$ 并使用 (π^t, r^t, o^t) 更新估计预言机 $\mathbf{Alg}_{\mathsf{Est}}$。

/* 第 i 轮的精化*/

有放回地均匀随机抽取一个大小为 L 的子集 $\mathcal{S}_i \subset \mathcal{E}_i$。初始化 $s_i^{\mathrm{tmp}} = \perp$。// 高概率下，$s_i^{\mathrm{tmp}}$ 将在 for 循环中更新。对于 $s \in \mathcal{S}_i$，执行初始化一个时域为 $J_i := |\mathcal{R}_i|/L$、失败概率为 δ、模型类为 $\mathcal{M}_i$ 的 $\mathsf{Alg}_{\mathsf{Est}}$ 实例。对于 $1 \le j \le J_i$，执行采样一个决策 $\pi_s^j \sim p^s$，并使用 (π_s^j, r_s^j, o_s^j) 更新 $\mathsf{Alg}_{\mathsf{Est}}$。获得估计值 $\widetilde{M}_s^j := \mathsf{Alg}_{\mathsf{Est}}\left(\{(\pi_s^k, r_s^k, o_s^k)_{k=1}^{j-1}\}\right)$，其中 $\widetilde{M}_s^j \in \mathsf{co}(\mathcal{M}_i)$。如果 $\sum_{k=1}^{j} \mathbb{E}_{\pi \sim p^s}\left[D_{\mathsf{H}}^2\left(\widehat{M}^s(\pi), \widetilde{M}_s^k(\pi)\right)\right] > \frac{J_i \varepsilon_i^2}{4}$，则定义 $J_{i,s} := j$，并跳出关于 j 的循环（即继续处理 $\mathcal{S}_i$ 的下一个取值）。如果 $j = J_i$，则设置 $s_i^{\mathrm{tmp}} = s$ 和 $J_{i,s} = J_i$。设置 $s_i^{\mathrm{tmp}} \neq \perp$，则设置 $s_i = s_i^{\mathrm{tmp}}$；否则任意选择 $s_i \in \mathcal{S}_i$。设置 $\widehat{M}_i := \frac{1}{J_i} \sum_{j=1}^{J_i} \widetilde{M}_{s_i}^j$ 和 $\widehat{p}_i = p^{s_i}$。对于 $\mathcal{R}_i$ 中任意剩余轮次 t，执行采取行动 $\pi^t \sim p^{s_i}$（奖励和观测可以忽略）。设置

$$\mathcal{M}_{i+1} := \left\{ M \in \mathcal{M} \mid \mathbb{E}_{\pi \sim \widehat{p}_i}\left[D_{\mathsf{H}}^2\left(M(\pi), \widehat{M}_i(\pi)\right)\right] \le \frac{\mathbf{Est}_{\mathsf{H}}(J_i, \delta)}{J_i} \right\}.$$

算法概述。算法 2 采用Foster 等人(2021年) 的从估计到决策原则，但与算法 1一样，针对受约束（而非偏移）$\widehat{}$DEC 做了大量修改。算法的核心是第 10 行，它在每一轮 t都会获得一个估计器 $M^t \in \mathrm{co}(\mathcal{M})$，然后通过求解定义遗憾 DEC 的极小极大问题来计算探索分布 p^t （公式$\widehat{}$(7)），其中代入了 M^t：

$$p^t := \arg\min_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{H}_{p,\varepsilon}(\widehat{M}^t)\cup\{\widehat{M}^t\}} \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)], \tag{33}$$

对于依赖于 t 的适当 $\varepsilon > 0$ 取值。

与算法 1一样，算法 2 需要克服的主要挑战是错误排除：每当 $\widehat{}M^\star \in \mathcal{H}_{p^t,\ \varepsilon}(M^t)$ 时，根据上述定义可立即得出

$$\mathbb{E}_{\pi^t\sim p^t}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)\right] \le \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\widehat{M}^t\}, \widehat{M}^t),$$

$\widehat{}$但如果 $M^\star \notin \mathcal{H}_{p^t,\ \varepsilon}(M^t)$会发生什么？对于 PAC，算法 1 采用的解决方案很简单：对于 $\mathbf{Alg}_{\mathsf{Est}}$，在线估计保证 $M^\star$ 至少在一半的轮次中被正确纳入，而我们只需找出其中一个被纳入的轮次即可。对于遗憾而言，这种推理已不再充分：我们不能简单地忽略 $M^\star$ 被排除的轮次，因为必须控制这些轮次所产生的遗憾。

轮次方案。 为了解决错误排除的问题，算法 2 将轮次 $1,\dots,T$划分为$1,\dots,N$个长度依次翻倍的轮次，每个轮次 i由一段连续的探索轮次 $\mathcal{E}_i \subseteq [T]$和细化轮次 $\mathcal{R}_i \subseteq [T]$组成。每个轮次的执行方式与算法 1类似，但会利用先前轮次收集的数据，以一种能够抵御错误排除的方式进行探索：

- 每个探索阶段都使用一系列探索分布来收集数据 $\{p^t\}_{t\in\mathcal{E}_i}$，这些分布通过$\widehat{}$使用估计模型求解 DEC $\{M^t\}_{t\in\mathcal{E}_i}$计算得到，依据公式（33）。然而，估计$\widehat{}$模型 M^t 被限制在$\mathrm{co}(\mathcal{M}_i)$之中，其中 $\mathcal{M}_i$ 是使用上一轮次的数据计算得到的置信集。

- 轮次 i 中细化阶段的目的是利用探索阶段生成的分布，为下一轮次计算置信集 $\mathcal{M}_{i+1}$，使其满足某个不变量；该不变量使我们能够将相对于置信集内模型的低遗憾转化为在 $M^\star$ 下的低遗憾。

为了更详细地描述轮次 $i \in [N]$ 中的各个阶段，我们定义

$$\alpha_i := C_0 \cdot \textsf{r-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M}) + 64\varepsilon_i,$$

对于一个常数 $C_0 > 1$，其取值在完整证明（附录 C）中给出。

$\widehat{}$**探索阶段**。对于探索阶段 $\mathcal{E}_i$中的每个步骤 t，我们从估计预言机获得一个估计器 $M^t \in \mathsf{co}(\mathcal{M}_i)$，并求解

$$p^t := \arg\min_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{H}_{p,\varepsilon_i}(\widehat{M}^t)\cup\{\widehat{M}^t\}} \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)]. \tag{34}$$

其中 $\mathcal{M}_i \subseteq \mathcal{M}$ 是第 $i-1$ 轮细化阶段产生的置信集。集合 $\mathcal{M}_i$ 的构造方式保证了 $M^\star \in \mathcal{M}_i$ 以高概率成立；此外，还可以证明以下定位性质$\widehat{}$（利用 $M^t \in \mathrm{co}(\mathcal{M}_i)$ 对所有 $t \in \mathcal{E}_i$均成立；参见引理 C.3的证明）：

$$f^{M^\star}(\pi_{M^\star}) \le f^{\widehat{M}^t}(\pi_{\widehat{M}^t}) + \frac{\alpha_{i-1}}{2} \quad \forall t \in \mathcal{E}_i. \tag{35}$$

条件 (35) 意味着，就遗憾而言，我们始终处于有利位置：

- 如果 $M^\star \in \mathcal{H}_{p^t,\ \varepsilon_i}(\widehat{M}^t)$——也就是说，真实模型并未被错误排除——根据定义，我们有 $\mathbb{E}_{\pi\sim p^t}[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)] \le \textsf{r-dec}^{\mathrm{c}}{}_{\varepsilon_i}(\mathcal{M}) \le \alpha_i$。

- 即使 $M^\star$ 被错误排除，式（35）也意味着，除去某些干扰项后，我们有（如引理 C.4所示）：

$$\mathbb{E}_{\pi\sim p^t}\left[f^{M^\star}(\pi_{M^\star})-f^{M^\star}(\pi)\right]\lesssim \mathbb{E}_{\pi\sim p^t}\left[f^{\widehat{M}^t}(\pi_{\widehat{M}^t})-f^{\widehat{M}^t}(\pi)\right]\lesssim \alpha_{i-1},$$

其中，后一个不等式用到了定义 (34)，由此可得

$$\mathbb{E}_{\pi\sim p^t}\left[f^{\widehat{M}^t}(\pi_{\widehat{M}^t})-f^{\widehat{M}^t}(\pi)\right]\le \textsf{r-dec}^{\textsf{c}}_{\varepsilon_i}(\mathcal{M}\cup\{\widehat{M}^t\},\widehat{M}^t)\le\alpha_i.$$

在这两种情况下，每轮产生的遗憾不超过 $O(\alpha_i)$。由于采用倍增轮次安排，所有探索轮次对遗憾的总贡献不超过 $\widetilde{O}\big(\textsf{r-dec}^{\textsf{c}}_{\varepsilon(T)}(\mathcal{M})\cdot T\big)$。

细化阶段。 为了使第 $i+1$ 轮次的探索阶段分析顺利进行，第 i 轮次的细化阶段必须构造一个置信集 $\mathcal{M}_{i+1}$，使其满足尺度为 α_i 的定位性质 (35) 。为此，我们采用与算法 1中利用阶段类似的方法。对于所有轮次 $t\in\mathcal{E}_i$，只要 $M^\star$未被错误排除，我们就能保证：i) $M^\star\in\mathcal{H}_{p^t,\ \varepsilon_i}(M^t)$，以及 ii) 所有$M\in\mathcal{H}_{p^t,\ \varepsilon_i}(M^t)$ 满足

$$\mathbb{E}_{\pi\sim p^t}\left[f^{M}(\pi_{M})-f^{M}(\pi)\right]\le \textsf{r-dec}^{\textsf{c}}_{\varepsilon_i}(\mathcal{M})\le\frac{\alpha_i}{4},$$

由此可以推出，在轮次 i 中，公式 (35)中的定位性质成立（该事实的证明利用了：对于所有 $t\in\mathcal{E}_{i+1}$，都有 $M^t\in\textsf{co}(\mathcal{M}_{i+1})$）。此外，预言机的估计保证（假设 3.2）意味着，对于至少一半的轮次 $t\in\mathcal{E}_i$，$M^\star$包含于 $\mathcal{H}_{p^t,\ \varepsilon_i}(M^t)$ 中。因此，为了构造$\mathcal{M}_{i+1}$，我们抽取少量轮次 t_1、……、$t_L\in\mathcal{E}_i$（第 12 行；对数数量的 L 即足够），并基于海林格距离执行测试，从集合 $\{p^{t_1},\dots,p^{t_L}\}$ 中识别出具有以下性质的“良好”分布 p^{t_ℓ}：$M^\star\in\mathcal{H}_{p^{t_\ell},\ \varepsilon_i}(M^{t_\ell})$。然后，我们设置 $\mathcal{M}_{i+1}=\mathcal{H}_{p^{t_\ell},\ \varepsilon_i}(M^{t_\ell})$，从而确保 $M^\star\in\mathcal{M}_{i+1}$，并满足定位性质。

主要结果。 现在给出E2D$^+$ (算法 2的主要遗憾保证。为尽可能简洁地陈述结果，我们假设遗憾决策估计系数满足一个温和的增长条件。

假设 3.3 (DEC 的正则性)。 类别 $\mathcal{M}$ 满足以下条件：存在常数 $C_{\mathrm{reg}}>1$，使得对所有 $\varepsilon\in(0,2)$，均有

$$\textsf{r-dec}^{\textsf{c}}_{\varepsilon}(\mathcal{M})\le C_{\mathrm{reg}}^2\cdot\textsf{r-dec}^{\textsf{c}}_{\varepsilon/C_{\mathrm{reg}}}(\mathcal{M}).$$

该条件表明，DEC 随参数 ε变化时不会收缩得过快。当 $\textsf{r-dec}^{\textsf{c}}_{\varepsilon}(\mathcal{M})\propto\varepsilon^\rho$ 对于 $\rho\le 2$成立时，该条件自动满足，其中 $\rho\le 1$ 对应于“$\sqrt{T}$-遗憾或更高”的情形；在更一般假设下成立的遗憾界见附录 C。

定理 3.2 (主要上界：遗憾)。 固定 $T\in\mathbb{N}$ 和 $\delta\in(0,1)$。假设 假设 *1.1*、*3.2*和 *3.3* 均成立，并令 $\mathbf{Est}_{\mathsf{H}}:=\mathbf{Est}_{\mathsf{H}}(T,\delta^2)$。则 算法 *2*在适当选取 $C_1>0$ 后，可确保其以至少 $1-\delta$ 的概率，

$$\mathbf{Reg}_{\mathsf{DM}}(T)\le\textsf{r-dec}^{\textsf{c}}_{\bar\varepsilon(T)}(\mathcal{M})\cdot O(T\log(T))+O\Big(\sqrt{T\log(1/\delta)\cdot\overline{\mathbf{Est}}_{\mathsf{H}}}\Big),$$

其中 $\bar\varepsilon(T):=C\cdot\sqrt{\frac{\text{估计值}_{\mathsf{H}}\cdot\log(1/\delta)}{T}}$，其中 $C>0$ 为数值常数。

该结果的证明推迟至附录 C。与 PAC 设定中一样，该上界与相应的下界（定理 2.2）相匹配，二者仅在受约束 DEC 的半径上有所差异（$\varepsilon(T)\approx\sqrt{\frac{1}{T}}$对应定理 2.2，而 $\varepsilon(T)\approx\sqrt{\mathbf{Est}_T/T}$对应定理 3.2），并且若无进一步假设，除了对数因子外，该界无法改进。

备注 3.1（放宽正则性条件）。 由定理 3.2 的证明可立即得出以下结论。假设不采用假设 3.3，而是假设存在某个函数 $\mathrm{r}(\varepsilon)$，使得对所有 $\varepsilon \in (0,2)$，均有：(1) $\text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \le \mathrm{r}(\varepsilon)$，且 (2) $\mathrm{r}(\varepsilon) \le C_{\mathrm{reg}}^2 \cdot \mathrm{r}(\varepsilon / C_{\mathrm{reg}})$。那么定理 3.2 中的遗憾界仍然成立，只需将 $\text{r-dec}^{\mathrm{c}}_{\varepsilon(T)}(\mathcal{M})$ 替换为 $\mathrm{r}(\varepsilon(T))$。</output>

在以下条件下的定理 3.2 示例：假设 $\mathcal{M}$ 有限且 $\mathbf{Est}_{\mathsf{H}}(T,\delta) \le O\big(\log(|\mathcal{M}|/\delta)\big)$，包括：

- 每当 $\text{r-dec}_{c\varepsilon}(\mathcal{M}) \propto \varepsilon \cdot \sqrt{C_{\text{概率}}}$，定理 3.2 给出 $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \le \widetilde{O}\big(\sqrt{C_{\text{概率}} \cdot T \cdot \log|\mathcal{M}|}\big)$。
- 每当 $\text{r-dec}_{c\varepsilon}(\mathcal{M}) \propto \varepsilon^{1-\rho}$ 对于 $\rho \in (0,1)$，定理 3.2 给出 $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \le \widetilde{O}\big(T^{\frac{1}{(1+\rho)^2}} \log^{\frac{1-\rho}{2}}|\mathcal{M}| + T^{\frac{1}{2}} \log^{\frac{1}{2}}|\mathcal{M}|\big)$。

4 决策估计系数：结构性质

第 2 节和第 3 节中的下界与上界以受约束决策估计系数表示，它们除了对模型估计误差 $\mathrm{Est}_{\mathsf{H}}(T,\delta)$ 的依赖外，均是紧的（回想一下，下界采用的尺度为 $\varepsilon(T) = \widetilde{\Omega}\big(\sqrt{1/T}\big)$，而上界采用的尺度为 $\varepsilon(T) = \widetilde{O}\big(\sqrt{\mathrm{Est}_{\mathsf{H}}(T,\delta)/T}\big)$）。自然会有人想知道，这些结果与 Foster 等人（2021，2022b）给出的下界和上界有何关系；后者以偏移 DEC 表示，乍看之下两者并不容易比较。为加深对此的理解，本节将阐述 DEC 的若干结构性质。

- 在第 4.1 节中，我们证明，对于 PAC，受约束 DEC 与偏移 DEC 几乎等价。对于遗憾，我们证明，始终可以用偏移 DEC 以紧确的方式界定受约束 DEC，但反过来一般并不成立。

- 在第 4.2 节中，我们证明，受约束的 DEC 隐式地施加了一种局部化，并揭示了局部化类的遗憾决策估计系数的受约束变体与偏移变体之间更紧密的关系。

第 4.3 节和第 4.4 节探讨参考模型的作用 $M \in \mathcal{M}^+$。

- 首先，在第 4.3 节中，我们证明，定义 $\text{r-dec}_{c\varepsilon}(\mathcal{M}) = \sup_{M \in \mathrm{co}(\mathcal{M})} \text{r-dec}_{c\varepsilon}(\mathcal{M} \cup \{M\}, M)$，其中纳入了相对于参考模型 M 的次优性，对于获得紧确的上界和下界至关重要。这与 PAC 有根本区别：我们证明，对于 PAC，只需采用定义 $\text{p-dec}_{c\varepsilon}(\mathcal{M}) = \sup_{M \in \mathrm{co}(\mathcal{M})} \text{p-dec}_{c\varepsilon}(\mathcal{M}, M)$ 即可，该定义未纳入相对于 M 的次优性。

- 接着，在第 4.4 节中，我们证明，允许使用任意的、可能不适当的参考模型 $M \in \mathcal{M}^+$，不会使 DEC 的值高于使用参考模型 $M \in \mathrm{co}(\mathcal{M})$ 所能达到的值。这凸显了凸性的重要作用。此外，我们还证明，对于纳入随机混合估计量的 DEC 变体，也存在类似的等价关系。

4.1 受约束 DEC 与偏移 DEC 之间的关系

显然，可以通过偏移 DEC 以紧致的方式界定受约束 DEC。为具体起见，聚焦于遗憾，我们可以使用拉格朗日乘子法证明，对于任意 $M \in \mathcal{M}$ 和 $\varepsilon > 0$，

$$
\begin{aligned}
\text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) &= \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \big\{ \mathbb{E}_{\pi \sim p}[g^M(\pi)] \mid \mathbb{E}_{\pi \sim p}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] \le \varepsilon^2 \big\} \\
&= \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \inf_{\gamma > 0} \big\{ \mathbb{E}_{\pi \sim p}[g^M(\pi)] - \gamma\big(\mathbb{E}_{\pi \sim p}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] - \varepsilon^2\big) \big\} \\
&\le \inf_{\gamma > 0} \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \big\{ \mathbb{E}_{\pi \sim p}[g^M(\pi)] - \gamma\big(\mathbb{E}_{\pi \sim p}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] - \varepsilon^2\big) \big\} \\
&= \inf_{\gamma > 0} \big\{ \text{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) + \gamma \varepsilon^2 \big\}. \qquad (36)
\end{aligned}
$$

同样的方法也能得到 PAC 的类似不等式。对于一般模型 $M \notin \mathcal{M}$，以下略宽松的公式 (36) 成立。[4]

命题 4.1. 对于所有 $M \in \mathcal{M}^{\mp}$ 和 $\varepsilon > 0$，我们有

$$\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \le 8 \cdot \inf_{\gamma > 0} \left\{ \text{r-dec}^{\text{o}}_{\gamma}(\mathcal{M}, \overline{M}) \vee 0 + \gamma \varepsilon^2 \right\} + 7\varepsilon. \tag{37}$$

示例包括：

- 每当 $\text{r-dec}^{\text{o}}_{\gamma}(\mathcal{M}) \propto C_{\text{概率}}\gamma$ 时，命题 4.1 便推出 $\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) \lesssim \varepsilon \sqrt{C_{\text{概率}}}$。在这种情况下，我们的结果和 Foster 等人 (2021年)中的界都给出 $\mathbb{E}[\mathbf{Reg}_{\text{DM}}(T)] \le \widetilde{O}\left(\sqrt{C_{\text{概率}} \cdot T \cdot \mathbf{Est}_{\text{H}}(T, \delta)}\right)$。
- 更一般地，对于某个 $\rho < 1$，只要 $\text{r-dec}^{\text{o}}_{\gamma}(\mathcal{M}) \propto \left(\frac{C_{\text{概率}}}{\gamma}\right)^{\rho}$，那么 命题 4.1 便推出 $\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) \lesssim C_{\text{概率}}^{\frac{\rho}{1+\rho}} \cdot \varepsilon^{\frac{2\rho}{1+\rho}}$。

接下来，我们将探讨受约束变体和偏移变体的 DEC 在何种情况下、以及在多大程度上能够以与 式 (36)相反的方向建立联系——这对遗憾和 PAC 均适用。

4.1.1 PAC 决策估计复杂度：受约束与偏移

对于 PAC 设定，以下结果表明，在大多数参数范围内，受约束 DEC 与偏移 DEC 除了对数因子外是等价的。

命题 4.2。 对于所有 $\varepsilon > 0$ 和 $M \in \overline{\mathcal{M}}^{+}$，有

$$\text{p-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \le \inf_{\gamma \ge 0} \left\{ \text{p-dec}^{\text{o}}_{\gamma}(\mathcal{M}, \overline{M}) \vee 0 + \gamma \varepsilon^2 \right\}. \tag{38}$$

另一方面，对于所有 $\gamma \ge 1$ 和 $M \in \overline{\mathcal{M}}^{+}$，令 $L := 2\lceil \log 2\gamma \rceil$，有

$$\text{p-dec}^{\text{o}}_{\gamma \cdot (4L+1)}(\mathcal{M}, \overline{M}) \le \frac{2}{\gamma} + \sup_{\varepsilon > 0} \left\{ \text{p-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) - \frac{\gamma \varepsilon^2}{4} \right\}. \tag{39}$$

只要 $\text{p-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) \propto \varepsilon^{\rho}$ 对于 $\rho \le 1$ 成立，那么根据公式 (38) 转换为偏移 DEC，再利用公式 (39) 转回受约束 DEC，只会损失对数因子。然而，在受约束 DEC 呈现 $\text{p-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) \propto C_{\text{prob}} \cdot \varepsilon^2$ 或 $\text{p-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) \propto \mathbb{I}\{\varepsilon \ge 1/\sqrt{C_{\text{prob}}}\}$ 形式的“快速”行为时，这一过程会造成损失（由于公式 (39) 中的 $\frac{1}{\gamma}$ 项），从而破坏获得快于 $\sqrt{T}$ 速率的可能性。这就是我们以受约束 DEC 的形式给出 PAC 结果，并采用针对受约束 DEC 专门设计的算法（算法 1），而不是直接改编 Foster 等人（2021）中基于偏移 DEC 的算法的原因；采取后一种方法并结合公式 (39) 不会得到快速速率。

4.1.2 遗憾 DEC：约束与偏移

鉴于 PAC 设定下约束 DEC 与偏移 DEC 近乎等价，人们或许会希望遗憾情形也存在类似的等价关系。然而，直接套用用于证明命题 4.2 的技巧，只能得到下面这个在定量上较弱的命题 4.1的逆命题。

命题 4.3。 对于所有 $\gamma > 0$ 和 $M \in \mathcal{M}^{\mp}$，

$$\text{r-dec}^{\text{o}}_{\gamma}(\mathcal{M}, \overline{M}) \le \text{r-dec}^{\text{c}}_{\gamma^{-1/2}}(\mathcal{M}, \overline{M}). \tag{40}$$

公式（40）中对偏移 DEC 的界并不令人满意，因为右侧存在尺度 $\varepsilon = \gamma^{-1/2}$。例如，在具有 A 个动作的多臂老虎机问题中，我们有 $\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) \propto \varepsilon\sqrt{A}$，并且

[4]将式（37）右侧替换为$\text{r-dec}^{\text{o}}_{\gamma}(\mathcal{M} \cup \{M\}, M)$的不等式，可通过将式（36）应用于类别 $\mathcal{M} \cup \{M\}$立即得到。不等式（37）更强，证明也更复杂。

$\text{r-dec}^{\text{o}}_{\gamma}(\mathcal{M}) \propto \frac{A}{\gamma}$，然而 公式（40）只能得到如下不等式：$\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) \lesssim \sqrt{\frac{A}{\gamma}}$。将其代入 Foster 等人的上界中，会导致次优的$A^{1/3}T^{2/3}$-型遗憾界（2021年）（参见公式（59））。遗憾的是，以下结果表明，命题 4.3一般而言是紧的，即使在 $M \in \mathcal{M}$的情况下也是如此。

命题 4.4. 对于所有 $\gamma \geq 1$，都存在一个模型类 $\mathcal{M} \subseteq \mathcal{M}^{+}$，使得对于所有 $\varepsilon \in (0, 1)$，

$$\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) = \sup_{\overline{M} \in \text{co}(\mathcal{M})} \text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \leq O\left(\varepsilon^2 \gamma^{1/2}\right),$$

因此，特别地，对于 $\varepsilon \leq \gamma^{-1/2}$，有 $\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) \leq O(\varepsilon)$。然而，存在 $\overline{M} \in \mathcal{M}$，使得

$$\text{r-dec}^{\text{o}}_{\gamma}(\mathcal{M}, \overline{M}) \geq \Omega\left(\gamma^{-1/2}\right).$$

这排除了存在比 公式（40）更紧的不等式的可能性，并表明约束与偏移 DEC 在遗憾方面具有本质不同的行为。

4.2 局部化

对于尺度参数 $\alpha > 0$ 和参考模型 $M \in \mathcal{M}^{+}$，定义 $\mathcal{M}$ 的以下局部化 子类：

$$\mathcal{M}_{\alpha}(\overline{M}) := \{M \in \mathcal{M} \mid f^{M}(\pi_{M}) \leq f^{\overline{M}}(\pi_{\overline{M}}) + \alpha\}. \tag{41}$$

关于遗憾的最紧上界和下界，在Foster 等人（2021年）中以局部化类别(41)的偏移 DEC 表述，其中 $\alpha > 0$ 的选择恰当，且取决于 T（精确表述见第 5 节）。基于受约束 DEC 得出的界避免了显式使用局部化，但接下来我们将说明，受约束 DEC会隐式地施加某种形式的局部化。

命题 4.5（PAC DEC 的局部化）。对于所有 $\varepsilon > 0$ 和 $M \in \mathcal{M}^{+}$，令 $\bar{\alpha}(\varepsilon) := \sqrt{3}\varepsilon + \text{p-dec}^{\text{c}}_{\sqrt{6}\varepsilon}(\mathcal{M}, \overline{M})$，则有

$$\text{p-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \leq \text{p-dec}^{\text{c}}_{\sqrt{3}\varepsilon}(\mathcal{M}_{\bar{\alpha}(\varepsilon)}(\overline{M}), \overline{M}).$$

对于遗憾也有类似的结果，但我们要求 DEC 满足比假设 3.3中的正则性条件略强的版本。

定义 4.1（DEC 的强正则性）。对于 $M \in \mathcal{M}^{+}$，如果存在常数 $C_{\text{reg}} \geq \sqrt{2}$ 和 $c_{\text{reg}} < C_{\text{reg}}$，使得对所有 $\varepsilon > 0$，受约束 *DEC* 相对于 M 满足强正则性条件，则称其满足强正则性条件。

$$\text{r-dec}^{\text{c}}_{C_{\text{reg}} \cdot \varepsilon}(\mathcal{M}, \overline{M}) \leq c^2_{\text{reg}} \cdot \text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}, \overline{M}). \tag{42}$$

若对所有 $\varepsilon > 0$ 都成立，则称受约束 *DEC* 相对于类别 $\mathcal{M}' \subseteq \mathcal{M}^{+}$ 满足强正则性：

$$\sup_{\overline{M} \in \mathcal{M}'} \text{r-dec}^{\text{c}}_{C_{\text{reg}} \cdot \varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \leq c^2_{\text{reg}} \cdot \sup_{\overline{M} \in \mathcal{M}'} \text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}). \tag{43}$$

只要 $\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}, M) \propto \varepsilon^{\rho}$ 对 $\rho < 2$ 成立，此条件便在 $C_{\text{reg}} = 2$ 和 $c_{\text{reg}} = 2^{\rho/2}$ 下满足。

命题 4.6 (遗憾决策估计系数的局部化)。给定 $M \in \mathcal{M}^{+}$，并假设强正则性条件 *(42)*相对于 M 成立。那么，对于所有 $\varepsilon > 0$，令 $\bar{\alpha}(\varepsilon) := C_{\text{reg}} \cdot \varepsilon + \text{r-dec}^{\text{c}}_{C_{\text{reg}} \cdot \varepsilon}(\mathcal{M}, M) \leq C^2_{\text{reg}} \cdot \left(\varepsilon + \text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}, M)\right)$，我们有

$$\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \leq C_{\text{loc}} \cdot \text{r-dec}^{\text{c}}_{C_{\text{reg}} \cdot \varepsilon}(\mathcal{M}_{\bar{\alpha}(\varepsilon)}(\overline{M}), \overline{M}),$$

其中 $C_{\text{loc}} := \left(\frac{1}{c^2_{\text{reg}}} - \frac{1}{C^2_{\text{reg}}}\right)^{-1}$。

注意，在 $\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}, M) \propto \varepsilon^{\rho}$ 对于某个常数 $\rho < 2$ 的情况下，选择 $C_{\text{reg}} = 2$ 和 $c_{\text{reg}} = 2^{\rho/2}$ 可得$C_{\text{loc}} = O(1)$。这些结果表明，受约束 DEC——无论针对 PAC 还是遗憾——都等价于（相差常数因子）局部化子类 $\mathcal{M}_{\alpha}(M)$ 的受约束 DEC，其中半径 α 取决于 DEC 本身的值。相较之下，偏移 DEC 不会自动施加任何形式的局部化，这也解释了为什么先前工作需要明确限制在局部化子类上。

4.2.1 受约束 DEC 与偏移 DEC：局部化类别的更紧等价

基于前篇中的见解，我们现在表明，对于局部化类，可以用受约束 DEC 以更紧致的方式界定遗憾对应的偏移 DEC，从而改进 命题 4.3。

命题 4.7。 给定 α、$\gamma>0$ 和 $M\in\overline{\mathcal{M}}^{+}$。对于所有 $\varepsilon>0$，我们有

$$\text{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\alpha}(\overline{M})\cup\{\overline{M}\},\overline{M})\le \text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M})+\max\left\{0,\ \alpha+\frac{1}{2\gamma}-\frac{\gamma\varepsilon^{2}}{2}\right\}, \tag{44}$$

这尤其推出

$$\text{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\alpha}(\overline{M})\cup\{\overline{M}\},\overline{M})\le \text{r-dec}^{\mathrm{c}}_{\sqrt{2\alpha/\gamma}}(\mathcal{M}\cup\{\overline{M}\},\overline{M})+\frac{1}{2\gamma}. \tag{45}$$

式（45）中的界$\text{r-dec}^{\mathrm{c}}_{\sqrt{1/\gamma}}(\mathcal{M})$将命题 4.3 中的$\text{r-dec}^{\mathrm{c}}_{\sqrt{\alpha/\gamma}}(\mathcal{M})$替换为 $\alpha\ll 1$时会有所改进。值得注意的是，该界足够强，因此将其与命题 4.6结合后，可以用局部化偏移 DEC 给受约束 DEC 建立上界，再以仅损失常数因子的方式回到受约束 DEC——至少在$\text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M})\gtrsim\varepsilon$的情况下是如此。下述结果运用这一方法，推导出受约束 DEC 与局部化偏移 DEC 的近似等价；我们也在命题 4.11的证明中采用了这一方法。

命题 4.8。只要 *(42)* 中的定义 *4.1*所述强正则性条件对 $M\in\mathcal{M}^{+}$成立，则对所有 $\varepsilon>0$，令 $\alpha(\varepsilon,\gamma):=\gamma\varepsilon^{2}$，

$$c_1\cdot\sup_{\gamma>c_3\varepsilon^{-1}}\text{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{c_2\cdot\alpha(\varepsilon,\gamma)}(\overline{M}),\overline{M})\le\text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M})\le c_1'\cdot\sup_{\gamma>c_3'\varepsilon^{-1}}\text{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{c_2'\cdot\alpha(\varepsilon,\gamma)}(\overline{M}),\overline{M})+c_4'\varepsilon, \tag{46}$$

其中 c_1、c_2、$c_3>0$ 是数值常数，而 c_1'、c_2'、c_3'、$c_4'>0$ 是仅依赖于 C_{reg} 和 c_{reg} 的常数。

鉴于这一结果，我们的上界（定理 3.2）可以理解为对先前工作的改进：通过实现尽可能紧的局部化半径（大致为 $\alpha=O(\gamma\varepsilon^{2})$ 而非 $\alpha=O(\gamma\varepsilon^{2}+\text{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}))$）。第 5 节给出了若干示例，说明这如何带来速率上的定量改进。

4.3 参考模型：次优性的作用

现在，我们来探讨参考模型 M 在决策估计系数定义中的作用。回顾一下，对于遗憾$\overline{\ }$，我们的上界和下界按以下量缩放

$$\begin{aligned}\text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M})&=\sup_{\overline{M}\in\mathrm{co}(\mathcal{M})}\text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M})\\&=\sup_{\overline{M}\in\mathrm{co}(\mathcal{M})}\inf_{p\in\Delta(\Pi)}\sup_{M\in\mathcal{M}\cup\{\overline{M}\}}\left\{\mathbb{E}_{\pi\sim p}[f^{M}(\pi_{M})-f^{M}(\pi)]\mid\mathbb{E}_{\pi\sim p}\left[D^{2}_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right]\le\varepsilon^{2}\right\}.\end{aligned} \tag{47}$$

通过对 $M\in\mathcal{M}\cup\{\overline{M}\}$ 取最大值，该定义迫使最小化方选择 $p\in\Delta(\Pi)$，使得在 M 下的次优性 $\mathbb{E}_{\pi\sim p}\left[f^{M}(\pi_{M})-f^{M}(\pi)\right]$较小。这有些违背直觉，因为 $\overline{M}\in\mathrm{co}(\mathcal{M})$ 不一定属于类别 $\mathcal{M}$，但我们的结果表明，$\text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M})$ 刻画了 $\mathcal{M}$ 的极小极大遗憾。事先看来，人们可能会认为，未考虑在 M 下的次优性的量 $\sup_{M\in\mathrm{co}(\mathcal{M})}\text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M},M)$ 是一种更自然的复杂度度量。接下来，我们将说明，$\overline{\ }$该量的行为与式 (47) 有根本不同，而且在 M 下纳入次优性对于刻画极小极大遗憾至关重要。

命题 4.9。 对于任意足够小的 $\varepsilon>0$，都存在一个模型类 $\mathcal{M}$，使得

$$\sup_{\overline{M}\in\mathrm{co}(\mathcal{M})}\text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M},\overline{M})\le c\cdot\varepsilon, \tag{48}$$

yet

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) = \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M}) \geq c'\cdot\varepsilon^{2/3}, \tag{49}$$

其中 c、$c'>0$ 是数值常数。

容易证明，对于选择 $\varepsilon=\varepsilon(T)\propto 1/\sqrt{T}$，命题 4.9中类别的最优遗憾为 $\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)]=\widetilde{\Theta}(T^{2/3})$。定理 2.2给出了这一结果，其尺度取决于 式 (49)中的量。然而，式 (48) 中的量错误地暗示了 $\sqrt{T}$ 型速率，而这种速率无法达到。

对于偏移 DEC，在 M 下次优性的作用更加微妙。可以证明，一般而言，$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}\cup\{M\},M) \gg \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},M)$，这与命题 4.9类似；但命题 4.1表明，后一个量足以作为$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{M\},M)$ 的上界。

虽然前面的讨论表明，要获得遗憾的紧确保证，必须考虑 $M\notin\mathcal{M}$ 下的次优性，但下面的结果表明，对于 PAC 而言，这一区别基本无关紧要，并由此引出定义 $\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) = \sup_{M\in\mathrm{co}(\mathcal{M})}\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M},M)$。

命题 4.10。对于所有 $M\in\mathcal{M}^{+}$ 以及 $\varepsilon>0$,

$$\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M}) \leq \textsf{p-dec}^{\mathrm{c}}_{\sqrt{3}\varepsilon}(\mathcal{M},\overline{M}) + 4\varepsilon. \tag{50}$$

4.4 参考模型：凸性与随机化的作用

现在，我们着重理解非适当参考模型 $M\notin\mathcal{M}$的作用。着眼于遗憾，我们的上界（定理 3.2）随

$$\textsf{r-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M}) = \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \textsf{r-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M}\cup\{\overline{M}\},\overline{M}), \tag{51}$$

在凸包中的所有可能参考模型上取最大值$\mathrm{co}(\mathcal{M})$。另一方面，我们的下界（定理 2.2）随

$$\sup_{\overline{M}\in\mathcal{M}^{+}} \textsf{r-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M}\cup\{\overline{M}\},\overline{M}) \geq \textsf{r-dec}^{\mathrm{c}}_{\bar{\varepsilon}(T)}(\mathcal{M}). \tag{52}$$

两个结果都允许使用非适当模型 $M\notin\mathcal{M}$，但量(52) 允许参考模型不受约束，因此其值先验上可能大于量(51)。这为何并不矛盾？接下来我们将证明，对于约束与偏移 DEC，像式（52）那样允许任意、不受约束的参考模型，只会使其值超过 $M\in\mathrm{co}(\mathcal{M})$ 所达到的值，且至多相差常数因子。

在陈述我们的结果之前，我们先提及一个相关的次要目标，即理解随机参考模型的作用。Foster 等人 (2021年) 针对随机（或混合）参考模型引入了决策估计系数的一种变体，其中 M 从分布 $\nu\in\Delta(\mathcal{M})$中抽取。我们将此复杂度度量的约束与偏移变体定义为 $\nu\in\Delta(\mathcal{M})$ 如下：

$$\textsf{r-dec}^{\mathrm{c},\mathrm{rnd}}_{\varepsilon}(\mathcal{M},\nu) = \inf_{p\in\Delta(\Pi)}\sup_{M\in\mathcal{M}}\left\{\mathbb{E}_{\pi\sim p}[f^{M}(\pi_{M})-f^{M}(\pi)] \mid \mathbb{E}_{\overline{M}\sim\nu}\mathbb{E}_{\pi\sim p}\left[D^{2}_{\mathsf{H}}\left(M(\pi),\overline{M}(\pi)\right)\right]\leq\varepsilon^{2}\right\}, \tag{53}$$

$$\textsf{r-dec}^{\mathrm{o},\mathrm{rnd}}_{\gamma}(\mathcal{M},\nu) = \inf_{p\in\Delta(\Pi)}\sup_{M\in\mathcal{M}}\mathbb{E}_{\pi\sim p}\left[f^{M}(\pi_{M})-f^{M}(\pi)-\gamma\cdot\mathbb{E}_{\overline{M}\sim\nu}\left[D^{2}_{\mathsf{H}}\left(M(\pi),\overline{M}(\pi)\right)\right]\right]. \tag{54}$$

Chen 等人 (2022) 的最新研究将 Foster 等人 (2021年) 的结果推广，给出了随 $\sup_{\nu\in\Delta(\mathcal{M})}\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\nu)$ 缩放的遗憾界，人们可能会希望它比 $\sup_{M\in\mathrm{co}(\mathcal{M})}\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},M)$ 更小（由于詹森不等式，它绝不会更大）。我们证明事实并非如此：无论是受约束情形还是偏移情形，随机化 DEC 都介于带有 $M\in\mathcal{M}^{+}$ 的 DEC 与带有 $M\in\mathrm{co}(\mathcal{M})$ 的 DEC 之间。

命题 4.11。 假设 假设 *A.1* 成立。对所有 $\gamma > 0$，我们有 ve

$$\sup_{\overline{M}\in\mathcal{M}^+} \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\overline{M}) \le \sup_{\nu\in\Delta(\mathcal{M})} \textsf{r-dec}^{\mathrm{o},\mathrm{rnd}}_{\gamma/4}(\mathcal{M},\nu) \le \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \textsf{r-dec}^{\mathrm{o}}_{\gamma/4}(\mathcal{M},\overline{M}). \tag{55}$$

此外，假设相对于 定义 *4.1*、式 *(43)*，强正则性条件成立。则对所有 $\mathcal{M}^+$， $\varepsilon > 0$，都有

$$\sup_{\overline{M}\in\mathcal{M}^+} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M}) \le c_1 \sup_{\nu\in\Delta(\mathcal{M})} \textsf{r-dec}^{\mathrm{c},\mathrm{rnd}}_{c_2\varepsilon}(\mathcal{M}\cup\{\overline{M}_\nu\},\nu) + c_3\varepsilon \tag{56}$$

$$\le c_1 \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \textsf{r-dec}^{\mathrm{c}}_{c_2\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M}) + c_3\varepsilon, \tag{57}$$

whe对于 $M_\nu \colon = \mathbb{E}_{M'\sim\nu}[M']$， c_1、 c_2、 $c_3 > 0$是仅取决于 C_{reg}、 $C_{\mathrm{loc}} >$的常数 $0.$

PAC 也有类似的等价关系；参见附录 D.2.2。该结果的主要推论如下。

- 由于允许任意参考模型 $M \in \mathcal{M}^+$ 不会使参考模型$M \in$co($\mathcal{M}$)上的值增大，因此无论是推导上界还是下界，都可以自由采用更方便的版本。

- 从统计视角来看，采用随机化估计量的 DEC 无法进一步收紧我们的结果，因为这种复杂度度量与采用 $M \in$co($\mathcal{M}$)的变体相比，至多只会大出常数因子。

顺便提一下，偏移 DEC 的等价关系 (55) 的证明直接来自极小极大定理以及平方海林格距离的凸性，但等价关系 (56) 的证明相当复杂，并且运用了第 4.2 节中建立的工具，以便在约束与偏移 DEC 之间来回转换。我们想知道是否存在更简单的证明。

5 相较先前工作的改进

在本节中，我们将运用第 4 节中开发的工具，证明定理 2.2和3.2 中的遗憾界始终优于先前工作中的结果（Foster 等人，2021年、2022b）。随后，我们将重点介绍一些能从我们的界中获得显著改进的具体模型类，并讨论其他相关工作。

先前工作的遗憾界。回顾一下，对于模型类 $\mathcal{M}$ 和参考模型 $M \in \mathcal{M}$，我们将 M 附近的局部化子奚定义为

$$\mathcal{M}_\alpha(\overline{M}) = \left\{M \in \mathcal{M} : f^{\overline{M}}(\pi_{\overline{M}}) \ge f^{M}(\pi_M) - \alpha\right\}. \tag{58}$$

其中 $\alpha > 0$ 是半径。为简单起见，聚焦于有限类，先前工作中最好的上界来自Foster 等人(2021年，其形式为

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \le \widetilde{O}(1)\cdot \min_{\gamma>0}\max\left\{\sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\underline{\Delta}(\gamma,T)}(\overline{M}),\overline{M})\cdot T,\ \gamma\cdot\log|\mathcal{M}|\right\}, \tag{59}$$

对于 $\Delta(\gamma,\ T) = \widetilde{O}\big(\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}) + \frac{\gamma}{T}\log|\mathcal{M}| + \gamma^{-1}\big)$。先前工作中最好的下界来自Foster 等人(2022b，定理 D.1)，适用于所有具有“次切比雪夫”尾部行为的算法，[5] 并且其量级为

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \ge \Omega(1)\cdot \max_{\gamma>\sqrt{C(T)T}} \sup_{\overline{M}\in\mathcal{M}} \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\underline{\Delta}(\gamma,T)}(\overline{M}),\overline{M})\cdot T, \tag{60}$$

其中 $C(T) := O(\log(T \wedge V(\mathcal{M})))$，且 $\underline{\Delta}(\gamma,T) := C(T)^{-1}\cdot\frac{\gamma}{T}$。

[5]次切比雪夫算法是指其均方根遗憾与期望遗憾处于同一数量级的算法。Foster 等人(2021年) 给出了不要求次切比雪夫尾部行为假设的下界，但这些结果依赖于如下形式的较小子类的 DEC： $\mathcal{M}^{\infty}_{\underline{\alpha}}(M) = \left\{M \in \mathcal{M} :\ |g^M(\pi) - g^M(\pi)| \le \alpha\ \ \forall\pi \in \Pi\right\}$，与公式（60）相比可能较松。

我们的改进。以下结果可由命题 4.8立即推出，这意味着在满足适当的正则性条件时，定理 3.2 和定理 2.2 中的上界和下界分别总是比公式（59） 和公式（60）中的保证更紧。

推论 5.1。 只要 定义 4.1中的强正则性条件对 $M \in \mathcal{M}^+$ 满足C_{loc}， $C_{\mathrm{reg}} = O(1)$，则对所有 $\overline{\varepsilon} > 0$ 和 $\gamma > 0$，

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \le O\big(\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\overline{\alpha}(\varepsilon,\gamma)}(\overline{M}), \overline{M}) \vee 0 + \gamma\varepsilon^2 + \varepsilon\big), \tag{61}$$

其中 $\overline{\alpha}(\varepsilon, \gamma) = O\big(\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \vee 0 + \gamma\varepsilon^2 + \gamma^{-1}\big)$。此外，对所有 $\varepsilon > 0$、 $\gamma \ge \Omega(\varepsilon^{-1})$ 和 $M \in \overline{\mathcal{M}}^+$，

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \ge \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\underline{\alpha}(\varepsilon,\gamma)}(\overline{M}), \overline{M}), \tag{62}$$

其中 $\underline{\alpha}(\varepsilon, \gamma) = \Omega(\gamma\varepsilon^2)$。

令 $\varepsilon(T) = \widetilde{O}\Big(\sqrt{\log_T^{|\mathcal{M}|}}\Big)$，应用公式（61）可得，定理 3.2 中的上界始终不超过公式（59）中的量，二者相差至多 T 和 $1/\delta$ 的对数因子。类似地，令 $\varepsilon(T) = \widetilde{\Omega}\Big(\sqrt{\frac{1}{T}}\Big)$，应用公式（62）可知，定理 2.2 中的下界始终不小于公式（60）中的量，二者相差至多 $\log(T)$ 因子和一个加性 $O(\sqrt{T})$ 项。定理 2.2 不仅随更大的复杂度度量而缩放，还具有以下特点：1）适用于任意算法，无需 Foster 等人（2022b）所用的次切比雪夫假设；2）允许使用非适当参考模型 $M \notin \mathcal{M}$。

现在，我们给出一些具体的模型类，在这些模型类中，我们的结果能够在速率上带来定量改进。第一个例子是一个模型类，其中我们的主要上界（定理 3.2）通过（隐式地）实现比 Foster 等人(2021年) 更紧的局部化半径，从而优于公式（59）。

例 5.1 （上界带来的改进）。 考虑一个由 $\mathcal{M}^{\alpha,\beta}$ 参数化的模型类，其中参数为 $\alpha \in (0, 1/2], \beta \in (0, 1)$，以及 $A \in \mathbb{N}$。

1. $\Pi = [A] \cup \{\pi_\circ\}$，其中 $\pi_\circ$ 是一个“揭示性”决策。
2. $\mathcal{O} = [A] \cup \{\perp\}$，其中 $\perp$ 是一个空符号。
3. 我们有 $\mathcal{M} = \{M_{\alpha,i}\}_{i\in[A]} \cup \{\widetilde{M}\}$。对于每个 $i \in [A]$，模型 $M_{\alpha,i} \in \mathcal{M}^{\alpha,\beta}$的奖励和观测定义如下：

   (a) 对于 $\pi \in [A]$, $f^{M_{\alpha,i}}(\pi) = \frac{1}{2} + \alpha \cdot \mathbb{1}\{\pi = i\}$，以及 $f^{M_{\alpha,i}}(\pi_\circ) = 0$。所有 $\pi \in \Pi$ 在 $r \sim M_{\alpha,i}(\pi)$ 下几乎必然满足 $r = f^{M_{\alpha,i}}(\pi)$。(b) 对于 $\pi \in [A]$，我们收到观测 $o = \perp$。选择 $\pi_\circ$ 后，以概率 β 得到观测 $o = i \in [A]$，以概率 $1 - \beta$ 得到 $o = \perp$。

4. 模型 M 定义如下：

   (a) 对所有 $\pi \in [A]$，我们有 $f^{\widetilde{M}}(\pi) = \frac{1}{2}$ 和 $f^{\widetilde{M}}(\pi_\circ) = 0$；对于所有 $\pi \in \Pi$，在 $r \sim \widetilde{M}(\pi)$ 下，$r = f^{\widetilde{M}}(\pi)$ 几乎必然成立。(b) 所有 $\pi \in [A]$ 几乎必然都具有 $o = \perp$。对于 $\pi_\circ$，我们以概率 $1 - \beta$ 观测到 $o = \perp$，并以概率 β 观测到$o \sim$均匀分布$([A])$。

令 $\mathcal{M} := \mathcal{M}^{\alpha_1,\beta} \cup \mathcal{M}^{\alpha_2,\beta}$，其中 $\alpha_1 = 1/2$、 $\alpha_2 \propto T^{-1/4}$、 $\beta \propto T^{-1/2}$ 和 $A \propto T^2$。则：

- 根据定理 3.2，E2D$^+$ 算法可达到 $\mathbb{E}[\mathbf{Reg}_{\mathrm{DM}}(T)] \le \widetilde{O}(\sqrt{T})$。
- 遗憾界在 公式（59） 按 $\widetilde{\Omega}(T^{5/8})$。

◁

下一个例子展示了一个模型类：对于该类，允许使用非适当参考模型后，我们的主要下界（定理 2.2）优于 Foster 等人(2021年)的结果 $M \notin \mathcal{M}$。

例5.2（下界的改进）。设 $A \in \mathbb{N}$ 和 $\Pi = \{1, \dots, A\}$。考虑由以下形式的模型组成的多臂老虎机模型类 $\mathcal{M} = \{M_1, \dots, M_A\}$

$$M_i(\pi) = \mathrm{Ber}(f_i(\pi)),$$

其中 $f_i(\pi) := \frac{1}{2} + \Delta \mathbb{I}\{\pi = i\}$。Foster 等人（2021年）证明，无论如何选择 $\Delta > 0$，对于所有 $\gamma > 0$ 和 $M \in \mathcal{M}$，$\text{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, M) \leq \frac{1}{\gamma}$，因此下界 (60) 至多只能给出 $\mathbb{E}[\mathrm{Reg}_{\mathrm{DM}}(T)] \geq \Omega(\sqrt{T})$。另一方面，选择 $\bar{M}(\pi) = \mathrm{Ber}(\frac{1}{2})$，它具有 $\bar{M} \notin \mathcal{M}$，很容易看出，只要 $\Delta \propto \varepsilon\sqrt{A}$，就有 $\text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \geq \text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \bar{M}) \geq \Omega(\varepsilon\sqrt{A})$。令 $\Delta \propto \varepsilon(T) \cdot \sqrt{A}$，由定理 2.2 可得

$$\mathbb{E}[\mathbf{Reg}_{\mathrm{DM}}(T)] \geq \widetilde{\Omega}(\sqrt{AT}),$$

这是最优的。这表明，一般而言，允许使用不适当参考模型 $M \notin \mathcal{M}$ 是获得紧下界所必需的。◁

5.1 其他相关工作

Chen 等人的并行研究(2022) 独立发现了 PAC 决策估计系数的偏移变体，并通过调整 Foster 等人(2021年的方法，为 PAC 样本复杂度给出了上界和下界。我们对遗憾和 PAC 的保证始终比这些结果更紧，这与我们相较于 Foster 等人(2021年) 所取得的改进类似（另见 第 4.4 节），但除此之外，我们的方法与之互为补充。

6 其他示例

最后，我们通过一些简要示例，展示受约束决策估计系数以及我们的上界和下界在一些值得关注的标准模型类中的表现。对于遗憾，Foster 等人(2021年) 给出了多臂老虎机和强化学习中若干典型模型的（局部）偏移 DEC 下界。结合这些结果与 推论 5.1，即可直接推导出受约束 DEC 的下界。同样，Foster 等人(2021年) 给出了相同示例的偏移 DEC 全局上界，通过 命题 4.1 即可立即得到受约束 DEC 的上界。这种方法为 Foster 等人(2021年) 考虑的所有示例中的受约束决策估计系数给出了上界和下界。我们在 表1中总结了这些结果，以及由此推得的遗憾下界；遗憾上界与之类似，但还取决于 **估计**$_{\mathsf{H}}(T, \delta)$。更多背景信息请参见 Foster 等人(2021年)。

设定	$\text{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M})$	下界（定理 2.2）
多臂老虎机	$\varepsilon\sqrt{A}$	$\sqrt{AT}$
有间隙的多臂老虎机	$\Delta\mathbb{I}\{\varepsilon > \Delta/\sqrt{A}\}$	A/Δ
线性老虎机	$\varepsilon\sqrt{d}$	$\sqrt{dT}$
利普希茨老虎机	$\varepsilon^{1-\frac{d}{d+2}}$	$T^{\frac{d+1}{d+2}}$
ReLU老虎机	$\mathbb{I}\{\varepsilon > 2^{-\Omega(d)}\}$	$2^{\Omega(d)}$
表格型强化学习	$\varepsilon\sqrt{HSA}$	$\sqrt{HSAT}$
线性马尔可夫决策过程	$\varepsilon\sqrt{d}$	$\sqrt{dT}$
线性 $Q^\star$强化学习	$\mathbb{I}\{\varepsilon \geq 2^{-\Omega(d)} \vee 2^{-\Omega(H)}\}$	$2^{\Omega(d)} \wedge 2^{\Omega(H)}$
线性 $Q^\star$确定性强化学习	$\mathbb{I}\{\varepsilon \geq 1/\sqrt{d}\}$	d

表1：通过受约束决策估计系数恢复的多臂老虎机和强化学习下界，其中 A = #动作、 Δ = 间隙、 d = 特征维度、 H = 回合时域以及S = #状态。数值常数和 log(T) 因子均已省略。

示例：多臂老虎机。现在，我们将更详细地概述上述下界方法，并以多臂老虎机为具体例子。Foster 等人（2021年）证明，当 $\mathcal{M}$ 是

由所有具有 $\Pi = \{1, \dots, A\}$ 且奖励服从伯努利分布的多臂老虎机实例组成的类别中，存在 $M \in \mathcal{M}$，使得对所有 $\gamma \geq c_1 \cdot A$，

$$\sup_{\overline{M} \in \mathcal{M}} \text{r-dec}^{\text{o}}_{\gamma}(\mathcal{M}_{\alpha_\gamma}(\overline{M}), \overline{M}) \geq c_2 \cdot \frac{A}{\gamma},$$

其中 $\alpha_\gamma := c_3 \cdot \frac{A}{\gamma}$，且 c_1，c_2，$c_3 > 0$ 均为数值常数。推论 5.1 意味着，对于所有 $\varepsilon > 0$ 和 $M \in \mathcal{M}^+$，

$$\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) \geq \sup_{\gamma > 0} \text{r-dec}^{\text{o}}_{\gamma}(\mathcal{M}_{\underline{\alpha}(\varepsilon,\gamma)}(\overline{M}), \overline{M}),$$

其中 $\underline{\alpha}(\varepsilon, \gamma) = c \cdot \gamma \varepsilon^2$，这里的数值常数 c 足够小。对于任意给定的 $\varepsilon > 0$，若我们令 $\gamma = c' \cdot A^{1/2}/\varepsilon$，其中常数 c' 足够大，则有 $\mathcal{M}_{\alpha_\gamma}(M) \subseteq \mathcal{M}_{\underline{\alpha}(\varepsilon,\gamma)}(M)$，由此可得

$$\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) \geq \Omega\left(\varepsilon\sqrt{A}\right)$$

对于所有 $\varepsilon \leq c'' \cdot A^{-1/2}$，其中 c'' 是一个足够小的常数。将这个 DEC 下界代入定理 2.2 可得如下形式的遗憾下界：$\mathbb{E}[\textbf{遗憾}_{\text{DM}}(T)] \geq \widetilde{\Omega}(\sqrt{AT})$。

致谢

感谢 Jian Qian、Sasha Rakhlin、Rob Schapire 和 Andrew Wagenmaker 提供的宝贵意见并参与讨论。

参考文献

Jacob Abernethy、Elad Hazan 和 Alexander Rakhlin。《暗中竞争：一种高效的老虎机线性优化算法》。发表于第 *21* 届学习理论年会（*COLT*）论文集，2008年。Alex Ayoub、Zeyu Jia、Csaba Szepesvari、Mengdi Wang 和 Lin Yang。《利用价值目标回归进行基于模型的强化学习》。发表于国际机器学习会议，第 463–474 页。PMLR，2020年。Sébastien Bubeck、Rémi Munos、Gilles Stoltz 和 Csaba Szepesvári。《X 臂老虎机》。机器学习研究期刊，12(5)，2011年。Sébastien Bubeck、Nicolo Cesa-Bianchi 和 Sham M Kakade。《迈向具有老虎机反馈的在线线性优化极小极大策略》。发表于学习理论会议，第 41–1 页。JMLR 研讨会与会议论文集，2012年。Sébastien Bubeck、Yin Tat Lee 和 Ronen Eldan。《用于老虎机凸优化的基于核的方法》。发表于第 *49* 届 *ACM SIGACT* 计算理论年度研讨会论文集，第 72–85 页，2017年。Fan Chen、Song Mei 和 Yu Bai。《具有决策估计系数的强化学习统一算法：无遗憾、PAC 与无奖励学习》。*arXiv* 预印本 *arXiv:2209.11745*，2022年。Varsha Dani、Thomas P Hayes 和 Sham Kakade。《在线优化中老虎机信息的代价》。2007年。Sarah Dean、Horia Mania、Nikolai Matni、Benjamin Recht 和 Stephen Tu。《线性二次调节器的样本复杂度》。计算数学基础，20(4)：633–679，2020年。Shi Dong、Benjamin Van Roy 和 Zhengyuan Zhou。《利用聚合状态实现可证明高效的强化学习》。*arXiv* 预印本 *arXiv:1912.06366*，2019年。David L Donoho 和 Richard C Liu。《收敛速率的几何化》。统计年鉴，1987年。David L Donoho 和 Richard C Liu。《收敛速率的几何化，II》。统计年鉴，第 633–667 页，1991a。David L Donoho 和 Richard C Liu。《收敛速率的几何化，III》。统计年鉴，第 668–701 页，1991b。Simon Du、Akshay Krishnamurthy、Nan Jiang、Alekh Agarwal、Miroslav Dudik 和 John Langford。《通过潜在状态解码，利用丰富观测实现可证明高效的强化学习》。发表于国际机器学习会议，第 1665–1674 页。PMLR，2019年。Simon S Du、Sham M Kakade、Jason D Lee、Shachar Lovett、Gaurav Mahajan、Wen Sun 和 Ruosong Wang。《双线性类：用于可证明泛化的结构框架》。国际机器学习会议，2021年。Richard Durrett。《概率论：理论与例题》。Duxbury Press，加利福尼亚州贝尔蒙特，第五版，2019年。ISBN 0-534-24318-5。Abraham D Flaxman、Adam Tauman Kalai 和 H Brendan McMahan。《老虎机设定下的在线凸优化：无梯度的梯度下降》。发表于第十六届 *ACM-SIAM* 离散算法研讨会论文集，第 385–394 页，2005年。Dylan J Foster、Sham M Kakade、Jian Qian 和 Alexander Rakhlin。《交互式决策的统计复杂度》。*arXiv* 预印本 *arXiv:2112.13487*，2021年。Dylan J Foster、Noah Golowich、Jian Qian、Alexander Rakhlin 和 Ayush Sekhari。《关于具有决策估计系数的无模型强化学习的注记》。*arXiv* 预印本 *arXiv:2211.14250*，2022a。Dylan J Foster、Alexander Rakhlin、Ayush Sekhari 和 Karthik Sridharan。《关于对抗式决策的复杂度》。*arXiv* 预印本 *arXiv:2206.13063*，2022b。

Nan Jiang、Akshay Krishnamurthy、Alekh Agarwal、John Langford 和 Robert E Schapire。低 Bellman 秩的上下文决策过程是 PAC 可学习的。载于国际机器学习会议，第 1704–1713 页，2017年。Chi Jin、Zhuoran Yang、Zhaoran Wang 和 Michael I Jordan。采用线性函数近似的强化学习具有可证明的高效性。载于学习理论会议，第 2137–2143 页，2020年。Chi Jin、Qinghua Liu 和 Sobhan Miryoosefi。Bellman eluder 维度：强化学习问题的全新丰富类别及样本高效算法。神经信息处理系统，2021年。Robert Kleinberg。连续臂老虎机问题的近乎紧确界。神经信息处理系统进展，17：697–704，2004年。Akshay Krishnamurthy、Alekh Agarwal 和 John Langford。基于丰富观测的 PAC 强化学习。载于神经信息处理系统进展，第 1840–1848 页，2016年。Tor Lattimore。改进零阶对抗老虎机凸优化的遗憾界。数理统计与学习，2(3)：311–334，2020年。Lihong Li。计算强化学习理论的统一框架。罗格斯大学新泽西州立大学新不伦瑞克分校，2009年。Timothy P Lillicrap、Jonathan J Hunt、Alexander Pritzel、Nicolas Heess、Tom Erez、Yuval Tassa、David Silver 和 Daan Wierstra。利用深度强化学习实现连续控制。*arXiv* 预印本 *arXiv:1509.02971*，2015年。Stefan Magureanu、Richard Combes 和 Alexandre Proutiere。Lipschitz 多臂老虎机：遗憾下界与最优算法。载于学习理论会议，第 975–999 页。PMLR，2014年。Volodymyr Mnih、Koray Kavukcuoglu、David Silver、Andrei A Rusu、Joel Veness、Marc G Bellemare、Alex Graves、Martin Riedmiller、Andreas K Fidjeland、Georg Ostrovski 等。通过深度强化学习实现人类水平的控制。《自然》，518(7540)：529，2015年。Aditya Modi、Nan Jiang、Ambuj Tewari 和 Satinder Singh。采用线性组合模型集成的强化学习样本复杂度。载于国际人工智能与统计学会议，第 2010–2020 页。PMLR，2020年。Daniel Russo 和 Benjamin Van Roy。Eluder 维度与乐观探索的样本复杂度。载于神经信息处理系统进展，第 2256–2264 页，2013年。David Silver、Aja Huang、Chris J Maddison、Arthur Guez、Laurent Sifre、George Van Den Driessche、Julian Schrittwieser、Ioannis Antonoglou、Veda Panneershelvam、Marc Lanctot 等。利用深度神经网络和树搜索掌握围棋。《自然》，529(7587)：484，2016年。Wen Sun、Nan Jiang、Akshay Krishnamurthy、Alekh Agarwal 和 John Langford。上下文决策过程中的基于模型的强化学习：PAC 界及相较于无模型方法的指数级改进。载于学习理论会议，第 2898–2933 页。PMLR，2019年。Alexandre B Tsybakov。《非参数估计导论》。Springer 出版公司，股份有限公司，2008年。Ruosong Wang、Russ R Salakhutdinov 和 Lin Yang。采用一般价值函数近似的强化学习：通过有界 eluder 维度实现可证明的高效方法。神经信息处理系统进展，33，2020年。Lin Yang 和 Mengdi Wang。利用线性可加特征进行样本最优参数化 Q 学习。载于国际机器学习会议，第 6995–7004 页。PMLR，2019年。Bin Yu。Assouad、Fano 和 Le Cam。载于《献给 *Lucien Le Cam* 的纪念文集》，第 423–435 页。Springer，1997年。Dongruo Zhou、Quanquan Gu 和 Csaba Szepesvari。针对线性混合马尔可夫决策过程的近乎极小极大最优强化学习。载于学习理论会议，第 4532–4576 页。PMLR，

2021年。

A 预备知识

A.1 极小极大定理

对于某些结构性结果，我们要求偏移决策估计系数（无论是遗憾变体还是 PAC 变体）等于其贝叶斯对应量。只要满足一些温和的拓扑条件，这便是极小极大定理的推论；请注意，通过将我们的目标函数写成

$$\mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) = \inf_{p\in\Delta(\Pi)} \sup_{\mu\in\Delta(\mathcal{M})} \mathbb{E}_{\pi\sim p, M\sim\mu}\left[f^{M}(\pi_{M}) - f^{M}(\pi) - \gamma\cdot D^{2}_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right],$$

因此，调用极小极大定理所需的唯一条件就是紧致性。为避免限定具体的技术条件，我们将其表述为一个假设。

假设 A.1 (极小极大交换)。 对于遗憾决策估计系数，我们有

$$\mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) = \underline{\mathsf{r\text{-}dec}}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) := \sup_{\mu\in\Delta(\mathcal{M})} \inf_{p\in\Delta(\Pi)} \mathbb{E}_{\pi\sim p, M\sim\mu}\left[f^{M}(\pi_{M}) - f^{M}(\pi) - \gamma\cdot D^{2}_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right]. \tag{63}$$

对于 *PAC DEC*，我们有

$$\mathsf{p\text{-}dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) = \underline{\mathsf{p\text{-}dec}}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) := \sup_{\mu\in\Delta(\mathcal{M})} \inf_{p,q\in\Delta(\Pi)} \mathbb{E}_{M\sim\mu}\left[\mathbb{E}_{\pi\sim p}[f^{M}(\pi_{M}) - f^{M}(\pi)] - \gamma\cdot\mathbb{E}_{\pi\sim q}\left[D^{2}_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right]\right]. \tag{64}$$

作为最简单的示例，假设 A.1 在以下情况下成立： $\mathcal{R}$ 有界且 Π 有限（参见 Foster 等人(2021年) 中的命题 4.2），但可以证明，该假设在更一般得多的条件下也成立。

B 第 2 节中省略的证明

B.1 遗憾下界的证明（定理 2.2）

本节中，我们证明定理 2.2。证明分为两部分：

- 在附录 B.1.1中，我们陈述并证明引理 B.1，这是一个与定理 2.2类似的下界，但仅适用于适当的参考模型（具体而言，下界随 $\sup_{M\in\mathcal{M}}\mathsf{r\text{-}dec}_{c\widetilde{\varepsilon}(T)}(\mathcal{M}, M)$ 缩放）。为 M。
- 在附录 B.1.2中，我们证明以下算法结果（引理 B.2）：对于任意类别 $\mathcal{M}$ 以及任意$M \in \mathcal{M}^{+}$（不一定属于 $\mathcal{M}$），如果存在一个算法，其相对于模型类 R 的遗憾至多为 $\mathcal{M}$，那么就存在一个算法，其遗憾至多为 $O(R\cdot\log T)$，且相对于模型类 $\mathcal{M}\cup\{M\}$。然后，我们将此结果与定理 2.2 结合，以证明引理 B.1。

顺带一提，本节的两部分方法也可用于推导 PAC 框架的下界，但我们采用了另一种方法，见第 2.2 节，因为这种方法得出的结果所含的对数因子更少。

B.1.1 适当参考模型的下界 ($M \in \mathcal{M}$)

本节证明引理 B.1，这是一个较弱的下界，与定理 2.2中陈述的下界类似，但将 DEC 替换为一个更小的量，并要求该量满足 $M \in \mathcal{M}$。下文将给出这个较弱版本。

引理 B.1。 令 $\widetilde{\varepsilon}(T) := c_1 \cdot \sqrt{\frac{1}{TC(T)}}$，其中 $c_1 > 0$ 是一个足够小的数值常数。对于所有满足条件的 $T \in \mathbb{N}$，

$$\sup_{\overline{M}\in\mathcal{M}} \mathsf{r\text{-}dec}^{\mathrm{c}}_{\widetilde{\varepsilon}(T)}(\mathcal{M}, \overline{M}) \geq 8\cdot\widetilde{\varepsilon}(T) \tag{65}$$

对于任意遗憾最小化算法，都存在一个属于 $\mathcal{M}$ 的模型，使得在该模型下，

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \geq \Omega(T) \cdot \sup_{\overline{M} \in \mathcal{M}} \mathsf{r\text{-}dec}^{\mathrm{c}}_{\widetilde{\varepsilon}(T)}(\mathcal{M}, \overline{M}). \tag{66}$$

我们指出，条件 (65) 可以放宽：将右侧的常数8替换为任何严格大于 1 的常数。

引理 引理 B.1的证明。 固定所考虑的算法，并令 $\mathbb{P}^M$ 表示当 M 为底层模型时， $\mathfrak{H}^T$ 的诱导分布。令 $\mathbb{E}^M$ 表示相应的期望，并令 $p_M := \mathbb{E}^M\left[\frac{1}{T}\sum_{t=1}^T p^t\right]$。定义 $\varepsilon := \widetilde{\varepsilon}(T) = \sqrt{\frac{c_1}{TC(T)}}$，其中常数 $c_1 > 0$将在下文给出。令 $\overline{M} \in \mathcal{M}$取值为使 $\mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, M)$ 最大的值，并定义 $\delta := \mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, M)$。

将范围限制为在 M 上表现不佳的模型。 若 $\mathbb{E}_{\pi\sim p_M}[g^M(\pi)] \geq \delta/10$，则根据$p_M$的定义，有 $\mathbb{E}^M[\mathbf{Reg}_{\mathsf{DM}}(T)] \geq T \cdot \delta/10$，从而完成引理的证明。因此，接下来我们可以假设 $\mathbb{E}_{\pi\sim p_M}[g^M(\pi)] < \delta/10$。

选择一个替代模型。 定义

$$M = \arg\max_{\overline{M}\in\mathcal{M}} \left\{\mathbb{E}_{\pi\sim p_{\overline{M}}}[f^M(\pi_M) - f^M(\pi)] \mid \mathbb{E}_{\pi\sim p_{\overline{M}}}\left[D^2_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right] \leq \varepsilon^2\right\}, \tag{67}$$

使得

$$\mathbb{E}_{\pi\sim p_{\overline{M}}}[g^M(\pi)] \geq \mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) = \delta. \tag{68}$$

令 $c_2 \in (0,1)$ 为固定值，并定义 $\mathcal{E} = \{\pi\colon g^M(\pi) \geq c_2 \cdot \delta\}$。回顾引理 A.13：根据 Foster 等人(2021年)，我们有

$$D^2_{\mathsf{H}}\left(\mathbb{P}^M, \mathbb{P}^{\overline{M}}\right) \leq C(T) \cdot T \cdot \mathbb{E}_{\pi\sim p_{\overline{M}}}\left[D^2_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right] \leq C(T) \cdot T \cdot \varepsilon^2,$$

这里提醒读者， $C(T) = O(\log(T \wedge V(\mathcal{M})))$。我们在 $\varepsilon = \widetilde{\varepsilon}(T)$ 的定义中选取常数 $c_1 \geq 0$，使其足够小，从而有 $D^2_{\mathsf{H}}\left(\mathbb{P}^M, \mathbb{P}^M\right) \leq 1/100$，因此 $D_{\mathsf{TV}}\left(\mathbb{P}^M, \mathbb{P}^M\right) \leq \frac{1}{10}$。
观测到，根据 $\mathcal{E}$的定义，我们有

$$\begin{aligned}\mathbb{E}_{\pi\sim p_M}[f^M(\pi_M) - f^M(\pi)] &\geq c_2\delta \cdot p_M(\mathcal{E}) \geq c_2\delta \cdot \left(p_{\overline{M}}(\mathcal{E}) - D_{\mathsf{TV}}\left(\mathbb{P}^M, \mathbb{P}^{\overline{M}}\right)\right) \\ &\geq c_2\delta \cdot \left(p_{\overline{M}}(\mathcal{E}) - 1/10\right).\end{aligned} \tag{69}$$

因此，只需证明 $p_M(\mathcal{E})$ 的下界为 $1/2$。

间隙下界。 我们现在计算

$$\begin{aligned}f^M(\pi_M) - f^{\overline{M}}(\pi_{\overline{M}}) &\geq \mathbb{E}_{\pi\sim p_{\overline{M}}}[g^M(\pi) - g^{\overline{M}}(\pi)] - \mathbb{E}_{\pi\sim p_{\overline{M}}}[|f^M(\pi) - f^{\overline{M}}(\pi)|] \\ &\geq \mathbb{E}_{\pi\sim p_{\overline{M}}}[g^M(\pi)] - \frac{\delta}{10} - \varepsilon \\ &\geq \frac{9}{10}\delta - \varepsilon,\end{aligned} \tag{70}$$

其中，第二个不等式利用了以下假设： $\mathbb{E}_{p_M}[g^M] < \delta/10$ 以及引理 D.1，最后一个不等式则利用了式 (68)。

最后，我们总结如下：

$$\begin{aligned} p_{\overline{M}}(\mathcal{E}^{\mathrm{c}}) \cdot \left(\frac{9}{10}\delta - \varepsilon\right) &\leq \mathbb{E}_{\pi \sim p_{\overline{M}}}[\mathbb{1}\{\mathcal{E}^{\mathrm{c}}\} \cdot (f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{\overline{M}}))] \\ &\leq \mathbb{E}_{\pi \sim p_{\overline{M}}}[\mathbb{1}\{\mathcal{E}^{\mathrm{c}}\} \cdot (g^{M}(\pi) - g^{\overline{M}}(\pi))] + \varepsilon \\ &\leq c_2\delta + \varepsilon, \end{aligned}$$

其中，第一个不等式用到了公式（70），第二个不等式用到了引理 D.1，第三个不等式则用到了对所有 $\pi \in \Pi$，$g^{M}(\pi) \geq 0$，以及当 $\overline{\pi} \in \mathcal{E}^{\mathrm{c}}$ 时 $g^{M}(\pi) < c_2\delta$。重新整理后可得，只要 $c_2 \leq 1/8$ 且 $\varepsilon \leq \delta/8$，就有 $p_{M}(\mathcal{E}_1^{\mathrm{c}}) \leq 1/2$。我们对 M 的选择结合增长条件（65），确保确实有 $\varepsilon \leq \delta/8$，因此通过公式（69）可得 $\mathbb{E}_{\pi \sim p_M}[g^{M}(\pi)] \geq \Omega(\delta)$，如所期望的那样。

□

B.1.2 从非适当（$M \in \mathcal{M}^{+}$）化简为适当（$M \in \mathcal{M}$）

本节中，我们考虑模型类和遗憾最小化算法的几种选择。为避免歧义，我们先引入一些额外的记号。回顾一下，在遗憾框架下，针对 T 时间步交互式决策问题的算法由一个序列 $p = (p^1, \ldots\ldots, p^T)$ 指定，其中对每个 $t \in [T]$，p^t 都是从 $(\Omega^{t-1}, \mathscr{F}^{t-1})$ 到 $(\Pi, \mathscr{P})$ 的概率核。给定算法 p，我们用 $\mathbb{P}^{M,\, p}[\cdot]$ 表示当 $M \in \mathcal{M}^{+}$ 是底层模型时，该算法在 $\mathfrak{H}^T$ 上诱导的概率分布，并用 $\mathbb{E}^{M,\, p}[\cdot]$ 表示相应的期望。采用这一记号，当底层模型为 $M \in \mathcal{M}^{+}$ 时，该算法的期望遗憾为 $\mathbb{E}^{M,\, p}[\mathbf{Reg}_{\mathsf{DM}}(T)]$。

下面的引理是本节的主要技术结果。它表明，任何最优值有界的模型 $M \in \mathcal{M}$，都可以加入模型类 $\mathcal{M}$，而不会显著增加极小极大遗憾。

引理 B.2. 固定时间 $T \in \mathbb{N}$ 和模型类 $\mathcal{M}$。设 $M \in \mathcal{M}^{+}$ 是任意模型，满足对所有 $M \in \mathcal{M}$，$f^{M}(\pi_{M}) \leq f^{M}(\pi_{M}) + \delta$，其中某个 $\delta > 0$成立。模型类 $\mathcal{M} \cup \{M\}$ 的极小极大遗憾有如下上界：

$$\inf_{(p')^1, \ldots, (p')^T} \sup_{M^\star \in \mathcal{M} \cup \{\overline{M}\}} \mathbb{E}^{M^\star, p'}[\mathbf{Reg}_{\mathsf{DM}}(T)] \leq C \log T \cdot \inf_{p^1, \ldots, p^T} \sup_{M^\star \in \mathcal{M}} \mathbb{E}^{M^\star, p}[\mathbf{Reg}_{\mathsf{DM}}(T)] + C \cdot (\sqrt{T} + \delta T). \quad (71)$$

其中 $C > 0$ 表示一个通用常数。

在证明 引理 B.2之前，我们先说明它如何推出 定理 2.2。

定理 2.2的证明。固定，并记 () (T)，其中常数 $c_1 \geq 0$按引理 B.1 中的方式选取（特别地，注意 $\sqrt{2}\varepsilon = \widetilde{\varepsilon}(T)$，其中 $\widetilde{\varepsilon}(T)$ 按引理 B.1中的定义给出）。令 $M \in \mathcal{M}^{+}$取值使$\mathsf{r\text{-}dec}_{\varepsilon}^{\mathrm{c}}(\mathcal{M} \cup \{M\}, M)$最大化，因此$\mathsf{r\text{-}dec}_{\varepsilon}^{\mathrm{c}}(\mathcal{M}) = \mathsf{r\text{-}dec}_{\varepsilon}^{\mathrm{c}}(\mathcal{M} \cup \{M\}, M)$。定义

$$\widetilde{\mathcal{M}} := \{M \in \mathcal{M} \cup \{\overline{M}\} \mid f^{M}(\pi_{\overline{M}}) \geq f^{\overline{M}}(\pi_{\overline{M}}) - \sqrt{2}\varepsilon\},$$

使得 $M \in \mathcal{M}$。然后根据 引理 D.2，我们有

$$\mathsf{r\text{-}dec}_{\varepsilon}^{\mathrm{c}}(\mathcal{M}) = \mathsf{r\text{-}dec}_{\varepsilon}^{\mathrm{c}}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \leq \mathsf{r\text{-}dec}_{\sqrt{2}\varepsilon}^{\mathrm{c}}(\widetilde{\mathcal{M}}, \overline{M}) + \sqrt{2}\varepsilon, \quad (72)$$

注意，对于所有 $M \in \widetilde{\mathcal{M}}$，我们都有 $f^{M}(\pi_{M}) \leq f^{M}(\pi_{M}) + \sqrt{2}\varepsilon \leq f^{M}(\pi_{M}) + \sqrt{2}\varepsilon$。因此，通过将引理 B.2 与 $\delta = \sqrt{2}\varepsilon$ 应用于类别 $\widetilde{\mathcal{M}} \backslash \{M\}$，我们可得

$$\begin{aligned} \inf_{(p')^1, \ldots, (p')^T} \sup_{M^\star \in \widetilde{\mathcal{M}}} \mathbb{E}^{M^\star, p'}[\mathbf{Reg}_{\mathsf{DM}}(T)] &\leq C \log T \cdot \inf_{p^1, \ldots, p^T} \sup_{M^\star \in \widetilde{\mathcal{M}} \backslash \{\overline{M}\}} \mathbb{E}^{M^\star, p}[\mathbf{Reg}_{\mathsf{DM}}(T)] + C \cdot (\sqrt{T} + \sqrt{2}\varepsilon T) \\ &\leq C \log T \cdot \inf_{p^1, \ldots, p^T} \sup_{M^\star \in \mathcal{M}} \mathbb{E}^{M^\star, p}[\mathbf{Reg}_{\mathsf{DM}}(T)] + C \cdot (\sqrt{T} + \sqrt{2}\varepsilon T), \quad (73) \end{aligned}$$

其中，第二个不等式由 $\mathcal{M} \backslash \{M\} \subset \mathcal{M}$ 得出。

接下来，我们将应用引理 B.1 于类别 $\widetilde{\mathcal{M}}$。为验证该类别满足公式（65）中的条件，我们注意到，正如上文所述，$\widetilde{\varepsilon}(T) = \sqrt{2}\varepsilon$，因此

$$\sup_{\overline{M}_0 \in \widetilde{\mathcal{M}}} \text{r-dec}^{\text{c}}_{\widetilde{\varepsilon}(T)}(\widetilde{\mathcal{M}}, \overline{M}_0) = \sup_{\overline{M}_0 \in \widetilde{\mathcal{M}}} \text{r-dec}^{\text{c}}_{\sqrt{2}\varepsilon}(\widetilde{\mathcal{M}}, \overline{M}_0) \geq \text{r-dec}^{\text{c}}_{\sqrt{2}\varepsilon}(\widetilde{\mathcal{M}}, \overline{M}) \geq \text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) - \sqrt{2}\varepsilon \geq 8\varepsilon,$$

其中，倒数第二个不等式使用了公式（72），而最后一个不等式使用了公式（15）中的$\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) \geq 10\varepsilon$的假设。引理 B.1 表明，对于某个通用常数 $c_2 > 0$,

$$\begin{aligned}\inf_{(p')^1,\ldots,(p')^T} \sup_{M^\star \in \widetilde{\mathcal{M}}} \mathbb{E}^{M^\star, p'}[\mathbf{Reg}_{\text{DM}}(T)] &\geq c_2 \cdot T \cdot \sup_{\overline{M}_0 \in \widetilde{\mathcal{M}}} \text{r-dec}^{\text{c}}_{\sqrt{2}\varepsilon}(\widetilde{\mathcal{M}}, \overline{M}_0) \\ &\geq c_2 \cdot T \cdot \text{r-dec}^{\text{c}}_{\sqrt{2}\varepsilon}(\widetilde{\mathcal{M}}, \overline{M}) \geq c_2 \cdot T \cdot (\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) - \sqrt{2}\varepsilon), \end{aligned} \tag{74}$$

其中，最后一个不等式使用了公式（72）。结合公式（74）和公式（73）可得

$$\inf_{p^1,\ldots,p^T} \sup_{M^\star \in \mathcal{M}} \mathbb{E}^{M^\star, p}[\mathbf{Reg}_{\text{DM}}(T)] \geq \frac{1}{C \log T} \cdot \left(c_2 \cdot T \cdot (\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) - \sqrt{2}\varepsilon) - C \cdot (\sqrt{T} + \sqrt{2}\varepsilon T) \right),$$

这意味着，对于某些常数 C'，C''，$c_2' > 0$，有

$$\inf_{p^1,\ldots,p^T} \sup_{M^\star \in \mathcal{M}} \mathbb{E}^{M^\star, p}[\mathbf{Reg}_{\text{DM}}(T)] \geq \frac{1}{\log T} \cdot \left(c_2' \cdot T \cdot (\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) - C' \cdot \varepsilon) - C'' \cdot \sqrt{T} \right).$$

只要 $\frac{1}{2} \cdot \text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) > C' \cdot \varepsilon$，就有

$$\inf_{p^1,\ldots,p^T} \sup_{M^\star \in \mathcal{M}} \mathbb{E}^{M^\star, p}[\mathbf{Reg}_{\text{DM}}(T)] \geq \frac{c_2' \cdot T}{2 \log T} \cdot \text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}) - \frac{C'' \cdot \sqrt{T}}{\log T},$$

如所需。

□

最后，我们证明引理 B.2。

引理 B.2 的证明。固定任意算法 $p = (p^1, \ldots, p^T)$。我们在算法 3 中定义一个修改后的算法 $p' = ((p')^1, \ldots, (p')^T)$。粗略地说，$p'$ 会多次运行 p，并在当前运行的平均奖励低于 $f^M(\pi_M)$ 太多时重新初始化 p。如果算法 p' 发现重新初始化 p 的次数超过 $\log(T)$，就会在所有剩余轮次中改为运行 π_M。证明的关键在于说明，对于属于 $\mathcal{M} \cup \{M\}$ 的模型，p' 的最坏情况遗憾不会比对于属于 $\mathcal{M}$ 的模型，p 的最坏情况遗憾大多少。

按照惯例，在 算法 3 的语境下，我们令 $\mathscr{F}^t$ 表示由 $\{(\pi^s, r^s, o^s)\}_{1 \leq s \leq t}$ 生成的 σ-代数。我们通过考虑以下关于 p' 的情况来界定算法 $M^\star \in \mathcal{M} \cup \{M\}$。

情况 1： $M^\star = \overline{M}$。令 $T_0 \in [T+1]$ 定义为使得决策 t 依据 π^t 第10 行的规则被选取的最小 $T+1$；若不存在这样的步骤，则取 $T+1$。根据构造，我们有

$$\sum_{t=1}^{T_0-1} (f^{\overline{M}}(\pi_{\overline{M}}) - r^t) = \sum_{t=1}^{T} \mathbb{1}\{t < T_0\} \cdot (f^{\overline{M}}(\pi_{\overline{M}}) - r^t) \leq (R+1) \cdot \lceil \log T \rceil. \tag{75}$$

算法 3 引理 B.2 证明中使用的算法 p' 1：参数：轮数 $T \in \mathbb{N}$。算法 $p = (p^1, \dots, p^T)$。2：初始化 $I = 1$、$T_1 = 1$ 和 $R = 4 \cdot \sup_{M^\star \in \mathcal{M}} \mathbb{E}^{M^\star, p}[\mathbf{Reg}_{\mathsf{DM}}(T)] + \delta T + 8\sqrt{T}$。3：对于 $1 \le t \le T$，执行 4：定义 $(p')^t(\cdot)$ 为分布 $p^{t-T_I+1}(\cdot \mid \{(\pi^s, r^s, o^s)\}_{s=T_I}^{t-1})$。5：抽取 $\pi^t \sim (p')^t$，并观测 (π^t, r^t, o^t)。6：如果 $\sum_{s=T_I}^{t} (f^M(\pi_M) - r^s) \ge R$，则 7：设定 $T_{I+1} := t+1$，然后递增 I。// 这会重新初始化 p。8：如果 $I > \lceil \log T \rceil$，则 9：跳出循环。10：对于剩余的时间步 t（如果有）：执行 $\pi^t := \pi_M$（即设定 $(p')^t = \mathbb{I}_{\pi_M}$）。

注意，对于每个 $t \in [T]$，变量 $\mathbb{1}\{t < T_0\}$ 关于 $\mathscr{F}^{t-1}$ 可测。因此，我们有

$$\begin{aligned}
\mathbb{E}^{\overline{M}, p'}[\mathbf{Reg}_{\mathsf{DM}}(T)] &= \mathbb{E}^{\overline{M}, p'}\left[\sum_{t=1}^{T} \left(f^{\overline{M}}(\pi_{\overline{M}}) - \mathbb{E}_{\pi^t \sim (p')^t}[f^{\overline{M}}(\pi^t)]\right)\right] \\
&= \mathbb{E}^{\overline{M}, p'}\left[\sum_{t=1}^{T} \mathbb{1}\{t < T_0\} \cdot \left(f^{\overline{M}}(\pi_{\overline{M}}) - \mathbb{E}_{\pi^t \sim (p')^t}[f^{\overline{M}}(\pi^t)]\right)\right] \\
&= \mathbb{E}^{\overline{M}, p'}\left[\sum_{t=1}^{T} \mathbb{1}\{t < T_0\} \cdot \left(f^{\overline{M}}(\pi_{\overline{M}}) - r^t\right)\right] + \mathbb{E}^{\overline{M}, p'}\left[\sum_{t=1}^{T} \mathbb{1}\{t < T_0\} \cdot \left(r^t - \mathbb{E}_{\pi^t \sim (p')^t}[f^{\overline{M}}(\pi^t)]\right)\right] \\
&\le (R+1) \cdot \log T,
\end{aligned}$$

其中，最后一个不等式用到了 公式 (75) 以及对每个 t，我们都有这一事实

$$\mathbb{E}\left[\mathbb{1}\{t < T_0\} \cdot \left(r^t - \mathbb{E}_{\pi^t \sim (p')^t}[f^{\overline{M}}(\pi^t)]\right) \mid \mathscr{F}^{t-1}\right] = \mathbb{1}\{t < T_0\} \cdot \mathbb{E}[r^t - \mathbb{E}_{\pi^t \sim (p')^t}\left[f^{\overline{M}}(\pi^t)\right] \mid \mathscr{F}^{t-1}] = 0.$$

因此，在 $M^\star = M$ 的情况下，我们已经验证了 公式 (71) 中关于 p' 的遗憾上界成立。

情况 2： $M^\star \in \mathcal{M}$。我们首先陈述并证明两个技术引理。

引理 B.3。 对于算法 p、任意模型 $M^\star \in \mathcal{M}^+$ 以及取值于 $[T]$ 的随机变量 τ（可能依赖于 $\mathfrak{H}^T$），有

$$\mathbb{E}^{M^\star, p}\left[\sum_{t=1}^{\tau} \mathbb{E}_{\pi^t \sim p^t}[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)]\right] \le \mathbb{E}^{M^\star, p}[\mathbf{Reg}_{\mathsf{DM}}(T)].$$

证明引理 B.3。 注意到

$$\begin{aligned}
&\mathbb{E}^{M^\star, p}[\mathbf{Reg}_{\mathsf{DM}}(T)] - \mathbb{E}^{M^\star, p}\left[\sum_{t=1}^{\tau} \mathbb{E}_{\pi^t \sim p^t}[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)]\right] \\
&= \mathbb{E}^{M^\star, p}\left[\sum_{t=1}^{T} \mathbb{1}\{\tau < t\} \cdot \mathbb{E}_{\pi^t \sim p^t}[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)]\right] \ge 0,
\end{aligned}$$

其中 w 我们使用了随机变量 $\mathbb{E}_{\pi^t\sim p^t}[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)]$ 是非负的这一事实 a.s. □

下一个引理涉及 算法 的一次运行 p 违反 第 6 行中 算法 3的条件的概率。

引理 B.4。 对于任意算法 $p=(p^1、\ldots\ldots、p^T)$ 和模型 $M^\star\in\mathcal{M}$，有

$$\mathbb{P}^{M^\star,p}\left(\exists t\le T\ :\ \sum_{s=1}^{t}(f^{\overline{M}}(\pi_{\overline{M}})-r^s)>R\right)\le\frac{1}{2},$$

其中 $R=4\cdot$上确界$_{M^\star\in\mathcal{M}}\mathbb{E}^{M^\star,\,p}[\mathbf{Reg}_{\mathrm{DM}}(T)]+\delta T+8\sqrt{T}$。

引理 B.4 的证明。 令 $\mathscr{F}^t$ 为由 $\{(\pi^s、r^s、o^s)\}_{s=1}^t$生成的 σ-代数。固定 $M^\star\in\mathcal{M}$并定义 $R_0:=\mathbb{E}^{M^\star,\,p}[\mathbf{Reg}_{\mathrm{DM}}(T)]$。根据马尔可夫不等式以及随机变量$\mathbb{E}_{\pi^t\sim p^t}[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)]$均为非负这一事实，有

$$\begin{aligned}&\mathbb{P}^{M^\star,p}\left(\sup_{t\le T}\sum_{s=1}^{t}\mathbb{E}_{\pi^s\sim p^s}[f^{M^\star}(\pi_{M^\star})-f^{M^\star}(\pi^s)]>4R_0\right)\\&=\mathbb{P}^{M^\star,p}\left(\sum_{t=1}^{T}\mathbb{E}_{\pi^t\sim p^t}[f^{M^\star}(\pi_{M^\star})-f^{M^\star}(\pi^t)]>4R_0\right)\le\frac{1}{4}.\end{aligned}\tag{76}$$

现在，定义 $X_0=0$和 $X_t=\sum_{s=1}^t(\mathbb{E}_{\pi^s\sim p^s}[f^{M^\star}(\pi^s)]-r^s)$对于 $t\in[T]$。注意，$(X_t)_{t\ge0}$是相对于滤过 $\mathscr{F}^t$的鞅。因此，根据 Durrett (2019年中的定理 4.5.1，可得

$$\mathbb{E}^{M^\star,p}\left[\sup_{t\le T}|X_t|^2\right]\le4\cdot\mathbb{E}^{M^\star,p}\left[\sum_{t=1}^{T}\mathbb{E}\left[(\mathbb{E}_{\pi^t\sim p^t}[f^{M^\star}(\pi^t)]-r^t)^2\mid\mathscr{F}^{t-1}\right]\right]\le4T,$$

其中，最后一个不等式用到了对所有 t， $|\mathbb{E}_{\pi^t\sim p^t}[f^{M^\star}(\pi^t)]-r^t|\le1$。根据詹森不等式和马尔可夫不等式可知，对于任意 $\lambda>0$,

$$\mathbb{P}^{M^\star,p}\left(\sup_{t\le T}|X_t|>2\lambda\sqrt{T}\right)\le\frac{1}{\lambda},$$

并通过选择 $\lambda=4$，可得

$$\mathbb{P}^{M^\star,p}\left(\sup_{t\le T}\sum_{s=1}^{t}\left(\mathbb{E}_{\pi^s\sim p^s}[f^{M^\star}(\pi^s)]-r^s\right)>8\sqrt{T}\right)\le\frac{1}{4}.\tag{77}$$

结合 公式（76）和公式（77），可得

$$\mathbb{P}^{M^\star,p}\left(\exists t\le T\ :\ \sum_{s=1}^{t}\mathbb{E}_{\pi^s\sim p^s}[f^{M^\star}(\pi_{M^\star})-r^s]>4R_0+8\sqrt{T}\right)\le\frac{1}{2}.$$

由于 $M^\star\in\mathcal{M}$，因此 $f^{M^\star}(\pi_{M^\star})\ge f^{M}(\pi_{M})-\delta$，由此可得

$$\mathbb{P}^{M^\star,p}\left(\exists t\le T\ :\ \sum_{s=1}^{t}\mathbb{E}_{\pi^s\sim p^s}[f^{\overline{M}}(\pi_{\overline{M}})-r^s]>4R_0+8\sqrt{T}+\delta T\right)\le\frac{1}{2},$$

如所愿。 □

现在继续证明引理 B.2。记 $L=\lceil\log T\rceil$，并将 I在算法 3 中的最终值记为 $I'\le L+1$。若 $I'\le L$，则令 $T_{I'+1}=\cdots=T_{L+1}=T+1$。注意，对每个 $\ell\in[L+1]$，$T_\ell-1$都是一个停时（因为事件 $\{T_\ell-1=t\}$相对于 $\mathscr{F}^t$对每个 $t\in[T]$都是可测的），因此 T_ℓ 也同样是一个停时。

Le引理 B.4 结合$(p')^t$ 在第 4 行中的定义可得，对每个 $\ell \in [L]$ ，

$$\mathbb{P}^{M^\star,p'}\left(T_{\ell+1} \le T \mid T_\ell \le T\right) = \mathbb{P}^{M^\star,p'}\left(\exists t \text{ s.t. } \; T-1 \ge t \ge T_\ell \text{ and } \sum_{s=T_\ell}^{t} (f^{\overline{M}}(\pi_{\overline{M}}) - r^s) > R \mid T_\ell \le T\right) \le \frac{1}{2}.$$

因此，由于事件 $\{T_\ell \le T\}$ 等于事件 $\{T_{\ell'} \le T \ \forall \ell' \le \ell\}$,

$$\begin{aligned}
\mathbb{P}^{M^\star,p'}\left(T_{L+1} \le T\right) &= \mathbb{P}^{M^\star,p'}\left(\forall \ell \le L+1,\ T_\ell \le T\right) \\
&= \prod_{\ell=1}^{L} \mathbb{P}^{M^\star,p'}\left(T_{\ell+1} \le T \mid T_{\ell'} \le T \ \forall \ell' \le \ell\right) \\
&= \prod_{\ell=1}^{L} \mathbb{P}^{M^\star,p'}\left(T_{\ell+1} \le T \mid T_\ell \le T\right) \\
&\le (1/2)^L \le 1/T.
\end{aligned}$$

此外，对于每个 $\ell \in [L]$，我们有

$$\begin{aligned}
\mathbb{E}^{M^\star,p'}\left[\sum_{t=T_\ell}^{T_{\ell+1}-1} \mathbb{E}_{\pi^t\sim(p')^t}[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)]\right] &= \mathbb{E}^{M^\star,p'}\left[\sum_{t=1}^{T} \mathbb{1}\left\{T_\ell \le t < T_{\ell+1}\right\} \cdot \mathbb{E}_{\pi^t\sim(p')^t}[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)]\right] \\
&\le \mathbb{E}^{M^\star,p}[\mathbf{Reg}_{\mathsf{DM}}(T)],
\end{aligned}$$

其中不等式 使用 引理 B.3 以及 p' 在 第 4 行 对于步骤 $T_\ell \le t < T_{\ell+1}$。由此可得 t

$$\begin{aligned}
&\mathbb{E}^{M^\star,p'}\left[\sum_{t=1}^{T} \mathbb{E}_{\pi^t\sim(p')^t}[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)]\right] \\
&\le T \cdot \mathbb{P}^{M^\star,p'}(T_{L+1} \le T) + \sum_{\ell=1}^{L} \mathbb{E}^{M^\star,p'}\left[\sum_{t=T_\ell}^{T_{\ell+1}-1} \mathbb{E}_{\pi^t\sim(p')^t}[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)]\right] \\
&\le 1 + L \cdot \mathbb{E}^{M^\star,p}[\mathbf{Reg}_{\mathsf{DM}}(T)],
\end{aligned}$$

这验证了 式 (71) 中所声称的遗憾上界。 □

C 第 3 节中省略的证明

本节将证明定理 3.2，该定理表明，算法 2 达到基于受约束 DEC 的遗憾界

C.1 遗憾上界的证明（定理 3.2）

为证明定 理 3.2，我们引入几个成功事件，它们将在整个分析过程中使用：

1. 对于每个 $i \in [N]$, $\mathscr{A}_i$ 表示以下所有不等式均成立的事件：

$$\forall s \in \mathcal{S}_i, \qquad \sum_{j=1}^{J_{i,s}} \mathbb{E}_{\pi\sim p^s}\left[D^2_{\mathsf{H}}\left(M^\star(\pi), \widetilde{M}^j_s(\pi)\right)\right] \le \mathbf{Est}_{\mathsf{H}}(J_i,\delta), \tag{78}$$

$$\mathbb{E}_{\pi\sim p^{s_i}}\left[D^2_{\mathsf{H}}\left(\widehat{M}_i(\pi), M^\star(\pi)\right)\right] \le \frac{\mathbf{Est}_{\mathsf{H}}(J_i,\delta)}{J_i}, \tag{79}$$

$$\mathbb{E}_{\pi\sim p^{s_i}}\left[D^2_{\mathsf{H}}\left(\widehat{M}^{s_i}(\pi), M^\star(\pi)\right)\right] \le \varepsilon_i^2, \tag{80}$$

其中我们回顾一下， s_i 是定义在第 23 行上的索引。

2. 对于每个 $i \in [N]$， $\mathscr{B}_i$ 表示如下事件：

$$\sum_{t \in \mathcal{E}_i} \mathbb{E}_{\pi \sim p^t} \left[D^2_{\mathsf{H}} \left(M^\star(\pi), \widehat{M}^t(\pi) \right) \right] \le \mathbf{Est}_{\mathsf{H}}(|\mathcal{E}_i|, \delta). \tag{81}$$

3. 对每个 $i \in [N]$， $\mathscr{C}_i$ 表示事件： $M^\star \in \mathcal{M}_i$。

4. 对每个 $i \in [N]$， $\mathscr{D}_i$ 表示存在某个 $s \in \mathcal{S}_i$ ，使得

$$\mathbb{E}_{\pi \sim p^s} \left[D^2_{\mathsf{H}} \left(M^\star(\pi), \widehat{M}^s(\pi) \right) \right] \le \frac{\varepsilon_i^2}{16}.$$

此外，我们定义

$$\mathscr{A} = \bigcap_{i \in [N]} \mathscr{A}_i, \ \mathscr{B} = \bigcap_{i \in [N]} \mathscr{B}_i, \ \mathscr{C} = \bigcap_{i \in [N]} \mathscr{C}_i, \text{ and } \mathscr{D} = \bigcap_{i \in [N]} \mathscr{D}_i.$$

我们还回顾以下记号，它将在整个证明中使用。

- 对于 $i \in [N]$ ，我们设

$$\alpha_i := C_0 \cdot \mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M}) + 64\varepsilon_i, \tag{82}$$

约定 $\alpha_0 = 1$。常数 $C_0 > 0$ 在 式 (82)中，以及常数 $C_1 > 0$在 算法 2中指定的值，都需要取得足够大；接下来我们将证明，取 $C_0 \ge 20$和 $C_1 \ge 128$ 就足够了。

C.1.1 技术引理

在证明定理 3.2之前，我们陈述并证明若干关于算法 2性能的技术引理。下述引理表明，事件 $\mathscr{A} \cap \mathscr{B} \cap \mathscr{C}$ 以高概率发生。

引理 C.1. 假设 $C_1 \ge 128$。事件 $\mathscr{A} \cap \mathscr{B} \cap \mathscr{C} \cap \mathscr{D}$ 发生的概率至少为 $1 - 3L\delta N$。

引理 引理 C.1的证明。 我们证明 $\mathbb{P}\left(\bigcap_{i' \le i} \mathscr{A}_{i'} \cap \mathscr{B}_{i'} \cap \mathscr{C}_{i'} \cap \mathscr{D}_{i'} \right) \ge 1 - 3L\delta i$ 对每个 $i \in [N]$ 都成立，方法是对 i进行归纳。固定 $i \in [N]$ ，并令我们在$\bigcap_{i' < i}(\mathscr{A}_{i'} \cap \mathscr{B}_{i'} \cap \mathscr{C}_{i'} \cap \mathscr{D}_{i'})$的条件下进行分析。

证明 $\mathscr{C}_i$ 成立。 事实 $\mathscr{A}_{i-1}$ 成立意味着 $\mathbb{E}_{\pi \sim \widehat{p}_{i-1}} \left[D^2_{\mathsf{H}} \left(M^\star(\pi) \ \widehat{M}_{i-1}(\pi) \right) \right] \le \mathbf{Est}_{\mathsf{H}}(J_{i-1}, \ \delta) \ J_{i-1}$ ，根据定义，这意味着 $M^\star \in \mathcal{M}_i$，即 $\mathscr{C}_i$ 成立。

证明 $\mathscr{B}_i$ 成立。 在 $\mathscr{C}_i$ 条件下，根据 假设 3.2 可知 $\mathscr{B}_i$ 成立的概率至少为 $1 - \delta$。

证明 $\mathscr{D}_i$ 成立。 注意，由于我们的参数设置，

$$\varepsilon_i^2 \ge 2^{-i} \cdot T \cdot \varepsilon_N^2 = 2^{-i} \cdot C_1 \cdot \mathbf{Est}_{\mathsf{H}}(T, \delta) \cdot L \ge \max \left\{ 32 \cdot \frac{\mathbf{Est}_{\mathsf{H}}(J_i, \delta)}{J_i}, 32 \cdot \frac{\mathbf{Est}_{\mathsf{H}}(|\mathcal{E}_i|, \delta)}{|\mathcal{E}_i|} \right\}, \tag{83}$$

其中我们用到了 $C_1 \ge 128$，并且 $\mathsf{Est}_{\mathsf{H}}(T, \delta) \ge \max\{\mathsf{Est}_{\mathsf{H}}(|\mathcal{E}_i|, \delta), \mathsf{Est}_{\mathsf{H}}(J_i, \delta)\}$。接着，由于 $\mathscr{B}_i$（即 $\widehat{\ }$(81)）成立，至少有 $|\mathcal{E}_i|/2$ 轮次 $t \in \mathcal{E}_i$ 满足 $M^\star \in \mathcal{H}_{\varepsilon_i/4, p^t}(M^t)$。由于 $|\mathcal{S}_i| = L \ge \log 1/\delta$，因此以至少 $1 - \delta$ 的概率，存在某个 $s \in \mathcal{S}_i$，记为 $s_i^\star$，使得 $M^\star \in \mathcal{H}_{\varepsilon_i/4, p^{s_i^\star}}(\widehat{M}^{s_i^\star})$。特别地，在给定 $\mathscr{B}_i$ 的条件下，事件 $\mathscr{D}_i$ 以至少 $1 - \delta$ 的概率成立。

证明 $\mathscr{A}_i$ **成立。** 接下来，根据 假设 3.2 以及海林格距离恒为非负这一事实，我们有：以至少 $1-L\delta$ 的概率，对所有 $s\in\mathcal{S}_i$,

$$\sum_{j=1}^{J_{i,s}}\mathbb{E}_{\pi\sim p^s}\left[D_{\mathsf{H}}^2\left(M^\star(\pi),\widetilde{M}_s^j(\pi)\right)\right]\le \mathbf{Est}_{\mathsf{H}}(J_i,\delta), \tag{84}$$

这验证了(78)成立。接下来，我们对 $\mathscr{D}_i$ 成立这一事件进行条件化。然后，将(84) 应用于$s=s_i^\star$，并利用 $s_i^\star$ 的定义（回想一下， $s_i^\star$在前文中定义为满足 $M^\star\in\mathcal{H}_{\varepsilon_i/4,p^{s_i^\star}}(\widehat{M}^{s_i^\star})$），可得

$$\begin{aligned}
\sum_{j=1}^{J_{i,s_i^\star}}\mathbb{E}_{\pi\sim p^{s_i^\star}}\left[D_{\mathsf{H}}^2\left(\widehat{M}^{s_i^\star}(\pi),\widetilde{M}_{s_i^\star}^j(\pi)\right)\right]\le&\sum_{j=1}^{J_{i,s_i^\star}}2\mathbb{E}_{\pi\sim p^{s_i^\star}}\left[D_{\mathsf{H}}^2\left(\widehat{M}^{s_i^\star}(\pi),M^\star(\pi)\right)\right]+2\mathbb{E}_{\pi\sim p^{s_i^\star}}\left[D_{\mathsf{H}}^2\left(\widehat{M}^\star(\pi),\widetilde{M}_{s_i^\star}^j(\pi)\right)\right]\\
\le&2J_{i,s_i^\star}\cdot\frac{\varepsilon_i^2}{16}+2\sum_{j=1}^{J_{i,s_i^\star}}\mathbb{E}_{\pi\sim p^{s_i^\star}}\left[D_{\mathsf{H}}^2\left(M^\star(\pi),\widetilde{M}_{s_i^\star}^j(\pi)\right)\right]\\
\le&\frac{J_i\varepsilon_i^2}{8}+2\cdot\mathbf{Est}_{\mathsf{H}}(J_i,\delta)\le\frac{3J_i\varepsilon_i^2}{16},
\end{aligned}$$

其中，最后一个不等式用到了式 (83)。根据 $J_{i,s_i^\star}$ 在 第 20 行 和 22上的定义，必然有 $J_{i,s_i^\star}=J_i$，因此 s_i^{tmp} 在 第 22 行至少被赋值一次。因此，在 第 23 行设置的 s_i 满足

$$\mathbb{E}_{\pi\sim p^{s_i}}\left[D_{\mathsf{H}}^2\left(\widehat{M}_i(\pi),M^\star(\pi)\right)\right]\le\frac{1}{J_i}\sum_{j=1}^{J_i}\mathbb{E}_{\pi\sim p^{s_i}}\left[D_{\mathsf{H}}^2\left(\widetilde{M}_{s_i}^j(\pi),M^\star(\pi)\right)\right]\le\frac{\mathbf{Est}_{\mathsf{H}}(J_i,\delta)}{J_i},$$

其中，第一个不等式用到了平方海灵格距离的凸性，第二个不等式则用到了 (84)以及 $J_{i,s_i}=J_i$这一事实。上述推导验证了 (79)；为验证 (80)，注意到由于 s_i^{tmp} 至少被赋值一次，

$$\sum_{j=1}^{J_i}\mathbb{E}_{\pi\sim p^{s_i}}\left[D_{\mathsf{H}}^2\left(\widehat{M}^{s_i}(\pi),\widetilde{M}_{s_i}^j(\pi)\right)\right]\le\frac{J_i\varepsilon_i^2}{4}. \tag{85}$$

因此，我们可以计算

$$\begin{aligned}
\mathbb{E}_{\pi\sim p^{s_i}}\left[D_{\mathsf{H}}^2\left(\widehat{M}^{s_i}(\pi),M^\star(\pi)\right)\right]\le&2\cdot\mathbb{E}_{\pi\sim p^{s_i}}\left[D_{\mathsf{H}}^2\left(\widehat{M}_i(\pi),\widehat{M}^{s_i}(\pi)\right)\right]+2\cdot\mathbb{E}_{\pi\sim p^{s_i}}\left[D_{\mathsf{H}}^2\left(\widehat{M}_i(\pi),M^\star(\pi)\right)\right]\\
\le&\frac{2}{J_i}\sum_{j=1}^{J_i}\mathbb{E}_{\pi\sim p^{s_i}}\left[D_{\mathsf{H}}^2\left(\widehat{M}^{s_i}(\pi),\widetilde{M}_{s_i}^j(\pi)\right)\right]+\frac{2}{J_i}\sum_{j=1}^{J_i}\mathbb{E}_{\pi\sim p^{s_i}}\left[D_{\mathsf{H}}^2\left(M^\star(\pi),\widetilde{M}_{s_i}^j(\pi)\right)\right]\\
\le&\frac{\varepsilon_i^2}{2}+\frac{2\cdot\mathbf{Est}_{\mathsf{H}}(J_i,\delta)}{J_i}\le\varepsilon_i^2,
\end{aligned}$$

其中，第二个不等式利用了平方海灵格距离的凸性，第三个不等式利用了 (84)对于 $s=s_i$ 以及 (85)，最后一个不等式则利用了 (83)。由上述推导可验证 (80)，我们得出，在 $\mathscr{D}_i$ 成立的条件下， $\mathscr{A}_i$ 以至少 $1-L\delta$的概率成立。

总结。 总而言之，在 $\bigcap_{i'<i}\mathscr{A}_{i'}\cap\mathscr{B}_{i'}\cap\mathscr{C}_{i'}\cap\mathscr{D}_{i'}$ 的条件下，我们证明了 $\mathscr{A}_i\cap\mathscr{B}_i\cap\mathscr{C}_i\cap\mathscr{D}_i$以概率 $1-2\delta-L\delta\ge1-3L\delta$ 成立。因此，归纳假设 $\mathbb{P}\left(\bigcap_{i'<i}\mathscr{A}_{i'}\cap\mathscr{B}_{i'}\cap\mathscr{C}_{i'}\right)\ge1-3L\delta(i-1)$ 蕴含 $\mathbb{P}\left(\bigcap_{i'\le i}\mathscr{A}_{i'}\cap\mathscr{B}_{i'}\cap\mathscr{C}_{i'}\cap\mathscr{D}_{i'}\right)\ge(1-3L\delta(i-1))\cdot(1-3L\delta)\ge1-3L\delta i$。

综上，我们得到 $\mathbb{P}(\mathscr{A}\cap\mathscr{B}\cap\mathscr{C}\cap\mathscr{D})\ge1-3L\delta N$。 □

引理 C.2 表明， $\widehat{p}_i$ 在算法 2 中计算出的分布具有较低的次优性，相对于$M^\star$。

引理 C.2 (精化策略的准确性)。 假设 $C_0 \geq 4$。那么，对于每个轮次 $i \in [N]$，在事件 $\mathscr{A}_i$ 下，分布 $\widehat{p}_i$ 满足

$$\mathbb{E}_{\pi \sim \widehat{p}_i}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)\right] \leq \frac{\alpha_i}{4}. \tag{86}$$

引理的证明 **引理 C.2**。 在事件 $\mathscr{A}_i$ 条件下，可得 (80)成立，特别地可写为 $M^\star \in \mathcal{H}_{\varepsilon_i,\, p^{s_i}}(M^{s_i})$。因此，根据 $\widehat{p}_i = p^{s_i}$ 在 第 23 行 中的选择以及 第 10 行中的定义，在 $\mathscr{A}_i$ 下，

$$\begin{aligned}\mathbb{E}_{\pi \sim p^{s_i}}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)\right] &\leq \sup_{M \in \mathcal{H}_{\varepsilon_i, p^{s_i}}(\widehat{M}^{s_i}) \cup \{\widehat{M}^{s_i}\}} \mathbb{E}_{\pi \sim p^{s_i}}\left[f^{M}(\pi_M) - f^{M}(\pi)\right] \\ &= \textsf{r-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M} \cup \{\widehat{M}^{s_i}\}, \widehat{M}^{s_i}) \leq \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \leq \frac{\alpha_i}{4},\end{aligned}$$

只要 $C_0 \geq 4$ 成立，最后一个不等式便可得出。 □

引理 C.3 将 $M^\star$ 下任意分布 $p \in \Delta(\Pi)$ 对应的次优性与任意模型$\overline{M} \in \mathrm{co}(\mathcal{M}_{i+1})$的次优性联系起来，依据是 M与 $M^\star$之间的距离。最终，我们对每个 $t \in \mathcal{E}_{i+1}$ 应用该引理，从而得出以下引理 引理 C.4。

引理 C.3 (与精化类中的模型比较)。 固定 $i \in [N]$。那么，对于所有 $M \in \mathrm{co}(\mathcal{M}_{i+1})$ 以及所有$p \in \Delta(\Pi)$，在事件 $\mathscr{A}_i$ 下，

$$\mathbb{E}_{\pi \sim p}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)\right] \leq \mathbb{E}_{\pi \sim p}\left[f^{M}(\pi_M) - f^{M}(\pi)\right] + \frac{\alpha_i}{2} + \sqrt{\mathbb{E}_{\pi \sim p}\left[D^2_{\mathsf{H}}(M(\pi), M^\star(\pi))\right]}. \tag{87}$$

引理的证明： **引理 C.3**。我们首先证明，在 $M^\star$ 下的最优值不超过任意 $M \in \mathrm{co}(\mathcal{M}_{i+1})$ 下的最优值。为此，首先注意到，根据 引理 C.2 （特别是，式 (86) 在轮次i成立这一事实），我们有 $\mathbb{E}_{\pi \sim \widehat{p}_i}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)\right] \leq \frac{\alpha_i}{4}$在 $\mathscr{A}_i$ 下成立。于是，对于任意 $M \in \overline{\mathcal{M}_{i+1}}$，在事件 $\mathscr{A}_i$ 发生时，

$$\begin{aligned}\mathbb{E}_{\pi \sim \widehat{p}_i}\left[f^{M^\star}(\pi_{M^\star}) - f^{M}(\pi)\right] &\leq \mathbb{E}_{\pi \sim \widehat{p}_i}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)\right] + \sqrt{\mathbb{E}_{\pi \sim \widehat{p}_i}\left[D^2_{\mathsf{H}}(M(\pi), M^\star(\pi))\right]} \\ &\leq \frac{\alpha_i}{4} + \sqrt{2\mathbb{E}_{\pi \sim \widehat{p}_i}\left[D^2_{\mathsf{H}}\Big(M(\pi), \widehat{M}_i(\pi)\Big)\right] + 2\mathbb{E}_{\pi \sim \widehat{p}_i}\left[D^2_{\mathsf{H}}\Big(\widehat{M}_i(\pi), M^\star(\pi)\Big)\right]} \\ &\leq \frac{\alpha_i}{4} + 2\sqrt{\frac{\mathbf{Est}_{\mathsf{H}}(J_i, \delta)}{J_i}} \leq \frac{\alpha_i}{2},\end{aligned} \tag{88}$$

其中倒数第二个不等式成立，是因为 $M \in \mathcal{M}_{i+1}$，并且根据事件 $\mathscr{A}_i$ 的假设（特别是，利用 (79)）；最后一个不等式成立，是因为 $2\sqrt{\mathbf{Est}_{\mathsf{H}}(J_i,\delta)\ J_i} \leq \frac{\alpha_i}{4}$，这是由我们对 $\alpha_i \geq 64\varepsilon_i$ 的选择以及 式 (83) 得出的。

现在固定任意一个 $M \in \mathrm{co}(\mathcal{M}_{i+1})$，并注意到我们可以将 $M = \mathbb{E}_{M' \sim \nu_M}[M']$ 写成某个 $\nu_M \in \Delta(\mathcal{M}_{i+1})$ 的形式。于是，对于所有 $\pi \in \Pi$， $f^M(\pi) = \mathbb{E}_{M' \sim \nu_M}[f^{M'}(\pi)]$,，由 式 (88) 可知（同样在 $\mathscr{A}_i$ 下）

$$f^{M^\star}(\pi_{M^\star}) - f^{M}(\pi_M) \leq \mathbb{E}_{\pi \sim \widehat{p}_i}\left[f^{M^\star}(\pi_{M^\star}) - f^{M}(\pi)\right] = \mathbb{E}_{M' \sim \nu_M}\mathbb{E}_{\pi \sim \widehat{p}_i}\left[f^{M^\star}(\pi_{M^\star}) - f^{M'}(\pi)\right] \leq \frac{\alpha_i}{2}.$$

对于任意 $M \in \mathrm{co}(\mathcal{M}_{i+1})$，现在我们有，在 $\mathscr{A}_i$ 下，

$$\begin{aligned}\mathbb{E}_{\pi \sim p}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)\right] &\leq \frac{\alpha_i}{2} + \mathbb{E}_{\pi \sim p}\left[f^{M}(\pi_M) - f^{M}(\pi)\right] + \mathbb{E}_{\pi \sim p}\left[\left|f^{M}(\pi) - f^{M^\star}(\pi)\right|\right] \\ &\leq \frac{\alpha_i}{2} + \mathbb{E}_{\pi \sim p}\left[f^{M}(\pi_M) - f^{M}(\pi)\right] + \sqrt{\mathbb{E}_{\pi \sim p}\left[D^2_{\mathsf{H}}(M(\pi), M^\star(\pi))\right]},\end{aligned}$$

如预期。 □

我们的最后一个技术引理，引理 C.4，对每一轮中执行的所有策略的次优性进行了界定 $\mathcal{E}_i$。需要确立这类结果，是遗憾框架与 PAC 框架之间的一个关键区别，也促使我们在算法 2的设计中作出许多选择。

引理 C.4 （“备用”遗憾保证）。 任取 $i \in [N]$。那么对于所有 $t \in \mathcal{E}_i$，在事件 $\mathscr{A}_{i-1}$ 发生时，我们有：

$$\mathbb{E}_{\pi\sim p^t}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)\right] \le \textsf{r-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M}\cup\{\widehat{M}^t\},\widehat{M}^t) + \frac{\alpha_{i-1}}{2} + \sqrt{\mathbb{E}_{\pi\sim p^t}\left[D^2_{\mathsf{H}}\left(\widehat{M}^t(\pi), M^\star(\pi)\right)\right]}.$$

引理 C.4 的证明。 任取 $t \in \mathcal{E}_i$。在 p^t 中的选择 第 10 行 确保了

$$\mathbb{E}_{\pi\sim p^t}[f^{\widehat{M}^t}(\pi_{\widehat{M}^t}) - f^{\widehat{M}^t}(\pi)] \le \textsf{r-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M}\cup\{\widehat{M}^t\},\widehat{M}^t). \tag{89}$$

接下来，在事件 $\mathscr{A}_{i-1}$ 发生时，我们有

$$\begin{aligned}
&\mathbb{E}_{\pi\sim p^t}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)\right] \\
&\le \mathbb{E}_{\pi\sim p^t}\left[f^{\widehat{M}^t}(\pi_{\widehat{M}^t}) - f^{\widehat{M}^t}(\pi)\right] + \frac{\alpha_{i-1}}{2} + \sqrt{\mathbb{E}_{\pi\sim p^t}\left[D^2_{\mathsf{H}}\left(\widehat{M}^t(\pi), M^\star(\pi)\right)\right]} \\
&\le \textsf{r-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M}\cup\{\widehat{M}^t\},\widehat{M}^t) + \frac{\alpha_{i-1}}{2} + \sqrt{\mathbb{E}_{\pi\sim p^t}\left[D^2_{\mathsf{H}}\left(\widehat{M}^t(\pi), M^\star(\pi)\right)\right]},
\end{aligned}$$

其中，第一个不等式利用了 引理 C.3 在轮次 $i-1$ 且 $p = p^t$ 和 $M = M^t$的条件下，以及 这一事实：$M^t \in \mathrm{co}(\mathcal{M}_i)$ 根据构造成立；第二个不等式则利用了 式 (89)。 □

C.1.2 定理 3.2 的证明

定理 3.2 的证明。 令我们以事件 $\mathscr{A}\cap\mathscr{B}\cap\mathscr{C}\cap\mathscr{D}$为条件，根据 引理 C.1，该事件以概率 $1 - 3L\delta = 1 - 3\lceil\log 1/\delta\rceil\cdot\delta$成立。固定 $i \in [N]$。我们按如下方式分析每个轮次 i 中的遗憾。

- 我们首先分析其中的轮次 $t \in \mathcal{E}_i$。根据 引理 C.4，在事件 $\mathscr{A}_{i-1}$ 下，我们有 e

$$\begin{aligned}
&\sum_{t\in\mathcal{E}_i}\mathbb{E}_{\pi\sim p^t}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi)\right] \\
&\le |\mathcal{E}_i|\cdot\left(\sup_{\overline{M}\in\mathrm{co}(\mathcal{M})}\textsf{r-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M}\cup\{\overline{M}\},\overline{M}) + \alpha_{i-1}\right) + \sum_{t\in\mathcal{E}_i}\sqrt{\mathbb{E}_{\pi\sim p^t}\left[D^2_{\mathsf{H}}\left(\widehat{M}^t(\pi), M^\star(\pi)\right)\right]} \\
&\le |\mathcal{E}_i|\cdot(\alpha_i + \alpha_{i-1}) + \sqrt{|\mathcal{E}_i|\cdot\sum_{t\in\mathcal{E}_i}\mathbb{E}_{\pi\sim p^t}\left[D^2_{\mathsf{H}}\left(\widehat{M}^t(\pi), M^\star(\pi)\right)\right]} \\
&\le 2\cdot|\mathcal{E}_i|\cdot\alpha_{i-1} + \sqrt{|\mathcal{E}_i|\cdot\mathbf{Est}_{\mathsf{H}}(|\mathcal{E}_i|,\delta)},
\end{aligned} \tag{90}$$

其中倒数第二个不等式由我们对 α_i 的选择得出，最后一个不等式则由 $\mathscr{B}_i$ 成立且 $\alpha_i \le \alpha_{i-1}$ 这一事实得出。

- 接下来，我们分析 $\mathcal{R}_i$ 中的轮次。首先分析在 第 17 行 对决策 $\pi^j_s \sim p^s$ 进行采样的那些轮次。为此，固定任意一个 $s \in \mathcal{S}_i$。我们首先注意到，根据 $J_{i,s}$ 的定义，

$$\begin{aligned}
\sum_{j=1}^{J_{i,s}}\sqrt{\mathbb{E}_{\pi\sim p^s}\left[D^2_{\mathsf{H}}\left(\widetilde{M}^j_s(\pi), \widehat{M}^s(\pi)\right)\right]} &\le \sqrt{J_{i,s}\cdot\sum_{j=1}^{J_{i,s}}\mathbb{E}_{\pi\sim p^s}\left[D^2_{\mathsf{H}}\left(\widetilde{M}^j_s(\pi), \widehat{M}^s(\pi)\right)\right]} \\
&\le \sqrt{J_{i,s}\cdot\left(\frac{J_i\varepsilon_i^2}{4} + 2\right)} \le \sqrt{2J_i} + J_i\varepsilon_i/2.
\end{aligned}$$

此外，由于事件 $\mathscr{A}_i$ 成立（特别是利用(78)），我们有

$$\begin{aligned}\sum_{j=1}^{J_{i,s}}\sqrt{\mathbb{E}_{\pi\sim p^s}\left[D^2_{\mathsf{H}}\left(\widetilde{M}^j_s(\pi),M^\star(\pi)\right)\right]}\leq&\sqrt{J_{i,s}\cdot\sum_{j=1}^{J_{i,s}}\mathbb{E}_{\pi\sim p^s}\left[D^2_{\mathsf{H}}\left(\widetilde{M}^j_s(\pi),M^\star(\pi)\right)\right]}\\\leq&\sqrt{J_{i,s}\cdot\mathbf{Est}_{\mathsf{H}}(J_i,\delta)}\leq\sqrt{J_i^2\varepsilon_i^2/32}\leq J_i\varepsilon_i,\end{aligned}$$

其中倒数第二个不等式用到了(83)。利用上述各式可得

$$\begin{aligned}&\sum_{j=1}^{J_{i,s}}\mathbb{E}_{\pi^j_s\sim p^s}\left[f^{M^\star}(\pi_{M^\star})-f^{M^\star}(\pi^j_s)\right]\\\leq&J_{i,s}\cdot\left(\textsf{r-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M})+\alpha_{i-1}+\sqrt{\mathbb{E}_{\pi\sim p^s}\left[D^2_{\mathsf{H}}\left(\widehat{M}^s(\pi),M^\star(\pi)\right)\right]}\right)\\\leq&2J_{i,s}\cdot\alpha_{i-1}+\sum_{j=1}^{J_{i,s}}\sqrt{2\mathbb{E}_{\pi\sim p^s}\left[D^2_{\mathsf{H}}\left(\widehat{M}^s(\pi),\widetilde{M}^j_s(\pi)\right)\right]}+\sqrt{2\mathbb{E}_{\pi\sim p^s}\left[D^2_{\mathsf{H}}\left(M^\star(\pi),\widetilde{M}^j_s(\pi)\right)\right]}\\\leq&2J_i\cdot\alpha_{i-1}+\frac{3J_i\varepsilon_i}{2}+\sqrt{2J_i}.\end{aligned}$$

接下来分析满足以下条件的轮次 $t\in\mathcal{R}_i$，其中 $\pi^t\sim p^{s_i}=\widehat{p}_i$ 在 第 25 行。由于 $\mathscr{A}_i$ 成立，由引理 C.2 可得 $\mathbb{E}_{\pi^t\sim\widehat{p}_i}[f^{M^\star}(\pi_{M^\star})-f^{M^\star}(\pi)]\leq\alpha_i/4$，这意味着此类轮次对遗憾的总贡献 $t\in\mathcal{R}_i$ 至多为 $|\mathcal{R}_i|\cdot\alpha_i/4$。因此，$\mathcal{R}_i$ 中各轮次对遗憾的总贡献上界如下:

$$\begin{aligned}\sum_{t\in\mathcal{R}_i}\mathbb{E}_{\pi\sim p^t}[f^{M^\star}(\pi_{M^\star})-f^{M^\star}(\pi)]\leq&\frac{|\mathcal{R}_i|\alpha_i}{4}+L\cdot\left(2J_i\alpha_i+\frac{3J_i\varepsilon_i}{2}+\sqrt{2J_i}\right)\\\leq&4|\mathcal{R}_i|\alpha_i+\sqrt{2L|\mathcal{R}_i|},\end{aligned}$$

其中，在第二个不等式中，我们用到了 $|\mathcal{R}_i|=J_i\cdot L$ 和 $3\varepsilon_i/2\leq\alpha_i$。

综上，在事件 $\mathscr{A}\cap\mathscr{B}\cap\mathscr{C}\cap\mathscr{D}$发生时，总遗憾的上界为

$$\sum_{t=1}^{T}\mathbb{E}_{\pi\sim p^t}\left[f^{M^\star}(\pi_{M^\star})-f^{M^\star}(\pi)\right]\leq\sum_{i=1}^{N}\left(4\alpha_{i-1}\cdot(|\mathcal{R}_i|+|\mathcal{E}_i|)+\sqrt{|\mathcal{E}_i|\cdot\mathbf{Est}_{\mathsf{H}}(|\mathcal{E}_i|,\delta)}+\sqrt{2L|\mathcal{R}_i|}\right).\tag{91}$$

现在我们化简公式（91）中的表达式。回顾一下，假设 3.3 表明，对所有 $\varepsilon>0$，

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M})\leq C^2_{\mathrm{reg}}\cdot\textsf{r-dec}^{\mathrm{c}}_{\varepsilon/C_{\mathrm{reg}}}(\mathcal{M}).$$

对每个 $i\in[N]$ 将此不等式应用总计 $\left\lceil\frac{\log(\varepsilon_i/\varepsilon_N)}{\log(C_{\mathrm{reg}})}\right\rceil$ 次，可得

$$\begin{aligned}\textsf{r-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M})\leq&C^2_{\mathrm{reg}}\cdot\left(\frac{\varepsilon_i}{\varepsilon_N}\right)^2\cdot\textsf{r-dec}^{\mathrm{c}}_{\varepsilon_N}(\mathcal{M})\\=&C^2_{\mathrm{reg}}\cdot2^{N-i}\cdot\textsf{r-dec}^{\mathrm{c}}_{\varepsilon_N}(\mathcal{M}).\end{aligned}$$

然后，根据对每个 $i \in [N]$ 选择 $\alpha_i = C_0 \cdot \text{r-dec}^{\text{c}}_{\varepsilon_i}(\mathcal{M}) + 64\varepsilon_i$，我们有

$$\begin{aligned}\sum_{i=1}^{N} \alpha_{i-1} \cdot 2^i &\le 64 \sum_{i=1}^{N} \varepsilon_{i-1} \cdot 2^i + C_0 \sum_{i=1}^{N} 2^i \cdot \text{r-dec}^{\text{c}}_{\varepsilon_{i-1}}(\mathcal{M}) \\ &\le 64 \sum_{i=1}^{N} \varepsilon_N \cdot \sqrt{2^{N+1+i}} + O\left(\sum_{i=1}^{N} 2^N \cdot \text{r-dec}^{\text{c}}_{\varepsilon_N}(\mathcal{M})\right) \\ &\le 128 \cdot \sqrt{C_1 \cdot \mathbf{Est}_{\mathsf{H}}(T,\delta) \cdot L} \cdot \sum_{i=1}^{N} \sqrt{2^{i+1}} + O\left(NT \cdot \text{r-dec}^{\text{c}}_{\varepsilon_N}(\mathcal{M})\right).\end{aligned}$$

因此，我们可以将公式（91）中的表达式上界为如下形式（利用$\mathbf{Est}_{\mathsf{H}}(|\mathcal{E}_i|,\delta) \le \mathbf{Est}_{\mathsf{H}}(T,\delta)$对每个 i）：

$$\begin{aligned}&\sum_{i=1}^{N} \left(4\alpha_{i-1} \cdot (|\mathcal{R}_i| + |\mathcal{E}_i|) + \sqrt{|\mathcal{E}_i| \cdot \mathbf{Est}_{\mathsf{H}}(|\mathcal{E}_i|,\delta)} + \sqrt{2L|\mathcal{R}_i|}\right) \\ \le& 4\sum_{i=1}^{N} \alpha_{i-1} \cdot 2^i + \sqrt{\mathbf{Est}_{\mathsf{H}}(T,\delta)} \sum_{i=1}^{N} \sqrt{2^i} + \sqrt{2L} \sum_{i=1}^{N} \sqrt{2^i} \\ \le& O\left(\sqrt{\mathbf{Est}_{\mathsf{H}}(T,\delta) \cdot L} \sum_{i=1}^{N} \sqrt{2^i} + \sqrt{L} \sum_{i=1}^{N} \sqrt{2^i} + NT \cdot \text{r-dec}^{\text{c}}_{\varepsilon_N}(\mathcal{M})\right) \\ \le& O\left(\sqrt{T\log(1/\delta) \cdot \mathbf{Est}_{\mathsf{H}}(T,\delta)} + T\log(T) \cdot \text{r-dec}^{\text{c}}_{\varepsilon_N}(\mathcal{M})\right),\end{aligned}$$

其中，最后一个不等式中用到了 $L = O(\log 1/\delta)$。通过将 δ重新缩放至 δ^2，并注意到根据构造，对于某个通用常数 $C > 0$，有 $\varepsilon_N \le C \cdot \sqrt{\mathbf{Est}_{\mathsf{H}}(T,1/\delta)\cdot\log T}^{\,1/\delta}$，即可完成证明。 □

D 第 4 节的证明与其他结果

D.1 技术引理

引理 D.1。 设 M 和 $\overline{M}$ 满足 $\mathcal{R} \subseteq [0,1]$。那么，对于所有 $\varepsilon \ge 0$ 和 $p \in \Delta(\Pi)$，若 $M \in \mathcal{H}_{p,\varepsilon}(\overline{M})$，则 en

$$\mathbb{E}_{\pi\sim p}\left[\left|f^{M}(\pi) - f^{\overline{M}}(\pi)\right|\right] \le \varepsilon. \tag{92}$$

证明引理 D.1。由于奖励属于 $[0,1]$, ，我们有

$$\mathbb{E}_{\pi\sim p}\left[\left|f^{M}(\pi) - f^{\overline{M}}(\pi)\right|\right] \le \mathbb{E}_{\pi\sim p}\left[D_{\mathsf{TV}}\left(M(\pi), \overline{M}(\pi)\right)\right] \le \sqrt{\mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right]} \le \varepsilon.$$

□

引理 D.2。固定一个模型类 $\mathcal{M}$。设 $M \in \overline{\mathcal{M}}^{+}$ 和 $\varepsilon > 0$ 给定，并令

$$\mathcal{M}' = \{M \in \mathcal{M} \mid f^{\overline{M}}(\pi_{\overline{M}}) \le f^{M}(\pi_{\overline{M}}) + \varepsilon\}.$$

那么我们有

$$\text{r-dec}^{\text{c}}_{\varepsilon/\sqrt{2}}(\mathcal{M}, \overline{M}) \le \text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}', \overline{M}) + \varepsilon.$$

引理的证明 引理 D.2。 令 $p \in \Delta(\Pi)$ 取得 $\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}', M)$ 的值，并令 $p' = \frac{1}{2}p + \frac{1}{2}\mathbb{I}_{\pi_M}$。令$\underline{M} \in \mathcal{H}_{p',\varepsilon/\sqrt{2}}(M) \subseteq \mathcal{H}_{p,\ \varepsilon}(\overline{M}) \cap \mathcal{H}_{\mathbb{I}_{\pi_M},\ \varepsilon}(\overline{M})$。我们声称 $M \in \mathcal{M}'$。事实上，

$$f^{\overline{M}}(\pi_{\overline{M}}) - f^{M}(\pi_{\overline{M}}) \le D_{\mathsf{H}}\left(\overline{M}(\pi_{\overline{M}}), M(\pi_{\overline{M}})\right) \le \varepsilon$$

根据 引理 D.1。由此可得

$$\sup_{M\in\mathcal{H}_{p',\varepsilon/\sqrt{2}}(\overline{M})}\mathbb{E}_{\pi\sim p}[f^M(\pi_M)-f^M(\pi)]\le\sup_{M\in\mathcal{H}_{p,\varepsilon}(\overline{M})\cap\mathcal{M}'}\mathbb{E}_{\pi\sim p}[f^M(\pi_M)-f^M(\pi)]\le\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}',\overline{M}),\tag{93}$$

因此

$$\sup_{M\in\mathcal{H}_{p',\varepsilon/\sqrt{2}}(\overline{M})}\mathbb{E}_{\pi\sim p'}[f^M(\pi_M)-f^M(\pi)]\le\frac{1}{2}\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}',\overline{M})+\frac{1}{2}\sup_{M\in\mathcal{H}_{p',\varepsilon/\sqrt{2}}(\overline{M})}[f^M(\pi_M)-f^M(\pi_{\overline{M}})].$$

为了对上面的最后一项进行界定，我们有

$$\begin{aligned}\sup_{M\in\mathcal{H}_{p',\varepsilon/\sqrt{2}}(\overline{M})}[f^M(\pi_M)-f^M(\pi_{\overline{M}})]&\le\sup_{M\in\mathcal{H}_{p',\varepsilon/\sqrt{2}}(\overline{M})}\left[f^M(\pi_M)-f^{\overline{M}}(\pi_{\overline{M}})\right]+\varepsilon\\&\le\sup_{M\in\mathcal{H}_{p',\varepsilon/\sqrt{2}}(\overline{M})}\mathbb{E}_{\pi\sim p}\left[f^M(\pi_M)-f^{\overline{M}}(\pi)\right]+\varepsilon\\&\le\sup_{M\in\mathcal{H}_{p',\varepsilon/\sqrt{2}}(\overline{M})}\mathbb{E}_{\pi\sim p}[f^M(\pi_M)-f^M(\pi)]+2\varepsilon\\&\le\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}',\overline{M})+2\varepsilon,\end{aligned}$$

其中，第一个和第三个不等式用到了引理 D.1，最后一个不等式则应用了公式（93）。

□

引理 D.3。对于模型类 $\mathcal{M}$ 和参考模型 $M\in\mathcal{M}^+$，定义

$$\widetilde{\textsf{p-dec}}^{\mathrm{c}}_{\varepsilon}(\mathcal{M},\overline{M})=\inf_{p,q\in\Delta(\Pi)}\sup_{M\in\mathcal{H}_{p,\varepsilon}(\overline{M})\cap\mathcal{H}_{q,\varepsilon}(\overline{M})}\mathbb{E}_{\pi\sim p}[f^M(\pi_M)-f^M(\pi)],\tag{94}$$

约定当 $\mathcal{H}_{p,\varepsilon}(M)\cap\mathcal{H}_{q,\varepsilon}(M)=\varnothing$ 时，上述值为零。对于所有 $M\in\mathcal{M}^+$ 和 $\varepsilon>0$，我们有

$$\widetilde{\textsf{p-dec}}^{\mathrm{c}}_{\varepsilon}(\mathcal{M},\overline{M})\le\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M},\overline{M})\le\widetilde{\textsf{p-dec}}^{\mathrm{c}}_{\sqrt{2}\varepsilon}(\mathcal{M},\overline{M}).\tag{95}$$

引理 D.3 的证明。第一个不等式显然成立。对于第二个不等式，我们有

$$\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M},\overline{M})\le\inf_{p,q\in\Delta(\Pi)}\sup_{M\in\mathcal{H}_{\frac{1}{2}p+\frac{1}{2}q,\varepsilon}(\overline{M})}\mathbb{E}_{\pi\sim p}[f^M(\pi_M)-f^M(\pi)],$$

注意到，对于任意的 $\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}$ 极小化器 q，代入 $q'=\frac{1}{2}p+\frac{1}{2}q$ 即可得到一个上界。结果由此成立，因为 $\mathcal{H}_{\frac{1}{2}p+\frac{1}{2}q,\ \varepsilon}(M)\subseteq\mathcal{H}_{p,\ \sqrt{2}\varepsilon}(M)\cap\mathcal{H}_{q,\ \sqrt{2}\varepsilon}(M)$。□

D.2 决策估计系数的其他性质

D.2.1 局部化

以下 结果是对 命题 4.7 的扩展，可处理随机化估计量 s.

命题 D.1。设给定 α、$\gamma>0$ 和 $\nu\in\Delta(\mathcal{M})$。令 $M_\nu=\mathbb{E}_{M'\sim\nu}[M']$。对于所有 $\varepsilon>0$，均有

$$\textsf{r-dec}^{\mathrm{o,rnd}}_{\gamma}(\mathcal{M}_\alpha(\overline{M}_\nu)\cup\{\overline{M}_\nu\},\nu)\le\textsf{r-dec}^{\mathrm{c,rnd}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}_\nu\},\nu)+\max\left\{0,\ \alpha+\frac{1}{2\gamma}-\frac{\gamma\varepsilon^2}{2}\right\}.\tag{96}$$

由此特别可得

$$\textsf{r-dec}^{\mathrm{o,rnd}}_{\gamma}(\mathcal{M}_\alpha(\overline{M}_\nu)\cup\{\overline{M}_\nu\},\nu)\le\textsf{r-dec}^{\mathrm{c,rnd}}_{\sqrt{2\alpha/\gamma}}(\mathcal{M}\cup\{\overline{M}_\nu\},\nu)+\frac{1}{2\gamma}.\tag{97}$$

命题证明 命题 D.1。 固定 $\nu \in \Delta(\mathcal{M})$ 和 $\varepsilon > 0$，并令 $p \in \Delta(\Pi)$ 为 $\mathsf{r\text{-}dec}^{\mathrm{c\ rnd}}_{\varepsilon}(\mathcal{M} \cup \{M_\nu\}, \nu)$ 的极小化器。任取 $M \in \mathcal{M}_\alpha(M_\nu) \cup \{M_\nu\}$。我们通过考虑两种情况来界定 p 下的遗憾。

情况 *1.* 若 $\mathbb{E}_{M\sim\nu}\mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\big(M(\pi)\ M(\pi)\big)\right] \le \varepsilon^2$，则根据该$\mathbb{E}_{\pi\sim p}\left[f^M(\pi_M) - f^M(\pi)\right] \le \mathsf{r\text{-}dec}^{\mathrm{c\ rnd}}_{\varepsilon}(\mathcal{M} \cup \{M_\nu\}, \nu)$ 的定义 $\mathsf{r\text{-}dec}^{\mathrm{c\ rnd}}_{\varepsilon}(\mathcal{M} \cup \{M_\nu\}, \nu)$ 可得。 —

情况 *2.* 对于第二种情况，假设 $\mathbb{E}_{M\sim\nu}\mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\big(M(\pi)\ M(\pi)\big)\right] > \varepsilon^2$—现在我们计算

$$\begin{aligned}
\mathbb{E}_{\pi\sim p}\left[f^M(\pi_M) - f^M(\pi)\right] &\le \alpha + \mathbb{E}_{\pi\sim p}\left[f^{\overline{M}_\nu}(\pi_{\overline{M}_\nu}) - f^M(\pi)\right] \\
&\le \alpha + \mathbb{E}_{\pi\sim p}\left[f^{\overline{M}_\nu}(\pi_{\overline{M}_\nu}) - f^{\overline{M}_\nu}(\pi)\right] + \frac{1}{2\gamma} + \frac{\gamma}{2}\cdot\mathbb{E}_{\pi\sim p}\left[(f^M(\pi) - f^{\overline{M}_\nu}(\pi))^2\right] \\
&\le \alpha + \mathsf{r\text{-}dec}^{\mathrm{c,rnd}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}_\nu\},\nu) + \frac{1}{2\gamma} + \frac{\gamma}{2}\cdot\mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}_\nu(\pi)\big)\right], \\
&\le \alpha + \mathsf{r\text{-}dec}^{\mathrm{c,rnd}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}_\nu\},\nu) + \frac{1}{2\gamma} + \frac{\gamma}{2}\cdot\mathbb{E}_{\overline{M}\sim\nu}\mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right],
\end{aligned}$$

其中第二个不等式使用了杨氏不等式，最后一个不等式则利用了平方海灵格距离的凸性。重新整理后，我们得到

$$\begin{aligned}
&\mathbb{E}_{\pi\sim p}\left[f^M(\pi_M) - f^M(\pi) - \gamma\cdot\mathbb{E}_{\overline{M}\sim\nu}\big[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\big]\right] \\
&\le \alpha + \mathsf{r\text{-}dec}^{\mathrm{c,rnd}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\nu) + \frac{1}{2\gamma} - \frac{\gamma}{2}\cdot\mathbb{E}_{\overline{M}\sim\nu}\mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right] \\
&\le \alpha + \mathsf{r\text{-}dec}^{\mathrm{c,rnd}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}_\nu\},\overline{M}_\nu) + \frac{1}{2\gamma} - \frac{\gamma\varepsilon^2}{2}.
\end{aligned}$$

回想一下， M可以是 $\mathcal{M}_\alpha(M_\nu) \cup \{M_\nu\}$中的任意模型，我们得到

$$\begin{aligned}
\mathsf{r\text{-}dec}^{\mathrm{o,rnd}}_{\gamma}(\mathcal{M}_\alpha(\overline{M}_\nu)\cup\{\overline{M}_\nu\},\nu) &\le \max\left\{\mathsf{r\text{-}dec}^{\mathrm{c,rnd}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}_\nu\},\nu),\ \alpha + \frac{1}{2\gamma} + \mathsf{r\text{-}dec}^{\mathrm{c,rnd}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}_\nu\},\nu) - \frac{\gamma\varepsilon^2}{2}\right\} \\
&= \mathsf{r\text{-}dec}^{\mathrm{c,rnd}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}_\nu\},\nu) + \max\left\{0,\ \alpha + \frac{1}{2\gamma} - \frac{\gamma\varepsilon^2}{2}\right\}.
\end{aligned}$$

□

D.2.2 凸性在 PAC DEC 中的作用

对于 $\nu \in \Delta(\mathcal{M})$，我们定义 PAC DEC 的“随机化”变体，类似于 第 4 节中引入的变体，具体如下：

$$\mathsf{p\text{-}dec}^{\mathrm{c,rnd}}_{\varepsilon}(\mathcal{M},\nu) = \inf_{p,q\in\Delta(\Pi)}\sup_{M\in\mathcal{M}}\left\{\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] \mid \mathbb{E}_{\overline{M}\sim\nu}\mathbb{E}_{\pi\sim q}\big[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\big] \le \varepsilon^2\right\}, \tag{98}$$

$$\mathsf{p\text{-}dec}^{\mathrm{o,rnd}}_{\gamma}(\mathcal{M},\nu) = \inf_{p,q\in\Delta(\Pi)}\sup_{M\in\mathcal{M}}\left\{\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] - \gamma\cdot\mathbb{E}_{\overline{M}\sim\nu}\mathbb{E}_{\pi\sim q}\big[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\big]\right\}. \tag{99}$$

以下结果给出了 公式 (55) 的 PAC 对应结果，见 命题 4.11。

命题 D.2。 假设 假设 *A.1* 成立。对于所有 $\gamma > 0$，有

$$\sup_{\overline{M}\in\mathcal{M}^+}\mathsf{p\text{-}dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\overline{M}) \le \sup_{\nu\in\Delta(\mathcal{M})}\mathsf{p\text{-}dec}^{\mathrm{o,rnd}}_{\gamma/4}(\mathcal{M},\nu) \le \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})}\mathsf{p\text{-}dec}^{\mathrm{o}}_{\gamma/4}(\mathcal{M},\overline{M}). \tag{100}$$

可以通过调整 公式 (56) 的证明，证明其 PAC 类比结果命题 4.11；我们不列出这一结果。

命题 D.2的证明。 设 $M \in \mathcal{M}^+$ 和 $\overline{\gamma} > 0$ 已给定。我们首先证明不等式 (100)。根据假设 A.1，有

$$\textsf{p-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) = \sup_{\mu \in \Delta(\mathcal{M})} \inf_{p,q \in \Delta(\Pi)} \mathbb{E}_{M\sim\mu}\big[\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] - \gamma \cdot \mathbb{E}_{\pi\sim q}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big]\big].$$

由于海林格距离满足三角不等式，对于所有 $\pi \in \Pi$，我们有

$$\begin{aligned}\mathbb{E}_{M,M'\sim\mu}\big[D^2_{\mathsf{H}}(M(\pi), M'(\pi))\big] &\le 2\,\mathbb{E}_{M\sim\mu}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] + 2\,\mathbb{E}_{M'\sim\mu}\big[D^2_{\mathsf{H}}\big(M'(\pi), \overline{M}(\pi)\big)\big] \\ &= 4\,\mathbb{E}_{M\sim\mu}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big].\end{aligned}$$

由此可得

$$\begin{aligned}\textsf{p-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) &\le \sup_{\mu \in \Delta(\mathcal{M})} \inf_{p,q \in \Delta(\Pi)} \mathbb{E}_{M\sim\mu}\Big[\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] - \frac{\gamma}{4} \cdot \mathbb{E}_{M'\sim\mu}\,\mathbb{E}_{\pi\sim q}\big[D^2_{\mathsf{H}}(M(\pi), M'(\pi))\big]\Big] \\ &\le \sup_{\nu \in \Delta(\mathcal{M})} \sup_{\mu \in \Delta(\mathcal{M})} \inf_{p,q \in \Delta(\Pi)} \mathbb{E}_{M\sim\mu}\Big[\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] - \frac{\gamma}{4} \cdot \mathbb{E}_{M'\sim\nu}\,\mathbb{E}_{\pi\sim q}\big[D^2_{\mathsf{H}}(M(\pi), M'(\pi))\big]\Big] \\ &\le \sup_{\nu \in \Delta(\mathcal{M})} \inf_{p,q \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \Big\{\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] - \frac{\gamma}{4} \cdot \mathbb{E}_{M'\sim\nu}\,\mathbb{E}_{\pi\sim q}\big[D^2_{\mathsf{H}}(M(\pi), M'(\pi))\big]\Big\} \\ &= \sup_{\nu \in \Delta(\mathcal{M})} \textsf{p-dec}^{\mathrm{o,rnd}}_{\gamma/4}(\mathcal{M}, \nu).\end{aligned}$$

詹森不等式进一步蕴含，$\sup_{\nu\in\Delta(\mathcal{M})} \textsf{p-dec}^{\mathrm{o\ rnd}}_{\gamma/4}(\mathcal{M}, \nu) \le \sup_{M\in\mathrm{co}(\mathcal{M})} \textsf{p-dec}^{\mathrm{o}}_{\gamma/4}(\mathcal{M}, M)$。 □

D.2.3 带贪心决策的 PACDEC

对于模型类 $\mathcal{M}$ 和参考模型 $M \in \mathcal{M}^+$，定义

$$\textsf{p-dec}^{\mathrm{c,greedy}}_{\varepsilon}(\mathcal{M}, \overline{M}) = \inf_{q \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \big\{f^M(\pi_M) - f^M(\pi_{\overline{M}}) \mid \mathbb{E}_{\pi\sim q}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] \le \varepsilon^2\big\},$$

约定当 $\mathcal{H}_{q,\,\varepsilon}(M) = \varnothing$ 时，上述值为零。

命题 D.3。 对于所有 $\varepsilon > 0$ 和 $M \in \mathcal{M}^{+}$，我们有

$$\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \le \textsf{p-dec}^{\mathrm{c,greedy}}_{\varepsilon}(\mathcal{M}, \overline{M}) \le \textsf{p-dec}^{\mathrm{c}}_{\sqrt{3}\varepsilon}(\mathcal{M}, \overline{M}) + 4\varepsilon. \tag{101}$$

命题 D.3 的证明。显然，$\textsf{p-dec}^{\mathrm{c\ greedy}}_{\varepsilon}(\mathcal{M}, M) \ge \textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, M)$，因此我们来证明第二个不等式。给定 $M \in \mathcal{M}$ 和 $\varepsilon > 0$，令 (p, q) 为 $\widetilde{\textsf{p-dec}}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, M)$ 的极小化器。定义 $q' = \frac{1}{3}q + \frac{1}{3}p + \frac{1}{3}\mathbb{I}_{\pi_M}$。注意，$\mathcal{H}_{q',\varepsilon/\sqrt{3}}(M) \subseteq \mathcal{H}_{q,\varepsilon}(M) \cap \mathcal{H}_{p,\varepsilon}(M) \cap \mathcal{H}_{\mathbb{I}_{\pi_M},\varepsilon}(M)$。因此，对于所有 $\overline{M} \in \mathcal{H}_{q',\varepsilon/\sqrt{3}}(M)$，我们有

$$\begin{aligned}f^M(\pi_M) - f^M(\pi_{\overline{M}}) &= f^M(\pi_M) - f^{\overline{M}}(\pi_{\overline{M}}) + (f^{\overline{M}}(\pi_{\overline{M}}) - f^M(\pi_{\overline{M}})) \\ &\le f^M(\pi_M) - f^{\overline{M}}(\pi_{\overline{M}}) + \varepsilon \\ &\le \mathbb{E}_{\pi\sim p}\big[f^M(\pi_M) - f^{\overline{M}}(\pi)\big] + \varepsilon \\ &\le \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] + 2\varepsilon \\ &\le \widetilde{\textsf{p-dec}}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) + 2\varepsilon \\ &\le \textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) + 2\varepsilon,\end{aligned}$$

其中第一个和第三个不等式用到了 引理 D.1，最后一个不等式用到了 引理 D.3。 □

D.3 省略的证明

命题 命题 4.1 的证明。 设 $M \in \mathcal{M}^+$ 已给定。我们首先在以下假设下证明一个更一般的结果：存在某个 $\delta > 0$，$f^M(\pi_M) \le f^M(\pi_{\overline{M}}) + \delta$ 对所有 $\overline{M} \in \mathcal{M}$：

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \le \delta + \inf_{\gamma > 0} \left\{ \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \vee 0 + 4\gamma\varepsilon^2 + (4\gamma)^{-1} \right\}. \tag{102}$$

然后，在证明的最后，我们将说明不失一般性，可以取 $\delta = O(\varepsilon)$。

给定 $\gamma > 0$，令 p_0 为 $\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, M)$ 的极小化器，使得

$$\sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p_0}\left[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big) \right] \le \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}).$$

令 $M^\star := \arg\min_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p_0}\left[D^2_{\mathsf{H}}\big(M(\pi),\ \overline{M}(\pi)\big) \right]$，并令 $\Delta^2 := \mathbb{E}_{\pi \sim p_0}\left[D^2_{\mathsf{H}}\big(M^\star(\pi),\ \overline{M}(\pi)\big) \right]$。

我们将通过采用分布来界定受约束 DEC，即 $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\overline{\mathcal{M}} \cup \overline{\{M\}}, M)$

$$p := (1-q) \cdot \mathbb{I}_{\pi_{\overline{M}}} + q \cdot p_0,$$

其中

$$q := \frac{2\varepsilon^2}{\Delta^2} \wedge 1.$$

在继续之前，我们先陈述一个基本的技术引理。

引理 D.4. 分布 p_0 满足

$$\mathbb{E}_{\pi \sim p_0}\left[f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi) \right] \le \delta + \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) + (4\gamma)^{-1} + 2\gamma\Delta^2. \tag{103}$$

我们通过考虑两种情况来界定 p 的受约束 DEC 值。

情况 *1:* $q = 1$. 如果 $q = 1$，那么 $\Delta^2 \le 2\varepsilon^2$，且 $p = p_0$。对于模型 $M \in \mathcal{H}_{p,\varepsilon}(\overline{M})$，$p_0$ 的定义意味着

$$\begin{aligned}
\mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi)] &\le \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) + \gamma \cdot \mathbb{E}_{\pi \sim p}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big) \right] \\
&\le \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) + \gamma\varepsilon^2.
\end{aligned}$$

对于模型 M，引理 D.4 意味着

$$\mathbb{E}_{\pi \sim p}\left[f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi) \right] \le \delta + \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) + (4\gamma)^{-1} + 4\gamma\varepsilon^2.$$

情况 *2:* $q < 1$. 如果 $q < 1$，那么对于所有 $M \in \mathcal{M}$，我们有

$$\begin{aligned}
\mathbb{E}_{\pi \sim p}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big) \right] &\ge q \cdot \mathbb{E}_{\pi \sim p_0}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big) \right] \\
&\ge q \cdot \mathbb{E}_{\pi \sim p_0}\left[D^2_{\mathsf{H}}\big(M^\star(\pi), \overline{M}(\pi)\big) \right] = 2\varepsilon^2 > \varepsilon^2,
\end{aligned}$$

其中，第二个不等式利用了 $M^\star$ 是 $\mathbb{E}_{\pi \sim p_0}\left[D^2_{\mathsf{H}}\big(M(\pi)\ M(\pi)\big) \right]$ 的最小值，而最后一个不等式利用了只要 $q < 1$，就有 $q = \frac{2\varepsilon^2}{\Delta^2}$。由此可得 $\mathcal{H}_{p,\ \varepsilon}(M) \cup \{M\} = \{M\}$，因此我们只需界定 M 下分布 p 的遗憾。为此，我们观测到

$$\begin{aligned}
\mathbb{E}_{\pi \sim p}\left[g^{\overline{M}}(\pi) \right] = q \cdot \mathbb{E}_{\pi \sim p_0}\left[g^{\overline{M}}(\pi) \right] &\le q \cdot \left(\delta + \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) + (4\gamma)^{-1} + 2\gamma\Delta^2 \right) \\
&\le \delta + \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \vee 0 + (4\gamma)^{-1} + q \cdot 2\gamma\Delta^2 \\
&= \delta + \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \vee 0 + (4\gamma)^{-1} + 4\gamma\varepsilon^2,
\end{aligned}$$

其中第一个不等式用到了引理 D.4，最后的等式则用到了 $q = \frac{2\varepsilon^2}{\Delta^2}$。

最后，我们已经证明了 我们已确立

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M})\le\delta+\inf_{\gamma>0}\left\{\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\overline{M})\vee 0+4\gamma\varepsilon^2+(4\gamma)^{-1}\right\}$$

每当 $f^{M}(\pi_{M})\le f^{M}(\pi_{M})+\delta$对所有 $M\in\mathcal{M}$。最后，我们借助引理 D.2 应用于类别$\mathcal{M}\cup\{M\}$，这意味着

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M})\le\textsf{r-dec}^{\mathrm{c}}_{\sqrt{2}\varepsilon}(\mathcal{M}'\cup\{\overline{M}\},\overline{M})+\sqrt{2}\varepsilon,\tag{104}$$

其中 $\mathcal{M}'=\left\{M\in\mathcal{M}\mid f^{M}(\pi_{M})\le f^{M}(\pi_{M})+\sqrt{2}\varepsilon\right\}$。将(102)应用于量$\textsf{r-dec}^{\mathrm{c}}_{\sqrt{2}\varepsilon}(\mathcal{M}'\cup\{M\},\ M)$并与(104)结合可得

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M})\le 2\sqrt{2}\varepsilon+\inf_{\gamma>0}\left\{\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\overline{M})\vee 0+8\gamma\varepsilon^2+(4\gamma)^{-1}\right\}.$$

为了稍微简化这一结果，我们考虑两种情况。如果$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\ M)\le(4\gamma)^{-1}$，那么选择 $\gamma=(4\varepsilon)^{-1}$可得 $\inf_{\gamma>0}\{\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\ M)+4\gamma\varepsilon^2+(4\gamma)^{-1}\}\le 4\varepsilon$。否则，我们有 $\inf_{\gamma>0}\{\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\ M)+4\gamma\varepsilon^2+(4\gamma)^{-1}\}\le\inf_{\gamma>0}\{2\cdot\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\ M)+8\gamma\varepsilon^2\}$。

□

引理的证明 引理 D.4。 观测到

$$\begin{aligned}\mathbb{E}_{\pi\sim p_0}\left[f^{\overline{M}}(\pi_{\overline{M}})-f^{\overline{M}}(\pi)\right]&\le\delta+\mathbb{E}_{\pi\sim p_0}\left[f^{M^\star}(\pi_{M^\star})-f^{\overline{M}}(\pi)\right]\\&\le\delta+\mathbb{E}_{\pi\sim p_0}\left[f^{M^\star}(\pi_{M^\star})-f^{M^\star}(\pi)\right]+\Delta,\end{aligned}$$

其中，第一个不等式利用了对所有 $M\in\mathcal{M}$，$f^{M}(\pi_{\overline{M}})\le f^{M}(\pi_{M})+\delta$这一事实，第二个不等式则利用了引理 D.1。此外，根据 p_0 的定义可知，$\mathbb{E}_{\pi\sim p_0}[f^{M^\star}(\pi_{M^\star})-f^{M^\star}(\pi)]\le\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\ M)+\gamma\Delta^2$，因此我们有

$$\begin{aligned}\mathbb{E}_{\pi\sim p_0}\left[f^{\overline{M}}(\pi_{\overline{M}})-f^{\overline{M}}(\pi)\right]&\le\delta+\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\overline{M})+\gamma\Delta^2+\Delta\\&\le\delta+\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\overline{M})+2\gamma\Delta^2+(4\gamma)^{-1}.\end{aligned}$$

□

命题的证明 命题 4.2。 我们首先证明不等式 (38)。令 $\varepsilon>0$ 和 $M\in\mathcal{M}^{+}$ 固定。利用拉格朗日乘子法，我们有

$$\begin{aligned}\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M},\overline{M})&=\inf_{p,q\in\Delta(\Pi)}\sup_{M\in\mathcal{M}}\left\{\mathbb{E}_{\pi\sim p}[g^{M}(\pi)]\mid\mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\left(M(\pi),\overline{M}(\pi)\right)\right]\le\varepsilon^2\right\}\\&=\inf_{p,q\in\Delta(\Pi)}\sup_{M\in\mathcal{M}}\max\left\{\inf_{\gamma\ge0}\left\{\mathbb{E}_{\pi\sim p}[g^{M}(\pi)]-\gamma\left(\mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\left(M(\pi),\overline{M}(\pi)\right)\right]-\varepsilon^2\right)\right\},0\right\}\\&\le\inf_{\gamma\ge0}\inf_{p,q\in\Delta(\Pi)}\sup_{M\in\mathcal{M}}\max\left\{\mathbb{E}_{\pi\sim p}[g^{M}(\pi)]-\gamma\left(\mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\left(M(\pi),\overline{M}(\pi)\right)\right]-\varepsilon^2\right),0\right\}\\&\le\inf_{\gamma\ge0}\left\{\textsf{p-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\overline{M})\vee 0+\gamma\varepsilon^2\right\}.\end{aligned}$$

现在我们证明不等式 (39)。令 $\gamma\ge1$ 和 $M\in\mathcal{M}^{+}$ 固定。对于整数 $i\ge0$，定义 $\varepsilon_i=2^{-i/2}$。对于每个 $i\ge0$，令 $(p_i,\ q_i)$ 表示以下表达式的一个极小化器：

$$\inf_{p_i,q_i\in\Delta(\Pi)}\sup_{M\in\mathcal{H}_{q_i,\varepsilon_i}(\overline{M})}\mathbb{E}_{\pi\sim p_i}[g^{M}(\pi)]=\textsf{p-dec}^{\mathrm{c}}_{\varepsilon_i}(\mathcal{M},\overline{M}).$$

回想一下 $L=2\lceil\log 2\gamma\rceil$，令

$$q=\frac{1}{2}\cdot\mathbb{I}_{\pi_{\overline{M}}}+\frac{q_0+\cdots+q_{L-1}}{4L}+\frac{p_0+\cdots+p_{L-1}}{4L},\quad\text{and}\quad p=\mathbb{I}_{\pi_{\overline{M}}}.$$

考虑任意一个 $M \in \mathcal{M}$。我们将给出该值的上界

$$\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] - 4\gamma L \cdot \mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right].$$

选择尽可能大的 $j \in \{0, \dots, L-1\}$，使得

$$\mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right] \le \frac{\varepsilon_j^2}{4L}. \tag{105}$$

如果不存在这样的索引 j，那么必有 $\mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\left(M(\pi)\ M(\pi)\right)\right] > 1/(4L)$。此时，我们有

$$\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] - 4\gamma L \cdot \mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right] \le 1 - \gamma \le 0 \le \mathsf{p\text{-}dec}^{\mathrm{c}}_0(\mathcal{M}, \overline{M}).$$

假设接下来 $0 \le j \le L-1$ 满足 (105) 的情况存在。如果 $j < L-1$，由于我们选择了 j 的最大可能值，因此有 $\mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\left(M(\pi)\ M(\pi)\right)\right] \ge \frac{\varepsilon_j^2}{8L}$。此外，无论$j \in \{0, 1, \dots, L-1\}$的值如何，根据 q的定义，我们有

$$\mathbb{E}_{\pi\sim q_j}\left[D^2_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right] \le 4L \cdot \mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right] \le \varepsilon_j^2,$$

and

$$\mathbb{E}_{\pi\sim p_j}\left[D^2_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right] \le 4L \cdot \mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right] \le \varepsilon_j^2,$$

也就是说，$M \in \mathcal{H}_{p_j,\ \overline{\varepsilon_j}}(M) \cap \mathcal{H}_{q_j,\ \overline{\varepsilon_j}}(M)$。由此可得

$$\begin{aligned} f^M(\pi_M) - f^{\overline{M}}(\pi_{\overline{M}}) &\le \mathbb{E}_{\pi\sim p_j}\left[f^M(\pi_M) - f^{\overline{M}}(\pi)\right] \\ &\le \mathbb{E}_{\pi\sim p_j}[f^M(\pi_M) - f^M(\pi)] + \varepsilon_j \\ &\le \mathsf{p\text{-}dec}^{\mathrm{c}}_{\varepsilon_j}(\mathcal{M}, \overline{M}) + \varepsilon_j, \end{aligned}$$

其中，第二个不等式利用了 $M \in \mathcal{H}_{p_j,\varepsilon_j}(M)$，最后一个不等式利用了 $M \in \mathcal{H}_{q_j,\varepsilon_j}(M)$。因此，我们可以计算

$$\begin{aligned} &\mathbb{E}_{\pi\sim p}\left[f^M(\pi_M) - f^M(\pi)\right] - 4\gamma L \cdot \mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right] \\ &\le f^M(\pi_M) - f^M(\pi_{\overline{M}}) - \frac{4\gamma L}{8L} \cdot \varepsilon_j^2 \cdot \mathbb{1}\{j < L-1\} \\ &\le f^{\overline{M}}(\pi_{\overline{M}}) - f^M(\pi_{\overline{M}}) + \mathsf{p\text{-}dec}^{\mathrm{c}}_{\varepsilon_j}(\mathcal{M}, \overline{M}) + \varepsilon_j - \frac{\gamma}{2}\varepsilon_j^2 \cdot \mathbb{1}\{j < L-1\} \\ &\le \frac{1}{2\gamma} + \frac{\gamma}{2} \cdot \left(f^{\overline{M}}(\pi_{\overline{M}}) - f^M(\pi_{\overline{M}})\right)^2 + \mathsf{p\text{-}dec}^{\mathrm{c}}_{\varepsilon_j}(\mathcal{M}, \overline{M}) + \varepsilon_j - \frac{\gamma}{2}\varepsilon_j^2 \cdot \mathbb{1}\{j < L-1\} \\ &\le \frac{1}{2\gamma} + \gamma \cdot \mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right] + \mathsf{p\text{-}dec}^{\mathrm{c}}_{\varepsilon_j}(\mathcal{M}, \overline{M}) + \varepsilon_j - \frac{\gamma}{2}\varepsilon_j^2 \cdot \mathbb{1}\{j < L-1\} \\ &\le \frac{1}{\gamma} + \gamma \cdot \mathbb{E}_{\pi\sim q}\left[D^2_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right] + \mathsf{p\text{-}dec}^{\mathrm{c}}_{\varepsilon_j}(\mathcal{M}, \overline{M}) + \varepsilon_j - \frac{\gamma}{2}\varepsilon_j^2, \end{aligned}$$

其中，最后一个不等式利用了 $\varepsilon_{L-1}^2 \le 1/\gamma^2$，因为 $L \ge 2\log(2\gamma)$。重新整理可得

$$\begin{aligned} &\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] - \gamma \cdot (4L+1) \cdot \mathbb{E}_{\pi\sim q}[D^2_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)] \\ &\le \frac{1}{\gamma} + \mathsf{p\text{-}dec}^{\mathrm{c}}_{\varepsilon_j}(\mathcal{M}, \overline{M}) + \varepsilon_j - \frac{\gamma}{2} \cdot \varepsilon_j^2 \\ &\le \frac{2}{\gamma} + \mathsf{p\text{-}dec}^{\mathrm{c}}_{\varepsilon_j}(\mathcal{M}, \overline{M}) - \frac{\gamma}{4} \cdot \varepsilon_j^2, \end{aligned} \tag{106}$$

如所需。 □

命题 命题 4.3的证明。 给定 $\gamma > 0$ 和 $M \in \overline{\mathcal{M}}^{+}$。固定稍后再选择的 $\varepsilon > 0$，并令 p 为 $\mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, M)$ 的极小化器。考虑 $M \in \mathcal{M}$ 的偏移 DEC 的值：

$$\mathbb{E}_{\pi\sim p}\left[f^{M}(\pi_{M}) - f^{M}(\pi) - \gamma \cdot D^{2}_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right].$$

我们考虑两种情况。首先，如果 $M \in \mathcal{H}_{p,\varepsilon}(\overline{M})$，显然有

$$\mathbb{E}_{\pi\sim p}\left[f^{M}(\pi_{M}) - f^{M}(\pi) - \gamma \cdot D^{2}_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right] \le \mathbb{E}_{\pi\sim p}[f^{M}(\pi_{M}) - f^{M}(\pi)] \le \mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M})。$$

另一方面，，如果 $M \notin \mathcal{H}_{p,\varepsilon}(M)$，则有 $\mathbb{E}_{\pi\sim p}\left[D^{2}_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right] \ge \varepsilon^{2}$，并且由于 $g^{M} \le 1$，可得

$$\mathbb{E}_{\pi\sim p}\left[f^{M}(\pi_{M}) - f^{M}(\pi) - \gamma \cdot D^{2}_{\mathsf{H}}\left(M(\pi), \overline{M}(\pi)\right)\right] \le 1 - \gamma\varepsilon^{2} \le 0$$

通过选择 $\varepsilon = \gamma^{-1/2}$。我们得出

$$\mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \le \mathsf{r\text{-}dec}^{\mathrm{c}}_{\gamma^{-1/2}}(\mathcal{M}, \overline{M}).$$

□

命题 4.4 的证明。 回顾 $\mathcal{M}^{\alpha,\beta}$ 中定义的模型类 例 5.1，其参数为$\alpha \in (0, 1/2], \beta \in (0,1)$ $A \in \mathbb{N}$。对于任意选择的 α, β 和 A，我们将在下文指定具体取值。引理 E.2 表明，对于所有 $\varepsilon > 0$,

$$\mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}^{\alpha,\beta}) = \sup_{\overline{M}\in\mathrm{co}(\mathcal{M}^{\alpha,\beta})} \mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}^{\alpha,\beta} \cup \{\overline{M}\}, \overline{M}) \le O\left(\frac{\varepsilon^{2}}{\beta}\right).$$

另一方面，引理 E.1 表明，对于 $M = \widetilde{M} \in \mathcal{M}^{\alpha,\beta}$ 的选择，对所有 $\gamma > 0$，都有

$$\mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}^{\alpha,\beta}, \widetilde{M}) \ge \frac{\alpha}{2 + 8\gamma\beta} - 4\gamma/A.$$

给定 $\gamma > 0$，我们选择 $\alpha = 1/2$, $\beta = 1/\sqrt{\gamma}$ 和 $A = 256\gamma^{2}/\beta$。那么，由此得到的模型类 $\mathcal{M} = \mathcal{M}^{\alpha,\beta}$ 满足 $\mathsf{r\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \le O(\varepsilon^{2}\gamma^{1/2})$，对所有 $\varepsilon > 0$ 均成立，然而

$$\sup_{\overline{M}\in\mathcal{M}} \mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \ge \frac{1}{4 + 16\gamma\beta} - \frac{4\gamma}{A} \ge \frac{1}{32} \cdot \left(\frac{1}{\gamma^{1/2}} \wedge 1\right) - \frac{4\gamma}{A} \ge \frac{1}{64} \cdot \left(\frac{1}{\gamma^{1/2}} \wedge 1\right) \ge \Omega(\gamma^{-1/2}).$$

□

证明 命题 4.5。 回顾 (94) 中 $\widetilde{\mathsf{p\text{-}dec}}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, M)$ 的定义。设给定 $\varepsilon > 0$ 和 $M \in \mathcal{M}^{+}$，并令 (p', q') 为 $\widetilde{\mathsf{p\text{-}dec}}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, M)$ 的极小化器。那么对所有 $M \in \mathcal{H}_{p',\varepsilon}(M) \cap \mathcal{H}_{q',\varepsilon}(M)$，都有

$$\begin{aligned}
\widetilde{\mathsf{p\text{-}dec}}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) &\ge \mathbb{E}_{\pi\sim p'}[f^{M}(\pi_{M}) - f^{M}(\pi)] \\
&\ge \mathbb{E}_{\pi\sim p'}\left[f^{M}(\pi_{M}) - f^{\overline{M}}(\pi)\right] - \varepsilon \\
&\ge f^{M}(\pi_{M}) - f^{\overline{M}}(\pi_{\overline{M}}) - \varepsilon,
\end{aligned}$$

其中第二个不等式用到了引理 D.1。也就是说，若定义 $\alpha = \varepsilon + \widetilde{\mathsf{p\text{-}dec}}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, M)$，则有$\mathcal{H}_{p',\varepsilon}(M) \cap \mathcal{H}_{q',\varepsilon}(M) \subseteq \mathcal{M}_{\alpha}(M)$。现在，令 $(\bar{p}, \bar{q})$ 为 $\mathsf{p\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}_{\alpha}(M), M)$ 的极小化器，并令 $\bar{q} = \frac{1}{3}q + \frac{1}{3}q' + \frac{1}{3}p'$。我们有

$$\begin{aligned}
\mathsf{p\text{-}dec}^{\mathrm{c}}_{\varepsilon/\sqrt{3}}(\mathcal{M}, \overline{M}) &\le \sup_{M\in\mathcal{H}_{\bar{q},\varepsilon/\sqrt{3}}(\overline{M})} \mathbb{E}_{\pi\sim p}[f^{M}(\pi_{M}) - f^{M}(\pi)] \\
&\le \sup_{M\in\mathcal{H}_{q,\varepsilon}(\overline{M})\cap\mathcal{H}_{p',\varepsilon}(\overline{M})\cap\mathcal{H}_{q',\varepsilon}(\overline{M})} \mathbb{E}_{\pi\sim p}[f^{M}(\pi_{M}) - f^{M}(\pi)] \\
&\le \sup_{M\in\mathcal{H}_{q,\varepsilon}(\overline{M})\cap\mathcal{M}_{\alpha}(\overline{M})} \mathbb{E}_{\pi\sim p}[f^{M}(\pi_{M}) - f^{M}(\pi)] \\
&= \mathsf{p\text{-}dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}_{\alpha}(\overline{M}), \overline{M}).
\end{aligned}$$

最后，利用 引理 D.3，我们有 $\alpha \le \varepsilon + \text{p-dec}^{\text{c}}_{\sqrt{2}\varepsilon}(\mathcal{M}, M)$。

□

证明 命题 4.6。 我们首先证明以下结果，该结果不需要任何正则性条件。

引理 D.5。 任取 $M \in \mathcal{M}^+$ 和 $\varepsilon > 0$，并令 $\alpha' := \varepsilon + \text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}, M)$。则对于任意常数$C_{\text{reg}} \ge \sqrt{2}$，有

$$\text{r-dec}^{\text{c}}_{\varepsilon/C_{\text{reg}}}(\mathcal{M}, \overline{M}) \le \frac{1}{C^2_{\text{reg}}} \text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) + \text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}_{\alpha'}(\overline{M}), \overline{M}).$$

证明 引理 D.5。 给定 $\varepsilon > 0$ 和 M，令 p 为使$\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}, M)$取到该值的分布。那么对于所有 $M \in \mathcal{H}_{p,\varepsilon}(M)$，都有

$$\begin{aligned}
\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) &\ge \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] \\
&\ge \mathbb{E}_{\pi\sim p}\left[f^M(\pi_M) - f^{\overline{M}}(\pi)\right] - \varepsilon \\
&\ge f^M(\pi_M) - f^{\overline{M}}(\pi_{\overline{M}}) - \varepsilon,
\end{aligned}$$

其中第二个不等式 中的等号用到了引理 D.1。因此，对于 $\alpha' := \varepsilon + \text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}, M)$，有 $\mathcal{H}_{p,\varepsilon}(\overline{M}) \subseteq \mathcal{M}_{\alpha'}(\overline{M})$. 现在，令 p' 为$\text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}_{\alpha'}(M)\ M)$ 的极小化器。令 $\bar{p} = \frac{1}{C^2_{\text{reg}}} p + \left(1 - \frac{1}{C^2_{\text{reg}}}\right) p'$。利用

$$\frac{1}{C^2_{\text{reg}}} \cdot \left(1 - \frac{1}{C^2_{\text{reg}}}\right)^{-1} \le 1$$

每当 $C_{\text{reg}} \ge \sqrt{2}$，我们就有

$$\begin{aligned}
\text{r-dec}^{\text{c}}_{\varepsilon/C_{\text{reg}}}(\mathcal{M}, \overline{M}) &\le \sup_{M \in \mathcal{H}_{\bar{p},\varepsilon/C_{\text{reg}}}(\overline{M})} \mathbb{E}_{\pi\sim\bar{p}}[f^M(\pi_M) - f^M(\pi)] \\
&\le \sup_{M \in \mathcal{H}_{p,\varepsilon}(\overline{M}) \cap \mathcal{H}_{p',\varepsilon}(\overline{M})} \mathbb{E}_{\pi\sim\bar{p}}[f^M(\pi_M) - f^M(\pi)] \\
&\le \frac{1}{C^2_{\text{reg}}} \sup_{M \in \mathcal{H}_{p,\varepsilon}(\overline{M})} \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] + \sup_{M \in \mathcal{H}_{p,\varepsilon}(\overline{M}) \cap \mathcal{H}_{p',\varepsilon}(\overline{M})} \mathbb{E}_{\pi\sim p'}[f^M(\pi_M) - f^M(\pi)] \\
&= \frac{1}{C^2_{\text{reg}}} \text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) + \sup_{M \in \mathcal{H}_{p,\varepsilon}(\overline{M}) \cap \mathcal{H}_{p',\varepsilon}(\overline{M})} \mathbb{E}_{\pi\sim p'}[f^M(\pi_M) - f^M(\pi)] \\
&\le \frac{1}{C^2_{\text{reg}}} \text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) + \sup_{M \in \mathcal{M}_{\alpha'}(\overline{M}) \cap \mathcal{H}_{p',\varepsilon}(\overline{M})} \mathbb{E}_{\pi\sim p'}[f^M(\pi_M) - f^M(\pi)] \\
&= \frac{1}{C^2_{\text{reg}}} \text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) + \text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}_{\alpha'}(\overline{M}), \overline{M}).
\end{aligned}$$

□

现在我们完成命题 4.6的证明。在假设的增长条件下，我们有$\text{r-dec}^{\text{c}}_{\varepsilon/C_{\text{reg}}}(\mathcal{M},\ M) \ge \frac{1}{c^2_{\text{reg}}} \text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M},\ M)$，因此重新整理引理 D.5 （其中 $\alpha' = \alpha(\varepsilon)$）的结果可得

$$\text{r-dec}^{\text{c}}_{\varepsilon/C_{\text{reg}}}(\mathcal{M}, \overline{M}) \le \left(\frac{1}{c^2_{\text{reg}}} - \frac{1}{C^2_{\text{reg}}}\right)^{-1} \cdot \text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}_{\alpha(\varepsilon)}(\overline{M}), \overline{M}).$$

该命题中的结果可通过将 ε 重新缩放为 $\varepsilon \cdot C_{\text{reg}}$ 得出。 □

为了后续使用，我们还证明了 命题 4.6的以下变体，该变体涉及类别$\mathcal{M} \cup \{M\}$。

命题 D.4 (遗憾决策估计系数的局部化；全局版本)。 考虑任意集合 $\mathcal{M}' \subseteq \mathcal{M}^+$，并假设相对于 $\mathcal{M}'$ 满足强正则性条件 *(43)* 。那么，对于所有 $\varepsilon > 0$，令 $\alpha(\varepsilon) := C_{\mathrm{reg}} \cdot \varepsilon + \sup_{M \in \mathcal{M}'} \textsf{r-dec}^{\mathrm{c}}_{C_{\mathrm{reg}} \cdot \varepsilon}(\mathcal{M} \cup \{M\}, M) \leq C_{\mathrm{reg}}^2 \cdot \big(\varepsilon + \sup_{M \in \mathcal{M}'} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{M\}, M)\big)$，我们有

$$\sup_{\overline{M} \in \mathcal{M}'} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \leq C_{\mathrm{loc}} \cdot \sup_{\overline{M} \in \mathcal{M}'} \textsf{r-dec}^{\mathrm{c}}_{C_{\mathrm{reg}} \cdot \varepsilon}(\mathcal{M}_{\alpha(\varepsilon)}(\overline{M}) \cup \{\overline{M}\}, \overline{M}),$$

其中 $C_{\mathrm{loc}} := \left(\frac{1}{c_{\mathrm{reg}}^2} - \frac{1}{C_{\mathrm{reg}}^2}\right)^{-1}$。

命题 D.4 的证明。 定义 $\alpha := \varepsilon + \sup_{M \in \mathcal{M}'} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{M\}, M)$。将 引理 D.5 应用于类别 $\mathcal{M} \cup \{M\}$，对每个 $M \in \mathcal{M}'$的取值，我们得到

$$\sup_{\overline{M} \in \mathcal{M}'} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon/C_{\mathrm{reg}}}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \leq \frac{1}{C_{\mathrm{reg}}^2} \sup_{\overline{M} \in \mathcal{M}'} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) + \sup_{\overline{M} \in \mathcal{M}'} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}_{\alpha}(\overline{M}) \cup \{\overline{M}\}, \overline{M}). \quad (107)$$

增长条件(43) 表明

$$\sup_{\overline{M} \in \mathcal{M}'} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon/C_{\mathrm{reg}}}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \geq \frac{1}{c_{\mathrm{reg}}^2} \sup_{\overline{M} \in \mathcal{M}'} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}).$$

对 (107) 重新整理可得

$$\sup_{\overline{M} \in \mathcal{M}'} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon/C_{\mathrm{reg}}}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \leq \left(\frac{1}{c_{\mathrm{reg}}^2} - \frac{1}{C_{\mathrm{reg}}^2}\right)^{-1} \cdot \sup_{\overline{M} \in \mathcal{M}'} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}_{\alpha}(\overline{M}) \cup \{\overline{M}\}, \overline{M}). \quad (108)$$

将 ε 替换为 $\varepsilon \cdot C_{\mathrm{reg}}$即可得到命题陈述中的结果。 □

命题 4.7 的证明。 固定 $M \in \mathcal{M}^+$ 和 $\varepsilon > 0$，并令 $p \in \Delta(\Pi)$ 为 $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{M\}, M)$ 的一个极小化器。

任取 $M \in \mathcal{M}_{\alpha}(M) \cup \{M\}$。通过考虑两种情况，我们对 p下的遗憾进行界定。

情况 *1*。 若 $\mathbb{E}_{\pi \sim p}\left[D^2_{\mathsf{H}}\big(M(\pi), M(\pi)\big)\right] \leq \varepsilon^2$，则 $M \in \mathcal{H}_{p,\varepsilon}(M)$，由此可得 $\mathbb{E}_{\pi \sim p}\left[f^M(\pi_M) - f^M(\pi)\right] \leq \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{M\}, M)$。

情况 *2*. 对于第二种情况，假设 $\mathbb{E}_{\pi \sim p}\left[D^2_{\mathsf{H}}\big(M(\pi), M(\pi)\big)\right] > \varepsilon^2$。现在我们计算

$$\begin{aligned}
\mathbb{E}_{\pi \sim p}\left[f^M(\pi_M) - f^M(\pi)\right] &\leq \alpha + \mathbb{E}_{\pi \sim p}\left[f^{\overline{M}}(\pi_{\overline{M}}) - f^M(\pi)\right] \\
&\leq \alpha + \mathbb{E}_{\pi \sim p}\left[f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi)\right] + \frac{1}{2\gamma} + \frac{\gamma}{2} \cdot \mathbb{E}_{\pi \sim p}\left[(f^M(\pi) - f^{\overline{M}}(\pi))^2\right] \\
&\leq \alpha + \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) + \frac{1}{2\gamma} + \frac{\gamma}{2} \cdot \mathbb{E}_{\pi \sim p}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right],
\end{aligned}$$

其中第二个不等式使用了杨氏不等式。重新整理后，我们得到

$$\begin{aligned}
\mathbb{E}_{\pi \sim p}\left[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] &\leq \alpha + \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) + \frac{1}{2\gamma} - \frac{\gamma}{2} \cdot \mathbb{E}_{\pi \sim p}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \\
&\leq \alpha + \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) + \frac{1}{2\gamma} - \frac{\gamma\varepsilon^2}{2}. \quad (109)
\end{aligned}$$

回想一下， M可以是 $\mathcal{M}_{\alpha}(M) \cup \{M\}$中的任意模型，我们得到

$$\begin{aligned}
\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\alpha}(\overline{M}) \cup \{\overline{M}\}, \overline{M}) &\leq \max\left\{\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}),\ \alpha + \frac{1}{2\gamma} + \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) - \frac{\gamma\varepsilon^2}{2}\right\} \\
&= \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) + \max\left\{0,\ \alpha + \frac{1}{2\gamma} - \frac{\gamma\varepsilon^2}{2}\right\}.
\end{aligned}$$

□

命题 命题 4.8的证明。 我们先给出受约束 DEC 的上界。令 $\varepsilon > 0$ 固定。根据 命题 4.6，我们有

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \le C_{\mathrm{loc}} \cdot \textsf{r-dec}^{\mathrm{c}}_{C_{\mathrm{reg}}\varepsilon}(\mathcal{M}_{\alpha}(\overline{M}) \cup \{\overline{M}\}, \overline{M}),$$

其中 $\alpha = C_{\mathrm{reg}}\varepsilon + \textsf{r-dec}^{\mathrm{c}}_{C_{\mathrm{reg}}\varepsilon}(\mathcal{M} \cup \{M\}, M) \le C_{\mathrm{reg}}\varepsilon + c^2_{\mathrm{reg}}\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{M\}, M) \le C^2_{\mathrm{reg}}\left(\varepsilon + \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, M)\right)$（参见 定义 4.1）。接下来，对所有 $\gamma > 0$，利用 命题 4.1，可得

$$\textsf{r-dec}^{\mathrm{c}}_{C_{\mathrm{reg}}\varepsilon}(\mathcal{M}_{\alpha}(\overline{M}) \cup \{\overline{M}\}, \overline{M}) \le 8 \cdot \left(\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\alpha}(\overline{M}), \overline{M}) \vee 0 + C^2_{\mathrm{reg}}\gamma\varepsilon^2\right) + 7C_{\mathrm{reg}}\varepsilon,$$

使得

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \le 8C_{\mathrm{loc}} \cdot \left(\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\alpha}(\overline{M}), \overline{M}) \vee 0 + C^2_{\mathrm{reg}}\gamma\varepsilon^2\right) + 7C_{\mathrm{loc}}C_{\mathrm{reg}}\varepsilon.$$

现在令

$$\gamma^{\star} = (16C^2_{\mathrm{reg}}C_{\mathrm{loc}})^{-1} \cdot \frac{\varepsilon + \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M})}{\varepsilon^2},$$

其满足 $\gamma^{\star} \ge \frac{1}{16C^2_{\mathrm{reg}}C_{\mathrm{loc}}\cdot\varepsilon}$ 并给出

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \le 8C_{\mathrm{loc}} \cdot \textsf{r-dec}^{\mathrm{o}}_{\gamma^{\star}}(\mathcal{M}_{\alpha}(\overline{M}), \overline{M}) \vee 0 + \frac{1}{2} \cdot \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \overline{M}, \overline{M}) + (7C_{\mathrm{loc}}C_{\mathrm{reg}} + 1/2)\varepsilon,$$

或者重新整理后，

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) \le 16C_{\mathrm{loc}} \cdot \textsf{r-dec}^{\mathrm{o}}_{\gamma^{\star}}(\mathcal{M}_{\alpha}(\overline{M}), \overline{M}) \vee 0 + 2(7C_{\mathrm{loc}}C_{\mathrm{reg}} + 1/2)\varepsilon.$$

此外，我们有

$$\alpha \le C^2_{\mathrm{reg}}(16C^2_{\mathrm{reg}}C_{\mathrm{loc}}) \cdot \gamma^{\star}\varepsilon^2 = C^2_{\mathrm{reg}}(16C^2_{\mathrm{reg}}C_{\mathrm{loc}}) \cdot \alpha(\varepsilon, \gamma^{\star}).$$

最后，我们通过对 $\gamma^{\star} \ge \frac{1}{16C^2_{\mathrm{reg}}C_{\mathrm{loc}}\cdot\varepsilon}$ 取最大值来得到一个上界。

对于受约束 DEC 的下界，根据命题 4.7 可立即得出，对所有$\varepsilon > 0$ 和 $\gamma > \sqrt{2} \cdot \varepsilon^{-1}$，令 $\alpha = \frac{\gamma\varepsilon^2}{4}$，

$$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_{\alpha}(\overline{M}), \overline{M}) \le \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) + \max\left\{\alpha + \frac{1}{2\gamma} - \frac{\gamma\varepsilon^2}{2}, 0\right\} = \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}),$$

对所有$\overline{M} \in \mathcal{M}^{+}$均成立。由于我们可以对 $\gamma \ge \sqrt{2}\varepsilon^{-1}$ 取最大值，因此得证。

□

证明命题 4.9。 考虑以下模型类 $\mathcal{M}$，其参数由 $\alpha \in (0, 1/2)$给出：

1. $\Pi = \mathbb{N} \cup \{\pi_{\circ}\}$。
2. 我们有 $\mathcal{M} = \{M_a\}_{a \in \mathbb{N}^{\circ}}$。对于每个 $a \in \mathbb{N}$，模型 $M_a \in \mathcal{M}$ 的奖励和观测定义如下：

   (a) 对于 $\pi \in \mathbb{N}$，$f^{M_a}(\pi) = \frac{1}{2} + \alpha \cdot \mathbb{1}\{\pi = a\}$，而 $f^{M_a}(\pi_{\circ}) = 0$。

   (b) 对所有 $\pi \in \Pi$，在 $r \sim M_a(\pi)$ 下，$r = f^{M_a}(\pi)$ 几乎必然成立。

   (c) 对于 $\pi \in \mathbb{N}$，我们收到观测 $o = \perp$。

   (d) 选择 $\pi_{\circ}$会得到观测 $o \in \{0,1\}^{\mathbb{N}}$；对于每个 $i \in \mathbb{N}$，都独立抽取 $o_i \sim \mathrm{Ber}(1/2 + \alpha \cdot \mathbb{1}\{a = i\})$（因此，我们有 $\mathcal{O} = \{0,1\}^{\mathbb{N}} \cup \{\perp\}$）。

上界。我们将证明存在常数 $c, C > 0$，使得对于 $\varepsilon > 0$,

$$\sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \le \alpha \cdot \mathbb{1}\left\{\varepsilon \ge \sqrt{c} \cdot \alpha\right\} \le C \cdot \varepsilon.$$

由于对所有 $\varepsilon \ge 0$（因为 $p = \overline{\mathbb{I}_1}$ 的选择满足对所有 $M \in \mathcal{M}$, $\mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi)] \le \alpha$），都有 $\sup^{M \in \mathrm{co}(\mathcal{M})}$ $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, M) \le \alpha$，因此只需证明，对于 $\varepsilon < \sqrt{c \cdot \alpha}$ 以及任意 $M \in \overline{\mathrm{co}(\mathcal{M})}$，都有 $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, M) = 0$。

给定 $M \in \mathrm{co}(\mathcal{M})$ 和 $\varepsilon \le 1/2$，我们可以将 $M(\pi)$ 写为 $= \mathbb{E}_{M' \sim \nu}[M'(\pi)]$，其中 $\nu \in \Delta(\mathcal{M})$。我们根据以下情况定义一个分布 $p \in \Delta(\Pi)$:

- 如果 ν 在每个模型 $M \in \mathcal{M}$ 上分配的概率质量至多为 $2/3$，我们定义 $p = \mathbb{I}_{\pi_\circ}$。
- 否则，存在唯一的 $a^\star \in [A]$ 选择，使得 $\nu(M_{a^\star}) \ge 2/3$；在这种情况下，我们定义$p = \mathbb{I}_{a^\star}$。

现在考虑任意模型 $M_a \in \mathcal{M}$。考虑上面的第一种情况，并记 $o \sim M_a(\pi_\circ)$ 和 $o \sim M(\pi_\circ)$。注意，$o_a \sim$服从伯努利分布 $\mathrm{Ber}(1/2 + \alpha)$，而对于某个 $\beta \le \frac{2}{3} \cdot \alpha$, $o_a \sim$服从伯努利分布 $\mathrm{Ber}(1/2 + \beta)$。由此可得

$$\mathbb{E}_{\pi \sim p}\left[D^2_{\mathsf{H}}\big(M_a(\pi), \overline{M}(\pi)\big)\right] = D^2_{\mathsf{H}}\big(M_a(\pi_\circ), \overline{M}(\pi_\circ)\big) \ge D^2_{\mathsf{H}}(\mathrm{Ber}(1/2 + \alpha), \mathrm{Ber}(1/2 + \beta)) \ge c \cdot \alpha^2,$$

其中 $c > 0$ 是一个数值常数。由于 $c\alpha^2 > \varepsilon^2$，可得 $M_a \notin \mathcal{H}_{p,\varepsilon}(M)$；由于 M_a 的选择是任意的，我们得出结论，在这种情况下，$\mathcal{H}_{p,\varepsilon}(M) = \varnothing$。

现在考虑上面的第二种情况。对于 $a = a^\star$，我们有 $\mathbb{E}_{\pi \sim p}\left[f^{M_a}(\pi_{M_a}) - f^{M_a}(\pi)\right] = 0$. 对于 $a \ne a^\star$，我们有 $\mathbb{P}_{r \sim M_a(a^\star)}(r \ne 1/2) = 0$，而 $\mathbb{P}_{r \sim M(a^\star)}(r \ne 1/2) \ge 2/3$。因此，

$$\mathbb{E}_{\pi \sim p}\left[D^2_{\mathsf{H}}\big(M_a(\pi), \overline{M}(\pi)\big)\right] = D^2_{\mathsf{H}}\big(M_a(a^\star), \overline{M}(a^\star)\big) \ge D^2_{\mathsf{H}}(\mathrm{Ber}(0), \mathrm{Ber}(2/3)) \ge 4/9 > \varepsilon^2, \tag{110}$$

这意味着 $M_a \notin \mathcal{H}_{p,\ \varepsilon}(M)$。

下界。选取 $A \ge 2$，并令 $M =$服从均匀分布 $\mathrm{Unif}(\{M_a\}_{a \in [A]})$。给定 $p \in \Delta(\Pi)$，令 $a = \arg\min^{a \in [A]} p(a)$，使得$p(a) \le 1/A$。我们观测到

$$\mathbb{E}_{\pi \sim p}[f^{M_a}(\pi_{M_a}) - f^{M_a}(\pi)] \ge \alpha(1 - 1/A) \ge \alpha/2$$

and

$$\mathbb{E}_{\pi \sim p}\big[D^2_{\mathsf{H}}\big(M_a(\pi), \overline{M}(\pi)\big)\big] \le p(\pi_\circ) \cdot D^2_{\mathsf{H}}\big(M_a(\pi_\circ), \overline{M}(\pi_\circ)\big) + 2p(a) + \sum_{i \in [A], i \ne a} p(i) D^2_{\mathsf{H}}\big(M_a(i), \overline{M}(i)\big).$$

对于所有 $i \ne a$，我们有 $D^2_{\mathsf{H}}\big(M_a(i)\ M(i)\big) \le D^2_{\mathsf{H}}$(伯努利分布(0), 伯努利分布$(1/A)) \le 2/A$，因此 $\sum_{i \in [A], i \ne a} p(i) D^2_{\mathsf{H}}\big(M_a(i)\ M(i)\big) \le 2/A$。只要 α 是一个充分小的数值常数，我们还可以利用平方海灵格距离的张量化性质得到,

$$\begin{aligned} D^2_{\mathsf{H}}\big(M_a(\pi_\circ), \overline{M}(\pi_\circ)\big) &\le D^2_{\mathsf{H}}(\mathrm{Ber}(1/2 + \alpha), \mathrm{Ber}(1/2 + \alpha/A)) + (A - 1) \cdot D^2_{\mathsf{H}}(\mathrm{Ber}(1/2), \mathrm{Ber}(1/2 + \alpha/A)) \\ &\le c \cdot \left(\alpha^2 + (A - 1) \cdot \frac{\alpha^2}{A^2}\right) \\ &\le C \cdot \alpha^2, \end{aligned}$$

其中 $C,\ c > 0$ 是数值常数。综上可得

$$\mathbb{E}_{\pi \sim p}\big[D^2_{\mathsf{H}}\big(M_a(\pi), \overline{M}(\pi)\big)\big] \le p(\pi_\circ) \cdot C\alpha^2 + 4/A.$$

我们选择足够大的 A，使得 $4/A \le \varepsilon^2/2$。现在需要考虑两种情况。

- 如果 $p(\pi_\circ) \le \frac{\varepsilon^2}{2C\alpha^2}$，那么 $M_a \in \mathcal{H}_{p,\ \varepsilon}(M)$，并且

$$\mathbb{E}_{\pi \sim p}[f^{M_a}(\pi_{M_a}) - f^{M_a}(\pi)] \ge \frac{\alpha}{2}.$$

- 如果情况并非如此，我们有

$$\mathbb{E}_{\pi\sim p}\big[f^{\overline{M}}(\pi_{\overline{M}})-f^{\overline{M}}(\pi)\big]\geq\frac{1}{2}p(\pi_\circ)\geq\frac{\varepsilon^2}{4C\alpha^2}\wedge 1.$$

结合这些情况，我们得出存在数值常数 C，$c>0$ 使得

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M})\geq C\cdot\alpha\mathbb{I}\big\{\varepsilon>c\cdot\alpha^{3/2}\big\}.$$

特别是，选择 $\alpha\propto\varepsilon^{2/3}$ 可得 $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{M\},\overline{M})\geq\Omega(\varepsilon^{2/3})$，而 $\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M},\overline{M})\leq O(\varepsilon)$。

□

证明 **命题 4.10的证明**。根据命题 D.3，我们有，对于所有 $\varepsilon>0$ 和 $M\in\mathcal{M}^+$，

$$\textsf{p-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M})\leq\textsf{p-dec}^{\mathrm{c\ greedy}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M})=\textsf{p-dec}^{\mathrm{c\ greedy}}_{\varepsilon}(\mathcal{M},\overline{M})\leq\textsf{p-dec}^{\mathrm{c}}_{\sqrt{3}\varepsilon}(\mathcal{M},\overline{M})+4\varepsilon.$$

□

命题 4.11 的证明。根据 假设 A.1，我们有

$$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\overline{M})=\sup_{\mu\in\Delta(\mathcal{M})}\inf_{p\in\Delta(\Pi)}\mathbb{E}_{\pi\sim p,M\sim\mu}\big[f^M(\pi_M)-f^M(\pi)-\gamma\cdot D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\big].$$

观测到 t hat，由于海林格距离满足三角不等式，对于所有 $\pi\in$，我们有 Π，

$$\begin{aligned}\mathbb{E}_{M,M'\sim\mu}\big[D^2_{\mathsf{H}}(M(\pi),M'(\pi))\big]&\leq 2\,\mathbb{E}_{M\sim\mu}\big[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\big]+2\,\mathbb{E}_{M'\sim\mu}\big[D^2_{\mathsf{H}}\big(M'(\pi),\overline{M}(\pi)\big)\big]\\&=4\,\mathbb{E}_{M\sim\mu}\big[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\big].\end{aligned}$$

因此，我们有

$$\begin{aligned}\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\overline{M})&\leq\sup_{\mu\in\Delta(\mathcal{M})}\inf_{p\in\Delta(\Pi)}\mathbb{E}_{\pi\sim p,M\sim\mu}\Big[f^M(\pi_M)-f^M(\pi)-\frac{\gamma}{4}\cdot\mathbb{E}_{M'\sim\mu}D^2_{\mathsf{H}}(M(\pi),M'(\pi))\Big]\\&\leq\sup_{\nu\in\Delta(\mathcal{M})}\sup_{\mu\in\Delta(\mathcal{M})}\inf_{p\in\Delta(\Pi)}\mathbb{E}_{\pi\sim p,M\sim\mu}\Big[f^M(\pi_M)-f^M(\pi)-\frac{\gamma}{4}\cdot\mathbb{E}_{M'\sim\nu}D^2_{\mathsf{H}}(M(\pi),M'(\pi))\Big]\\&\leq\sup_{\nu\in\Delta(\mathcal{M})}\inf_{p\in\Delta(\Pi)}\sup_{M\in\mathcal{M}}\mathbb{E}_{\pi\sim p}\Big[f^M(\pi_M)-f^M(\pi)-\frac{\gamma}{4}\cdot\mathbb{E}_{M'\sim\nu}D^2_{\mathsf{H}}(M(\pi),M'(\pi))\Big]\\&=\sup_{\nu\in\Delta(\mathcal{M})}\textsf{r-dec}^{\mathrm{o,rnd}}_{\gamma/4}(\mathcal{M},\nu).\end{aligned}$$

对于(55)中的第二个不等式，由平方海灵格距离的凸性可立即得到 $\sup^{\nu\in\Delta(\mathcal{M})}\textsf{r-dec}^{\mathrm{o\ rnd}}_{\gamma}(\mathcal{M},\nu)\leq\sup_{M\in\mathrm{co}(\mathcal{M})}\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},M)$。

现在我们证明 (56)。设 $\varepsilon>0$ 给定。由于相对于 $\mathcal{M}^+$ 满足强正则性条件，命题 D.4 与 $\mathcal{M}'=\mathcal{M}^+$ 蕴含

$$\sup_{\overline{M}\in\mathcal{M}^+}\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{\overline{M}\},\overline{M})\leq C_{\mathrm{loc}}\cdot\sup_{\overline{M}\in\mathcal{M}^+}\textsf{r-dec}^{\mathrm{c}}_{C_{\mathrm{reg}}\cdot\varepsilon}(\mathcal{M}_\alpha(\overline{M})\cup\{\overline{M}\},\overline{M}),$$

其中 $\alpha:=C^2_{\mathrm{reg}}\cdot(\varepsilon+\text{上确界}_{M\in\mathcal{M}^+}\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}\cup\{M\},M))$。现在，令

$$\widetilde{\mathcal{M}}_\alpha(\overline{M})=\big\{M\in\mathcal{M}\mid f^M(\pi_M)\leq f^{\overline{M}}(\pi_{\overline{M}})+\alpha,f^{\overline{M}}(\pi_{\overline{M}})\leq f^M(\pi_{\overline{M}})+\alpha\big\}.$$

使用引理 D.2，结合以下事实：$\alpha\geq C^2_{\mathrm{reg}}\cdot\varepsilon\geq\sqrt{2}C_{\mathrm{reg}}\cdot\varepsilon$ 因为 $C_{\mathrm{reg}}\geq\sqrt{2}$，我们有

$$\sup_{\overline{M}\in\mathcal{M}^+}\textsf{r-dec}^{\mathrm{c}}_{C_{\mathrm{reg}}\cdot\varepsilon}(\mathcal{M}_\alpha(\overline{M})\cup\{\overline{M}\},\overline{M})\leq\sup_{\overline{M}\in\mathcal{M}^+}\textsf{r-dec}^{\mathrm{c}}_{\sqrt{2}C_{\mathrm{reg}}\cdot\varepsilon}(\widetilde{\mathcal{M}}_\alpha(\overline{M})\cup\{\overline{M}\},\overline{M})+\sqrt{2}C_{\mathrm{reg}}\varepsilon.$$

设 M 为 $\mathcal{M}^+$ 中使上方右侧取最大值的模型，并令 $\mathcal{M}' = \mathcal{M}_\alpha(M)$。设 $\gamma > 0$ 为固定值。利用命题 4.1，我们有

$$\textsf{r-dec}^{\mathrm{c}}_{\sqrt{2}C_{\mathrm{reg}}\varepsilon}(\mathcal{M}' \cup \{\widetilde{M}\}, \widetilde{M}) \le 8 \sup_{\overline{M}\in\mathcal{M}^+} \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}', \overline{M}) \vee 0 + 16C_{\mathrm{reg}}^2\gamma\varepsilon^2 + 7\sqrt{2}C_{\mathrm{reg}}\varepsilon.$$

根据公式 (55)，我们有

$$\sup_{\overline{M}\in\mathcal{M}^+} \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}', \overline{M}) \le \sup_{\nu\in\Delta(\mathcal{M}')} \textsf{r-dec}^{\mathrm{o,rnd}}_{\gamma/4}(\mathcal{M}', \nu).$$

考虑任意的 $\nu \in \Delta(\mathcal{M}')$，并令 $M_\nu := \mathbb{E}_{M'\sim\nu}[M'] \in \mathrm{co}(\mathcal{M}') \subseteq \mathrm{co}(\mathcal{M})$。观测到，如果 $M \in \mathcal{M}' = \mathcal{M}_\alpha(M)$，那么

$$f^M(\pi_M) \le f^{\widetilde{M}}(\pi_{\widetilde{M}}) + \alpha \le \mathbb{E}_{M'\sim\nu}\left[f^{M'}(\pi_{\widetilde{M}})\right] + 2\alpha \le \max_{\pi\in\Pi} \mathbb{E}_{M'\sim\nu}\left[f^{M'}(\pi)\right] + 2\alpha = f^{\overline{M}_\nu}(\pi_{\overline{M}_\nu}) + 2\alpha.$$

因此，$\mathcal{M}' \subseteq \mathcal{M}_{2\alpha}(\overline{M}_\nu)$，且我们有如下上界

$$\sup_{\nu\in\Delta(\mathcal{M}')} \textsf{r-dec}^{\mathrm{o,rnd}}_{\gamma/4}(\mathcal{M}', \nu) \le \sup_{\nu\in\Delta(\mathcal{M})} \textsf{r-dec}^{\mathrm{o,rnd}}_{\gamma/4}(\mathcal{M}_{2\alpha}(\overline{M}_\nu), \nu).$$

对于任意 $\nu \in \Delta(\mathcal{M})$，命题 D.1 表明

$$\textsf{r-dec}^{\mathrm{o,rnd}}_{\gamma/4}(\mathcal{M}_{2\alpha}(\overline{M}_\nu), \nu) \le \textsf{r-dec}^{\mathrm{c,rnd}}_{4\sqrt{\alpha/\gamma}}(\mathcal{M} \cup \{\overline{M}_\nu\}, \nu) + \frac{2}{\gamma}.$$

综合以上结果可得，对所有 $\gamma > 0$，

$$\begin{aligned}\sup_{\overline{M}\in\mathcal{M}^+} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) &\le C_{\mathrm{loc}} \cdot \left(8 \sup_{\nu\in\Delta(\mathcal{M})} \textsf{r-dec}^{\mathrm{c,rnd}}_{4\sqrt{\alpha/\gamma}}(\mathcal{M} \cup \{\overline{M}_\nu\}, \nu) + 8\sqrt{2}C_{\mathrm{reg}}\varepsilon + 16\gamma C_{\mathrm{reg}}^2\varepsilon^2 + \frac{16}{\gamma}\right) \\ &\le C_{\mathrm{loc}} \cdot \left(8 \sup_{\nu\in\Delta(\mathcal{M})} \textsf{r-dec}^{\mathrm{c,rnd}}_{4\sqrt{\alpha/\gamma}}(\mathcal{M} \cup \{\overline{M}_\nu\}, \nu) + 24C_{\mathrm{reg}}^2\gamma\varepsilon^2 + \frac{24}{\gamma}\right).\end{aligned}$$

我们取 $\gamma = \frac{1}{24C_{\mathrm{loc}}C_{\mathrm{reg}}^4} \cdot \frac{\alpha}{\varepsilon^2}$。由于 $\varepsilon^2/\alpha \le \varepsilon$，因此得到

$$\begin{aligned}\sup_{\overline{M}\in\mathcal{M}^+} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) &\le c_1 \cdot \sup_{\nu\in\Delta(\mathcal{M})} \textsf{r-dec}^{\mathrm{c,rnd}}_{c_2\varepsilon}(\mathcal{M} \cup \{\overline{M}_\nu\}, \nu) + \frac{1}{2C_{\mathrm{reg}}^2}\alpha + c_4\varepsilon, \\ &\le c_1 \cdot \sup_{\nu\in\Delta(\mathcal{M})} \textsf{r-dec}^{\mathrm{c,rnd}}_{c_2\varepsilon}(\mathcal{M} \cup \{\overline{M}_\nu\}, \nu) + \frac{1}{2} \sup_{\overline{M}\in\mathcal{M}^+} \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M} \cup \{\overline{M}\}, \overline{M}) + c_3\varepsilon,\end{aligned}$$

其中 c_1，c_2，c_3，$c_4 > 0$ 是仅取决于 C_{reg} 和 C_{loc}的常数。整理后得到公式 (56)中的第一个不等式；再根据詹森不等式即可得到第二个不等式。

□

E 第 5 节中省略的证明

推论推论 5.1的证明。下界可直接由命题 4.8推出，因此下面证明上界。固定 $\varepsilon > 0$。根据命题 4.6，我们有

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M}) \le C_{\mathrm{loc}} \cdot \textsf{r-dec}^{\mathrm{c}}_{C_{\mathrm{reg}}\varepsilon}(\mathcal{M}_\alpha(\overline{M}), \overline{M}),$$

其中 $\alpha = C_{\mathrm{reg}}^2 \cdot \left(\varepsilon + \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \overline{M})\right)$。回顾一下，我们假设 C_{reg}，$C_{\mathrm{loc}} = O(1)$。因此，对于所有 $\gamma > 0$ 固定，利用命题 4.1，我们有

$$\textsf{r-dec}^{\mathrm{c}}_{C_{\mathrm{reg}}\varepsilon}(\mathcal{M}_\alpha(\overline{M}), \overline{M}) \le O\left(\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_\alpha(\overline{M}), \overline{M}) \vee 0 + \gamma\varepsilon^2 + \varepsilon\right).$$

命题 4.1还给出了

$$\alpha = O(\varepsilon + \text{r-dec}^{\text{c}}_{\varepsilon}(\mathcal{M}, \overline{M})) \le O(\varepsilon + \text{r-dec}^{\text{o}}_{\gamma}(\mathcal{M}, \overline{M}) \vee 0 + \gamma\varepsilon^2) \le O(\text{r-dec}^{\text{o}}_{\gamma}(\mathcal{M}, \overline{M}) \vee 0 + \gamma\varepsilon^2 + \gamma^{-1}) = \alpha(\varepsilon, \gamma),$$

其中，第二个不等式来自算术-几何平均不等式。这就证明了该结论。 □

例 5.1 的证明。

我们首先给出公式（59）中该量的下界，然后证明E2D$^+$ 的遗憾上界。

偏移 *DEC* 的下界以及由 公式（*59*） 得到的遗憾界。 我们首先给出类别 $\mathcal{M}^{\alpha,\ \beta}$ 的偏移 DEC 的一个基本下界。

引理 E.1。 设给定 $\alpha \in (0、\ 1/4)、\ \beta \in (0、\ 1)$ 以及 $A \ge 2$ 。对于所有 $\gamma > 0$,

$$\text{r-dec}^{\text{o}}_{\gamma}(\mathcal{M}^{\alpha,\beta}, \widetilde{M}) = \text{r-dec}^{\text{o}}_{\gamma}(\mathcal{M}^{\alpha,\beta}_{\alpha}(\widetilde{M}), \widetilde{M}) \ge \frac{\alpha}{2 + 8\gamma\beta} - 4\gamma/A.$$

现在我们证明该量的一个下界

$$R := \min_{\gamma>0} \max\left\{ \sup_{\overline{M} \in \text{co}(\mathcal{M})} \text{r-dec}^{\text{o}}_{\gamma}(\mathcal{M}_{\overline{\Delta}(\gamma,T)}(\overline{M}), \overline{M}) \cdot T,\ \gamma \cdot \log|\mathcal{M}| \right\}$$

出现在 公式（59）中。我们首先对局部化半径给出下界

$$\overline{\Delta}(\gamma, T) = \Omega\Big(\frac{\gamma}{T} + \text{r-dec}^{\text{o}}_{\gamma}(\mathcal{M})\Big).$$

我们取 $\alpha_1 = 1/2$ 和 $A = T^2$，从而只要 T 是一个充分大的常数， 引理 E.1 便给出

$$\overline{\Delta}(\gamma, T) \ge \Omega\Big(\frac{\gamma}{T} + \text{r-dec}^{\text{o}}_{\gamma}(\mathcal{M}^{\alpha_1,\beta}, \widetilde{M})\Big) \ge \Omega\left(\frac{\gamma}{T} + \frac{1}{1+\gamma\beta} - 4\gamma/A\right) \ge \Omega\left(\frac{\gamma}{T} + \frac{1}{1+\gamma\beta}\right) \ge \Omega\left(\sqrt{\frac{1}{\beta T}} \wedge 1\right).$$

因此，只要 $\beta \ge 1/T$，令

$$\alpha_2 = c \cdot \sqrt{\frac{1}{\beta T}},$$

其中 c 是一个足够小的数值常数，那么无论如何选择 γ ，

$$\mathcal{M}^{\alpha_2,\beta}(\widetilde{M}) \subseteq \mathcal{M}_{\overline{\Delta}(\gamma,T)}(\widetilde{M}),$$

and

$$R \ge \min_{\gamma>0} \max\left\{ \text{r-dec}^{\text{o}}_{\gamma}(\mathcal{M}^{\alpha_2,\beta}(\widetilde{M}), \widetilde{M}) \cdot T,\ \gamma \right\}.$$

再次应用 引理 E.1，可得

$$R \ge \Omega\left(\min_{\gamma>0}\left\{\frac{\alpha_2 T}{1+\gamma\beta} + \gamma\right\}\right) \ge \Omega\left(\alpha_2 T \wedge \sqrt{\frac{\alpha_2 T}{\beta}}\right) \ge \Omega\Big(\beta^{-1/2}T^{1/2} \wedge \beta^{-3/4}T^{1/4}\Big).$$

令 $\beta = T^{-1/2}$，则有

$$R \ge \Omega(T^{5/8}),$$

这正是所需。

受约束 *DEC* 和 *E2D*$^+$ 的上界。现在，我们通过 E2D$^+$ 的 定理3.2 来界定遗憾。我们首先对受约束 DEC 作界。

引理 E.2。 设 $\beta \in (0,1)$ 已给定，并设 $\mathcal{M}^{\mathrm{all}} := \cup_{\alpha \in (0,1/2]} \mathcal{M}^{\alpha,\beta}$。则对所有 $\varepsilon > 0$,

$$\textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \le \textsf{r-dec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}^{\mathrm{all}}) \le O\left(\frac{\varepsilon^2}{\beta}\right).$$

令 $\beta \propto T^{-1/2}$ 和 $A \propto T^2$ 如前文所述。将 引理 E.2 中的界代入 定理 3.2 （参见 公式（91）），得到

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{DM}}(T)] \le \widetilde{O}\left(\frac{\bar{\varepsilon}(T)^2}{\beta} \cdot T + \sqrt{T}\right) = \widetilde{O}\left(\sqrt{T}\right),$$

因为采用通常的估计预言机选择时，我们可以取 $\varepsilon(T) \le \widetilde{O}\left(\sqrt{\frac{\log|\mathcal{M}|}{T}}\right)$，且 $\log|\mathcal{M}| \le O(\log(A)) \le O(\log(T))$。

□

引理 E.1 的证明。首先注意到 $\mathcal{M}^{\alpha,\beta} = \mathcal{M}^{\alpha,\alpha}(\widetilde{M})$，因为对所有 $\pi \in [A]$，$f^{\widetilde{M}}(\pi) = 1/2$，且所有 $M \in \mathcal{M}^{\alpha,\beta}$都满足 $f^{M}(\pi_M) \le 1/2 + \alpha$。
现在我们给出偏移 DEC 值的下界。考虑任意分布 $p \in \Delta(\Pi)$，并令 $i := \arg\min_{i \in [A]} p(i)$，使得 $p(i) \le 1/A$。我们有

$$\mathbb{E}_{\pi \sim p}\left[f^{M_{\alpha,i}}(\pi_{M_{\alpha,i}}) - f^{M_{\alpha,i}}(\pi)\right] \ge \alpha \cdot (1 - 1/A - p(\pi_\circ)) + \frac{1}{4}p(\pi_\circ) \ge \frac{\alpha}{2},$$

因为 $\alpha \le 1/4$ 和 $A \ge 2$。现在我们通过以下方式界定海林格距离

$$\mathbb{E}_{\pi \sim p}\left[D^2_{\mathsf{H}}\left(M_{\alpha,i}(\pi), \widetilde{M}(\pi)\right)\right] \le p(\pi_\circ) \cdot D^2_{\mathsf{H}}\left(M_{\alpha,i}(\pi_\circ), \widetilde{M}(\pi_\circ)\right) + 2p(i).$$

注意到 $D^2_{\mathsf{H}}\left(M_{\alpha,i}(\pi_\circ), \widetilde{M}(\pi_\circ)\right) \le 2\beta$ 和 $p(i) \le 1/A$，因此

$$\mathbb{E}_{\pi \sim p}\left[D^2_{\mathsf{H}}\left(M_{\alpha,i}(\pi), \widetilde{M}(\pi)\right)\right] \le 2\beta \cdot p(\pi_\circ) + 2/A.$$

结合目前为止的计算可得

$$\mathbb{E}_{\pi \sim p}\left[f^{M_{\alpha,i}}(\pi_{M_{\alpha,i}}) - f^{M_{\alpha,i}}(\pi) - \gamma \cdot D^2_{\mathsf{H}}\left(M_{\alpha,i}(\pi), \widetilde{M}(\pi)\right)\right] \ge \frac{\alpha}{2} - 2\gamma\beta p(\pi_\circ) - 4\gamma/A.$$

另一方面，通过选择 $M = \widetilde{M}$，我们有

$$\mathbb{E}_{\pi \sim p}\left[f^{\widetilde{M}}(\pi_{\widetilde{M}}) - f^{\widetilde{M}}(\pi) - \gamma \cdot D^2_{\mathsf{H}}\left(\widetilde{M}(\pi), \widetilde{M}(\pi)\right)\right] = \frac{1}{2}p(\pi_\circ),$$

因此

$$\begin{aligned}\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}^{\alpha,\beta}, \widetilde{M}) &\ge \min_{p \in \Delta(\Pi)} \max\left\{\frac{\alpha}{2} - 2\gamma\beta p(\pi_\circ), \frac{1}{2}p(\pi_\circ)\right\} - 4\gamma/A, \\ &\ge \frac{\alpha}{2 + 8\gamma\beta} - 4\gamma/A.\end{aligned}$$

□

引理 E.2 的证明。 设 $M \in \mathrm{co}(\mathcal{M}^{\mathrm{all}})$ 和 $\varepsilon \le 1/10$ 已给定。假设 $25\frac{\varepsilon^2}{\beta} \le 1/2$，否则结论显然成立。

令 $i = \arg\max_{i\in[A]} \mathbb{P}_{o\sim M(\pi_\circ)}(o = i)$，并设 $p = (1 - 25\frac{\varepsilon^2}{\beta})q + 25\frac{\varepsilon^2}{\beta}\mathbb{I}_{\pi_\circ}$，其中 $q \in \Delta([A])$ 是另一个分布，其取值稍后确定。我们首先注意到，若 $M_{\alpha,\ j} \in \mathcal{M}^{\alpha,\ \beta} \subseteq \mathcal{M}^{\mathrm{all}}$ 对于某些 $\alpha > 0$ 和 $j \neq i$ 成立，那么由于 $\mathbb{P}_{o\sim M(\pi_\circ)}(o = j) \leq \beta/2$，我们有

$$D^2_{\mathsf{H}}\big(M_{\alpha,j}(\pi_\circ), \overline{M}(\pi_\circ)\big) \geq \Big(\sqrt{\mathbb{P}_{o\sim M_{\alpha,j}(\pi_\circ)}(o = j)} - \sqrt{\mathbb{P}_{o\sim \overline{M}(\pi_\circ)}(o = j)}\Big)^2 \geq (\sqrt{\beta} - \sqrt{\beta/2})^2 \geq \frac{\beta}{20},$$

这里我们用到了 $\mathbb{P}_{o\sim M_{\alpha,j}(\pi_\circ)}(o \neq \perp) = \mathbb{P}_{o\sim M(\pi_\circ)}(o \neq \perp) = \beta$，因为 $M \in \mathrm{co}(\mathcal{M}^{\mathrm{all}})$。因此，无论如何选择 $q \in \Delta([A])$，都有 $M_{\alpha,j} \notin \mathcal{H}_{p,\varepsilon}(M)$，因为

$$\mathbb{E}_{\pi\sim p}\big[D^2_{\mathsf{H}}\big(M_{\alpha,j}(\pi), \overline{M}(\pi)\big)\big] \geq \frac{25\beta}{20\beta}\varepsilon^2 > \varepsilon^2.$$

因此，若我们定义 $\mathcal{M}'_i = \{M_{\alpha,i} \in \mathcal{M}^{\mathrm{all}} \mid \alpha \in (0, 1/2]\}$，则有 $\mathcal{H}_{p,\varepsilon}(M) \cup \{M\} \subseteq \mathcal{M}'_i \cup \{\widetilde{M}\} \cup \{M\}$，接下来只需选择 q，使其在所有这些模型上的遗憾都很小。注意，无论如何选择 q，$\mathbb{E}_{\pi\sim p}\big[g^{\widetilde{M}}\big] \leq 25\frac{\varepsilon^2}{\beta}$ 都成立，因此接下来我们将注意力限制在 M 和 $\mathcal{M}'_i$ 上。

令 $M^\star = \arg\min_{M\in\mathcal{M}'_i} D^2_{\mathsf{H}}\big(M(i), \overline{M}(i)\big)$。我们将证明

$$f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(i) \leq D^2_{\mathsf{H}}\big(M^\star(i), \overline{M}(i)\big). \tag{111}$$

为证明这一事实，首先注意到，若 $\pi_M = i$，则 (111) 显然成立。否则，令 $\nu_M \in \Delta(\mathcal{M}^{\mathrm{all}})$ 满足 $M(\pi) = \mathbb{E}_{M'\sim\nu_M}[M'(\pi)]$ 对所有 $\pi \in \Pi$ 成立，然后

$$f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(i) = \max_{\pi\in[A]} \mathbb{E}_{M'\sim\nu_{\overline{M}}}\big[f^{M'}(\pi) - f^{M'}(i)\big] \leq \frac{1}{2}\cdot\mathbb{P}_{M'\sim\nu_{\overline{M}}}\left(M' \notin \mathcal{M}'_i\right) \leq \frac{1}{2}\cdot\mathbb{P}_{r\sim\overline{M}(i)}\left(r = 1/2\right), \tag{112}$$

其中，最后一个不等式成立是因为所有模型 $M' \in \mathcal{M}^{\mathrm{all}}\backslash\mathcal{M}'_i$ 在 $r \sim M'(i)$ 时几乎必然满足 $r = 1/2$。回顾一个基本事实：对于所有事件 A 以及分布 $\mathbb{P}$ 和 $\mathbb{Q}$。

$$\frac{(\mathbb{P}(A) - \mathbb{Q}(A))^2}{\mathbb{P}(A) + \mathbb{Q}(A)} \leq 2D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}). \tag{113}$$

由于 $M^\star \in \mathcal{M}'_i$，我们有 $\mathbb{P}_{r\sim M^\star(i)}(r = 1/2) = 0$，因此，利用 (113)，可得

$$\mathbb{P}_{r\sim\overline{M}(i)}(r = 1/2) \leq 2D^2_{\mathsf{H}}\big(\overline{M}(i), M^\star(i)\big),$$

并结合(112)可得(111)。

接下来，我们通过令 $q(i) = \frac{4\varepsilon^2}{D^2_{\mathsf{H}}(M^\star(i), \overline{M}(i))} \wedge 1$，来选择 $q \in \Delta([A])$；并令 $q(\pi_M) \equiv 1 - q(i)$。我们考虑两种情况

- 若 $q(i) = 1$，则显然对所有 $M \in \mathcal{M}'_i$，$\mathbb{E}_{\pi\sim p}[g^M(\pi)] \leq 25\frac{\varepsilon^2}{\beta}$。此外，

$$\frac{4\varepsilon^2}{D^2_{\mathsf{H}}\big(M^\star(i), \overline{M}(i)\big)} \geq 1,$$

  因此 (111)意味着

$$f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(i) \leq 4\varepsilon^2.$$

  由此可得 $\mathbb{E}_{\pi\sim p}\big[g^M(\pi)\big] \leq 25\frac{\varepsilon^2}{\beta} + \big(f^M(\pi_M) - f^M(i)\big) \leq 25\frac{\varepsilon^2}{\beta} + 4\varepsilon^2$。

- 如果 $q(i) < 1$，那么对于所有 $M \in \mathcal{M}'_i$，

$$\mathbb{E}_{\pi\sim p}\big[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\big] \geq \frac{1}{2}q(i)D^2_{\mathsf{H}}\big(M(i), \overline{M}(i)\big) \geq \frac{1}{2}q(i)D^2_{\mathsf{H}}\big(M^\star(i), \overline{M}(i)\big) = 2\varepsilon^2,$$

因此 $M \notin \mathcal{H}_{p,\ \varepsilon}(M)$。由此可得 $\mathcal{H}_{p,\ \varepsilon}(\overline{M}) \cap \mathcal{M}_i' = \varnothing$。剩下的只需界定 M 下的遗憾，我们如下进行：

$$\begin{aligned}\mathbb{E}_{\pi\sim p}[g^{\overline{M}}(\pi)] \le q(i)(f^{\overline{M}}(\pi_{\overline{M}}) - f^{\overline{M}}(\pi)) + 25\frac{\varepsilon^2}{\beta} &\le q(i)\cdot D_{\mathsf{H}}^2\big(M^\star(i), \overline{M}(i)\big) + 25\frac{\varepsilon^2}{\beta} \\ &= 4\varepsilon^2 + 25\frac{\varepsilon^2}{\beta}.\end{aligned}$$

综合以上各种情况，我们得出

$$\sup_{\overline{M}\in\mathrm{co}(\mathcal{M}^{\mathrm{all}})} \textsf{r-dec}_\varepsilon^{\mathrm{c}}(\mathcal{M}^{\mathrm{all}} \cup \{\overline{M}\}, \overline{M}) \le O\left(\frac{\varepsilon^2}{\beta}\right).$$

□