arXiv:2603.05739v1 [cs.LG] 5 Mar 2026

重新审视 Best-of-N 在推理时对齐中的（次）最优性

VedSriraman*[1] 和 Adam Block†[1,2]

[1]计算机科学系，哥伦比亚大学[2]电气工程系，哥伦比亚大学

摘要

Best-of-N (BoN) 采样是一种广泛使用的语言模型推理时对齐方法：从参考模型中采样出 N 个候选响应，再根据学习得到的奖励模型预测的奖励，选出得分最高的响应。尽管它在实践中应用广泛，近期理论研究指出，该方法在统计上并非最优，且容易受到奖励劫持的影响；所谓奖励劫持，是指模型利用学习得到的奖励模型的弱点，在并未真正提升性能的情况下获得较高的估计奖励。我们在更贴近实践的假设下重新探讨这一问题。具体而言，有别于早期聚焦于期望真实奖励的分析——这一指标在许多实际场景中可能缺乏意义——我们研究推理时对齐如何影响胜率这一基于成对比较的指标，它更符合奖励模型在实践中的训练和评估方式。我们证明，只要对参考模型和学习得到的奖励模型的质量作出最低限度的要求，经过适当调优的 BoN 在计算和统计意义上都能最优地实现高胜率，这在一定程度上解释了它在实践中广泛取得成功的原因。由于在这一设定下 BoN 仍容易受到奖励劫持的影响，我们提出了一种简单实用的变体，该方法能够以可证明的方式消除奖励劫持，同时保持统计性能最优。最后，我们证明，若以胜率为考量，先前的方法可证明是次优 的，这凸显了在分析推理时对齐方法时选择恰当目标的重要性。

1 引言

近年来，推理时计算已成为语言模型（LM）的一项关键新扩展维度，并推动了数学推理、编码能力和可靠性方面的重大进展 [5, 12, 15]。以往研究指出，扩展推理时计算主要有两种方法：(i) 并行扩展，即模型并行生成多个输出，并从中选择一个返回给用户；(ii)串行扩展，即模型生成长得多的输出，从而能够求解计算量更大的问题 [22]。尽管这两种方法都已展现出显著的实证成效，但并行扩展尤其具有吸引力，因为它简单、易于实现，而且无需额外训练即可利用现有模型；许多实证结果也表明，仅靠并行扩展就能显著提升性能 [5, 11, 13]。此外，目前用于训练

*ved@cs.columbia.edu
†adam.block@columbia.edu

利用强化学习训练模型以发挥串行扩展优势的方法，在很大程度上依赖并行扩展，以便在训练过程中估计奖励信号 [12]。因此，理解并行扩展的理论基础，是理解现代语言模型能力的重要一步。

并行扩展机制中出现的关键问题是：如何从一组 N 独立采样的候选回答中，最佳地选择最终返回的输出。最初的实证方法侧重于可验证场景，在此类场景中，可以根据提示 x 和候选回答 y，将模型输出直接与已知的真实奖励$r^\star(x,\ y)$进行评分；在这种情况下，可以选择得分最高的候选回答 [5]。尽管这种方法能够带来强大的前沿性能，但它并不适用于绝大多数现实场景，因为在这些场景中无法进行此类验证。相反，如今的标准做法是使用某种学习得到的代理奖励 $\hat{r}(x,\ y)$，同样根据提示和候选回答来选择最佳候选回答 [21, 23, 29, 30, 33]；这种方法形成了广为人知的*Best-of-N*（BoN）算法，该算法返回 $\hat{y}_N \in \text{argmax}_{y_i} \hat{r}(x,\ y_i)$。尽管 BoN 算法已在各种场景中展现出强大的实证性能，并且已有一些相关理论研究[13, 14],，但我们对其行为的理解仍存在关键空白。

在 Huang 等人 [14], 的研究中，作者针对这样一种情形对 BoN 算法进行了理论分析：在用于生成候选项的采样分布 $\pi_{\text{ref}}(\cdot|x)$ 下，学习得到的奖励 $\hat{r}$ 与真实奖励 $r^\star$ 之间的均方误差较小 y_i；并证明，只要 $\pi^\star$ 在适当意义上接近 π_{ref}，与某个“优质”语言模型 $\pi^\star$ 相比，该算法就能取得合理的性能。然而，关键的是，他们证明了在这种情形下，BoN 算法是次优 的；若要达到最优性能，就必须改用他们提出的一种更复杂（也更繁琐）的 χ^2正则化 BoN 算法变体。这种次优性源于 BoN 的另一项缺陷：当候选项数量 N 较大时，算法会陷入奖励劫持，这是古德哈特定律 [10] 的一种表现，即算法会选择奖励 $\hat{r}$ 很高、但根据真实奖励 $r^\star$ [9]实际上质量较低的候选项。χ^2 正则化 BoN 算法能够通过惩罚在采样分布 π_{ref}下出现概率较低的候选项，减轻这种奖励劫持行为，从而提升性能。尽管这是一个颇具吸引力的理论结果，但这种次优性在实证上的相关性尚不明确，因为在实践中，经过最优调优的 BoN 算法大多优于更复杂的 χ^2正则化变体：要么在统计上表现更好，要么在计算上更高效（达到相同性能所需的样本更少）。

Huang 等人 [14]的理论与实践之间的差距，部分可以用该论文作者的认识来解释：在奖励为二元值这一特殊情形下（推理任务中经常如此，输出非对即错），只要候选响应的数量 N 设置得当，BoN 算法实际上就是最优的。[1] 另一方面，在许多情况下，例如对话或创意写作等开放式生成任务，奖励并非二元值，而是会根据响应质量呈现较大范围的取值。在这类情形下，尚不清楚 BoN 算法究竟是最优还是次优；如果是次优，那么更复杂的变体在实践中能否胜过它也尚不明确。此外，在任何情形下，Huang 等人 [14] 使用的均方误差指标是否是衡量学习得到的奖励模型质量的最恰当方式 $\hat{r}$，以及采样分布 π_{ref} 与目标语言模型 $\pi^\star$ 之间恰当的“接近”概念应当是什么，也仍不清楚。此外，该研究考察的性能指标是期望真实奖励；而在许多实际情形中，这一指标往往意义不大，此时相对于某种比较策略的胜率（参见（1））是更自然的评估方式 [20]。

在这项工作中，我们旨在通过更全面地从理论上理解 BoN 算法来解决这些尚未解答的问题。我们认为，这种理解更贴合该算法实际应用时的情形，

[1]从技术上讲，在那项工作中，作者允许比较策略具有一致有界的似然比，但由于现代语言模型的生成越来越长，在没有二元奖励的情况下，这一假设不太可能对合理的 $\pi^\star$ 成立。

同时，我们提出了一种新算法，它在理论上达到最优，并且能够以可证明的方式缓解奖励劫持。我们的出发点是观察到，无论是奖励模型的均方误差指标，还是目标语言模型下的期望真实奖励指标，都不足以全面理解 BoN 算法的性能。事实上，在语言模型的奖励并非二元值的大多数情况下，奖励的“真实标准”是胜率，即某个预言式对齐来源（人工标注员或能力强大的语言模型）偏好返回的响应而非离线参考响应的比例 [7, 8, 20, 24]。此外，大多数奖励模型都以类似的方式训练：通过对候选响应进行成对比较，而不是直接针对某个真实奖励值进行回归；后者在实践中很少能获得 [27]。事实上，当从未对此进行直接监督时，所谓真实奖励 $r^\star$ 是否存在本身就是一个值得探讨的哲学问题 [4, 24]。大多数实证研究出于实用目的采用 Bradley-Terry模型，但期望奖励仍然是一种多少有些人为的构造。

此外，虽然平方误差在学习理论中很自然，但在 BoN 算法的语境下使用平方误差却有些牵强，因为它不像算法本身那样具备尺度不变性：将奖励模型 $\hat{r}$ 乘以一个正数不会改变 BoN 算法的行为，却会改变均方误差。相反，更自然的奖励失配概念应符合 $\hat{r}$ 在 BoN 中的使用方式，即单调变换不应影响我们的质量指标。最后，我们通过考虑一种更一般的质量概念—— $\mathcal{E}_M$-散度 [25], ——来解决经 χ^2正则化的 BoN 算法所具有的临时性；Block 和 Polyanskiy [2] 对这一概念进行了深入研究，而它还与近期提出的覆盖度 [6, 13], 概念密切相关。后者已被有效用于理解语言模型在各种场景下的表现。[2]

本文有两个主要结果，第一个如下。

定理 1 （定理 3 和4 的非正式表述）。若 $\hat{r}$ 在 $r^\star$ 的成对胜率误差方面较为接近（参见 (2)），则使用 $\hat{r}$ 并对 N 进行调优的 *BoN* 算法可实现最优性能。

这一结果表明，当采用一种实践中常用的恰当奖励模型误差概念时，BoN 算法实际上是最优的，这在一定程度上与 Huang 等人 [14] 所体现的实践直觉相悖。事实上，尽管这并未与不同形式化设定下的差异构成形式上的矛盾，我们的结果表明，Huang 等人 [14] 指出的次优性是其假设所导致的表象，而非 BoN 算法的根本局限。

另一方面，当 N 较大时可能发生奖励劫持这一根本洞见依然成立。因此，我们进一步提出一种新的 $\mathcal{E}_M$ 正则化 BoN 算法，该算法通过 $\mathcal{E}_M$-散度，直接惩罚在采样分布 π_{ref} 下出现概率较低的候选项。我们的第二个主要结果如下：

定理 2 （定理 5的非正式表述）。如果 $\hat{r}$ 在 $r^\star$ 方面与其接近，即成对胜率误差较小，那么使用 $\mathcal{E}_M$正则化的 *BoN* 算法（参见(6)），并采用 $\hat{r}$ 且经过适当调优的正则化，则该算法既是最优的，也能实现有理论保证的性能，且性能不会随 N而下降。

这一结果表明，通过使用 $\mathcal{E}_M$正则化，我们可以缓解奖励劫持。与先前旨在应对奖励过度优化的方法相比，我们提出的算法具有一项关键优势：实现简单，无需在线估计、大量额外训练，也无需承担 BoN 算法本身之外的显著计算开销 [14, 16, 17]。此外，我们在命题 2 中证明， χ^2正则化的 Huang 等人的 BoN 算法 [14] 可能会差得任意多

[2]形式化地说，覆盖度定义为 $\text{Cov}_M(\pi^\star\|\pi_{\text{ref}}) = \mathbb{P}_{y\sim\pi^\star}\left(\frac{d\pi^\star}{d\pi_{\text{ref}}}(y) \geq M\right)$。不难看出， $\mathcal{E}_M(\pi^\star\|\pi_{\text{ref}}) = \text{Cov}_M(\pi^\star\|\pi_{\text{ref}}) - M\cdot\mathbb{P}_{y\sim\pi_{\text{ref}}}\left(\frac{d\pi^\star}{d\pi_{\text{ref}}}(y) \geq M\right)$ 的上界是覆盖度。

在胜率遗憾方面优于我们的 $\mathcal{E}_M$-正则化 BoN 算法，这表明以往方法在这一设定下并不充分。

我们现在进入第 2 节，介绍并说明我们的形式化设定，同时阐明它与先前工作所考虑的设定有何不同，然后在第 3 节中给出 BoN 算法的结果。接着，我们在第 4 节中引入覆盖率正则化 BoN 算法，并在第 5 节中讨论若干关键证明技巧。尽管本文大部分内容考虑的是一种简化设定，其中定义胜率的比较策略等于 π_{ref}，但我们在第 6 节中将结果推广到胜率相对于任意比较策略定义的情形。最后，我们在第 7 节中讨论相关工作与未来方向。

2 预备知识与问题设定

在本节中，我们介绍推理时对齐框架，并形式化地陈述所考虑的问题。我们首先描述与语言模型和奖励模型交互的计算模型，然后讨论适当的模型质量概念。

2.1 推理时对齐框架

我们的框架受到 Huang 等人 [13, 14], 的启发，但我们所考虑的期望性质有所不同。全文中，我们令 $\mathcal{Y}$ 表示可能响应的空间，$\mathcal{X}$ 表示可能提示或上下文的空间。语言模型 是任意从提示到响应分布的映射 $\pi: \mathcal{X} \to \Delta(\mathcal{Y})$；由于与强化学习（RL）的联系，我们交替使用语言模型和策略这两个术语。为便于分析，我们假设存在一个真实奖励函数 $r^\star: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$，用于衡量给定提示 x时响应 y 的质量；在实践中，该奖励函数是未知的，只能通过某个学习得到的奖励模型 $\widehat{r}: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$来访问。为简单起见，我们始终假设对某个已知的 $R_{\max} > 0$，有 $0 \le r^\star(x, y)$ $\widehat{r}(x, y) \le R_{\max}$。与 Huang 等人 [14], 一样，我们将考虑一种简化设定，其中提示 x 始终固定，因此为清晰起见将其从记号中省略，即我们写 $\pi_{\text{ref}}(\cdot)$ 而非 $\pi_{\text{ref}}(\cdot|x)$，对 $r^\star$和 $\widehat{r}$亦然。这样做并不损失一般性，因为如有需要，我们可以在分析结束时对提示的边缘分布取期望。

我们假设学习者可以使用以下计算模型。这一假设的动机是，vLLM [18] 等经验性语言模型推理引擎能够快速从大语言模型中采样，并并行评估奖励模型。

定义 1（采样与评估）。 学习者可以使用一个参考模型 π_{ref}，并从中独立采样响应 $y \sim \pi_{\text{ref}}$。此外，对于任意采样得到的响应 y，学习者可以查询奖励模型 $\widehat{r}$ 以获取奖励值 $\widehat{r}(y)$。学习者的效率以从 π_{ref}中抽取的样本总数 N来衡量。

学习者的目标是利用这种采样与评估访问来生成响应 $\hat{y}$，使其获得较高的真实奖励 $r^\star(\hat{y})$，同时尽可能少地使用来自 π_{ref} 的样本。为了评估产生响应 $\hat{y} \sim \widehat{\pi}$ 的候选策略 $\widehat{\pi}$，自然的做法是通过遗憾将它们与某个能获得较高真实奖励的 比较策略 $\pi^\star$ 进行比较。虽然 Huang 等人 [14] 从期望真实奖励的角度考虑遗憾，即试图最小化 $\mathbb{E}_{y\sim\pi^\star}[r^\star(y)] - \mathbb{E}_{y\sim\widehat{\pi}}[r^\star(y)]$,，但我们注意到，在许多实际场景中，真实奖励要么是 (i) 二元的，要么是 (ii) 只能通过（可能含噪声的）成对比较获得。在前一种情况下，期望真实奖励会大幅简化为一个搜索问题，并且在 Huang 等人 [14] 中得到了严密分析。然而，在后一种情况下，奖励的期望数值则不那么有意义

因为学习者永远无法访问 $r^\star$ 本身。相反，我们考虑由 $r^\star$ 诱导的学习策略相对于比较策略 π_{ref} 的 胜率，其定义为：

$$R^{r^\star}(\pi) = \mathbb{P}_{y\sim\pi, y'\sim\pi_{\text{ref}}}(r^\star(y) > r^\star(y')) + \frac{1}{2}\cdot\mathbb{P}_{y\sim\pi, y'\sim\pi_{\text{ref}}}(r^\star(y) = r^\star(y'))\,. \tag{1}$$

更一般地， $r^\star$ 可以替换为另一个奖励函数，例如 $\widehat{r}$，此时 (1) 将定义该奖励函数下的胜率。我们指出，虽然 (1) 使用我们据以抽取候选样本的基础模型 π_{ref} 作为比较策略，但原则上也可以针对任意比较分布 q 评估胜率。后一种设定在实践中很有意义，因为它对应于 AlpacaEval 式的流程 [7, 8]，其中 q 对应于 GPT-4 这类强比较策略。为使阐述简单明了，我们将先给出 $q = \pi_{\text{ref}}$ 这一特殊情形下的结果，随后再说明，对于一般的 q，在 q 与 π_{ref} 满足自然的接近性假设时，同样的论证也能得出定性上相似的结果，详见第 6 节。显然，当所有奖励函数都是二元函数时，期望奖励与胜率的概念一致（相差一个常数因子）。给定比较策略 $\pi^\star$，我们希望最小化候选策略 $\widehat{\pi}$ 的遗憾，即 $\text{Reg}(\widehat{\pi}) = R^{r^\star}(\pi^\star) - R^{r^\star}(\widehat{\pi})$。我们提出的根本问题是：

在采样与评估框架中，为了相对于给定的比较策略 $\pi^\star$ 实现较低的遗憾，需要从 π_{ref} 采集多少个样本（这一数量既充分又必要）？我们又应如何选择要返回的输出？

为了回答这个问题，我们需要更好地理解奖励模型 $\widehat{r}$ 在何种意义上接近真实奖励 $r^\star$，以及参考策略 π_{ref} 在何种意义上是比较策略$\pi^\star$的良好代理。下面我们将依次讨论这两个问题。

2.2 $\widehat{r}$ 在何种意义上接近 $r^\star$?

我们的首要目标是形式化学习得到的奖励模型 $\widehat{r}$ 接近真实奖励$r^\star$的含义。关键在于，奖励模型 $\widehat{r}$ 是在“典型”输出上训练的；我们将其形式化为假设：对于从 π_{ref}中抽取的样本，$\widehat{r}$ 接近 $r^\star$。虽然 Huang 等人 [14] 假设平方误差,$\varepsilon^2_{\text{RM}}(\widehat{r}) = \mathbb{E}_{y\sim\pi_{\text{ref}}}\left[(\widehat{r}(y) - r^\star(y))^2\right]$ 很小，但我们认为，在当前设定下，这并非最恰当的度量，原因有二。首先，平方误差不是尺度不变的；也就是说，将 $\widehat{r}$ 乘以一个正数，不会改变任何使用 $\widehat{r}$ 选择输出的算法的行为，却会改变平方误差。其次，当奖励只能通过成对比较获取时，假设平方误差很小并不太自然。事实上，这里有一个值得探讨的哲学问题：如果奖励只能通过成对比较获取，那么除了其所蕴含的序数信息之外，是否存在任何有意义的真实数值奖励函数 $r^\star$，并不明确。数值奖励函数通常通过在成对比较数据上、基于 Bradley-Terry模型 [4, 27], 运行最大似然估计（MLE）来训练；众所周知，这种方法的样本复杂度会随最大奖励值[3] [28, 35]呈指数级扩展。相反，我们建议通过学习得到的奖励模型 $\widehat{r}$ 相对于 $r^\star$的成对胜率误差 来衡量其质量：

$$\varepsilon_{\text{pw}}(\widehat{r}) := \mathbb{E}_{y,y'\sim\pi_{\text{ref}}}[|\phi_{\widehat{r}}(y,y') - \phi_{r^\star}(y,y')|] \quad \text{and} \quad \phi_r(y,y') := \mathbf{1}\{r(y) > r(y')\} + \tfrac{1}{2}\mathbf{1}\{r(y) = r(y')\} \tag{2}$$

这里，$\varepsilon_{\text{pw}}(\widehat{r})$量化了 $\widehat{r}$ 与 $r^\star$ 在从 π_{ref}采样的两个“典型”响应的结果上不一致的程度。具体而言，绝对差 $|\phi_{\widehat{r}}(y,y') - \phi_{r^\star}(y,y')|$ 会消失，当 $\widehat{r}$ 与 $r^\star$

[3]事实上，它会自然出现，因为简单计算表明，最大逆费 Fisher 信息以类似的方式缩放。

就 y 与 y'之间的成对比较结果达成一致时；取值为 $1/2$，当其中一个奖励函数出现平局而另一个没有时；并对 $\hat{r}$ 与 $r^\star$ 之间颠倒胜负结果的不一致施加全部惩罚1。因此，$\varepsilon_{\mathrm{pw}}(\hat{r})$ 的这一定义与胜率概念紧密对齐，后者对严格获胜的奖励高于平局。这种与胜率泛函的对齐（该泛函同样将平局计为一半，且取值介于0 与1之间）呼应了我们在第1节中关于胜率的讨论，并将在后续分析中被证明是重要的。

这一指标显然具有尺度不变性，也更贴近实践中奖励模型的训练方式 [19, 24, 27], ，这表明与 ε^2 RM 相比，(2) 可能是一个更合理的、值得期待能够加以控制的量。在 Bradley-Terry模型中，简单分析表明，这两个量之间最紧密的关系随最大奖励值呈指数级变化。这意味着，当我们使用来自 π_{ref} 的有限个样本进行学习时，可能会期待 $\varepsilon_{\mathrm{pw}}(\hat{r}) \ll \varepsilon^2$ RM$(\hat{r})$。然而，一般而言，正如我们在附录A中所示，$\varepsilon_{\mathrm{pw}}(\hat{r})$ 与 ε^2 RM$(\hat{r})$ 无法相互比较。事实上，$\varepsilon_{\mathrm{pw}}$ 的尺度不变性立即表明，存在这样的例子：$\varepsilon_{\mathrm{pw}} = 0$，但 ε^2 RM 可以任意大。要求 ε^2 RM$(\hat{r})$ 较小，比要求 $\varepsilon_{\mathrm{pw}}(\hat{r})$ 较小更强的另一层含义是，我们无法从较小的 $\varepsilon_{\mathrm{pw}}(\hat{r})$ 推导出以期望真实奖励衡量的较小遗憾（参见附录A中的命题 3）；这与较小的 $\varepsilon_{\mathrm{pw}}(\hat{r})$ [14] 形成鲜明对比。正如我们上文所述，在许多情况下，胜率是更自然的奖励概念。

2.3 在何种意义上 π_{ref} 是一个好的参考模型?

除了 $\hat{r}$ 作为 $r^\star$的代理的质量之外，我们还必须考虑 π_{ref}的质量。直观而言，如果 π_{ref} 在高奖励输出上所赋予的质量远少于比较策略 $\pi^\star$，那么 N 必须非常大，才能获得与比较策略相当的性能。因此，任何样本复杂度的概念都必须依赖于 $\pi^\star$ 与 π_{ref}之间某种差异的测度。尽管先前的工作假定 $\mathrm{D}_{\chi^2}(\pi^\star\|\pi_{\mathrm{ref}})$很小——我们记得$\mathrm{D}_{\chi^2}(\nu\|\mu) = \mathbb{E}\left[(d\nu/d\mu - 1)^2\right]$ 是两个分布 ν 与 μ之间的 χ^2-散度——但这一假设多少有些特设。事实上，$\mathrm{D}_{\chi^2}(\pi^\star\|\pi_{\mathrm{ref}})$很小这一要求直接源于 $\varepsilon^2{}_{\mathrm{RM}}(\hat{r})$很小这一假设，而这是由于$\chi^2$-散度与平方误差之间的对偶性；因此，尽管在该语境下这很自然，但并没有先验理由偏爱 χ^2-散度而非任何其他差异测度。我们转而提议通过 $\mathcal{E}_M$-散度这一基本概念来度量这种差异 [2, 25], ，其中对任意 $M \geq 0$，我们令

$$\mathcal{E}_M(\pi^\star\|\pi_{\mathrm{ref}}) = \mathbb{E}_{y\sim\pi_{\mathrm{ref}}}\left[(d\pi^\star/d\pi_{\mathrm{ref}}(y) - M)_+\right]. \tag{3}$$

虽然 $\mathcal{E}_M$-散度有许多应用，[25, 26], 但它与本文的相关性源于 Block 和 Polyanskiy [2] 证明了它与近似采样问题之间的密切联系：为了从目标分布 $\pi^\star$ 中生成样本，π_{ref}并使结果处于容差范围内（以全变差距离衡量），ε任何算法至少需要从 M 中获取样本，π_{ref}其中$\mathcal{E}_M(\pi^\star\|\pi_{\mathrm{ref}}) \leq \varepsilon$。在 定义 1的采样与评估框架中，这样的刻画十分自然，因为对于我们的设定而言，一个自然的目标上限是：利用来自 $\pi^\star$ 的样本近似采样 π_{ref}，然后返回采样得到的输出，前提是学习者能够获得关于 $\pi^\star$ 的知识，而不仅仅是由 $\hat{r}$所提供的知识。此外，$\mathcal{E}_M$-散度还与近期提出的覆盖度 [6],概念密切相关，该概念已被有效应用于从多种角度理解语言模型。覆盖度衡量两个分布之间的尾部差异，并定义为对于 $M \geq 0$：

$$\mathrm{Cov}_M(\pi^\star\|\pi_{\mathrm{ref}}) = \mathbb{P}_{y\sim\pi^\star}\left(d\pi^\star/d\pi_{\mathrm{ref}}(y) \geq M\right).$$

显然，对于所有 $M \geq 0$，$\mathcal{E}_M(\pi^\star\|\pi_{\mathrm{ref}}) \leq \mathrm{Cov}_M(\pi^\star\|\pi_{\mathrm{ref}})$。更一般地，$\mathcal{E}_M$-散度是 f-散度的一种特殊情形，[26],其定义适用于凸的 f：$\mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$其中 $f(1) = f'(1) = 0$ 为

$$\mathrm{D}_f(\nu\|\mu) = \mathbb{E}_{y\sim\mu}\left[f\left(d\nu/d\mu(y)\right)\right].$$

这一类包括许多广为人知的散度，例如 KL散度、全变差距离以及χ^2-散度，而 Block 和 Polyanskiy 的一个关键结果 [2] 表明，对于任意此类 f-散度，都有$\mathcal{E}_M(\pi^\star\|\pi_{\mathrm{ref}}) \le {}^{M\cdot \mathrm{D}_f(\pi^\star\|\pi_{\mathrm{ref}})}/_{f(M)}$。

明确了奖励模型和参考模型质量的这些概念后，我们现在可以进一步明确核心问题：

> 在采样与评估框架中，为了相对于给定的比较策略 $\pi^\star$ 实现较低的遗憾，需要从 π_{ref} 中抽取多少个样本才足够？所需样本数与成对胜率误差 $\varepsilon_{\mathrm{pw}}(\widehat{r})$ 及 $\mathcal{E}_M(\pi^\star\|\pi_{\mathrm{ref}})$之间有何关系？

3 使用 Best-of-N 实现最优性能

本节开始回答第 2 节末尾提出的问题：我们将给出采样与评估框架中最简单直接的算法 Best-of-N (BoN) 所能达到的胜率遗憾上界：

$$\widehat{\pi}_{\mathrm{BoN}} = \operatorname*{argmax}_{y\in\{y_1,\dots,y_N\}} \widehat{r}(y), \quad y_i \sim \pi_{\mathrm{ref}}. \tag{4}$$

(4) 中的 BoN 算法由 [29], 引入，它简单地假定 $r^\star = \widehat{r}$，并从 N 个来自参考策略 π_{ref}的独立同分布抽样中，返回预测奖励最高的样本。尽管 BoN 容易出现过度优化，因为 $\widehat{r}$ 在参考策略 $r^\star$ 所抽取的典型输出上可能无法很好地近似 $\widehat{\pi}_{\mathrm{BoN}}$，即使它在 π_{ref}所抽取的典型输出上能够做到这一点，但该算法在实践中已得到广泛应用。我们对其性能有如下上界。

定理 3. 对于任意 $N \ge 1$、参考模型 π_{ref}、比较策略 $\pi^\star \ll \pi_{\mathrm{ref}}$、真实奖励 $r^\star$ 和学习得到的奖励模型 $\widehat{r}$，有

$$R^{r^\star}(\pi^\star) - R^{r^\star}(\widehat{\pi}_{\mathrm{BoN}}) \lesssim N\cdot\varepsilon_{\mathrm{pw}}(\widehat{r})\cdot\log\left({}^{1}/_{\varepsilon_{\mathrm{pw}}(\widehat{r})}\right) + \mathcal{E}_{N/\log(1/\varepsilon_{\mathrm{pw}}(\widehat{r}))}\left(\pi^\star\|\pi_{\mathrm{ref}}\right). \tag{5}$$

上界 (5) 中有两项。第一项量化了奖励黑客行为发生的程度：随着 N 增大，算法更有可能选择 $\widehat{r}$ 与 $r^\star$ 存在差异的输出。第二项量化了这样一种程度：即使在 $\widehat{r} = r^\star$ 的情况下，BoN 算法也能够仅使用来自 π_{ref} 的样本，找到与比较策略 $\pi^\star$ 的输出相当的高奖励输出；由于 Block 和 Polyanskiy [2] 概述的 $\mathcal{E}_N$ 与近似拒绝采样之间的联系，$\mathcal{E}_N$-散度自然成为衡量 $\pi^\star$ 与 π_{ref} 之间差异的相关测度。应用该工作中的一个关键引理（将 $\mathcal{E}_M$-散度与 f-散度联系起来；参见第 2 节），我们可以进一步将 (5) 上界为$N\cdot\varepsilon_{\mathrm{pw}}(\widehat{r}) + {}^{N\cdot\mathrm{D}_f(\pi^\star\|\pi_{\mathrm{ref}})}/_{f(N)}$；在 χ^2-散度这一特殊情形下，通过调整 N，我们得到遗憾的最优速率（相差一个对数因子）$\sqrt{\varepsilon_{\mathrm{pw}}(\widehat{r})\cdot\chi^2(\pi^\star\|\pi_{\mathrm{ref}})}$，这至少在表面上与 Huang 等人 [14], 的结果相似，不过我们强调，此处考虑的是胜率，而非期望奖励。然而，在二元奖励这一特殊情形下，期望奖励成为胜率的一个特殊情形（参见第 2 节），因此，上述上界在此设定下复现了 Huang 等人 [14]的结果。

与期望奖励的情形不同，在期望奖励下，BoN 的性能在统计意义上是次优的；我们发现，事实上，对于胜率遗憾，BoN 在常数因子范围内是最优的，如下界所示。

定理 4。 对于任意无原子 π_{ref}、采样与评估框架中的任意 $\widehat{\pi}_{\mathcal{A}}$，以及任意 $\varepsilon < {}^{1}/_{2}$，都存在满足 $\varepsilon_{\mathrm{pw}}(\widehat{r}) \le \varepsilon$ 的奖励函数 $\widehat{r}, r^\star$，以及一个比较策略 $\pi^\star \ll \pi_{\mathrm{ref}}$，使得

$$R^{r^\star}(\pi^\star) - R^{r^\star}(\widehat{\pi}_{\mathcal{A}}) \gtrsim \inf_M \left\{M\cdot\varepsilon + \mathcal{E}_M(\pi^\star\|\pi_{\mathrm{ref}})\right\}.$$

这一结果表明，与期望奖励设定不同，在采样与评估框架中，BoN 在胜率遗憾方面实际上达到了对数因子内的最优，因此也有助于解释尽管 BoN 简单，却为何在实践中得到广泛应用。这一观察对 Huang 等人得出的结论提出了温和的质疑： [14]他们指出，在以期望奖励为优化目标时，BoN 在统计意义上是次优的，必须采用更复杂的算法才能达到最优性能。然而，正如我们反复强调的那样，我们认为在许多实际场景中，胜率往往比期望奖励更自然。因此，认为 BoN 因统计上次优而应被弃用的说法，或许有些言过其实。除了上述统计下界，我们还给出了一个互补的计算下界。

命题1。 对于任意 $M \geq 2$ 和 $\delta < 1/2$，存在奖励函数 $r^\star = \widehat{r}$ 和策略 $\pi^\star, \pi_{\mathrm{ref}}$，满足 $\mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}}) = 0$，使得在采样与评估框架内，任何达到遗憾$R^{r^\star}(\pi^\star) - R^{r^\star}(\widehat{\pi}) \leq \delta$的策略 $\widehat{\pi}$都必须从 π_{ref}中至少使用 $N = \widetilde{\Omega}(M)$ 个样本。

这个下界在计算意义上成立：如果我们不关心计算效率（在采样与评估框架中由 N 量化），那么利用 $\widehat{\pi}_{\mathrm{BoN}}$就能轻而易举地实现较小的遗憾，因为 $\widehat{r} = r^\star$；另一方面，如果我们想实现较小的遗憾，那么所需样本数量就有一个很强的下界。我们强调，这个下界与定理3的上界相匹配，因此，在采样与评估框架内，对于胜率遗憾而言，BoN 在计算和统计上都是最优的。

4 实现最优性能而不发生奖励劫持

上一节中，我们看到，式（N）给出的朴素 Best-of-4算法在经最优调优的 N下，对于胜率遗憾而言在计算和统计上均为最优，但它仍容易受到奖励劫持的影响。也就是说，随着 N增大，该算法更有可能选择 $\widehat{r}$与 $r^\star$存在分歧的输出，导致性能随 N非单调变化。这种情况虽可以理解，却并不理想，因为在实践中，我们希望能够增大 N 来提升性能，而不必担心对奖励代理 $\widehat{r}$过拟合。本节将介绍一种新算法，即 $\mathcal{E}_M$正则化 *Best-of-*N。该算法能够减轻奖励劫持，并实现随 N 单调变化的性能，同时在统计和计算上仍然最优。

直观而言，我们希望策略能够返回高奖励输出，同时又不会偏离参考策略 π_{ref}太远，因为在 π_{ref} 下不太可能出现的输出更容易受到奖励劫持的影响。上一节的分析和结果表明，衡量与 π_{ref}之间差异的恰当方式，是采用 $\mathcal{E}_M$-散度，并选择适当的 M，其定义见式（3）。受此观察启发，我们考虑以下变分问题：

$$\pi_M \in \operatorname*{argmax}_{\pi} \mathbb{E}_\pi \left[\widehat{r}(y)\right] - R_{\max} \cdot \mathcal{E}_M(\pi \| \pi_{\mathrm{ref}}), \tag{6}$$

我们将其解称为 $\mathcal{E}_M$正则化 *Best-of-*N 策略。需要注意的是，(6) 是否是值得考虑的目标，或许并非一目了然；的确，虽然 Huang 等人 [14] 在期望奖励最大化的背景下研究过一个类似目标，其中将 $\mathcal{E}_M$ 替换为 χ^2，但在我们的设定中，我们关注的并非期望奖励本身，而是胜率。另一个潜在问题是，在采样与评估框架中，难以从 (6) 中的变分问题直接进行采样，因为尚不清楚如何从最优策略 π_M中抽取样本。不过，以下结果消除了第二个疑虑，它表明 (6) 的最优策略具有一种特别简单的形式。

引理 *1*。令 λ 为 π_{ref}下 $\widehat{r}$ 的（$1-1/M$）分位数，即满足 $\mathbb{P}(\widehat{r}(y) \geq \lambda) = 1/M$的任意值。那么，求解 *(6)* 的最优策略 π_M为 $\pi_M(\cdot) = \pi_{\text{ref}}(\cdot \mid \widehat{r}(\cdot) \geq \lambda)$。

备注 1。 人们也可以考虑（6）的另一种变体，其中 $R_{\max}$ 被替换为一般的缩放参数 $\beta > 0$。若 $\beta \geq R_{\max}$，则最终会得到相同的解；但若 $\beta < R_{\max}$，则解可能会变得更加复杂。

引理 1 的证明见 D.1 节，我们在其中利用拉格朗日乘子和凸性，推导出最优策略的一个极为简洁的刻画。注意， π_M 可在采样与评估框架内近似采样，即抽取 N 个来自 π_{ref} 的样本，并从排名前$\lceil N/M \rceil$ 个样本中均匀抽取一个样本，其依据为 $\widehat{r}$，同时采用随机打破平局（参见 附录 D）；事实上，该采样过程与 π_M 之间的全变差距离以 $1/N$ 为界，并且 $\widehat{\pi}_M$可解释为针对 $N \lesssim M$执行 BoN，但随着 N 的增长而保持单调性。我们应当强调，与 Huang 等人 [14]的方法相比，$\widehat{\pi}_M$ 要简单得多，且不需要任何额外的估计或拒绝采样。对于该变体的性能，我们有如下界，并将其称为 $\widehat{\pi}_M$。

定理 5。对于任意 M、$N \geq 1$、参考模型 π_{ref}、比较策略 $\pi^\star \ll \pi_{\text{ref}}$、真实奖励 $r^\star$ 和学习得到的奖励模型 $\widehat{r}$，均有

$$R^{r^\star}(\pi^\star) - R^{r^\star}(\widehat{\pi}_M) \lesssim \mathcal{E}_M(\pi^\star \| \pi_{\text{ref}}) + M \cdot \varepsilon_{\text{pw}} + \frac{1}{N}, \tag{7}$$

请注意，(7) 中的上界与定理 3中 BoN 的上界非常相似，主要区别在于缩放参数 N 现在与正则化参数 M相互独立。因此，通过适当调整 M，我们可以在保持 N单调性的同时，实现定理 3 所确立的性能前沿。除了单调性之外，单调算法的表现与 BoN 相似；我们注意到，它仍然需要调整正则化参数 M，因此我们认为，在不要求单调性的情况下，BoN 仍将是占主导地位的对齐算法。

本节最后，我们指出，唯一已有的可证明单调算法——Huang 等人 [14], 提出的 χ^2-正则化 Best-of-N 算法——在以胜率遗憾为目标时，可能无法达到最优性能。具体而言，我们有以下结果。

命题 2 (π_β^χ与 π_M之间的分离)。 对于每个 $c > 1$，都存在策略 π_{ref}、$\pi^\star$、奖励函数 $\widehat{r}$、$r^\star$ 和 $\varepsilon_{\text{pw}} \ll 1$，使得 $\varepsilon_{\text{pw}}(\widehat{r}) \leq \varepsilon_{\text{pw}}$，并且满足

$$\inf_{\beta > 0} R^{r^\star}(\pi^\star) - R^{r^\star}(\pi_\beta^\chi) \geq c \cdot \left(\inf_{M > 0} R^{r^\star}(\pi^\star) - R^{r^\star}(\pi_M) \right).$$

该结果的证明将在D.3 节给出，它表明 χ^2正则化的 Best-of-N算法的表现可能远逊于我们提出的算法。除了命题 2所展示的糟糕表现外，我们还强调，本节提出的算法相较于 π_β^χ 以及 soft-BoN 等替代方案 [17, 31]，（使用悲观奖励估计运行 BoN）和其他替代方案 [16] 的主要优势之一在于其简单性：最优策略采用简单的顶部位数量选择器形式，在实践中易于实现和采样，无需在线估计或额外的计算开销。

5 证明技巧

接下来，我们将简要证明关于 BoN 在胜率遗憾下的性能的主要结果（定理 3 和 4），以及 $\mathcal{E}_M$-正则化算法的一个上界（定理 5）。

5.1 定理3 的证明概要

为了比较 $\widehat{\pi}_{\mathrm{BoN}}$ 与任意比较策略 $\pi^\star$ 在真实奖励函数 $r^\star$下的性能，我们需要克服两个主要挑战：(1) 如果 N 过小，且 $\pi^\star$ 与π_{ref}相差甚远，那么由于无法找到可与$\pi^\star$的高奖励输出相媲美的输出，我们可能会付出遗憾代价；(2) 如果 N 过大，那么我们可能会过度优化 $\widehat{r}$，并选出 $\widehat{r}$ 与 $r^\star$ 意见不一致的输出。这种权衡体现在上界（5）中的两项上，它们源于以下对任意 $M > 0$均成立的遗憾分解：遗憾 $R^{r^\star}(\pi^\star) - R^{r^\star}(\widehat{\pi}_{\mathrm{BoN}})$ 等于以下三项之和，

$$\underbrace{R^{r^\star}(\pi^\star) - R^{\widehat{r}}(\pi^\star_M)}_{\text{(i)}} + \underbrace{R^{\widehat{r}}(\pi^\star_M) - R^{\widehat{r}}(\widehat{\pi}_{\mathrm{BoN}})}_{\text{(ii)}} + \underbrace{R^{\widehat{r}}(\widehat{\pi}_{\mathrm{BoN}}) - R^{r^\star}(\widehat{\pi}_{\mathrm{BoN}})}_{\text{(iii)}}, \tag{8}$$

其中，$\pi^\star_M$ 是一个中间比较策略，定义为在满足以下约束的策略：相对于 $\pi^\star$ 的密度比 π_{ref} 几乎必然不超过 M 的条件下，使其与 $\pi^\star$ 之间的全变差距离最小。Huang 等人 [14], 曾提出过类似的分解，但在我们的设定中，关键挑战在于我们考虑的是胜率遗憾而非期望奖励遗憾，因此需要采用不同的技术来控制式（8）中的各项。

第（i）项。我们进一步将其分解为

$$R^{r^\star}(\pi^\star) - R^{\widehat{r}}(\pi^\star_M) = R^{r^\star}(\pi^\star) - R^{r^\star}(\pi^\star_M) + R^{r^\star}(\pi^\star_M) - R^{\widehat{r}}(\pi^\star_M)$$

并分别控制由此得到的两个差值。对于第一个差值，我们注意到胜率有界，因此关于全变差距离是 1-利普希茨的，所以根据 Block 和 Polyanskiy [2]，可以用 $\mathcal{E}_M(\pi^\star\|\pi_{\mathrm{ref}})$ 来控制从 $\pi^\star$ 到 $\pi^\star_M$ 的损失。对于第二个差值，我们利用以下事实：$\pi^\star_M$ 相对于 π_{ref} 的似然比几乎必然不超过 M，从而通过测度变换，使用 $M \cdot \varepsilon_{\mathrm{pw}}(\widehat{r})$ 控制 $\pi^\star_M$ 下 $\widehat{r}$ 与 $r^\star$ 之间的成对胜率误差。

术语 (ii)。这里，我们再次采用 Block 和 Polyanskiy [2] 提出的近似拒绝采样形式体系，将选择规则 定义为策略 $\pi^\star_{M,\mathrm{rej}}$，它仅使用来自 π_{ref} 的 N 独立同分布抽样来近似 $\pi^\star_M$。我们再次将术语 (ii) 进一步分解为

$$R^{\widehat{r}}(\pi^\star_M) - R^{\widehat{r}}(\widehat{\pi}_{\mathrm{BoN}}) = R^{\widehat{r}}(\pi^\star_M) - R^{\widehat{r}}(\pi^\star_{M,\mathrm{rej}}) + R^{\widehat{r}}(\pi^\star_{M,\mathrm{rej}}) - R^{\widehat{r}}(\widehat{\pi}_{\mathrm{BoN}})$$

并分别控制由此得到的两个项。对于第一个差值，我们使用 Block 和 Polyanskiy [2] 的近似拒绝采样分析，证明$\pi^\star_{M,\mathrm{rej}}$ 对 $\pi^\star_M$ 的全变差距离至多为关于 N/M 的指数小项；根据有界性，这一结论同样可推广到胜率。对于第二项，我们注意到，给定来自 π_{ref} 的任意固定批次 N 样本 $\widehat{r}$ 后，BoN 只需选择 $\widehat{r}$ 值最高的输出，因此在 $\widehat{r}$ 下，它在该批次上优于任何其他选择规则$\pi^\star_{M,\mathrm{rej}}$，包括$\pi^\star_{M,\mathrm{rej}}$。尽管这一结论直观上成立，我们仍在 附录 C 中给出严格证明。

项 (iii)。最后，我们使用与项 (i) 类似的测度变换论证来控制项 (iii)。注意到 BoN 相对于 π_{ref} 的密度比几乎处处至多为 N，因此 $\widehat{\pi}_{\mathrm{BoN}}$ 下 $\widehat{r}$ 与 $r^\star$ 之间的成对胜率误差至多为 $N \cdot \varepsilon_{\mathrm{pw}}(\widehat{r})$。

结论。结合项 (i)、(ii) 和 (iii) 的界可得，对于任意 $M > 0$,

$$R^{r^\star}(\pi^\star) - R^{r^\star}(\widehat{\pi}_{\mathrm{BoN}}) \lesssim (M+N) \cdot \varepsilon_{\mathrm{pw}}(\widehat{r}) + \mathcal{E}_M(\pi^\star\|\pi_{\mathrm{ref}}) + \exp(-N/M).$$

令 $M = N/\log(1/\varepsilon_{\mathrm{pw}}(\widehat{r}))$，即可得到所需的上界。

5.2 定理证明概要4

此前关于期望奖励遗憾 [14] 的下界并不能直接用于胜率遗憾，因为这些下界依赖于构造这样的奖励函数：它们在一小部分输出上有所不同，而这部分输出在 π_{ref}下概率很低、但期望奖励很高，因此不会显著影响胜率。事实上，正是由于此类构造不会严重影响胜率，BoN 才能在胜率遗憾方面达到最优性能，这与期望奖励遗憾的情况形成对比。对于固定的 $\widehat{\pi}$、 π_{ref} 和 $\widehat{r}$，我们构造了一个由可测子集 $A \subseteq \mathcal{Y}$ 参数化的$r_A^\star$ 族，使得 $r_A^\star$ 仅在 A 上与 $\widehat{r}$ 不同，即 $r^\star(y) = \widehat{r}(y) + \mathbb{I}y \in A$。计算可得

$$\varepsilon_{\text{pw}}(\widehat{r}) \leq \pi_{\text{ref}}(A) \cdot (1 - \pi_{\text{ref}}(A)) \leq \pi_{\text{ref}}(A),$$

因此，通过选择 A，使其 π_{ref}-测度至多为 ε，我们可以确保成对胜率误差较小。经过计算可知，适当选择 $\widehat{r}$ 后，我们可以给出以下下界

$$R^{r^\star}(\pi^\star) - R^{r^\star}(\widehat{\pi}) \gtrsim \pi^\star(A) - \widehat{\pi}(A)$$

对于该族中的任意 $r_A^\star$ 均成立。因此，若在 $\pi_{\text{ref}}(A)$ 较小的条件下，我们能确保 $\pi^\star(A)$ 较大，同时 $\widehat{\pi}(A)$ 较小，则遗憾会很大。我们应用马尔可夫不等式证明

$$\widehat{\pi}\left(\frac{d\widehat{\pi}}{d\pi_{\text{ref}}} \leq 2\right) \geq \frac{1}{2}$$

然后利用 π_{ref} 的无原子性证明存在一个 A，使得 $\pi_{\text{ref}}(A) \leq \varepsilon$ 且 $\widehat{\pi}(A) \leq 2\varepsilon$。接着，我们选择 $\pi^\star$，使其在 A 上满足 $d\pi^\star/d\pi_{\text{ref}} = M$，并在其他地方取某个较小常数，以确保 $\pi^\star$ 是一个有效的概率测度。适当选择 M 后，我们可以利用该实例建立下界。

5.3 证明概要：定理5

我们以与定理 3 的证明类似的方式开始，通过考虑遗憾分解，即遗憾$R^{r^\star}(\pi^\star) - R^{r^\star}(\widehat{\pi}_M)$ 等于

$$R^{r^\star}(\pi^\star) - R^{r^\star}(\widehat{\pi}_M) = \underbrace{\left(R^{r^\star}(\pi^\star) - R^{\widehat{r}}(\pi_M^\star)\right)}_{\text{(i)}} + \underbrace{\left(R^{\widehat{r}}(\pi_M^\star) - R^{\widehat{r}}(\widehat{\pi}_M)\right)}_{\text{(ii)}} + \underbrace{\left(R^{\widehat{r}}(\widehat{\pi}_M) - R^{r^\star}(\widehat{\pi}_M)\right)}_{\text{(iii)}},$$

其中 $\pi_M^\star$ 与第 5.1 节中所定义的相同。因此，第 (i) 项的处理方式与 BoN 证明完全相同，其贡献为 $\mathcal{E}_M(\pi^\star \| \pi_{\text{ref}}) + M \cdot \varepsilon_{\text{pw}}(\widehat{r})$。类似地，第 (iii) 项也以与 BoN 证明相同的方式处理，只是我们有 $\widehat{\pi}_M$ 的密度比至多为 M，相对于 π_{ref} 几乎必然成立（不随 N增大，这与 BoN 不同）。接下来仍需控制第 (ii) 项。我们现在进行分解

$$R^{\widehat{r}}(\pi_M^\star) - R^{\widehat{r}}(\widehat{\pi}_M) = R^{\widehat{r}}(\pi_M^\star) - R^{\widehat{r}}(\pi_M) + R^{\widehat{r}}(\pi_M) - R^{\widehat{r}}(\widehat{\pi}_M),$$

其中，π_M 从 $\widehat{r}(y)$ 的顶部 $1/M$ 分位数中均匀抽取样本，适用于 $y \sim \pi_{\text{ref}}$。根据 BoN 分析中的类似论证（参见 D.2 节），第一个差值非正。为界定第二个差值，我们显式计算经验分位数策略 $\widehat{\pi}_M$ 和总体分位数策略 π_M 的胜率：经过平局裁决后，N 个样本的 $\widehat{r}$ 秩是独立同分布的均匀分布 $[0, 1]$, ，因此 $\widehat{\pi}_M$ 的胜率可由显式次序统计量公式给出，得到 $R^{\widehat{r}}(\widehat{\pi}_M) \geq 1 - 1/2M - 1/N$；而 $R^{\widehat{r}}(\pi_M) = 1 - 1/2M$，所以该差值至多为 $1/N$。结论得证。

6 针对任意策略的胜率性能

在本节中，我们将 BoN 和 $\mathcal{E}_M$-正则化 BoN 的性能保证扩展到以下情形：胜率是针对任意比较策略 q（而非公式（1）中使用的基线 $q = \pi_{\text{ref}}$）进行评估，如公式（1）中所述，参见第 2.1 节。这种一般性与常见的评估流程相符——例如，在 AlpacaEval 风格基准 [7, 8] 中，人们通常会测量相对于 GPT-4 这类强比较策略的胜率。形式上，对于任意定义在 q 上的 $\mathcal{Y}$ 比较测度 r和任意奖励函数，我们将相对于 q 的胜率定义为

$$R_q^r(\pi) := \mathbb{E}_{y\sim\pi,\, y'\sim q}[\phi_r(y, y')] \quad \text{and} \quad \phi_r(y, y') := \mathbf{1}\{r(y) > r(y')\} + \tfrac{1}{2}\,\mathbf{1}\{r(y) = r(y')\}.$$

该定义推广了公式（1）中的胜率（对应于 $q = \pi_{\text{ref}}$），方法是将 $y' \sim q$作为基线比较策略。在本节中，我们始终假设 q 被 π_{ref} ($q \ll \pi_{\text{ref}}$支配，且满足逐点密度比界

$$w_q(y) := \frac{dq}{d\pi_{\text{ref}}}(y) \le L \quad \pi_{\text{ref}}\text{-a.s.}, \tag{9}$$

对于某个常数 $L \ge 1$。该条件使我们能够将针对 q 的胜率与相对于 π_{ref} 定义的量（如成对误差和超额质量）联系起来，采用与第 5.1 节和第 5.3 节相同的测度变换步骤，但需乘以因子 L。也可以采用较弱的散度控制，例如 $\mathcal{E}_L(q\|\pi_{\text{ref}})$，但我们使用（9）以尽量简化计算。我们继续通过 π_{ref} 来衡量相对于

$$\varepsilon_{\text{pw}}(\widehat{r}) \;=\; \mathbb{E}_{y,y'\sim\pi_{\text{ref}}}\left[\left|\phi_{\widehat{r}}(y, y') - \phi_{r^\star}(y, y')\right|\right].$$

以下结果将我们的 BoN 保证从基线比较 $q = \pi_{\text{ref}}$ 推广到满足（ q 的任意比较测度9）。

定理 6。假设 π_{ref} 和 q 满足 (9)，并令 $\pi^\star \ll \pi_{\text{ref}}$。那么对于任意奖励模型 $r^\star$ 和 $\widehat{r}$，以及任意 $N \ge 1$, 都有

$$R_q^{r^\star}(\pi^\star) - R_q^{r^\star}(\widehat{\pi}_{\text{BoN}}) \lesssim L \cdot N \cdot \varepsilon_{\text{pw}}(\widehat{r}) \cdot \log\left(1/\varepsilon_{\text{pw}}(\widehat{r})\right) + \mathcal{E}_{N/\log(1/\varepsilon_{\text{pw}}(\widehat{r}))}\left(\pi^\star\|\pi_{\text{ref}}\right).$$

当 $q = \pi_{\text{ref}}$ 时，定理 6 推得定理 3；更一般地，我们通过均匀密度比界，对参考模型 π_{ref} 与比较策略 q 之间的不匹配付出遗憾代价。这一因子之所以出现，是因为遗憾分解以及随后进行的测度变换步骤，与 $q = \pi_{\text{ref}}$ 情形完全相同（参见第 5.1 节），唯一的区别在于，利用 (9)，通过针对 $y' \sim \pi_{\text{ref}}$ 的期望来控制针对 $y' \sim q$ 的期望。其余各项可通过 Block 和 Polyanskiy [2]，中相同的近似拒绝采样论证来界定，因此 BoN 在胜率遗憾意义下仍然具有统计和计算最优性，但仍容易受到通过 ε_{pw} 实施的奖励作弊影响。如第 4 节所述，经过 $\mathcal{E}_M$ 正则化的 BoN 克服了这一问题，并使缩放参数 N 与正则化参数 M 解耦，从而为 $q = \pi_{\text{ref}}$ 给出关于 N 单调的保证；在下一个定理中，我们将在 (9) 的条件下，将针对 $\widehat{\pi}_M$ 的类似保证推广到一般的 q。

定理 7. 对于任意 M、$N \ge 1$、参考模型 π_{ref}、比较策略 $\pi^\star \ll \pi_{\text{ref}}$、真实奖励 $r^\star$ 和学习得到的奖励模型 $\widehat{r}$，以下结论成立：

$$R_q^{r^\star}(\pi^\star) - R_q^{r^\star}(\widehat{\pi}_M) \lesssim L \cdot \mathcal{E}_M(\pi^\star\|\pi_{\text{ref}}) + L \cdot M \cdot \varepsilon_{\text{pw}}(\widehat{r}) + \sqrt{\frac{M}{N}}.$$

再次通过设置 $L = 1$，我们从定理 5 中恢复了 $q = \pi_{\text{ref}}$ 的情形，但现在引入了一个按 $\sqrt{M/N}$ 衰减的额外项。我们强调，这纯粹是计算层面的现象，因为取 $N \uparrow \infty$ 时，我们可以恢复上文 Best-of-N 中相同的统计速率。在 $q = \pi_{\text{ref}}$ 这一特殊情形下，$\widehat{r}$ 下 top-$1/M$ 分位数策略的胜率可以通过显式的次序统计量计算得出，从而对 $1/N$ 呈现快速依赖。对于一般的 q，同样的 top-分位数选择步骤必须通过分析 q 如何在截断值附近的 $\widehat{r}$ 级集合上分配概率质量来处理；若无额外结构，q 可能会将其概率质量集中在分位数阈值附近的狭窄分数区间内，使胜率对有限批次 N 样本所导致的经验截断值微小波动高度敏感。因此，我们通过经验分位数的集中性来控制这一步骤，由此得到 N 中较慢的速率。最后，我们指出，式 (9) 中的密度比界只是将相对于 π_{ref} 定义的量与针对 q 的胜率联系起来的一种便捷方法。我们将这一有趣的问题留待未来研究：在一般的 q 设置下，这类项在多大程度上是必要的，以及能否在对 q 施加额外正则性假设后恢复更快的速率。

7 讨论与相关工作

在本文中，我们研究了采样与评估框架下的推理时对齐问题，重点关注广泛使用的 Best-of-N (BoN) 算法。与以往将期望真实奖励作为关注目标的工作不同，我们认为，在使用成对比较训练和评估奖励模型的实际场景中，胜率往往是更自然的目标。我们证明，在采样与评估框架下，为了获得较高的胜率，BoN 在统计和计算方面都是最优的，这在一定程度上解释了它尽管简单，却在实践中广受欢迎。由于即使采用胜率目标，BoN 仍容易受到奖励劫持的影响，我们提出了一种简单的变体，该变体能够在保持统计性能最优的同时，从理论上消除奖励劫持。我们的工作为未来研究提出了若干有趣的方向。首先，我们可以探究是否存在一种统一范式，能够将期望奖励目标和胜率目标涵盖为特例，从而更好地理解在何种条件下（经过最优调优的）BoN 是最优的，以及何时需要更复杂的算法。其次，尽管我们将采样与评估框架作为推理时对齐的一种自然抽象并加以研究，但值得探究的是，在允许自适应采样或与参考模型进行其他形式交互的更一般框架中，能否实现类似的统计和计算保证。最后，尽管我们提出的 BoN 变体从理论上消除了奖励劫持，但由于依赖信息论保证，它在实践中仍可能过于保守；如何超越此类保证，并利用学习得到的表征所提供的丰富结构来改进推理时对齐，仍是一个值得研究的问题。接下来，我们将更详细地讨论若干相关工作。

推理时对齐。 推理时对齐是一种流行的框架，无需昂贵的重新训练 [11, 13, 23, 24,29, 34]，即可使大语言模型与人类偏好对齐。推理时和训练时对齐通常都针对推理、知识或价值对齐方面的能力 [5, 12, 15, 24], ，并已成功应用于各种任务。前一类领域通常涉及二元奖励，即答案非真即假；后一类则通常涉及更主观的人类偏好，通常通过胜率 [7, 8, 20, 24]来衡量。我们的工作聚焦于后一种情形，在这种情况下，胜率比期望奖励更适合作为优化目标。

Best-of-N。 当学习者可以使用真实标签验证器时，Brown 等人 [5] 证明，只需对基础模型生成的样本执行 Best-of-N （BoN）程序，并选择输出

真实奖励最高的输出，就能有效地在多种任务中生成高质量输出。然而在实践中，真实奖励无法获取，因此会使用学习得到的奖励模型，从基础模型生成的样本中进行选择 [1, 24, 29]。尽管使用学习得到的奖励模型进行 BoN 已得到广泛应用，但它容易受到奖励劫持的影响，这一问题已得到广泛研究 [9,30]。已有多种临时性的经验缓解方法，包括在表示空间中进行正则化 [16]，利用经验不确定性量化的悲观型 BoN，以及 BoN 的 KL 正则化变体 [17, 31]。这些方法都没有提供显著优于 BoN 的理论保证。在理论方面，最相关的工作来自 Huang 等人 [14], ，他们分析了期望奖励目标下的 BoN，并证明由于其容易受到奖励过度优化的影响，BoN 通常在统计意义上是次优的。随后，他们提出了一种更复杂的算法，通过借鉴近似拒绝采样的思路，实现了最优的统计性能 [2]。与他们的工作不同，我们证明，针对胜率目标，BoN 实际上在统计和计算意义上都是最优的；我们认为，在实践中这一目标往往更加自然。因此，尽管概念上相似，我们的结果对 BoN 的评价比 Huang 等人的结果更为积极 [14] ，而且我们的技术（尤其是下界）也有显著不同。

参考文献

[1] Yuntao Bai、Andy Jones、Kamal Ndousse、Amanda Askell、Anna Chen、Nova DasSarma、Dawn Drain、Stanislav Fort、Deep Ganguli、Tom Henighan 等。基于人类反馈的强化学习训练有用且无害的助手。*arXiv* 预印本 *arXiv:2204.05862*，2022年。[2] Adam Block 与 Yury Polyanskiy。近似拒绝采样的样本复杂度及其在平滑在线学习中的应用。载于第三十六届学习理论年会，第 228–273 页。PMLR，2023年。[3] Stephen Boyd 与 Lieven Vandenberghe。凸优化。剑桥大学出版社，2004年。[4] Ralph Allan Bradley 与 Milton E Terry。不完全区组设计的秩分析：I. 成对比较方法。《*Biometrika*》，39(3/4):324–345，1952年。[5] Bradley Brown、Jordan Juravsky、Ryan Ehrlich、Ronald Clark、Quoc V Le、Christopher Ré 与 Azalia Mirhoseini。大型语言猴子：通过重复采样扩展推理计算。*arXiv* 预印本 *arXiv:2407.21787*，2024年。[6] Fan Chen、Audrey Huang、Noah Golowich、Sadhika Malladi、Adam Block、Jordan T Ash、Akshay Krishnamurthy 与 Dylan J Foster。覆盖原则：预训练如何促成后训练。*arXiv* 预印本 *arXiv:2510.15020*，2025年。[7] Yann Dubois、Xuechen Li、Rohan Taori、Tianyi Zhang、Ishaan Gulrajani、Jimmy Ba、Carlos Guestrin、Percy Liang 与 Tatsunori B. Hashimoto。Alpacafarm：面向从人类反馈中学习的方法的模拟框架，2023年。[8] Yann Dubois、Balázs Galambosi、Percy Liang 与 Tatsunori B Hashimoto。长度受控的 AlpacaEval：消除自动评估器偏差的简单方法。*arXiv* 预印本 *arXiv:2404.04475*，2024年。[9] Leo Gao、John Schulman 与 Jacob Hilton。奖励模型过度优化的缩放定律。载于国际机器学习会议，第 10835–10866 页。PMLR，2023年。

[10] 查尔斯·古德哈特。货币理论与实践：英国的经验。1984。

[11] Lin Gui、Cristina Gârbacea 和 Victor Veitch。面向大语言模型的 Bonbon 对齐及 best-of-n 采样的甜蜜之处。《神经信息处理系统进展》，37：2851–2885，2024年。[12] Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Ruoyu Zhang、Runxin Xu、Qihao Zhu、Shi- rong Ma、Peiyi Wang、Xiao Bi 等。DeepSeek-R1：通过强化学习激励大语言模型的推理能力。arXiv 预印本 arXiv:2501.12948，2025年。[13] Audrey Huang、Adam Block、Dylan J Foster、Dhruv Rohatgi、Cyril Zhang、Max Simchowitz、Jordan T Ash 和 Akshay Krishnamurthy。语言模型的自我改进：锐化机制。发表于第十三届国际学习表征会议，2025年。[14] Audrey Huang、Adam Block、Qinghua Liu、Nan Jiang、Akshay Krishnamurthy 和 Dylan J Foster。Best-of-N 是否真的最好？推理时对齐中的覆盖度、扩展与最优性。发表于第四十二届国际机器学习会议，2025年。[15] Aaron Jaech、Adam Kalai、Adam Lerer、Adam Richardson、Ahmed El-Kishky、Aiden Low、Alec Helyar、Aleksander Madry、Alex Beutel、Alex Carney 等。OpenAI o1 系统卡。arXiv 预印本 arXiv:2412.16720，2024年。[16] Yuu Jinnai、Tetsuro Morimura、Kaito Ariu 和 Kenshi Abe。用于语言模型对齐的、以最小贝叶斯风险目标为基础的正则化 best-of-n 采样。发表于《2025年美洲计算语言学协会分会会议论文集：人类语言技术》（第1卷：长篇论文），第9321–9347页，2025年。[17] Hadi Khalaf、Claudio Mayrink Verdun、Alex Oesterling、Himabindu Lakkaraju 和 Flavio du Pin Calmon。大语言模型中的推理时奖励破解。arXiv 预印本 arXiv:2506.19248，2025年。[18] Woosuk Kwon、Zhuohan Li、Siyuan Zhuang、Ying Sheng、Lianmin Zheng、Cody Hao Yu、Joseph E. Gonzalez、Hao Zhang 和 Ion Stoica。利用 PagedAttention 高效管理大语言模型服务的内存。发表于 ACM SIGOPS 第29届操作系统原理研讨会论文集，2023年。[19] Nathan Lambert、Jacob Morrison、Valentina Pyatkin、Shengyi Huang、Hamish Ivison、Faeze Brah- man、Lester James V Miranda、Alisa Liu、Nouha Dziri、Shane Lyu 等。Tulu 3：推动开放语言模型后训练的前沿。arXiv 预印本 arXiv:2411.15124，2024年。[20] Xuechen Li、Tianyi Zhang、Yann Dubois、Rohan Taori、Ishaan Gulrajani、Carlos Guestrin、Percy Liang 和 Tatsunori B. Hashimoto。AlpacaEval：一种指令遵循模型自动评估器。https://github.com/tatsu-lab/alpaca_eval，2023年5月。[21] Yujia Li、David Choi、Junyoung Chung、Nate Kushman、Julian Schrittwieser、Rémi Leblond、Tom Eccles、James Keeling、Felix Gimeno、Agustin Dal Lago 等。利用 AlphaCode 生成竞赛级代码。《科学》，378(6624)：1092–1097，2022年。[22] Parsa Mirtaheri、Ezra Edelman、Samy Jelassi、Eran Malach 和 Enric Boix-Adsera。让我想想！一条很长的思维链，其价值可能相当于指数级数量的短思维链。arXiv 预印本 arXiv:2505.21825，2025年。

[23] Reiichiro Nakano、Jacob Hilton、Suchir Balaji、Jeff Wu、Long Ouyang、Christina Kim、Christopher Hesse、Shantanu Jain、Vineet Kosaraju、William Saunders 等。《WebGPT：借助浏览器并结合人类反馈回答问题》。*arXiv* 预印本 *arXiv:2112.09332*，2021年。[24] Long Ouyang、Jeffrey Wu、Xu Jiang、Diogo Almeida、Carroll Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray 等。《通过人类反馈训练语言模型遵循指令》。神经信息处理系统进展，35：27730–27744，2022年。[25] Yury Polyanskiy。《信道编码：非渐近基本极限》。普林斯顿大学，2010年。[26] Yury Polyanskiy 和 Yihong Wu。《信息论：从编码到学习》。剑桥大学出版社，2025年。[27] Rafael Rafailov、Archit Sharma、Eric Mitchell、Christopher D Manning、Stefano Ermon 和 Chelsea Finn。《直接偏好优化：你的语言模型其实是一个奖励模型》。神经信息处理系统进展，36：53728–53741，2023年。[28] Corby Rosset、Ching-An Cheng、Arindam Mitra、Michael Santacroce、Ahmed Awadallah 和 Tengyang Xie。《直接纳什优化：利用一般偏好教语言模型自我改进》。*arXiv* 预印本 *arXiv:2404.03715*，2024年。[29] Nisan Stiennon、Long Ouyang、Jeffrey Wu、Daniel Ziegler、Ryan Lowe、Chelsea Voss、Alec Radford、Dario Amodei 和 Paul F Christiano。《通过人类反馈学习摘要》。神经信息处理系统进展，33：3008–3021，2020年。[30] Benedikt Stroebl、Sayash Kapoor 和 Arvind Narayanan。《推理扩展缺陷：不完美验证器下大语言模型重采样的局限》。*arXiv* 预印本 *arXiv:2411.17501*，2024年。[31] Claudio Mayrink Verdun、Alex Oesterling、Himabindu Lakkaraju 和 Flavio P Calmon。《用于模型对齐的软 Best-of-N 采样》。*arXiv* 预印本 *arXiv:2505.03156*，2025年。[32] John 冯·诺伊曼。《与随机数字相关的各种技术》。《冯·诺依曼文集》，5：768–770，1963年。[33] Xuezhi Wang、Jason Wei、Dale Schuurmans、Quoc Le、Ed Chi、Sharan Narang、Aakanksha Chowdhery 和 Denny Zhou。《自洽性提升语言模型中的思维链推理》。*arXiv* 预印本 *arXiv:2203.11171*，2022年。[34] Yidong Wang、Zhuohao Yu、Zhengran Zeng、Linyi Yang、Cunxiang Wang、Hao Chen、Chaoya Jiang、Rui Xie、Jindong Wang、Xing Xie 等。《PandaLM：用于大语言模型指令微调优化的自动评估基准》。*arXiv* 预印本 *arXiv:2306.05087*，2023年。[35] Tengyang Xie、Dylan J Foster、Akshay Krishnamurthy、Corby Rosset、Ahmed Awadallah 和 Alexander Rakhlin。《探索性偏好优化：利用隐式 Q* 近似实现样本高效的人类反馈强化学习》。*arXiv* 预印本 *arXiv:2405.21046*，2024年。

目录

1 引言 **1**

2 预备知识与问题设定 4
2.1 推理时对齐框架 .42.2 从什么意义上说 $\hat{r}$ 接近 $r^\star$? .52.3 从什么意义上说 π_{ref} 是一个好的参考模型? .6

3 Best-of-N 7 的最优性能

4 实现无奖励劫持的最优性能 8

5 证明技巧 95.1 定理 3 的证明概要 .105.2 定理 4 的证明概要 .115.3 定理 5 的证明概要 .11

6 相对于任意策略的胜率性能 12

7 讨论与相关工作 13A 对不同误差与性能概念的考察 17A.1 奖励模型误差比较的证明 .18A.2 命题 3 的证明 .20**B 辅助引理 21C Best-of-N 24**C.1 定理 3 的证明 .24C.2 定理 4 的证明 .28C.3 命题1 的证明 .30

D 单调算法的证明 31
D.1 $\mathcal{E}_M$-正则化 BoN 策略的推导 .32D.2 定理 5 的证明 .34D.3 命题 2 的证明 .38

E 针对任意比较策略的胜率证明 46
E.1 针对 q 的胜率迁移引理.47E.2 定理 6 的证明 .47E.3 定理 7 的证明 .49

对不同误差与性能概念的探究

在本节中，我们证明，第 2.2 节中定义的成对误差与平方误差通常不可比较。具体而言，引理 2 表明，较小的 ε_{pw} 并不意味着较小的 $\varepsilon^2_{\text{奖励模型}}$，而引理 3 则表明反之亦然。此外，我们给出命题 3的证明。该命题表明，对于所有问题，均无法根据第 2.3 节中引入的成对误差和超额质量散度一致地界定期望奖励遗憾。

A.1 奖励模型误差比较的证明

引理 2。 固定一个基准策略 π_{ref}。对于所有 $c > 0$，都存在无界奖励模型 $\widehat{r}$ 和 $r^\star$，满足

$$\varepsilon_{\text{pw}}(\widehat{r}) = 0 \qquad \textit{but} \qquad \varepsilon^2_{\text{RM}}(\widehat{r}, r^\star) \geq c.$$

证明。 回顾成对结果

$$\phi_r(y, y') = \mathbf{1}\{r(y) > r(y')\} + \frac{1}{2}\,\mathbf{1}\{r(y) = r(y')\}.$$

以及成对误差

$$\varepsilon_{\text{pw}}(\widehat{r}) := \mathbb{E}_{y,y' \sim \pi_{\text{ref}}}\left[|\phi_{\widehat{r}}(y, y') - \phi_{r^\star}(y, y')|\right].$$

关键观察是，成对错排损失在严格递增变换下保持不变：若对于严格递增的 ψ， $\widehat{r}(y) = \psi(r^\star(y))$，则对于独立同分布的 $y, y' \sim \pi_{\text{ref}}$，

$$\phi_{\widehat{r}}(y, y') = \phi_{r^\star}(y, y'),$$

因此 $\varepsilon_{\text{pw}}(\widehat{r}) = 0$。将其应用于严格递增映射 $\psi(t) = ct$ (对任意 $c > 0$)，即$\widehat{r}(y) = c\,r^\star(y)$，可得对于每个 $\varepsilon > 0$， $\varepsilon_{\text{pw}}(\widehat{r}) = 0 \leq \varepsilon$。同时，

$$\varepsilon^2_{\text{RM}}(\widehat{r}, r^\star) = \mathbb{E}_{y\sim\pi_{\text{ref}}}\big[(\widehat{r}(y) - r^\star(y))^2\big] = \mathbb{E}_{y\sim\pi_{\text{ref}}}\big[\,((c-1)r^\star(y))^2\,\big] = (c-1)^2\,\mathbb{E}_{y\sim\pi_{\text{ref}}}\big[r^\star(y)^2\big].$$

在 π_{ref} 下选择任意二阶矩非零的 $r^\star$，并令 c 任意增大，即可得到所需结果。

□

引理 3. 设 $R_{\max} > 0$，并任取 $p \in (0,\ 1)$。对每一个 $\varepsilon > 0$，存在响应空间 $\mathcal{Y}$、定义在 $\mathcal{Y}$ 上的参考策略 π_{ref}、真实奖励 $r^\star(\cdot) :\ \mathcal{Y} \to [0, R_{\max}]$，以及奖励模型 $\widehat{r}$，使得

$$\varepsilon^2_{\text{RM}}(\widehat{r}, r^\star) = \mathbb{E}_{y\sim\pi_{\text{ref}}}\big[(\widehat{r}(y) - r^\star(y))^2\big] = \varepsilon^2, \qquad \varepsilon_{\text{pw}}(\widehat{r}) \ \geq\ \frac{(1-p)^2}{4},$$

且该奖励具有非平凡尺度：

$$\text{Var}_{y\sim\pi_{\text{ref}}}\big[r^\star(y)\big] \ \geq\ p(1-p)\,(R_{\max} - \varepsilon/4)^2\,.$$

特别地，当 p、$R_{\max}$ 和 $\varepsilon \to 0$ 固定时，均方误差趋于零，而 $\varepsilon_{\text{pw}}(\widehat{r})$ 保持为 $\Omega(1)$，且 方差$(r^\star)$ 保持为$\Omega(R^2_{\max})$。

证明. 固定 $\varepsilon > 0$。定义响应空间

$$\mathcal{Y} \ :=\ \{H\} \ \cup\ [0, \varepsilon/2].$$

将参考策略 π_{ref} 定义为：以概率 p 抽取 $y = H$；以概率 $1-p$ 抽取 $y = u$，其中 $u \sim$ 服从均匀分布$[0, \varepsilon/2]$。真实奖励定义如下

$$r^\star(x, H) = R_{\max}, \qquad r^\star(x, u) = u \quad \text{for } u \in [0, \varepsilon/2].$$

设 $\gamma \sim$ 服从 Rademacher 分布(± 1)，且与 $y \sim \pi_{\text{ref}}$独立，并定义

$$\widehat{r}(y) \ :=\ r^\star(y) + \varepsilon\,\gamma,$$

使得 $\widehat{r}$ 在给定 y的条件下是随机的。对于成对比较，对 (y, y') 使用独立的 $\gamma, \gamma' \overset{\text{iid}}{\sim} \text{Rad}(\pm 1)$。

均方误差。给定 y，我们有 $\widehat{r}(y) - r^\star(y) = \varepsilon\gamma$，因此

$$\mathbb{E}\big[(\widehat{r}(y) - r^\star(y))^2 \mid y\big] = \varepsilon^2,$$

因此 $\varepsilon^2{}_{\mathrm{RM}}(\widehat{r},\ r^\star) = \varepsilon^2$。

成对误差下界。回顾成对结果

$$\phi_r(y, y') = \mathbf{1}\{r(y) > r(y')\} + \frac{1}{2}\mathbf{1}\{r(y) = r(y')\},$$

以及成对误差

$$\varepsilon_{\mathrm{pw}}(\widehat{r}) = \mathbb{E}_{y,y'\sim\pi_{\mathrm{ref}}}\Big[\big|\phi_{r^\star}(y, y') - \phi_{\widehat{r}}(y, y')\big|\Big].$$

设 $y,\ y'\overset{\text{iid}}{\sim}\pi_{\mathrm{ref}}$，并考虑事件

$$L := \{y \neq H\} \cap \{y' \neq H\},$$

其概率为 $\mathbb{P}(L) = (1-p)^2$。在 L 上，我们可以写成 $y = u, y' = u'$，其中 $u,\ u'\overset{\text{iid}}{\sim}$ 均匀分布 $[0, \varepsilon/2]$，并取相互独立的 $\gamma,\ \gamma'\overset{\text{iid}}{\sim}\mathrm{Rad}(\pm 1)$。定义真实差距和预测差距

$$\Delta := r^\star(y) - r^\star(y') = u - u', \qquad \widehat{\Delta} := \widehat{r}(y) - \widehat{r}(y') = \Delta + \varepsilon(\gamma - \gamma').$$

在 L上，由连续性可得 $\mathbb{P}(\Delta = 0) = 0$，且 Δ 也是连续的，因此 $\mathbb{P}(\Delta = 0) = 0$。因此，几乎必然在$L$上，

$$\phi_{r^\star}(y, y') = \mathbf{1}\{\Delta > 0\}, \qquad \phi_{\widehat{r}}(y, y') = \mathbf{1}\{\widehat{\Delta} > 0\},$$

因此，在 L 上，$\big|\phi_{r^\star}(y,\ y') - \phi_{\widehat{r}}(y,\ y')\big| = \mathbf{1}\{\mathbf{1}\{\Delta > 0\} \neq \mathbf{1}\{\widehat{\Delta} > 0\}\}$。若 $\gamma = \gamma'$，则 $\widehat{\Delta} = \Delta$，因此绝对差为 0。若 $\gamma \neq \gamma'$（其发生概率为 $1/2$），则 $\varepsilon(\gamma - \gamma') \in \{\pm 2\varepsilon\}$ 支配 Δ，因为$|\Delta| \leq \varepsilon/2$。因此，在 $L \cap \{\gamma \neq \gamma'\}$ 上，

$$\mathrm{sign}(\widehat{\Delta}) = \mathrm{sign}(\gamma - \gamma').$$

此外，$\mathrm{sign}(\gamma - \gamma')$ 是对称的，并且与 $\mathrm{sign}(\Delta)$ 独立，因此

$$\mathbb{E}\left[|\phi_{r^\star}(y, y') - \phi_{\widehat{r}}(y, y')| \;\middle|\; L,\ \gamma \neq \gamma'\right] = \mathbb{P}\left(\mathbf{1}\{\Delta > 0\} \neq \mathbf{1}\{\widehat{\Delta} > 0\} \;\middle|\; L,\ \gamma \neq \gamma'\right) = \frac{1}{2}.$$

因此,

$$\varepsilon_{\mathrm{pw}}(\widehat{r}) \ \geq\ \mathbb{P}(L) \cdot \mathbb{P}(\gamma \neq \gamma') \cdot \frac{1}{2} = (1-p)^2 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{(1-p)^2}{4}.$$

奖励方差（尺度）。令 $B := \mathbf{1}\{y = H\} \sim$ 服从伯努利分布 $\mathrm{Bernoulli}(p)$。根据全方差定律,

$$\mathrm{Var}(r^\star(y)) = \mathrm{Var}(\mathbb{E}[r^\star(y) \mid B]) + \mathbb{E}[\mathrm{Var}(r^\star(y) \mid B)].$$

我们有 $\mathbb{E}[r^\star(y) \mid B = 1] = R_{\max}$ 和 $\mathbb{E}[r^\star(y) \mid B = 0] = \mathbb{E}[u] = \varepsilon/4$，因此

$$\mathrm{Var}(\mathbb{E}[r^\star(y) \mid B]) = p(1-p)\left(R_{\max} - \varepsilon/4\right)^2.$$

此外，$\mathrm{Var}(r^\star(y) \mid B = 1) = 0$ 和 $\mathrm{Var}(r^\star(y) \mid B = 0) = \mathrm{Var}(u) = \varepsilon^2/48$，因此 $\mathbb{E}[\mathrm{Var}(r^\star(y) \mid B)] = (1-p)\varepsilon^2/48 \geq 0$。因此

$$\mathrm{Var}(r^\star(y)) \geq p(1-p)\left(R_{\max} - \varepsilon/4\right)^2.$$

□

A.2 命题 3 的证明

为了构造期望奖励遗憾的下界，我们令 $\widehat{r}$ 对一个高分区域 S 赋予相同的最大得分，该区域的 π_{ref}-质量为 $1/M$，这样任何对齐算法 $\mathcal{A}$ 都无法区分 S 内的各点，只能选择 S 上的某种分布。接着，我们定义 $r^\star$，使其将全部价值放在 S 中 $\mathcal{A}$ 最不倾向选择的部分（同时保持 ε_{pw} 和 $\mathcal{E}_M(\pi^\star\|\pi_{\text{ref}})$ 较小），从而导致较大的真实奖励遗憾。为完整起见，我们先重述定理，再给出证明。

命题 3。 固定阈值 $M \geq 2$。对于任意推理时算法 $\mathcal{A}$ 以及每个 $\varepsilon \in (0,\ 1/M)$，都存在一个具有无界奖励的实例 $(\pi_{\text{ref}},\ \widehat{r},\ r^\star)$ 以及一个比较策略 $\pi^\star \ll \pi_{\text{ref}}$，使得

$$\varepsilon_{\text{pw}} \leq \varepsilon \qquad \textit{and} \qquad \mathcal{E}_M(\pi^\star\|\pi_{\text{ref}}) = \varepsilon M,$$

而对于每个目标差距 $\gamma > 0$，

$$\mathbb{E}_{\pi^\star}[r^\star] - \mathbb{E}_{\pi_{\mathcal{A}}}[r^\star] \ \geq\ \gamma. \tag{10}$$

取 $\varepsilon \to 0$ 后，可以使遗憾任意大，同时满足 $\varepsilon_{\text{pw}} \to 0$ 和$\mathcal{E}_M(\pi^\star\|\pi_{\text{ref}}) \to 0$。

证明。 定义 $\alpha := 1/M$。我们构造 π_{ref}，使其为无原子测度，并支撑在两个互不相交的可测区域 B和 S上，其中 $\pi_{\text{ref}}(S) = \alpha$且 $\pi_{\text{ref}}(B) = 1-\alpha$。具体而言，令 B和 S为互不相交的可测集，且各自配备一个非原子概率测度。定义奖励模型

$$\widehat{r}(y) = 0 \text{ for } y \in B, \qquad \widehat{r}(y) = R_{\max} \text{ for } y \in S.$$

令 $\pi_{\mathcal{A}}$表示在输入 $(\pi_{\text{ref}},\ \widehat{r})$ 上运行 $\mathcal{A}$ 时输出的（可能经过随机化的）策略。固定 $\varepsilon \in (0,\ \alpha)$。由于 $\widehat{r}$ 在 S上为常数，该算法没有基于分数的信号来区分 S 中的各点，因此可能会任意集中于 S的某个子集上。因此，我们以对抗方式选择区域 $P \subseteq S$，使其为一个质量为 ε、且尽可能多地涵盖 $\pi_{\mathcal{A}}$质量的子集：

$$P \in \arg\sup\Big\{\pi_{\mathcal{A}}(A) : A \subseteq S,\ \pi_{\text{ref}}(A) = \varepsilon\Big\}.$$

注意，由于 π_{ref} 是无原子的，且 $\pi_{\text{ref}}(S) = \alpha$，因此在 $[0,\alpha]$ 中存在 S 的可测子集，使其具有任意指定的 π_{ref}-质量。不一定存在精确的最大化者；任何使上确界仅相差任意小松弛量的 P 都足够。此外，也定义 $C := S \setminus P$，从而 $\pi_{\text{ref}}(C) = \alpha - \varepsilon$。由于 P 在 S 的 ε-质量子集中是（近似）最大的，其 $\pi_{\mathcal{A}}$-质量至少等于 S 中占 ε/α 比例部分的平均 $\pi_{\mathcal{A}}$-质量：

$$\pi_{\mathcal{A}}(P) \ \geq\ \frac{\pi_{\text{ref}}(P)}{\pi_{\text{ref}}(S)}\,\pi_{\mathcal{A}}(S) \ =\ (\varepsilon M)\,\pi_{\mathcal{A}}(S).$$

现在，我们可以通过设置以下内容，使真实奖励在 S 的其余部分上达到最大：

$$r^\star(y) := c\,\widehat{r}(y)\,\mathbf{1}\{y \in C\} = \begin{cases} cR_{\max}, & y \in C, \\ 0, & y \in B \cup P, \end{cases}$$

对于稍后选定的某个 $c > 0$，使得 $\pi^\star := \pi_{\text{ref}}(\cdot\,|\,C)$，进而 $\mathbb{E}_{\pi^\star}[r^\star] = cR_{\max}$。接下来，我们计算$\mathcal{E}_M(\pi^\star\|\pi_{\text{ref}})$。密度比 $w^\star = \frac{d\pi^\star}{d\pi_{\text{ref}}}$等于 $\frac{1}{\pi_{\text{ref}}(C)} = \frac{1}{\alpha-\varepsilon}$在 C上，~~其他地方为 0。由于~~ $M = 1/\alpha$，我们有 $w^\star > M$在 C上，因此

$$\begin{aligned} \mathcal{E}_M(\pi^\star\|\pi_{\text{ref}}) &= \mathbb{E}_{\pi_{\text{ref}}}\big[(w^\star - M)_+\big] \\ &= \pi_{\text{ref}}(C)\Big(\frac{1}{\pi_{\text{ref}}(C)} - M\Big) \\ &= \varepsilon M. \end{aligned}$$

我们还必须确保成对误差很小。回顾成对结果

$$\phi_r(y, y') = \mathbf{1}\{r(y) > r(y')\} + \tfrac{1}{2}\,\mathbf{1}\{r(y) = r(y')\}$$

以及成对误差

$$\varepsilon_{\mathrm{pw}}(\widehat{r}) = \mathbb{E}_{y,y' \sim \pi_{\mathrm{ref}}}\left[\left|\phi_{r^\star}(y, y') - \phi_{\widehat{r}}(y, y')\right|\right].$$

在我们的构造中，$\widehat{r}$在 S上为常数，且 $r^\star$在 B、P、C的每一个上都为常数。因此，除恰好(y, y')中的一个落入 P而另一个落入 $B \cup C$的情形外，$\phi_{r^\star}(y, y') = \phi_{\widehat{r}}(y, y')$。此外，当 $y \sim P$与$y' \sim C$（或反之）时，我们有 $\phi_{\widehat{r}}(y, y')$ $= \frac{1}{2}$（在 $\widehat{r}$下为平局），而 $\phi_{r^\star}(y, y') \in \{0, 1\}$（在 $r^\star$下为严格比较）；当 $y \sim P$与 $y' \sim B$（或反之）时，我们有 $\phi_{r^\star}(y, y') = \frac{1}{2}$（在 $r^\star$下为平局），而$\phi_{\widehat{r}}(y, y') \in \{0, 1\}$（在 $\widehat{r}$下为严格比较）。在所有这些情形中，$|\phi_{r^\star}(y, y') - \phi_{\widehat{r}}$ $(y, y')| = \frac{1}{2}$。因此，

$$\begin{aligned}
\varepsilon_{\mathrm{pw}}(\widehat{r}) &= \tfrac{1}{2} \cdot \left(\mathbb{P}\left(y \in P,\, y' \notin P\right) + \mathbb{P}\left(y' \in P,\, y \notin P\right)\right) \\
&= \tfrac{1}{2} \cdot 2\,\pi_{\mathrm{ref}}(P)\left(1 - \pi_{\mathrm{ref}}(P)\right) \\
&= \varepsilon(1 - \varepsilon) \le \varepsilon.
\end{aligned}$$

上文中，我们选择了 $P \subseteq S$，使其满足 $\pi_{\mathrm{ref}}(P) = \varepsilon$，从而

$$\pi_{\mathcal{A}}(P) \;\ge\; \frac{\pi_{\mathrm{ref}}(P)}{\pi_{\mathrm{ref}}(S)}\,\pi_{\mathcal{A}}(S) \;=\; (\varepsilon M)\,\pi_{\mathcal{A}}(S),$$

接着，由于 $r^\star$ 在 $C = S \setminus P$上为 $cR_{\max}$，在其他位置均为 0，

$$\begin{aligned}
\mathbb{E}_{\pi_{\mathcal{A}}}[r^\star] &= cR_{\max}\,\pi_{\mathcal{A}}(C) \\
&= cR_{\max}\big(\pi_{\mathcal{A}}(S) - \pi_{\mathcal{A}}(P)\big) \\
&\le cR_{\max}\,\pi_{\mathcal{A}}(S)\,(1 - \varepsilon M) \\
&\le cR_{\max}(1 - \varepsilon M).
\end{aligned}$$

因此，

$$\mathbb{E}_{\pi^\star}[r^\star] - \mathbb{E}_{\pi_{\mathcal{A}}}[r^\star] \ge cR_{\max}\varepsilon M.$$

最后，对于任意目标 $\gamma > 0$，选择 $c := \gamma/(R_{\max}\varepsilon M)$即可得到（10）。

□

B 辅助引理

本节中的引理给出了贯穿后续证明的关键辅助结果。引理 4量化了在不同奖励模型之间迁移时所产生的胜率损失，并为处理上界中的遗憾分解提供了良好直觉。成对误差与胜率定义之间的对齐在此处发挥了作用：我们可以将对成对结果分歧的控制转化为对胜率差异的控制。引理 5展示了我们的胜率度量与固定奖励模型下策略所达到的期望分位数之间的关系：$\widehat{r}$。

引理 4。 固定参考策略 π_{ref}、真实奖励 $r^\star$ 和学习得到的奖励模型 $\widehat{r}$。对于任意具有密度比 w_π 的策略$\pi \ll \pi_{\mathrm{ref}}$：$= {}^{d\pi}/_{d\pi_{\mathrm{ref}}}$ 以及任意 $M \ge 1$，

$$\left|R^{r^\star}(\pi) - R^{\widehat{r}}(\pi)\right| \;\le\; M \cdot \varepsilon_{\mathrm{pw}}(\widehat{r}) \;+\; \mathcal{E}_M(\pi \| \pi_{\mathrm{ref}}).$$

证明。 对于任意评分函数 r，我们定义成对结果

$$\phi_r(y, y') = \mathbf{1}\{r(y) > r(y')\} + \frac{1}{2}\mathbf{1}\{r(y) = r(y')\} \in \left\{0, \frac{1}{2}, 1\right\}.$$

那么，根据定义，胜率泛函可写为

$$R^r(\pi) = \mathbb{E}_{y\sim\pi,\, y'\sim\pi_{\mathrm{ref}}}\left[\phi_r(y, y')\right].$$

此外，回顾成对误差

$$\varepsilon_{\mathrm{pw}}(\widehat{r}) = \mathbb{E}_{y,y'\sim\pi_{\mathrm{ref}}}\left[|\phi_{r^\star}(y, y') - \phi_{\widehat{r}}(y, y')|\right].$$

为了界定胜率差异，我们首先注意到

$$\begin{aligned}\left|R^{r^\star}(\pi) - R^{\widehat{r}}(\pi)\right| &= \left|\mathbb{E}_{y\sim\pi,\, y'\sim\pi_{\mathrm{ref}}}\left[\phi_{r^\star}(y, y') - \phi_{\widehat{r}}(y, y')\right]\right| \\ &\le \mathbb{E}_{y\sim\pi,\, y'\sim\pi_{\mathrm{ref}}}\left[|\phi_{r^\star}(y, y') - \phi_{\widehat{r}}(y, y')|\right].\end{aligned}$$

简记为 $\ell(y,\ y') = |\phi_{r^\star}(y,\ y') - \phi_{\widehat{r}}(y,\ y')|$。我们可以利用 w_π 改变测度：

$$\mathbb{E}_{y\sim\pi,\, y'\sim\pi_{\mathrm{ref}}}\left[\ell(y, y')\right] = \mathbb{E}_{y\sim\pi_{\mathrm{ref}},\, y'\sim\pi_{\mathrm{ref}}}\left[w_\pi(y)\,\ell(y, y')\right].$$

那么，我们可以分解 $w_\pi(y) = (w_\pi(y) \wedge M) + (w_\pi(y) - M)_+$，并使用 $\ell \in [0, 1]$：

$$\begin{aligned}\mathbb{E}_{y,y'\sim\pi_{\mathrm{ref}}}\left[w_\pi(y)\ell(y, y')\right] &= \mathbb{E}_{y,y'\sim\pi_{\mathrm{ref}}}\left[(w_\pi(y) \wedge M)\,\ell(y, y')\right] + \mathbb{E}_{y,y'\sim\pi_{\mathrm{ref}}}\left[(w_\pi(y) - M)_+\,\ell(y, y')\right] \\ &\le M \cdot \mathbb{E}_{y,y'\sim\pi_{\mathrm{ref}}}\left[\ell(y, y')\right] + \mathbb{E}_{y\sim\pi_{\mathrm{ref}}}\left[(w_\pi(y) - M)_+\right] \\ &= M \cdot \varepsilon_{\mathrm{pw}}(\widehat{r}) + \mathcal{E}_M(\pi\|\pi_{\mathrm{ref}}).\end{aligned}$$

□

引理 5。 固定一个奖励模型 $\widehat{r}:\ \mathcal{Y} \to \mathbb{R}$。定义中点累积分布函数变换

$$\widetilde{F}_{\widehat{r}}(t)\ :=\ \frac{1}{2}\left(F_{\widehat{r}}(t^-) + F_{\widehat{r}}(t)\right) \in [0, 1].$$

其中

$$F_{\widehat{r}}(t)\ :=\ \mathbb{P}_{y'\sim\pi_{\mathrm{ref}}}\left(\widehat{r}(y') \le t\right), \qquad F_{\widehat{r}}(t^-)\ :=\ \mathbb{P}_{y'\sim\pi_{\mathrm{ref}}}\left(\widehat{r}(y') < t\right).$$

那么，对于任意策略 π，

$$R^{\widehat{r}}(\pi)\ =\ \mathbb{E}_{y\sim\pi}\Big[\widetilde{F}_{\widehat{r}}\left(\widehat{r}(y)\right)\Big], \tag{11}$$

此外，令 $V \sim$ 均匀分布 $[0, 1]$ 与所有其他变量独立，并定义随机均匀化评分

$$\tilde{u}(y)\ :=\ F_{\widehat{r}}(\widehat{r}(y)^-)\ +\ V \cdot \left(F_{\widehat{r}}(\widehat{r}(y)) - F_{\widehat{r}}(\widehat{r}(y)^-)\right) \in [0, 1]. \tag{12}$$

那么：

1. 对于任意 π, $R^{\widehat{r}}(\pi) = \mathbb{E}_{y\sim\pi,V}\big[\tilde{u}(y)\big]$。

2. 如果 $y' \sim \pi_{\mathrm{ref}}$，那么 $\tilde{u}(y')\ \sim$ 均匀分布 $[0, 1]$。

证明。根据全期望定律,

$$R^{\widehat{r}}(\pi) = \mathbb{E}_{y\sim\pi}\Big[\mathbb{P}_{y'\sim\pi_{\mathrm{ref}}}\left(\widehat{r}(y) > \widehat{r}(y') \mid y\right) + \frac{1}{2}\,\mathbb{P}_{y'\sim\pi_{\mathrm{ref}}}\left(\widehat{r}(y) = \widehat{r}(y') \mid y\right)\Big].$$

固定 y。那么

$$\mathbb{P}_{y'\sim\pi_{\mathrm{ref}}}\left(\widehat{r}(y) > \widehat{r}(y') \mid y\right) = \mathbb{P}_{y'\sim\pi_{\mathrm{ref}}}\left(\widehat{r}(y') < \widehat{r}(y)\right) = F_{\widehat{r}}(\widehat{r}(y)^{-}),$$

and

$$\begin{aligned}\mathbb{P}_{y'\sim\pi_{\mathrm{ref}}}\left(\widehat{r}(y) = \widehat{r}(y') \mid y\right) &= \mathbb{P}_{y'\sim\pi_{\mathrm{ref}}}\left(\widehat{r}(y') \le \widehat{r}(y)\right) - \mathbb{P}_{y'\sim\pi_{\mathrm{ref}}}\left(\widehat{r}(y') < \widehat{r}(y)\right)\\ &= F_{\widehat{r}}(\widehat{r}(y)) - F_{\widehat{r}}(\widehat{r}(y)^{-}).\end{aligned}$$

将这两个恒等式代入前面的式子，得到（11）。
接下来，令 $V \sim$ 均匀分布$[0,1]$ 相互独立。根据（12）和 $\mathbb{E}[V] = \frac{1}{2}$,

$$\begin{aligned}\mathbb{E}\big[\tilde{u}(y) \mid y\big] &= F_{\widehat{r}}(\widehat{r}(y)^{-}) \;+\; \mathbb{E}[V]\cdot\big(F_{\widehat{r}}(\widehat{r}(y)) - F_{\widehat{r}}(\widehat{r}(y)^{-})\big)\\ &= F_{\widehat{r}}(\widehat{r}(y)^{-}) + \frac{1}{2}\cdot\big(F_{\widehat{r}}(\widehat{r}(y)) - F_{\widehat{r}}(\widehat{r}(y)^{-})\big)\,.\end{aligned}$$

对两边取 $\mathbb{E}_{y\sim\pi}$，并与（11）比较，得到 $R^{\widehat{r}}(\pi) = \mathbb{E}_{y\sim\pi,\;V}[\tilde{u}(y)]$。
最后，当 $y' \sim \pi_{\mathrm{ref}}$ 时，我们证明 $\tilde{u}(y')$ ∼均匀分布$[0,1]$。令 $X := \widehat{r}(y')$ 并写 e

$$A := F_{\widehat{r}}(X^{-}), \qquad B := F_{\widehat{r}}(X),$$

因此，给定 X，$\tilde{u}(y') = A + V(B - A)$ 在区间 $[A, B]$ 上服从均匀分布。任取 $u \in [0,1]$，并定义广义逆

$$t_u \;:=\; \inf\{t \in \mathbb{R} :\; F_{\widehat{r}}(t) \ge u\}.$$

根据 t_u 的定义，我们有 $F_{\widehat{r}}(t_u^{-}) \le u \le F_{\widehat{r}}(t_u)$ 以及

$$\mathbb{P}(X < t_u) = F_{\widehat{r}}(t_u^{-}), \qquad \mathbb{P}(X = t_u) = F_{\widehat{r}}(t_u) - F_{\widehat{r}}(t_u^{-}).$$

现在根据 X进行分解:

- 若 $X < t_u$，则 $B = F_{\widehat{r}}(X) \le F_{\widehat{r}}(t_u^{-}) \le u$，因此 $\mathbb{P}(\tilde{u} \le u \mid X) = 1$。
- 若 $X > t_u$，则 $A = F_{\widehat{r}}(X^{-}) \ge F_{\widehat{r}}(t_u) \ge u$，因此 $\mathbb{P}(\tilde{u} \le u \mid X) = 0$。
- 若 $X = t_u$，则根据构造，$\tilde{u}$ 在 $[F_{\widehat{r}}(t_u^{-})\;\; F_{\widehat{r}}(t_u)]$, 上服从均匀分布，因此

$$\mathbb{P}(\tilde{u} \le u \mid X = t_u) = \frac{u - F_{\widehat{r}}(t_u^{-})}{F_{\widehat{r}}(t_u) - F_{\widehat{r}}(t_u^{-})}.$$

因此,

$$\begin{aligned}\mathbb{P}(\tilde{u} \le u) &= \mathbb{P}(X < t_u)\cdot 1 \;+\; \mathbb{P}(X = t_u)\cdot\frac{u - F_{\widehat{r}}(t_u^{-})}{F_{\widehat{r}}(t_u) - F_{\widehat{r}}(t_u^{-})}\\ &= F_{\widehat{r}}(t_u^{-}) \;+\; \big(F_{\widehat{r}}(t_u) - F_{\widehat{r}}(t_u^{-})\big)\cdot\frac{u - F_{\widehat{r}}(t_u^{-})}{F_{\widehat{r}}(t_u) - F_{\widehat{r}}(t_u^{-})} \;=\; u.\end{aligned}$$

由于这对所有 $u \in [0,1]$ 都成立，我们得出 $\tilde{u}(y')$ ∼均匀分布$[0,1]$。 □

C Best-of-N 的证明

本节给出了第 3 节中 Best-of-N 的保证的证明，第 3 节。本节的结果建立在 第 5.1 节中提到的近似拒绝采样保证之上，组织如下：

1. C.1 节 根据 $\mathcal{E}_M$-散度，为 BoN 的遗憾给出一个一般上界；该散度在 第 2.3 节中引入。在 C.2 节 中，我们给出遗憾的一般下界。
2. 在 C.3 节 中，我们给出一个下界，以展示 BoN 的计算效率。

下文定理 3 的证明用到了 Block 和 Polyanskiy [2] 的近似拒绝采样分析。简要回顾一下：给定基分布 π_{ref} 和密度比为 $w := d\pi/d\pi_{\text{ref}}$的目标策略 $\pi \ll \pi_{\text{ref}}$，经典拒绝采样会抽取 $y \sim \pi_{\text{ref}}$，并以概率 $\min\{w(y)/M, 1\}$接受它；若 $w \le M$ π_{ref}-几乎处处成立，则接受样本恰好服从 π（冯·诺伊曼 [32]）。当 w 无统一上界时，同样的接受／拒绝规则会截断目标分布：在 $w(y) > M$ 的区域，质量无法完全实现，实际上会在 M这一水平被“截断”。执行该过程进行 N 次提议（若未接受则返回默认提议），并令 M 足够大，可得到策略 π_R，其在全变差意义下近似于 π，误差由截断质量$\mathcal{E}_M(\pi\|\pi_{\text{ref}})$加上一个指数级小的未能接受项控制（2；见引理 8）。

C.1 定理 3 的证明

证明。 如第 5.1 节所述，为了在真实奖励 $r^\star$下，将 $\widehat{\pi}_{\text{BoN}}$ 与任意比较策略 $\pi^\star$ 进行比较， $r^\star$我们需要权衡两种相互竞争的影响：如果 N 太小，BoN 可能无法找到与 $\pi^\star$偏好的响应相当的响应；如果 N 太大，BoN 可能会对 $\widehat{r}$ 过拟合，并挑选出 $\widehat{r}$ 与$r^\star$ 不一致的输出。这两种影响恰好对应于上界（5）中的两项，并通过以下遗憾分解体现出来。该分解对任意 $M > 0$都成立：遗憾 $R^{r^\star}(\pi^\star) - R^{r^\star}(\widehat{\pi}_{\text{BoN}})$ 等于以下三项之和：

$$\underbrace{R^{r^\star}(\pi^\star) - R^{\widehat{r}}(\pi^\star_M)}_{\text{(T1)}} + \underbrace{R^{\widehat{r}}(\pi^\star_M) - R^{\widehat{r}}(\widehat{\pi}_{\text{BoN}})}_{\text{(T2)}} + \underbrace{R^{\widehat{r}}(\widehat{\pi}_{\text{BoN}}) - R^{r^\star}(\widehat{\pi}_{\text{BoN}})}_{\text{(T3)}},$$

其中， $\pi^\star_M$ 是一个中间比较策略，它是在全变差意义下对 $\pi^\star$ 最接近的 M截断近似：

$$\pi^\star_M \in \operatorname*{argmin}_{\pi \ll \pi_{\text{ref}}:\, \frac{d\pi}{d\pi_{\text{ref}}} \le M \ \pi_{\text{ref}}\text{-a.s.}} \text{TV}(\pi^\star, \pi). \tag{13}$$

为界定 (T1)，我们可以进一步分解

$$R^{r^\star}(\pi^\star) - R^{\widehat{r}}(\pi^\star_M) = R^{r^\star}(\pi^\star) - R^{r^\star}(\pi^\star_M) + R^{r^\star}(\pi^\star_M) - R^{\widehat{r}}(\pi^\star_M)$$

并分别控制由此产生的两个项。对于第一个差值，我们注意到胜率有界，因此 1-利普希茨连续于全变差距离（它可以写成中点累积分布函数的期望，见 ~累积分布函数 $F_{\widehat{r}}(\widehat{r}(y)) \in [0,1]$中的定义），所以我们可以根据 引理 5，用 $\pi^\star$到 $\pi^\star_M$的 $\mathcal{E}_M(\pi^\star\|\pi_{\text{ref}})$来控制损失。Block 和 Polyanskiy [2]指出： 引理 6 表明，任何密度比有界的策略，其与 $\pi^\star$的全变差距离都不可能小于 $\mathcal{E}_M(\pi^\star\|\pi_{\text{ref}})$；而 引理 7 则证明，全变差

如上所定义的 $\pi^\star$与 $\pi_M^\star$之间的距离小于 $\mathcal{E}_M(\pi^\star\|\pi_{\mathrm{ref}})$，因此 TV($\pi^\star$，$\pi_M^\star$) $= \mathcal{E}_M(\pi^\star\|\pi_{\mathrm{ref}})$，于是

$$R^{r^\star}(\pi^\star) - R^{r^\star}(\pi_M^\star) \;\le\; \mathrm{TV}(\pi^\star, \pi_M^\star) = \mathcal{E}_M(\pi^\star\|\pi_{\mathrm{ref}}).$$

对于第二个差值，我们利用以下事实：$\pi_M^\star$ 相对于 π_{ref} 的似然比几乎必然以 M为上界，从而控制 $\pi_M^\star$ 下 $\widehat{r}$ 与 $r^\star$之间的成对胜率误差。具体而言，我们有$w_{\pi_M^\star}(y) \le M$ π_{ref}-几乎必然成立，因此 $\mathcal{E}_M(\pi_M^\star\|\pi_{\mathrm{ref}}) = 0$并应用引理 4，取 $\pi = \pi_M^\star$并进行截断 M，得到

$$R^{r^\star}(\pi_M^\star) - R^{\widehat{r}}(\pi_M^\star) \le M\varepsilon_{\mathrm{pw}} + \mathcal{E}_M(\pi_M^\star\|\pi_{\mathrm{ref}}) = M\varepsilon_{\mathrm{pw}}.$$

综合以上结果，我们有

$$(\mathrm{T1}) = R^{r^\star}(\pi^\star) - R^{\widehat{r}}(\pi_M^\star) \le \mathcal{E}_M(\pi^\star\|\pi_{\mathrm{ref}}) + M\varepsilon_{\mathrm{pw}}.$$

为了界定 (T2)，我们引入另一个中间比较对象，以便利用 $\widehat{\pi}_{\mathrm{BoN}}$在固定批次的 π_{ref}中抽取的 N 个独立同分布样本上、针对 $\widehat{r}$ 的最优性。采用 Block 和 Polyanskiy [2], 提出的近似拒绝采样形式，我们将选择规则定义为策略 $\pi_{M,\ \mathrm{rej}}^\star$ 它仅使用来自 π_{ref}的 N 次独立同分布抽样来近似$\pi_M^\star$；该规则返回第一个被接受的 y_i，若没有样本被接受，则返回 y_1。我们再次将 (T2) 进一步分解为

$$R^{\widehat{r}}(\pi_M^\star) - R^{\widehat{r}}(\widehat{\pi}_{\mathrm{BoN}}) = R^{\widehat{r}}(\pi_M^\star) - R^{\widehat{r}}(\pi_{M,\mathrm{rej}}^\star) + R^{\widehat{r}}(\pi_{M,\mathrm{rej}}^\star) - R^{\widehat{r}}(\widehat{\pi}_{\mathrm{BoN}})$$

并分别控制由此产生的两个项。对于第一个差值，我们使用 Block 和 Polyanskiy [2] 的近似拒绝采样分析来证明，$\pi_{M,\ \mathrm{rej}}^\star$ 在总变差距离上近似于 $\pi_M^\star$，误差至多为关于 N/M的指数小项；由有界性可知，这一结论同样可传递至胜率。因此

$$\begin{aligned} R^{\widehat{r}}(\pi_M^\star) - R^{\widehat{r}}(\pi_{M,\mathrm{rej}}^\star) &\le \mathrm{TV}\left(\pi_M^\star, \pi_{M,\mathrm{rej}}^\star\right) \\ &\le \mathcal{E}_M(\pi_M^\star\|\pi_{\mathrm{ref}}) + \frac{1}{2}\exp\left(-\frac{N}{M}\left(1 - \mathcal{E}_M(\pi_M^\star\|\pi_{\mathrm{ref}})\right)\right) \\ &= \frac{1}{2}\exp(-N/M), \end{aligned}$$

其中，第二个不等式由引理 8推出，最后一个等式成立是因为 $\pi_M^\star$ 是 M-截断的，因此 $\mathcal{E}_M($ $\pi_M^\star\|\pi_{\mathrm{ref}}) = 0$。为界定第二个差值，我们注意到，在给定来自$N$ 的任意固定批次 π_{ref}样本的条件下，BoN 只需选择 $\widehat{r}$ 值最高的输出，因此在 $\widehat{r}$下，它在同一批次上优于任何其他选择规则，包括$\pi_{M,\ \mathrm{rej}}^\star$。具体而言，我们首先应用引理 5~，将 $\widehat{r}$ 下的胜率改写为 $F_{\widehat{r}}(\widehat{r}(\cdot))$的期望，从而得到

$$R^{\widehat{r}}(\pi_{M,\mathrm{rej}}^\star) - R^{\widehat{r}}(\widehat{\pi}_{\mathrm{BoN}}) = \mathbb{E}_{y\sim\pi_{M,\mathrm{rej}}^\star}\left[\widetilde{F}_{\widehat{r}}(\widehat{r}(y))\right] - \mathbb{E}_{y\sim\widehat{\pi}_{\mathrm{BoN}}}\left[\widetilde{F}_{\widehat{r}}(\widehat{r}(y))\right].$$

现在，我们可以依据 $\widehat{r}$下 $\widehat{\pi}_{\mathrm{BoN}}$ 的最优性。下文中，我们用 $\mathbb{E}_{\widehat{\mathcal{Y}}_N\sim\pi_{\mathrm{ref}}}[\cdot]$ 表示对从 π_{ref} 抽取的 N 个样本 $\mathcal{Y}_N$ 求期望；给定 $\mathcal{Y}_N$，我们用 $\mathbb{E}_{\pi_{M,\ \mathrm{rej}}^\star|\widehat{\mathcal{Y}}_N}[\cdot]$表示在给定算法抽取的集合 $\mathcal{Y}_N$ 后，通过运行近似拒绝采样得到的$\widehat{\ }$响应的期望；我们以类似方式定义 $\mathbb{E}_{\widehat{\pi}_{\mathrm{BoN}}|\widehat{\mathcal{Y}}_N}[\cdot]$。于是，我们可以利用期望的线性性，从而得到

$$\mathbb{E}_{\widehat{\mathcal{Y}}_N\sim\pi_{\mathrm{ref}}}\left[\mathbb{E}_{y\sim\pi_{M,\mathrm{rej}}^\star|\widehat{\mathcal{Y}}_N}\left[\widetilde{F}_{\widehat{r}}\left(\widehat{r}(y)\right)\right] - \mathbb{E}_{y\sim\widehat{\pi}_{\mathrm{BoN}}|\widehat{\mathcal{Y}}_N}\left[\widetilde{F}_{\widehat{r}}\left(\widehat{r}(y)\right)\right]\right]$$

将两个算法抽取的集合 $\mathcal{Y}_N$ 配对，并比较在固定的 N 个响应下，$\widehat{\pi}_{\mathrm{BoN}}$ 和 $\pi^{\star}_{M,\ \mathrm{rej}}$ 的性能。接着，由于对于任意固定的 $\mathcal{Y}_N$，BoN 策略总是选择在 $\widehat{r}$ 下具有最大值的响应，且 $F_{\widehat{r}}$ 单调不减，因此对于所有 $y \in \mathcal{Y}_N$，都有 $F_{\widehat{r}}(\widehat{r}(y)) \leq \max_{i\in[N]} F_{\widehat{r}}(\widehat{r}(y_i))$。因此，

$$R^{\widehat{r}}(\pi^{\star}_{M,\mathrm{rej}}) - R^{\widehat{r}}(\widehat{\pi}_{\mathrm{BoN}}) \leq 0.$$

结合这些步骤可得

$$(\mathrm{T2}) = R^{\widehat{r}}(\pi^{\star}_{M}) - R^{\widehat{r}}(\widehat{\pi}_{\mathrm{BoN}}) \leq \frac{1}{2}\exp(-N/M).$$

最后，我们使用与（T1）中类似的测度变换论证来控制（T3）。在再次应用引理 4 之前，我们首先证明 BoN 相对于 N的密度比至多为 π_{ref}，因此在 $\widehat{\pi}_{\mathrm{BoN}}$ 下，$\widehat{r}$ 与 $r^{\star}$ 之间的成对胜率误差至多为 $N \cdot \varepsilon_{\mathrm{pw}}(\widehat{r})$。令 $\widehat{\mathcal{Y}}_N := (y_1, \ldots\ldots, y_N) \in \mathcal{Y}^N$ 表示从 N 中进行的 π_{ref} 独立同分布抽样序列，并记为

$$\pi_{\mathrm{ref}}(\widehat{\mathcal{Y}}_N) := \prod_{i=1}^{N} \pi_{\mathrm{ref}}(y_i)$$

由于在给定 $\mathcal{Y}_N$ 的条件下，BoN 算法通过固定的平局裁决规则确定性地输出 $\mathcal{Y}_N$ 中的单个元素，因此我们可以将 $\widehat{\pi}_{\mathrm{BoN}}$ 写成闭式形式：

$$\widehat{\pi}_{\mathrm{BoN}}(y) = \sum_{\widehat{\mathcal{Y}}_N \in \mathcal{Y}^N} \pi_{\mathrm{ref}}(\widehat{\mathcal{Y}}_N) \cdot \mathbf{1}\{y \in \widehat{\mathcal{Y}}_N\} \cdot \mathbf{1}\Big\{\widehat{r}(y) \geq \widehat{r}(y')\ \ \forall\, y' \in \widehat{\mathcal{Y}}_N\Big\}. \tag{14}$$

接下来，通过对 $\mathcal{Y}_N$ 进行边缘化，并从 $\{1, \ldots, N\}$ 中均匀分布地采样一个索引，来表示 $\pi_{\mathrm{ref}}(y)$：

$$\pi_{\mathrm{ref}}(y) = \sum_{\widehat{\mathcal{Y}}_N \in \mathcal{Y}^N} \pi_{\mathrm{ref}}(\widehat{\mathcal{Y}}_N) \cdot \sum_{y' \in \widehat{\mathcal{Y}}_N} \frac{\mathbf{1}\{y' = y\}}{N}. \tag{15}$$

结合（14）和（15），对于任意 y 满足 $\pi_{\mathrm{ref}}(y) > 0$，我们有

$$\begin{aligned}
\frac{\widehat{\pi}_{\mathrm{BoN}}(y)}{\pi_{\mathrm{ref}}(y)} &= N \cdot \frac{\sum_{\widehat{\mathcal{Y}}_N \in \mathcal{Y}^N} \pi_{\mathrm{ref}}(\widehat{\mathcal{Y}}_N) \cdot \mathbf{1}\{y \in \widehat{\mathcal{Y}}_N\} \cdot \mathbf{1}\Big\{\widehat{r}(y) \geq \widehat{r}(y')\ \forall\, y' \in \widehat{\mathcal{Y}}_N\Big\}}{\sum_{\widehat{\mathcal{Y}}_N \in \mathcal{Y}^N} \pi_{\mathrm{ref}}(\widehat{\mathcal{Y}}_N) \cdot \sum_{y' \in \widehat{\mathcal{Y}}_N} \mathbf{1}\{y' = y\}} \\
&\leq N \cdot \frac{\sum_{\widehat{\mathcal{Y}}_N \in \mathcal{Y}^N} \pi_{\mathrm{ref}}(\widehat{\mathcal{Y}}_N) \cdot \mathbf{1}\{y \in \widehat{\mathcal{Y}}_N\}}{\sum_{\widehat{\mathcal{Y}}_N \in \mathcal{Y}^N} \pi_{\mathrm{ref}}(\widehat{\mathcal{Y}}_N) \cdot \mathbf{1}\{y \in \widehat{\mathcal{Y}}_N\}} \\
&= N.
\end{aligned}$$

因此，$\widehat{\pi}_{\mathrm{BoN}}/\pi_{\mathrm{ref}} \leq N$ 逐点成立，并且根据定义，

$$\mathcal{E}_N(\widehat{\pi}_{\mathrm{BoN}} \| \pi_{\mathrm{ref}}) = \mathbb{E}_{y\sim\pi_{\mathrm{ref}}}\left[\left(\frac{\widehat{\pi}_{\mathrm{BoN}}(y)}{\pi_{\mathrm{ref}}(y)} - N\right)_{+}\right] = 0.$$

现在我们可以将引理 4 应用于 $\pi = \widehat{\pi}_{\mathrm{BoN}}$ 并进行截断 $M = N$，使得

$$(\mathrm{T3}) \leq N\varepsilon_{\mathrm{pw}} + \mathcal{E}_N(\widehat{\pi}_{\mathrm{BoN}} \| \pi_{\mathrm{ref}}) = N\varepsilon_{\mathrm{pw}}.$$

综合以上所有界，我们得到

$$R^{r^{\star}}(\pi^{\star}) - R^{r^{\star}}(\widehat{\pi}_{\mathrm{BoN}}) \lesssim \mathcal{E}_M(\pi^{\star} \| \pi_{\mathrm{ref}}) + (M+N)\varepsilon_{\mathrm{pw}} + \exp(-N/M).$$

令 $M = N/\log(1/\varepsilon_{\mathrm{pw}}(\widehat{r}))$，并注意到对于此类 M，有 $\exp(-N/M) = \varepsilon_{\mathrm{pw}}(\widehat{r})$，即可得到所需结果。

□

引理 6 (Block 和 Polyanskiy 中的引理 32 [2])。 固定 $N \geq 1$并假设 $\pi^\star \ll \pi_{\mathrm{ref}}$ 的密度比为$w^\star$：$= {}^{d\pi^\star}\!/_{d\pi_{\mathrm{ref}}}$。 那么，对于任意 $\pi \ll \pi_{\mathrm{ref}}$ 其密度比 w： $= {}^{d\pi}\!/_{d\pi_{\mathrm{ref}}}$ 满足 $w \leq M$ π_{ref}-几乎必然，有

$$\mathrm{TV}(\pi, \pi^\star) \;\geq\; \mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}}).$$

引理 7. 固定 $N \geq 1$ ，并假设 $\pi^\star \ll \pi_{\mathrm{ref}}$ 的密度比为 $w^\star$： $= {}^{d\pi^\star}\!/_{d\pi_{\mathrm{ref}}}$。 则存在一个策略$\pi_M^\star \ll \pi_{\mathrm{ref}}$，使得 $\frac{d\pi_M^\star}{d\pi_{\mathrm{ref}}}(y) \leq M$ π_{ref}-几乎必然成立，且

$$\mathrm{TV}(\pi_M^\star, \pi^\star) = \mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}}).$$

证明. 定义截断密度

$$w_M(y) \;:=\; \min\{w^\star(y),\, M\}.$$

利用 $w^\star = w_M + (w^\star - M)_+$逐点成立以及 $\mathbb{E}_{y\sim\pi_{\mathrm{ref}}}[w^\star(y)] = 1$，可得

$$\mathbb{E}_{y\sim\pi_{\mathrm{ref}}}[w_M(y)] \;=\; 1 - \mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}}).$$

因此， w_M的总质量为 $1 - \mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}})$，但当 $\mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}}) > 0$时它尚不是有效的概率密度，故必须在确保密度比仍然有界的同时补回超额质量。定义 M下的松弛量：

$$S \;:=\; \mathbb{E}_{y\sim\pi_{\mathrm{ref}}}\big[(M - w^\star(y))_+\big].$$

由于$(M - w^\star)_+ = M - \min\{w^\star, M\} = M - w_M$逐点成立，我们有

$$S = M - \mathbb{E}_{y\sim\pi_{\mathrm{ref}}}[w_M(y)] = M - (1 - \mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}})) \geq \mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}}).$$

如果 $\mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}}) = 0$，则我们令 $g \equiv 0$和 $\pi_M^\star = \pi^\star$，即可完成。否则，我们定义

$$g(y) \;:=\; \mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}}) \cdot \frac{(M - w^\star(y))_+}{S}.$$

那么， $0 \leq g(y) \leq (M - w^\star(y))_+$ π_{ref}-几乎处处成立，且 $\mathbb{E}_{y\sim\pi_{\mathrm{ref}}}[g(y)] = \mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}})$。现在我们通过以下方式定义 $\pi_M^\star \ll \pi_{\mathrm{ref}}$

$$\frac{d\pi_M^\star}{d\pi_{\mathrm{ref}}}(y) \;:=\; w_M(y) + g(y).$$

我们可以验证归一化成立：

$$\mathbb{E}_{y\sim\pi_{\mathrm{ref}}}\left[\frac{d\pi_M^\star}{d\pi_{\mathrm{ref}}}(y)\right] = \mathbb{E}_{y\sim\pi_{\mathrm{ref}}}[w_M(y)] + \mathbb{E}_{y\sim\pi_{\mathrm{ref}}}[g(y)] = (1 - \mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}})) + \mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}}) = 1.$$

此外，逐点密度比界仍然成立：如果 $w^\star(y) > M$ 那么 $w_M(y) = M$ 并且 $(M - w^\star(y))_+ = 0$，因此 $g(y) = 0$并且 $w_M(y) + g(y) = M$；如果 $w^\star(y) \leq M$那么 $w_M(y) = w^\star(y)$ 并且 $g(y) \leq M - w^\star(y)$，因此 $w_M(y) + g(y) \leq M$。所以 $\mathcal{E}_M(\pi_M^\star \| \pi_{\mathrm{ref}}) = 0$。最后，我们使用密度计算全变差距离：

$$\mathrm{TV}(\pi^\star, \pi_M^\star) = \frac{1}{2}\,\mathbb{E}_{y\sim\pi_{\mathrm{ref}}}\Big[\Big|w^\star(y) - \big(w_M(y) + g(y)\big)\Big|\Big].$$

逐点地，

$$\left|w^\star(y) - \big(w_M(y) + g(y)\big)\right| = (w^\star(y) - M)_+ + g(y),$$

因为在 $\{w^\star > M\}$ 上，我们有 $w_M = M$ 和 $g = 0$；而在 $\{w^\star \le M\}$ 上，我们有 $w_M = w^\star$。取期望可得

$$\begin{aligned}\mathbb{E}_{y\sim\pi_{\mathrm{ref}}}\Big[\left|w^\star(y) - \big(w_M(y) + g(y)\big)\right|\Big] &= \mathbb{E}_{y\sim\pi_{\mathrm{ref}}}[(w^\star(y) - M)_+] + \mathbb{E}_{y\sim\pi_{\mathrm{ref}}}[g(y)] \\ &= 2\mathcal{E}_M(\pi^\star\|\pi_{\mathrm{ref}}),\end{aligned}$$

因此，全变差$(\pi^\star, \pi^\star_M)$ $= \mathcal{E}_M(\pi^\star\|\pi_{\mathrm{ref}})$，这便得到了所述上界。 □

以下引理（其内容最初改编自 Block 和 Polyanskiy [2]中的定理 3）给出了 Block 和 Polyanskiy [2] 所详述的近似拒绝采样过程的样本复杂度上界。关于这一结果的详细讨论，请参阅 Huang 等人 [14] 。

引理 8 (Huang 等人 [14]中的引理 D.4)。 对于任意有效策略 $\pi \in \Delta(\mathcal{Y})$以及 N、$M > 0$，令 $\pi_R \in \Delta(\mathcal{Y})$ 表示通过运行近似拒绝采样所诱导的响应分布。那么

$$\mathrm{TV}(\pi, \pi_R) \le \mathcal{E}_M(\pi\|\pi_{\mathrm{ref}}) + \frac{1}{2}\exp\left(-\frac{N\cdot(1 - \mathcal{E}_M(\pi\|\pi_{\mathrm{ref}})}{M}\right).$$

C.2 定理4 的证明

我们首先重述为完整起见陈述该定理，随后给出证明 .

定理 8. 对于任意无原子 π_{ref}、采样与评估框架中的任意 $\widehat{\pi}_{\mathcal{A}}$ 以及任意 $\varepsilon < 1/2$，存在奖励函数 $\widehat{r}, r^\star$ 满足 $\varepsilon_{\mathrm{pw}}(\widehat{r}) \le \varepsilon$，以及一个比较策略 $\pi^\star \ll \pi_{\mathrm{ref}}$，使得

$$R^{r^\star}(\pi^\star) - R^{r^\star}(\widehat{\pi}_{\mathcal{A}}) \ge k\cdot\inf_{M\ge 1}\left\{M\cdot\varepsilon + \mathcal{E}_M(\pi^\star\|\pi_{\mathrm{ref}})\right\},$$

其中 $k > 0$为某个普适常数。

证明。 我们分三步构造一个困难实例。首先，找出一个高概率区域 S，在该区域上$\widehat{\pi}$ 对 π_{ref} 的加权程度至多超过一个常数因子；在 S 内部，我们划出一个小集合 A，使其 π_{ref}质量与成对误差预算 ε 同阶，其中 $\widehat{\pi}$ $(A) \lesssim \varepsilon$。接着，我们选择一个简单的奖励函数对 $(\widehat{r}, r^\star)$，使算法在满足成对误差预算的同时，无法可靠地区分 A 与 A^c。然后，我们让 $\pi^\star$在 A上将 π_{ref}的权重提高 $M^\star$倍，从而迫使 $\pi^\star(A) \approx M^\star\varepsilon$。在我们的实例中，遗憾与 $\pi^\star(A) - \widehat{\pi}(A)$成正比，因此我们将把该遗憾与天际线目标联系起来。

首先，我们构造一个小集合 A，使得 $\widehat{\pi}$ 在其上的质量很小。固定任意一个采样与评估算法 $\widehat{\pi}$。令$\hat{}w(y)$ $:= d\widehat{\pi}/d\pi_{\mathrm{ref}}(y)$，使得 $\mathbb{E}_{\pi_{\mathrm{ref}}}[\hat{}w] = 1$。根据马尔可夫不等式，

$$\mathbb{P}_{y\sim\pi_{\mathrm{ref}}}\left(\hat{w}(y) > 2\right) \le \frac{1}{2},$$

因此 $S := \{\hat{}w \le 2\}$满足 $\pi_{\mathrm{ref}}(S) \ge \frac{1}{2}$。令

$$\varepsilon' := \min\{\varepsilon, 1/M^\star\} \in (0, \tfrac{1}{2}].$$

由于 π_{ref}是无原子的，且 $\varepsilon' \le \pi_{\mathrm{ref}}(S)$，选取一个可测的 $A \subseteq S$，使得 $\pi_{\mathrm{ref}}(A) = \varepsilon'$。那么

$$\widehat{\pi}(A) = \int_A \hat{w}\, d\pi_{\mathrm{ref}} \le 2\,\pi_{\mathrm{ref}}(A) = 2\varepsilon'. \tag{16}$$

我们以这种方式设置 ε' 有两个原因：首先，我们在下文构造的比较策略会在A上放置密度 $M^\star$，因此可行性要求 $\pi^\star(A) = M^\star\,\pi_{\mathrm{ref}}(A) \le 1$，即 $\pi_{\mathrm{ref}}(A) \le 1/M^\star$。此外，我们将通过选择不大于 ε 的 $\pi_{\mathrm{ref}}(A)$ 来满足成对误差预算，接下来会看到这一点。定义

$$r^\star(y) := \mathbf{1}\{y \in A\}, \qquad \widehat{r}(y) \equiv 0.$$

直观地说， $r^\star$ 使 A 成为唯一理想的区域，而常数 $\widehat{r}$ 无法提供关于是否属于 A的信息，因此，对 A 的任何可靠偏好都只能来自偶然，而非信息。对于任意奖励函数 r，回顾成对结果

$$\phi_r(y, y') = \mathbf{1}\{r(y) > r(y')\} + \tfrac{1}{2}\,\mathbf{1}\{r(y) = r(y')\},$$

以及成对误差

$$\varepsilon_{\mathrm{pw}}(\widehat{r}) = \mathbb{E}_{y,y' \sim \pi_{\mathrm{ref}}}\left[\,|\phi_{r^\star}(y, y') - \phi_{\widehat{r}}(y, y')|\,\right].$$

由于 $\widehat{r}$是常数， $\phi_{\widehat{r}}(y,\ y') \equiv \frac{1}{2}$。在 $r^\star = \mathbf{1}_A$下， $\phi_{r^\star}(y,\ y') \in \{0, \frac{1}{2}, 1\}$且 $\left|\phi_{r^\star}(y,\ y') - \frac{1}{2}\right| \equiv \frac{1}{2}$当且仅当 $(y,\ y')$ 中恰有一个属于 A。因此

$$\varepsilon_{\mathrm{pw}}(\widehat{r}) = \frac{1}{2} \cdot \mathbb{P}_{y,y' \sim \pi_{\mathrm{ref}}}\left(\mathbf{1}_A(y) \ne \mathbf{1}_A(y')\right) = \frac{1}{2} \cdot 2\varepsilon'(1 - \varepsilon') \le \varepsilon' \le \varepsilon.$$

接下来，我们构造 $\pi^\star$ 并评估天际线目标。通过以下方式定义 $\pi^\star \ll \pi_{\mathrm{ref}}$

$$w^\star(y) := \frac{d\pi^\star}{d\pi_{\mathrm{ref}}}(y) = \begin{cases} M^\star, & y \in A, \\ c, & y \notin A, \end{cases} \qquad c := \frac{1 - M^\star\varepsilon'}{1 - \varepsilon'} \ge 0 \quad (\text{since } \varepsilon' \le 1/M^\star).$$

注意，由于 $M^\star \ge 1$ 和 $\varepsilon' \in$（$\mathbf{0}$，$\mathbf{1}$），我们有 $1 - M^\star\varepsilon' \le 1 - \varepsilon'$，因此 $c \le 1$。于是 $\mathbb{E}_{\pi_{\mathrm{ref}}}[w^\star] = M^\star\varepsilon' + c\,(1 - \varepsilon') = 1$，所以 $\pi^\star$ 是一个有效策略，并且

$$\pi^\star(A) = \int_A w^\star\, d\pi_{\mathrm{ref}} = M^\star\varepsilon'. \tag{17}$$

此外，对于任意 $M \in [c, M^\star]$,

$$\mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}}) = \mathbb{E}_{\pi_{\mathrm{ref}}}\left[(w^\star - M)_+\right] = (M^\star - M)\pi_{\mathrm{ref}}(A) = (M^\star - M)\varepsilon'.$$

因此，对于这样的 M,

$$M\varepsilon + \mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}}) = M\varepsilon + (M^\star - M)\varepsilon' = M^\star\varepsilon' + M(\varepsilon - \varepsilon'). \tag{18}$$

接下来，我们可以显式计算胜率和遗憾。固定任意策略 π，并记 $\pi_{\mathrm{ref}}(A) = \varepsilon'$。利用 $r^\star = \mathbf{1}_A$下胜率的定义:

$$\begin{aligned} R^{r^\star}(\pi) &= \mathbb{P}(r^\star(y) > r^\star(y')) + \tfrac{1}{2}\,\mathbb{P}(r^\star(y) = r^\star(y')) \\ &= \pi(A)(1 - \pi_{\mathrm{ref}}(A)) + \frac{1}{2}\left(\pi(A)\,\pi_{\mathrm{ref}}(A) + (1 - \pi(A))(1 - \pi_{\mathrm{ref}}(A))\right) \\ &= \frac{1}{2}(1 - \pi_{\mathrm{ref}}(A)) + \frac{1}{2}\,\pi(A). \end{aligned}$$

对 $\pi = \pi^\star$ 与 $\pi = \widehat{\pi}$ 作差可消去平局质量项，从而得到

$$R^{r^\star}(\pi^\star) - R^{r^\star}(\widehat{\pi}) = \frac{1}{2}\left(\pi^\star(A) - \widehat{\pi}(A)\right). \tag{19}$$

将 (19) 与 (17) 和 (16) 结合可得

$$R^{r^\star}(\pi^\star) - R^{r^\star}(\widehat{\pi}) \geq \frac{1}{2}\left(M^\star\varepsilon' - 2\varepsilon'\right) = \frac{M^\star - 2}{2}\,\varepsilon'. \tag{20}$$

最后，我们将遗憾与天际线目标联系起来。我们考虑两种情形。

情形 *1*： $M^\star\varepsilon \leq 1$。那么 $\varepsilon' = \varepsilon$，由 (18) 可得，对所有 $M \in [c, M^\star]$,

$$M\varepsilon + \mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}}) = M^\star\varepsilon.$$

特别地,

$$\inf_{M \geq 1}\left\{M\varepsilon + \mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}})\right\} = M^\star\varepsilon.$$

事实上，对于 $M > M^\star$， $\mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}}) = 0$因此 $M\varepsilon \geq M^\star\varepsilon$；于是，对 $M \geq 1$取下确界等于 $M^\star\varepsilon$。结合 (20)，对于 $M^\star \geq 4$,

$$R^{r^\star}(\pi^\star) - R^{r^\star}(\widehat{\pi}) \geq \frac{M^\star - 2}{2}\,\varepsilon \geq \frac{1}{4}\,M^\star\varepsilon = \frac{1}{4}\inf_{M \geq 1}\left\{M\varepsilon + \mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}})\right\}.$$

情形 *2*： $M^\star\varepsilon > 1$。那么 $\varepsilon' = 1/M^\star$，因此 (20) 得到

$$R^{r^\star}(\pi^\star) - R^{r^\star}(\widehat{\pi}) \geq \frac{M^\star - 2}{2M^\star} \geq \frac{1}{4},$$

其中 $M^\star \geq 4$。另一方面，对于任意 $M \geq 1$，我们有 $\mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}}) = \mathbb{E}_{\pi_{\mathrm{ref}}}[(w^\star - M)_+] \leq \mathbb{E}_{\pi_{\mathrm{ref}}}[w^\star] = 1$，因此使用 $\varepsilon \leq \frac{1}{2}$，

$$\inf_{M \geq 1}\left\{M\varepsilon + \mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}})\right\} \leq \varepsilon + 1 \leq \frac{3}{2}.$$

因此 $\frac{1}{4} \geq \frac{1}{6}\inf_{M \geq 1}\{M\varepsilon + \mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}})\}$。令 $k = 6$服从均匀分布，即可完成证明。 □

C.3 命题1 的证明

证明。固定 $M \geq 2$。我们构造一个困难实例，其中最优策略必须将全部权重放在 π_{ref} 的一个子集上，并证明在采样与评估框架中，任何算法都需要数量级为 M 的样本，才能从该子集中采样并避免遭受常数遗憾。令 $\mathcal{Y} = [0, 1]$，并令 π_{ref} 服从均匀分布 $\mathrm{Unif}[0, 1]$。定义

$$I := \left[1 - \frac{1}{M},\, 1\right], \qquad \pi_{\mathrm{ref}}(I) = \frac{1}{M}.$$

令 $\hat{r}(y) = r^\star(y) := \mathbf{1}\{y \in I\}$，则 $\varepsilon_{\mathrm{pw}}(\hat{r},\ r^\star) = 0$。令 $\pi^\star := \pi_{\mathrm{ref}}(\cdot \mid I)$，使得 $\frac{d\pi^\star}{d\pi_{\mathrm{ref}}}(y) = M\mathbf{1}\{y \in I\}$，因此 $\mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}}) = 0$。此外，在 $\pi^\star$ 下，$r^\star(y) = 1$ 几乎处处成立；在 π_{ref} 下，$r^\star(y') = 1$ 以 $1/M$ 的概率成立，因此

$$\begin{aligned} R^{r^\star}(\pi^\star) &= \mathbb{P}_{y \sim \pi^\star, y' \sim \pi_{\mathrm{ref}}}(r^\star(y) > r^\star(y')) + \frac{1}{2} \cdot \mathbb{P}_{y \sim \pi^\star, y' \sim \pi_{\mathrm{ref}}}(r^\star(y) = r^\star(y')) \\ &= \left(1 - \frac{1}{M}\right) + \frac{1}{2} \cdot \frac{1}{M} \\ &= 1 - \frac{1}{2M}. \end{aligned} \tag{21}$$

令 y_1、……、y_N 独立同分布$\sim \pi_{\text{ref}}$ 为样本，并令 $\hat{y}$ 为 $\mathcal{A}$ 的输出。根据采样与评估框架，$\hat{y} \in \{y_1, \ldots, y_N\}$ 几乎必然成立。此外，定义事件 $E := \{\exists i \in [N] : y_i \in I\}$。在 E^c 上，对所有 i，都有$y_i \notin I$，因此 $r^\star(\hat{y}) = 0$ 确定性成立，故

$$R^{r^\star}(\pi_{\mathcal{A}} \mid E^c) = \frac{1}{2} \cdot \mathbb{P}_{y' \sim \pi_{\text{ref}}}(r^\star(y') = 0) = \frac{1}{2}\left(1 - \frac{1}{M}\right) = \frac{1}{2} - \frac{1}{2M}.$$

根据（21），对 E^c 我们得到

$$R^{r^\star}(\pi^\star) - R^{r^\star}(\pi_{\mathcal{A}} \mid E^c) = \frac{1}{2}.$$

由于遗憾非负，取期望可得

$$\mathbb{E}\left[R^{r^\star}(\pi^\star) - R^{r^\star}(\pi_{\mathcal{A}})\right] \geq \mathbb{P}(E^c) \cdot \frac{1}{2} = \frac{1}{2}\left(1 - \frac{1}{M}\right)^N.$$

若期望遗憾至多为 δ，则由上式可得

$$\left(1 - \frac{1}{M}\right)^N \leq 2\delta.$$

当 $\delta < \frac{1}{2}$ 时，右侧为 < 1，因此对两边取对数可得

$$N \log\left(1 - \frac{1}{M}\right) \leq \log(2\delta) \quad \Longrightarrow \quad N \geq \frac{\log(\frac{1}{2\delta})}{\log(\frac{M}{M-1})}.$$

对于 $M \geq 2$，我们可以应用 $x \leq -\log(1-x) \leq \frac{x}{1-x}$，其中 $x = 1/M$，从而得到

$$N \gtrsim M \log\left(\tfrac{1}{\delta}\right).$$

□

D 单调算法的证明

本节给出了第 4 节中保证的证明，组织如下：

1. D.1 节 推导了 $\mathcal{E}_M$-正则化变分问题的最优策略，该问题在第 4 节中介绍。

2. D.2 节 根据$\mathcal{E}_M$-散度（在第 2.3 节中介绍）给出了单调算法遗憾的一般上界，与遗憾的天际线下界相匹配。

3. 在 D.3 节 我们给出了 $\mathcal{E}_M$正则化 BoN 与唯一另一个可证明的单调算法之间的分离结果：即 χ^2 正则化变分问题的解，如 Huang 等人所分析的 [14]。重要的是，这凸显了先前方法在胜率遗憾目标下的弱点。

D.1 EM 正则化 BoN 策略的推导

在本节中，我们固定一个可测评分函数 $s:\ \mathcal{Y}\to[0,1]$，其定义在有限的 $\mathcal{Y}$ 上；类似论证适用于任意有界评分范围和无限响应空间。考虑 $\mathcal{E}_M$正则化变分问题，其由 $\beta>0$参数化：

$$\max_{\pi\ll\pi_{\mathrm{ref}}}\Big\{\mathbb{E}_\pi[s(y)]-\beta\,\mathcal{E}_M(\pi\|\pi_{\mathrm{ref}})\Big\}$$

这等价于关于 w 的泛函优化：

$$\max_{w\ge0,\ \mathbb{E}_{y\sim\pi_{\mathrm{ref}}}[w(y)]=1}\Big\{\mathbb{E}_{\pi_{\mathrm{ref}}}\big[w(y)s(y)\big]-\beta\,\mathbb{E}_{\pi_{\mathrm{ref}}}\big[(w(y)-M)_+\big]\Big\}. \tag{22}$$

引理 9. 固定 $M\ge2$ 和 $\beta\ge1$。那么存在一个优化器 $w^\star$ 满足 (22)，且满足 $0\le w^\star(y)\le M\pi_{\mathrm{ref}}$-几乎处处成立，因此 $\mathcal{E}_M(\pi^\star\|\pi_{\mathrm{ref}})=0$在最优点处成立（其中 $\pi^\star:=w^\star\,\pi_{\mathrm{ref}}$）。特别地，(22)等价于

$$\max_{w:\ 0\le w\le M,\ \mathbb{E}_{y\sim\pi_{\mathrm{ref}}}[w(y)]=1}\ \mathbb{E}_{\pi_{\mathrm{ref}}}[w(y)s(y)], \tag{23}$$

或者，回到 π，

$$\max_{\pi\ll\pi_{\mathrm{ref}}:\ d\pi/d\pi_{\mathrm{ref}}\le M}\ \mathbb{E}_\pi[s].$$

证明。 定义原始目标函数

$$J(w)\ :=\ \mathbb{E}_{\pi_{\mathrm{ref}}}\big[w(y)s(y)\big]-\beta\,\mathbb{E}_{\pi_{\mathrm{ref}}}\big[(w(y)-M)_+\big],\qquad\text{s.t. } w\ge0,\ \mathbb{E}_{\pi_{\mathrm{ref}}}[w]=1.$$

由于 $\mathcal{Y}$ 是有限的，这是一个有限维凸规划：惩罚项是凹函数，因此 J 关于 w 是凹的，且约束是仿射的。Slater 条件成立，因为 $w\equiv1$ 严格可行（$w(y)>0$ 对所有 y 和 $\mathbb{E}_{\pi_{\mathrm{ref}}}[w]=1$ 均成立），因此强对偶成立。接下来，我们为约束 $\mathbb{E}_{\pi_{\mathrm{ref}}}[w]=1$ 引入一个标量拉格朗日乘子 $\alpha\in\mathbb{R}$。拉格朗日函数为

$$\begin{aligned}\mathcal{L}(w,\alpha)\ &=\ \mathbb{E}_{\pi_{\mathrm{ref}}}\big[w(y)s(y)-\beta(w(y)-M)_+\big]-\alpha\big(\mathbb{E}_{\pi_{\mathrm{ref}}}[w]-1\big)\\ &=\ \alpha+\mathbb{E}_{\pi_{\mathrm{ref}}}\big[w(y)(s(y)-\alpha)-\beta(w(y)-M)_+\big].\end{aligned}$$

固定 α 时, 对 w 的最大化可按 y 分解：对于每个 y，我们在 $u\ge$上进行最大化 0,

$$\phi_y(\alpha)\ :=\ \max_{u\ge0}\Big\{u\,(s(y)-\alpha)-\beta(u-M)_+\Big\}.$$

这是一个一维分段线性问题。令 $s=s(y)$,

$$u(s-\alpha)-\beta(u-M)_+=\begin{cases}u(s-\alpha), & 0\le u\le M,\\ u(s-\alpha-\beta)+\beta M, & u\ge M.\end{cases}$$

因此：

- 若对某个 y 有 $s-\alpha-\beta>0$，则 $\phi_y(\alpha)=+\infty$（$u\ge M$ 分支的斜率为正），因此对偶函数为无穷大。因此，任何对偶最优 $\alpha^\star$ 都必须满足

$$s(y)-\alpha^\star-\beta\le0\quad\text{for all } y\in\mathcal{Y}.$$

• 在此条件下， $u \geq M$ 分支的斜率非正，因此可以选择其最大化点为$u = M$。于是，对于每个 y，都存在一个使 $\phi_y(\alpha^\star)$ 最大化的点，且该点位于 $[0, M]$。换言之：对于任何对偶值有限的 α，内层最大化都可以限制在 $0 \leq u \leq M$ 上。

根据强对偶，存在一个原始优化器 $w^\star$，它在某个 $\alpha^\star$ 处达到（有限的）对偶最优值；根据上述逐点结论，我们可以选择 $w^\star(y) \in [0, M]$，使其对每个 y 都成立。对于这样的 $w^\star$，对所有 y，有 $(w^\star(y) - M)_+ = 0$，因此 $\mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}}) = 0$，并且

$$J(w^\star) = \mathbb{E}_{\pi_{\mathrm{ref}}}[w^\star s].$$

因此，(22) 的最优值等于 (23) 的最优值；也就是说，我们可以（不失一般性地）将注意力限制在密度比满足 $0 \leq w \leq M$ π_{ref}-a.s. 的策略上；此外，在这个受限类中存在一个优化器。

□

有了这一简化，我们可以直接研究等价的密度比规划 (23)。由于该问题是凸的且强对偶成立，我们可以通过 KKT 条件刻画优化器；这些条件是最优性的充要条件 [3]。现在我们应用这些条件推导最优策略。在下述命题中，为简单起见，我们假设存在一个阈值 $\alpha^\star$，使得 $\mathbb{P}_{y\sim\pi_{\mathrm{ref}}}(s(y) \geq \alpha^\star) = 1/M$ 恰好成立；否则，可以在边界集合 $\{y\colon s(y) = \alpha^\star\}$ 上随机决定是否纳入，以使所选取的 π_{ref}-质量恰好为 $1/M$，且不影响最优性。

命题 4。 考虑 (23) 并假设根据 引理 *9* 存在优化器。固定 $M \geq 2$。则存在一个阈值 $\alpha^* \in \mathbb{R}$，使得某个最优密度比满足

$$w^\star(y) = M\,\mathbf{1}\{s(y) \geq \alpha^*\} \qquad (\pi_{\mathrm{ref}}\text{-}a.s.),$$

其中 α^* 可选取为使得 $\mathbb{P}_{y\sim\pi_{\mathrm{ref}}}(s(y) \geq \alpha^\star) = 1/M$。等价地，最优策略为

$$\pi_M(\cdot) = \pi_{\mathrm{ref}}(\cdot \mid s(\cdot) \geq \alpha^*),$$

即得分最高的 $1/M$ 阈值得分策略。

证明。 我们通过拉格朗日函数/KKT 条件求解 (23)。为等式约束 $\alpha \in \mathbb{R}$引入一个标量乘子 $\mathbb{E}_{y\sim\pi_{\mathrm{ref}}}[w(y)] = 1$，并为不等式约束 $\mu(y) \geq 0$ 和 $\nu(y) \geq 0$ 分别引入逐点乘子 $w(y) \geq 0$ 和 $w(y) \leq M$。拉格朗日函数为

$$\begin{aligned}\mathcal{L}(w; \alpha, \mu, \nu) := \mathbb{E}_{y\sim\pi_{\mathrm{ref}}}[w(y)\, s(y)] - \alpha\big(\mathbb{E}_{y\sim\pi_{\mathrm{ref}}}[w(y)] - 1\big) \\ + \mathbb{E}_{y\sim\pi_{\mathrm{ref}}}[\mu(y)\, w(y)] + \mathbb{E}_{y\sim\pi_{\mathrm{ref}}}[\nu(y)\,(M - w(y))].\end{aligned}$$

最优点处的 KKT 条件 $(w^\star, \alpha^\star, \mu^\star, \nu^\star)$ 为：

(i) 原始可行性： $0 \leq w^\star(y) \leq M$ 且 $\mathbb{E}_{y\sim\pi_{\mathrm{ref}}}[w^\star(y)] = 1$。 *(ii)* 对偶可行性： $\mu^\star(y) \geq 0$ 且 $\nu^\star(y) \geq 0$ π_{ref}-几乎必然 *(iii)* 驻点条件（逐点）：对 π_{ref}-几乎必然 y,

$$s(y) - \alpha^\star + \mu^\star(y) - \nu^\star(y) = 0. \tag{24}$$

(iv) 互补松弛条件：

$$\mu^\star(y)\, w^\star(y) = 0, \qquad \nu^\star(y)\,(M - w^\star(y)) = 0 \qquad (\pi_{\mathrm{ref}}\text{-a.s.}).$$

现在我们推导阈值结构。固定 y，并考虑三种情况。

情况 1： $0 < w^\star(y) < M$。此时，互补松弛条件要求 $\mu^\star(y) = 0$ 以及 $\nu^\star(y) = 0$，而式（24）给出$s(y) = \alpha^\star$。
情形 2： $w^\star(y) = 0$。此时 $\mu^\star(y) \geq 0$ 可能非零，而 $\nu^\star(y) = 0$ （因为 $M - w^\star(y) = M > 0$由松弛性蕴含 $\nu^\star(y) = 0$）。因此 (24) 变为 $s(y) - \alpha^\star + \mu^\star(y) = 0$，故 $s(y) \leq \alpha^\star$。
情形 3： $w^\star(y) = M$。此时 $\nu^\star(y) \geq 0$ 可能非零，而 $\mu^\star(y) = 0$ （因为 $w^\star(y) > 0$ 蕴含$\mu^\star(y) = 0$）。因此 (24) 变为 $s(y) - \alpha^\star - \nu^\star(y) = 0$，故 $s(y) \geq \alpha^\star$。
综合这三种情形，可得逐点刻画：

$$w^\star(y) = \begin{cases} M, & s(y) > \alpha^\star, \\ 0, & s(y) < \alpha^\star, \\ \text{any value in } [0, M], & s(y) = \alpha^\star. \end{cases}$$

因此，优化器的密度比满足

$$w^\star(y) = M\,\mathbf{1}\{s(y) \geq \alpha^\star\} \qquad (\pi_{\text{ref}}\text{-a.s.}).$$

最后，施加归一化条件 $\mathbb{E}_{y\sim\pi_{\text{ref}}}[w^\star(y)] = 1$:

$$1 = \mathbb{E}_{y\sim\pi_{\text{ref}}}[w^\star(y)] = M\,\mathbb{P}_{y\sim\pi_{\text{ref}}}\big(s(y) \geq \alpha^\star\big),$$

因此 $\mathbb{P}_{y\sim\pi_{\text{ref}}}(s(y) \geq \alpha^\star) = 1/M$。于是，诱导策略恰好为 $\pi_M = \pi_{\text{ref}}(\cdot \mid s \geq \alpha^*)$。 □

D.2 定理 5 的证明

证明。固定 $M \geq 1$，并令 $k := \lceil N/M \rceil$。回顾一下，$\widehat{\pi}_M$ 是随机输出y_{out}的（无条件）分布，其生成方式如下：采样 $y_1, \ldots, y_N$，它们独立同分布于$\sim \pi_{\text{ref}}$，并采样独立同分布于$[0,1]$均匀分布的 $V_1, \ldots, V_N \sim$；按照字典序（$\widehat{r}(y_i)$ V_i）选出排名前k 的样本，并令 I为其索引；输出 y_J，其中 J 在给定$\widehat{\ }(\mathcal{Y}_N, V_{1:N})$的条件下于 I 上均匀分布。为了比较 $\widehat{\pi}_M$与任意比较策略 $\pi^\star$的性能，我们采用与 C.1 节中定理 3 的证明类似的方法，考虑如下遗憾分解：

$$R^{r^\star}(\pi^\star) - R^{r^\star}(\widehat{\pi}_M) = \underbrace{\big(R^{r^\star}(\pi^\star) - R^{\widehat{r}}(\pi^\star_M)\big)}_{\text{(T1)}} + \underbrace{\big(R^{\widehat{r}}(\pi^\star_M) - R^{\widehat{r}}(\widehat{\pi}_M)\big)}_{\text{(T2)}} + \underbrace{\big(R^{\widehat{r}}(\widehat{\pi}_M) - R^{r^\star}(\widehat{\pi}_M)\big)}_{\text{(T3)}}.$$

与定理 3的证明中一样， $\pi^\star_M$ 是一个中间比较策略，定义为使到 $\pi^\star$ 的全变差距离最小化的策略，并满足其相对于 π_{ref} 的密度比几乎必然以 M 为上界。为了界定(T1)，我们可以进一步分解为

$$R^{r^\star}(\pi^\star) - R^{\widehat{r}}(\pi^\star_M) = R^{r^\star}(\pi^\star) - R^{\widehat{r}}(\pi^\star) + R^{\widehat{r}}(\pi^\star) - R^{\widehat{r}}(\pi^\star_M)$$

并分别控制这两项。对于第一项差值，我们可以直接援引引理 4，该引理指出，对于任意策略 π,

$$\big|R^{r^\star}(\pi) - R^{\widehat{r}}(\pi)\big| \;\leq\; M\,\varepsilon_{\text{pw}} + \mathcal{E}_M(\pi \| \pi_{\text{ref}}).$$

将其应用于 $\pi = \pi^\star$ 可得

$$R^{r^\star}(\pi^\star) - R^{\widehat{r}}(\pi^\star) \leq M\,\varepsilon_{\text{pw}} + \mathcal{E}_M(\pi^\star \| \pi_{\text{ref}}).$$

对于第二个差值，和定理 3的证明中一样，我们注意到胜率有界，因此1- $\widetilde{\ }$关于全变差距离是利普希茨的（它可以写成 $F_{\widehat{r}}(\widehat{r}(y)) \in [0,1]$ 的期望引理 5所示），

因此，我们可以根据 Block 和 Polyanskiy [2] 的结果，用 $\mathcal{E}_M(\pi^\star\|\pi_{\mathrm{ref}})$ 来控制从 $\pi^\star$ 到 $\pi_M^\star$ 的损失：引理 6 表明，任何密度比有界的策略与 $\pi^\star$ 的全变差距离都不可能小于 $\mathcal{E}_M(\pi^\star\|\pi_{\mathrm{ref}})$ 本身；引理 7 则证明，以上定义的 $\pi^\star$ 与 $\pi_M^\star$ 之间的全变差距离小于 $\mathcal{E}_M(\pi^\star\|\pi_{\mathrm{ref}})$，因此 $\mathrm{TV}(\pi^\star,\pi_M^\star) = \mathcal{E}_M(\pi^\star\|\pi_{\mathrm{ref}})$，进而

$$R^{\widehat{r}}(\pi^\star) - R^{\widehat{r}}(\pi_M^\star) \le \mathrm{TV}(\pi^\star,\pi_M^\star) = \mathcal{E}_M(\pi^\star\|\pi_{\mathrm{ref}}).$$

综合以上结果，我们得到

$$(\mathrm{T1}) = R^{r^\star}(\pi^\star) - R^{\widehat{r}}(\pi_M^\star) \lesssim \mathcal{E}_M(\pi^\star\|\pi_{\mathrm{ref}}) + M\varepsilon_{\mathrm{pw}}.$$

接下来，为了界定 (T2)，我们再次将其进一步分解为

$$R^{\widehat{r}}(\pi_M^\star) - R^{\widehat{r}}(\widehat{\pi}_M) = R^{\widehat{r}}(\pi_M^\star) - R^{\widehat{r}}(\pi_M) + R^{\widehat{r}}(\pi_M) - R^{\widehat{r}}(\widehat{\pi}_M).$$

这里，我们引入另一个中间比较对象 π_M，它是以下问题的解： $\mathcal{E}_M$-正则化变分问题（6），也是 $\widehat{\pi}_M$ 的“理想化”版本，它从$1/M$ 在 $\widehat{r}$下的顶部6) 中的解，见命题 4 ，令 $s=\widehat{r}$，并将所得阈值记为 λ_M ，从而 $\mathbb{P}_{y\sim\pi_{\mathrm{ref}}}(s(y) \ge \alpha^\star) = 1/M$ 严格成立。因此，优化器具有如下形式的有界密度比:

$$w_M^\star(y) = M\,\mathbf{1}\{\widehat{r}(y)\ge\lambda_M\},$$

π_{ref}-几乎处处成立，且 λ_M是–$\widehat{r}$在 π_{ref}下的（$1-\frac{1}{M}$）分位数，因此最优策略是顶部-$1/M$阈值策略

$$\pi_M(\cdot) = \pi_{\mathrm{ref}}\left(\cdot \mid \widehat{r}(\cdot)\ge\lambda_M\right).$$

现在来看第一个差项，回忆一下，根据引理 5，我们可以将 $R^{\widehat{r}}(\pi) = \mathbb{E}_{y\sim\pi}[\widetilde{F}_{\widehat{r}}(\widehat{r}(y))]$ 改写为。由于 $\widetilde{F}_{\widehat{r}}(\cdot)$ 单调不减， π_M （由命题 4 得出，且 $s=\widehat{r}$）在密度比不超过 M 的策略类中最大化 $R^{\widehat{r}}(\pi)$。又因 $\pi_M^\star$ 也受此约束，故可得$R^{\widehat{r}}(\pi_M^\star) \le R^{\widehat{r}}(\pi_M)$，因此

$$R^{\widehat{r}}(\pi_M^\star) - R^{\widehat{r}}(\pi_M) \le 0.$$

我们现在通过直接比较理想 top-$1/M$ 阈值策略 π_M 与我们的实现 $\widehat{\pi}_M$的胜率来界定第二个差分。分析 $R^{\widehat{r}}(\widehat{\pi}_M)$ 的一种便捷方法是：当 $\widehat{r}$ 可能存在并列时，“忽略 $\widehat{r}$ 的尺度”，并改用一个随机分位数，使其在π_{ref}下诱导严格全序。令 $F_{\widehat{r}}$与 $F_{\widehat{r}}(\cdot^-)$ 如 引理 5 中所设，并给定辅助随机种子 $V\sim$ 均匀分布$[0,1]$ 独立于y，我们定义

$$\tilde{u}(y;V) := F_{\widehat{r}}(\widehat{r}(y)^-) + V\cdot\left(F_{\widehat{r}}(\widehat{r}(y)) - F_{\widehat{r}}(\widehat{r}(y)^-)\right) \in [0,1]. \tag{25}$$

事实上， $\tilde{u}(y;V)$ 正是随机化的累积分布函数值：它关于 $\widehat{r}(y)$ 严格递增，而在平局 $\widehat{r}(y) = t$时，它退化为对 t 处累积分布函数跳跃的一次独立线性插值，因此按 $(\widehat{r},V)$ 排序与按 $\tilde{u}$ 排序相同。于是，对于 $y'\sim\pi_{\mathrm{ref}}$与 $V'\sim$均匀分布$[0,1]$ 独立，引理 5 表明$\tilde{u}(y';V')\sim$均匀分布$[0,1]$,，并且 $\widehat{r}$ 下的胜率可以实现为这些随机分位数的严格比较：对于任意（可能随机化的）输出 y,

$$\mathbb{P}\left(\widehat{r}(y) > \widehat{r}(y')\right) + \frac{1}{2}\mathbb{P}\left(\widehat{r}(y) = \widehat{r}(y')\right) = \mathbb{P}\left(\tilde{u}(y;V) > \tilde{u}(y';V')\right).$$

（事实上，若 $\hat{r}(y) \neq \hat{r}(y')$，则该不等式因单调性而与 $\tilde{u}$ 一致；而若 $\hat{r}(y) = \hat{r}(y')$，则 $\tilde{u}(y; V) > \tilde{u}(y'; V')$ 当且仅当 $V > V'$，其发生概率为 $1/2$。）回顾 $\widehat{\pi}_M$ 采用证明开头所述的均匀破同分规则，在 (25) 下，按 $(\hat{r}(y_i)\ V_i)$ 排序等价于按$\tilde{u}_i := \tilde{u}(y_i; V_i)$ 排序，且 $\tilde{u}_1$，……，$\tilde{u}_N$独立同分布$\sim$均匀分布$[0, 1]$。因此，我们借助引理 11 中的均匀次序统计量计算 $R^{\hat{r}}(\widehat{\pi}_M)$，由此得到下界

$$R^{\hat{r}}(\widehat{\pi}_M) \geq 1 - \frac{1}{2M} - \frac{1}{N+1}. \tag{26}$$

接下来，我们计算理想顶层$1/M$分位数策略π_M的胜率。在π_M下，抽样结果 $y \sim \pi_M$位于 $\hat{r}$的上侧 π_{ref}，其质量为 $1/M$，因此面对一个独立的 $y' \sim \pi_{\text{ref}}$时，只要 $\hat{r}(y) > \hat{r}(y')$ 就会获胜（其发生概率为 $1 - 1/M$）；当y' 也落入同一尾部时恰好平局（概率为 $1/M$），根据我们的1 2 胜率约定，其贡献因子为引理 10将这一直觉形式化，并给出

$$R^{\hat{r}}(\pi_M) = 1 - \frac{1}{2M}, \tag{27}$$

假设 $\mathbb{P}_{\pi_{\text{ref}}}(\hat{r} \geq \lambda_M) = 1/M$具有可行性，正如我们在命题 4中所做的那样。结合（26）和（27）可得

$$R^{\hat{r}}(\pi_M) - R^{\hat{r}}(\widehat{\pi}_M) \lesssim \frac{1}{N}.$$

因此，我们有

$$(\mathrm{T2}) = R^{\hat{r}}(\pi_M^\star) - R^{\hat{r}}(\widehat{\pi}_M) \lesssim \frac{1}{N}.$$

最后，为了界定(T3)，我们首先证明 $\mathcal{E}_M(\widehat{\pi}_M \| \pi_{\text{ref}}) = 0$，然后将引理 4应用于 $\widehat{\pi}_M$。令 $A \in \mathcal{F}$为可测函数。在条件为$(\mathcal{Y}_N,\ V_{1:N})$时，算法输出 $y_{\text{out}} = y_J$，其服从 $J \sim$均匀分布$(I(\mathcal{Y}_N,\ V_{1:N}))$，因此

$$\begin{aligned}
\widehat{\pi}_M(A) &= \mathbb{E}_{\widehat{\mathcal{Y}}_N \sim \pi_{\text{ref}},\, V_{1:N} \sim \mathrm{Unif}[0,1]^N} \left[\mathbb{P}\left(y_{\text{out}} \in A \,\middle|\, \widehat{\mathcal{Y}}_N, V_{1:N} \right) \right] \\
&= \mathbb{E}_{\widehat{\mathcal{Y}}_N, V_{1:N}} \left[\frac{1}{k} \sum_{i \in I(\widehat{\mathcal{Y}}_N, V_{1:N})} \mathbf{1}\{y_i \in A\} \right] \\
&\leq \mathbb{E}_{\widehat{\mathcal{Y}}_N, V_{1:N}} \left[\frac{1}{k} \sum_{i=1}^{N} \mathbf{1}\{y_i \in A\} \right] \\
&= \frac{N}{k}\, \pi_{\text{ref}}(A).
\end{aligned}$$

因此，$\widehat{\pi}_M \ll \pi_{\text{ref}}$ 及其密度比满足

$$w_{\widehat{\pi}_M}(y) := \frac{d\widehat{\pi}_M}{d\pi_{\text{ref}}}(y) \leq \frac{N}{k} \quad \pi_{\text{ref}}\text{-a.s.}$$

由于 $k = \lceil N/M \rceil$，我们有 $k \geq N/M$，因此 $\frac{N}{k} \leq M$。从而，$w_{\widehat{\pi}_M} \leq M$ π_{ref}-几乎处处成立，所以

$$\mathcal{E}_M(\widehat{\pi}_M \| \pi_{\text{ref}}) = \mathbb{E}_{y \sim \pi_{\text{ref}}} \left[(w_{\widehat{\pi}_M}(y) - M)_+ \right] = 0.$$

应用引理 4 于 $\pi = \widehat{\pi}_M$ 可得

$$(\mathrm{T3}) \le M\,\varepsilon_{\mathrm{pw}} + \mathcal{E}_M(\widehat{\pi}_M \| \pi_{\mathrm{ref}}) = M\,\varepsilon_{\mathrm{pw}}.$$

结合前述所有界可得

$$R^{r^\star}(\pi^\star) - R^{r^\star}(\widehat{\pi}_M) \;\lesssim\; \mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}}) \;+\; M\,\varepsilon_{\mathrm{pw}} \;+\; \frac{1}{N}.$$

□

引理 10. 令 $\lambda \in \mathbb{R}$ 和 $A_\lambda := \{y : \widehat{r}(y) \ge \lambda\}$ 满足 $p_\lambda := \mathbb{P}_{y\sim\pi_{\mathrm{ref}}}(A_\lambda)$。令 $\tilde{\pi}_\lambda := \pi_{\mathrm{ref}}(\cdot \mid A_\lambda)$。则

$$R^{\widehat{r}}(\tilde{\pi}_\lambda) = 1 - \frac{p_\lambda}{2}.$$

证明。 固定 λ，并令 $A_\lambda = \{y : \widehat{r}(y) \ge \lambda\}$、$p_\lambda = \pi_{\mathrm{ref}}(A_\lambda)$ 和 $\tilde{\pi}_\lambda = \pi_{\mathrm{ref}}(\cdot \mid A_\lambda)$。那么，对于 $\tilde{\pi}_\lambda$，$\widehat{r}(y) \ge \lambda$始终成立。为了计算 $\tilde{\pi}_\lambda$ 的胜率，我们可以根据事件 $\widehat{r}(y') \ge \lambda$ 是否发生进行分类：

$$\begin{aligned}
R^{\widehat{r}}(\tilde{\pi}_\lambda) &= \mathbb{P}_{y\sim\tilde{\pi}_\lambda, y'\sim\pi_{\mathrm{ref}}}\left(\widehat{r}(y) > \widehat{r}(y')\right) + \frac{1}{2}\,\mathbb{P}_{y\sim\tilde{\pi}_\lambda, y'\sim\pi_{\mathrm{ref}}}\left(\widehat{r}(y) = \widehat{r}(y')\right) \\
&= \mathbb{P}\left(\widehat{r}(y') < \lambda\right) + \mathbb{P}\left(\widehat{r}(y) > \widehat{r}(y'),\, \widehat{r}(y') \ge \lambda\right) + \frac{1}{2}\,\mathbb{P}\left(\widehat{r}(y) = \widehat{r}(y'),\, \widehat{r}(y') \ge \lambda\right) \\
&= (1 - p_\lambda) + \mathbb{P}\left(\widehat{r}(y) > \widehat{r}(y'),\, \widehat{r}(y') \ge \lambda\right) + \frac{1}{2}\,\mathbb{P}\left(\widehat{r}(y) = \widehat{r}(y'),\, \widehat{r}(y') \ge \lambda\right).
\end{aligned}$$

接下来，我们以事件 $\{\widehat{r}(y) \ge \lambda,\ \widehat{r}(y') \ge \lambda\}$ 为条件；该事件由上方方程最后两项中的事件蕴含。令

$$q := \mathbb{P}\left(\widehat{r}(y) > \widehat{r}(y') \,\middle|\, \widehat{r}(y) \ge \lambda,\, \widehat{r}(y') \ge \lambda\right), \qquad t := \mathbb{P}\left(\widehat{r}(y) = \widehat{r}(y') \,\middle|\, \widehat{r}(y) \ge \lambda,\, \widehat{r}(y') \ge \lambda\right).$$

根据可交换性，事件 $\{\widehat{r}(y) > \widehat{r}(y')\}$ 和 $\{\widehat{r}(y') > \widehat{r}(y)\}$ 具有相同的条件概率，并且它们与平局事件一起构成条件样本空间的一个划分，因此 $2q + t = 1$，这意味着 $q + \frac{1}{2}t = \frac{1}{2}$。因此，根据贝叶斯定理，

$$\begin{aligned}
\mathbb{P}\left(\widehat{r}(y) > \widehat{r}(y'),\, \widehat{r}(y') \ge \lambda\right) + \frac{1}{2}\,\mathbb{P}\left(\widehat{r}(y) = \widehat{r}(y'),\, \widehat{r}(y') \ge \lambda\right) &= \mathbb{P}\left(\widehat{r}(y) \ge \lambda,\, \widehat{r}(y') \ge \lambda\right) \cdot \left(q + \frac{1}{2}t\right) \\
&= \frac{1}{2}\,\mathbb{P}\left(\widehat{r}(y) \ge \lambda,\, \widehat{r}(y') \ge \lambda\right).
\end{aligned}$$

最后，请注意，

$$\begin{aligned}
\mathbb{P}\left(\widehat{r}(y) \ge \lambda,\, \widehat{r}(y') \ge \lambda\right) &= \mathbb{P}_{y\sim\tilde{\pi}_\lambda, y'\sim\pi_{\mathrm{ref}}}\left(y \in A_\lambda,\, y' \in A_\lambda\right) \\
&= \frac{1}{p_\lambda}\mathbb{P}_{y\sim\pi_{\mathrm{ref}}, y'\sim\pi_{\mathrm{ref}}}\left(y \in A_\lambda,\, y' \in A_\lambda\right) \\
&= p_\lambda.
\end{aligned}$$

代入后得到

$$R^{\widehat{r}}(\tilde{\pi}_\lambda) = 1 - \frac{p_\lambda}{2}.$$

□

引理 11. 固定 $M \geq 1$ 和整数 $N \geq 1$，并令 $k := \lceil N/M \rceil$。那么，在 $\widehat{r}$ 下，$\widehat{\pi}_M$ 相对于 π_{ref} 的胜率为

$$R^{\widehat{r}}(\widehat{\pi}_M) \;=\; 1 - \frac{k+1}{2(N+1)}.$$

证明。令 $F_{\widehat{r}}$ 和 $F_{\widehat{r}}(\cdot^-)$ 如 引理 5所述，并定义随机分位数

$$\tilde{u}(y; V) \;:=\; F_{\widehat{r}}(\widehat{r}(y)^-) + V\left(F_{\widehat{r}}(\widehat{r}(y)) - F_{\widehat{r}}(\widehat{r}(y)^-)\right).$$

令 $\tilde{u}_i := \tilde{u}(y_i; V_i)$。 引理 5 给出的随机概率积分变换论证表明

$$\tilde{u}_1, \ldots, \tilde{u}_N \overset{\text{iid}}{\sim} \mathrm{Unif}[0,1].$$

此外，根据构造，按 $(\widehat{r}(y_i)\ V_i)$ 对索引排序，等价于按 $\tilde{u}_i$ 排序（因为 $\widehat{r}$ 中的并列值由 V_i 打破，这恰好对应于在每个累积分布函数跳跃内部进行插值）。

令 y_{out} 表示算法输出，并定义 $\tilde{u}_{\mathrm{out}} := \tilde{u}(y_{\mathrm{out}}; V_J)$。引入一个独立的“对手”抽样 $y' \sim \pi_{\mathrm{ref}}$，以及一个与所有变量独立的 $V' \sim$ 均匀分布$[0,1]$，并记 $\tilde{u}' := \tilde{u}(y'; V')$ $\sim$均匀分布$[0,1]$。于是

$$\begin{aligned} R^{\widehat{r}}(\widehat{\pi}_M) &= \mathbb{P}\left(\widehat{r}(y_{\mathrm{out}}) > \widehat{r}(y')\right) + \frac{1}{2}\,\mathbb{P}\left(\widehat{r}(y_{\mathrm{out}}) = \widehat{r}(y')\right) \\ &= \mathbb{P}\left(\tilde{u}_{\mathrm{out}} > \tilde{u}'\right) = \mathbb{E}[\tilde{u}_{\mathrm{out}}], \end{aligned}$$

其中最后一步使用了与 $\tilde{u}_{\mathrm{out}}$ 独立的 $\tilde{u}' \sim$ 均匀分布$[0,1]$。

现在以已实现的批次 $(\tilde{u}_1, \ldots, \tilde{u}_N)$ 为条件，并令 $\tilde{u}_{(1)} \leq \cdots \leq \tilde{u}_{(N)}$ 为顺序统计量。由于 J 在前 k 个索引上服从均匀分布（在 $\tilde{u}$-排序下），

$$\mathbb{E}\big[\tilde{u}_{\mathrm{out}} \mid \tilde{u}_1, \ldots, \tilde{u}_N\big] = \frac{1}{k} \sum_{j=N-k+1}^{N} \tilde{u}_{(j)}.$$

T取期望，并利用 $\mathbb{E}[\tilde{u}_{(j)}] = \frac{j}{N+1}$ 关于均匀分布$[0,1]$ 顺序统计量的结果，可得

$$R^{\widehat{r}}(\widehat{\pi}_M) = \mathbb{E}[\tilde{u}_{\mathrm{out}}] = \frac{1}{k} \sum_{j=N-k+1}^{N} \mathbb{E}[\tilde{u}_{(j)}] = \frac{1}{k} \sum_{j=N-k+1}^{N} \frac{j}{N+1} = 1 - \frac{k+1}{2(N+1)}.$$

由于 $\left|\frac{k}{N+1} - \frac{1}{M}\right| \leq \frac{2}{N+1}$，当 $k = \lceil N/M \rceil$ 时，我们实际上有

$$1 - \frac{1}{2M} - \frac{1}{N+1} \;\leq\; R^{\widehat{r}}(\widehat{\pi}_M) \;\leq\; 1 - \frac{1}{2M} + \frac{1}{N+1}.$$

□

D.3 证明命题2

我们首先描述 h　　困难实例构造，然后再陈述正式的分离结果和证明。

困难实例设定。 固定一个目标常数 $c > 1$，并选择参数 $k \geq \sqrt{2c}$ 和 $\delta \in (0,\ 1)$，满足$\delta \geq 1 - \frac{1}{2c}$。对于任意 $\varepsilon \in (0,\ \frac{1}{4k^2}]$，定义 $\alpha := k\sqrt{\varepsilon}$ 和 $M := 1/\alpha$，并令参考分布 π_{ref} 支持三个结果 $\{B, C, P\}$，其概率质量为

$$\pi_{\mathrm{ref}}(B) = 1 - \alpha, \qquad \pi_{\mathrm{ref}}(C) = \alpha - \varepsilon, \qquad \pi_{\mathrm{ref}}(P) = \varepsilon, \tag{28}$$

因此 $\varepsilon < \alpha \leq 1/2$ 对所有 $\varepsilon \leq \frac{1}{4k^2}$ 均成立）。定义评分函数 $\hat{r}:\ \{B, C, P\} \to [0, 1]$ 为

$$\hat{r}(B) = 0, \qquad \hat{r}(C) = 1 - \delta, \qquad \hat{r}(P) = 1, \tag{29}$$

以及真实奖励 $r^\star:\ \{B, C, P\} \to [0, 1]$ 为

$$r^\star(B) = \tfrac{1}{2}, \qquad r^\star(C) = 1, \qquad r^\star(P) = 0.$$

回顾一下，对于奖励函数 r，成对结果为

$$\phi_r(y, y') = \mathbf{1}\{r(y) > r(y')\} + \frac{1}{2}\,\mathbf{1}\{r(y) = r(y')\},$$

以及成对奖励模型误差

$$\varepsilon_{\mathrm{pw}} = \mathbb{E}_{y,y' \sim \pi_{\mathrm{ref}}}\left[\left|\phi_{\hat{r}}(y, y') - \phi_{r^\star}(y, y')\right|\right].$$

在我们的构造中，$\hat{r}$ 和 $r^\star$ 仅在 $y = y'$ 时产生平局，此时 $\phi_{\hat{r}}(y, y') = \phi_{r^\star}(y, y') = \frac{1}{2}$；因此 $|\phi_{\hat{r}}(y, y') - \phi_{r^\star}(y, y')| = \mathbf{1}\{\phi_{\hat{r}}(y, y') \neq \phi_{r^\star}(y, y')\}$。由于 $\hat{r}(B) < \hat{r}(C) < \hat{r}(P)$，而 $r^\star(P) < r^\star(B) < r^\star(C)$，严格次序分歧恰好发生在一次抽取为 P 而另一次抽取属于 $\{B, C\}$ 时。因此，

$$\varepsilon_{\mathrm{pw}} = \mathbb{P}(y = P, y' \neq P) + \mathbb{P}(y \neq P, y' = P) = 2\,\pi_{\mathrm{ref}}(P)\,(1 - \pi_{\mathrm{ref}}(P)) = 2\varepsilon(1 - \varepsilon). \tag{30}$$

由于上述表达式在 $(0,\ 1/2)$ 上递增，且我们的设定假设 $\varepsilon \leq \frac{1}{4k^2}$，因此可得

$$0 < \varepsilon_{\mathrm{pw}} \leq \frac{1}{2k^2}\left(1 - \frac{1}{4k^2}\right).$$

最后，要比较的两个策略分别是 $\hat{r}$ 下的 top-α 分位数策略，以及 χ^2-正则化变分问题的一族解（由 β 参数化）。

证明概要。 在陈述正式结果之前，我们先直观说明这个例子为何会产生分离。该构造产生了一个参考质量 ε 极小的虚假最高分尖峰 P：它在 $\hat{r}$ 下看起来最好，却在 $r^\star$ 下最差；同时还存在一个质量为 $\alpha - \varepsilon$、接近顶部的结果 C，它才是真正最优的。π_M 必须将其质量分散到 $P \cup C$ 上，因此它只需承担与不可避免的污染比例 ε/α 成正比的遗憾。相比之下，χ^2-正则化解会按照分数差 $(\hat{r} - \lambda)_+$ 的比例重新加权，因此只要 P 的分数哪怕略高于 C（由 δ 控制），它就会过度集中于尖峰 P，从而产生一个无法在 β 上统一调节消除的额外遗憾项。因此，为了增大固定 c 所规定的两个策略之间的常数因子胜率遗憾差，我们希望通过减小 δ、增大 top-α 分位数 α 的总质量来提高分数差，且二者都作为 c 的函数。

命题 5。 对于任意目标常数 $c > 1$，存在满足 $k \geq \sqrt{2c}$ 和 $\delta \geq 1 - \frac{1}{2c}$ 的 $(k,\ \delta)$，以及由 $\varepsilon_{\mathrm{pw}_0} := \frac{1}{2k^2}(1 - \frac{1}{4k^2})$ 定义的 $\varepsilon_{\mathrm{pw}_0} = \varepsilon_{\mathrm{pw}_0}(k)$，使得对于任意 $\varepsilon_{\mathrm{pw}} \in (0,\ \varepsilon_{\mathrm{pw}_0}]$，都存在一个符合上述设定形式的困难实例，使得

$$\inf_{\beta > 0} R^{r^\star}(\pi^\star) - R^{r^\star}(\pi_\beta^\chi) \geq c \cdot \left(\inf_{M > 0} R^{r^\star}(\pi^\star) - R^{r^\star}(\pi_M)\right).$$

上述 $\varepsilon_{\mathrm{pw}}$ 的取值范围由基于我们构造的约束 $\varepsilon \le \frac{1}{4k^2}$–所蕴含。此外，对于任意目标 $\varepsilon_{\mathrm{pw}} \in (0,\ \varepsilon_{\mathrm{pw}_0}]$，我们可以通过对 30 式求逆，选择唯一的 ε。由于 $\varepsilon_{\mathrm{pw}} \asymp \varepsilon$，要实现更大的 c，就需要选择更小的 ε（因此更小的 $\varepsilon_{\mathrm{pw}}$），以保持 $\alpha = k\sqrt{\varepsilon} \le \frac{1}{2}$。在下面的证明中，我们首先求出每个策略的显式形式，然后比较性能并得出分离结果。

证明。 固定 $\beta > 0$。回顾 Huang 等人 [14] 的结果， χ^2-正则化解具有如下逐点形式

$$\pi_\beta^\chi(y) \;=\; \pi_{\mathrm{ref}}(y)\, w_\beta(y), \qquad w_\beta(y) \;:=\; \mathrm{relu}\big(\beta^{-1}(\widehat{r}(y) - \lambda)\big) \;=\; \frac{(\widehat{r}(y) - \lambda)_+}{\beta},$$

其中， λ 的选取满足 $\sum_y \pi_{\mathrm{ref}}(y) w_\beta(y) = 1$，即

$$\mathbb{E}_{y\sim\pi_{\mathrm{ref}}}\big[(\widehat{r}(y) - \lambda)_+\big] \;=\; \beta. \tag{31}$$

施加归一化条件。 对于我们的实例，将 $\widehat{r}$ 取自（29），并代入 π_{ref} 权重（28），归一化条件（31）变为

$$(1-\alpha)\,(0-\lambda)_+ \;+\; (\alpha-\varepsilon)\,(1-\delta-\lambda)_+ \;+\; \varepsilon\,(1-\lambda)_+ \;=\; \beta. \tag{32}$$

为方便起见，我们将 $\widehat{r}$ 在此实例上的参考均值定义为:

$$\mu \;:=\; \mathbb{E}_{\pi_{\mathrm{ref}}}[\widehat{r}] \;=\; (\alpha-\varepsilon)(1-\delta) \;+\; \varepsilon \;=\; \alpha(1-\delta) + \varepsilon\delta.$$

求解 $\lambda(\beta)$。 在计算 π_β^χ 的显式形式之前，我们根据 β的取值，确定策略的三种 λ 情形。我们根据32）中 λ 相对于 0、 $1-\delta$和 1的位置，分情况求解（

情形 I: $\lambda \in [1-\delta,\ 1)$。 那么 $(1-\delta-\lambda)_+ = 0$ 和 $(0-\lambda)_+ = 0$，而 $(1-\lambda)_+ = 1-\lambda$，s o

$$\beta = \varepsilon(1-\lambda) \quad\Longrightarrow\quad \lambda \;=\; 1 - \beta/\varepsilon.$$

当且仅当 $\lambda \ge 1-\delta$ 时，此情形是自洽的，即 $1-\beta/\varepsilon \ge 1-\delta$ 等价地

$$\beta \le \varepsilon\delta.$$

情形 II: $\lambda \in [0,\, 1-\delta)$。 那么 $(0-\lambda)_+ = 0$，而 $(1-\delta-\lambda)_+ = 1-\delta-\lambda$ 和 $(1-\lambda)_+ = 1-\lambda$ 均成立，因此

$$\beta = (\alpha-\varepsilon)(1-\delta-\lambda) + \varepsilon(1-\lambda) = \mu - \alpha\lambda.$$

因此

$$\lambda \;=\; \frac{\mu-\beta}{\alpha}.$$

自洽性要求 $\lambda < 1-\delta$ 和 $\lambda \ge 0$。第一个是

$$\frac{\mu-\beta}{\alpha} < 1-\delta \iff \mu-\beta < \alpha(1-\delta) \iff \beta > \mu - \alpha(1-\delta) = \varepsilon\delta,$$

第二个是 $\mu-\beta\geq 0 \iff \beta\leq\mu$。因此，情形 II 对于

$$\varepsilon\delta \;<\; \beta \;\leq\; \mu.$$

情形 III： $\lambda<0$。因此，三个修正线性单元项均处于激活状态：$(0-\lambda)_+=-\lambda$、$(1-\delta-\lambda)_+=1-\delta-\lambda$、$(1-\lambda)_+=1-\lambda$，所以

$$\beta=(1-\alpha)(-\lambda)+(\alpha-\varepsilon)(1-\delta-\lambda)+\varepsilon(1-\lambda)=\mu-\lambda.$$

自洽性 $\lambda<0$恰好为

$$\beta>\mu.$$

求出 π_β^χ 的显式形式。回顾 $\pi_\beta^\chi=\pi_{\mathrm{ref}}\cdot w_\beta$ 其中

$$w_\beta(y)=\frac{(\widehat{r}(y)-\lambda)_+}{\beta},$$

以及由前述情形分析给出的 $\lambda=\lambda(\beta)$。现在记录每个 β-区间内得到的 π_β^χ。

区间 I（$\beta\leq\varepsilon\delta$）：此处 $\lambda=1-\beta/\varepsilon\in[1-\delta,1)$，因此只有 P处于激活状态:

$$w_\beta(P)=\frac{1-\lambda}{\beta}=\frac{\beta/\varepsilon}{\beta}=\frac{1}{\varepsilon},$$

而 $w_\beta(C)=w_\beta(B)=0$。因此

$$\pi_\beta^\chi(P)=1,\qquad \pi_\beta^\chi(C)=0,\qquad \pi_\beta^\chi(B)=0. \tag{33}$$

区间 II ($\varepsilon\delta<\beta\leq\mu$)：此处 $\lambda=(\mu-\beta)/\alpha\in[0,1-\delta)$，因此 B未激活，而 C，P处于激活状态:

$$w_\beta(C)=\frac{1-\delta-\lambda}{\beta}=\frac{\beta-\varepsilon\delta}{\alpha\beta},\qquad w_\beta(P)=\frac{1-\lambda}{\beta}=\frac{\alpha-\mu+\beta}{\alpha\beta}.$$

因此

$$\pi_\beta^\chi(B)=0,\qquad \pi_\beta^\chi(C)=(\alpha-\varepsilon)\left(\frac{\beta-\varepsilon\delta}{\alpha\beta}\right),\qquad \pi_\beta^\chi(P)=\varepsilon\left(\frac{\alpha-\mu+\beta}{\alpha\beta}\right). \tag{34}$$

区间 III ($\beta>\mu$)：此处 $\lambda=\mu-\beta<0$，因此三个动作均处于激活状态。对于 B和 C，我们有

$$w_\beta(B)=\frac{-\lambda}{\beta}=1-\frac{\mu}{\beta},\qquad w_\beta(C)=\frac{1-\delta-\lambda}{\beta}=1+\frac{(1-\delta)-\mu}{\beta},$$

对于 P，我们有

$$w_\beta(P)=\frac{1-\lambda}{\beta}=1+\frac{1-\mu}{\beta}.$$

因此，对于 B和 C,

$$\pi_\beta^\chi(B)=(1-\alpha)\left(1-\frac{\mu}{\beta}\right),\quad \pi_\beta^\chi(C)=(\alpha-\varepsilon)\left(1+\frac{(1-\delta)-\mu}{\beta}\right), \tag{35}$$

而对于 P,

$$\pi_\beta^\chi(P)=\varepsilon\left(1+\frac{1-\mu}{\beta}\right).$$

利用 π_β^χ 在这三种情形下的显式表达式，接下来我们将精确计算胜率。

计算 $R^{r^\star}(\pi_\beta^\chi)$。对于每个区域 $t \in \{B, C, P\}$，回顾胜率的定义如下：

$$v(t) := \mathbb{P}_{y' \sim \pi_{\text{ref}}}(r^\star(t) > r^\star(y')) + \frac{1}{2}\mathbb{P}_{y' \sim \pi_{\text{ref}}}(r^\star(t) = r^\star(y')).$$

根据构造，$r^\star(C) = 1$、$r^\star(B) = \frac{1}{2}$、$r^\star(P) = 0$，因此只有在 $y' = t$ 时才会出现平局。于是，

$$\begin{aligned}
v(C) &= \mathbb{P}(y' \in \{B, P\}) + \frac{1}{2}\mathbb{P}(y' = C) = (1 - \alpha + \varepsilon) + \frac{1}{2}(\alpha - \varepsilon) = 1 - \frac{\alpha - \varepsilon}{2}, \\
v(B) &= \mathbb{P}(y' = P) + \frac{1}{2}\mathbb{P}(y' = B) = \varepsilon + \frac{1 - \alpha}{2}, \\
v(P) &= \frac{1}{2}\mathbb{P}(y' = P) = \frac{\varepsilon}{2}.
\end{aligned}$$

因此，对于任何支持于 $\{B, C, P\}$上的策略 π，

$$\begin{aligned}
R^{r^\star}(\pi) &= \mathbb{P}_{y \sim \pi,\, y' \sim \pi_{\text{ref}}}(r^\star(y) > r^\star(y')) + \frac{1}{2}\mathbb{P}_{y \sim \pi,\, y' \sim \pi_{\text{ref}}}(r^\star(y) = r^\star(y')) \\
&= \sum_{t \in \{B, C, P\}} \pi(t)\, v(t) = \pi(C)\, v(C) + \pi(B)\, v(B) + \pi(P)\, v(P).
\end{aligned} \tag{36}$$

现在代入 π_β^χ 在每个区间中的质量，其取值见（33）–（35）。

区间 I（$\beta \le \varepsilon\delta$）。利用（33）和（36），可得 $\pi_\beta^\chi(P) = 1$，因此

$$R^{r^\star}(\pi_\beta^\chi) = v(P) = \frac{\varepsilon}{2}. \tag{37}$$

区间 II（$\varepsilon\delta < \beta \le \mu$）。这里 $\pi_\beta^\chi(B) = 0$ 根据（34），因此由（36）可得，

$$\begin{aligned}
R^{r^\star}(\pi_\beta^\chi) &= \pi_\beta^\chi(C)\, v(C) + \pi_\beta^\chi(P)\, v(P) \\
&= v(P) + \pi_\beta^\chi(C)\,(v(C) - v(P)).
\end{aligned}$$

取值为 $v(P) = \varepsilon/2$ 和 $v(C) - v(P) = 1 - \alpha/2$ 时，将 $\pi_\beta^\chi(C)$ 中的显式（34）代入可得

$$R^{r^\star}(\pi_\beta^\chi) = \frac{\varepsilon}{2} + \left(1 - \frac{\alpha}{2}\right)(\alpha - \varepsilon)\left(\frac{\beta - \varepsilon\delta}{\alpha\beta}\right). \tag{38}$$

区间 III（$\beta > \mu$）。将显式的 π_β^χ 代入式（35）和式（36），得到

$$\begin{aligned}
R^{r^\star}(\pi_\beta^\chi) = (\alpha - \varepsilon)&\left(1 + \frac{(1 - \delta) - \mu}{\beta}\right)\left(1 - \frac{\alpha - \varepsilon}{2}\right) \\
&+ (1 - \alpha)\left(1 - \frac{\mu}{\beta}\right)\left(\varepsilon + \frac{1 - \alpha}{2}\right) + \varepsilon\left(1 + \frac{1 - \mu}{\beta}\right)\left(\frac{\varepsilon}{2}\right).
\end{aligned} \tag{39}$$

计算 $R^{r^\star}(\pi_M)$。对于这个实例（其中 $\alpha > \varepsilon$），在 $\widehat{r}$ 下的 top-α 集合恰好是 $\{P, C\}$，因此

$$\pi_M = \pi_{\text{ref}}(\cdot \mid \{P, C\}), \qquad \pi_M(C) = \frac{\alpha - \varepsilon}{\alpha} = 1 - \frac{\varepsilon}{\alpha}, \qquad \pi_M(P) = \frac{\varepsilon}{\alpha}, \qquad \pi_M(B) = 0.$$

使用 $v(P)$ 和 $v(C)$ 并代入式（36），得到

$$\begin{aligned} R^{r^\star}(\pi_M) &= \pi_M(C)\, v(C) + \pi_M(P)\, v(P) \\ &= \left(1 - \frac{\varepsilon}{\alpha}\right)\left(1 - \frac{\alpha - \varepsilon}{2}\right) + \frac{\varepsilon}{\alpha} \cdot \frac{\varepsilon}{2}. \end{aligned}$$

等价地，由于 $v(C) - v(P) = 1 - \alpha/2$，这可以更直观地写成

$$R^{r^\star}(\pi_M) = \frac{\varepsilon}{2} + \left(1 - \frac{\alpha}{2}\right)\pi_M(C) = \frac{\varepsilon}{2} + \left(1 - \frac{\alpha}{2}\right)\left(1 - \frac{\varepsilon}{\alpha}\right), \tag{40}$$

这种改写方式在比较策略的遗憾时会很有用。现在我们已经得到每个策略的胜率，可以计算相对于该实例上最优策略的遗憾。

计算胜率遗憾。由于 C 在 $r^\star$ 下严格最优，我们将基准策略取为 $\pi^\star := \delta_C$。因此，

$$R^{r^\star}(\pi^\star) = R^{r^\star}(\delta_C) = v(C) = 1 - \frac{\alpha - \varepsilon}{2}.$$

为简洁起见，我们将胜率遗憾定义为

$$\mathrm{Reg}(\pi) \;:=\; R^{r^\star}(\delta_C) - R^{r^\star}(\pi) = v(C) - R^{r^\star}(\pi). \tag{41}$$

的遗憾。利用（40）和 $v(C) = 1 - \frac{\alpha-\varepsilon}{2} = 1 - \frac{\alpha}{2} + \frac{\varepsilon}{2}$，-顶部位数量选择器的遗憾为

$$\begin{aligned} \mathrm{Reg}(\pi_M) &= v(C) - R^{r^\star}(\pi_M) \\ &= \left(1 - \frac{\alpha}{2} + \frac{\varepsilon}{2}\right) - \left(\frac{\varepsilon}{2} + \left(1 - \frac{\alpha}{2}\right)\left(1 - \frac{\varepsilon}{\alpha}\right)\right) \\ &= \left(1 - \frac{\alpha}{2}\right)\frac{\varepsilon}{\alpha}. \end{aligned} \tag{42}$$

遗憾（π_β^χ）：分区间。我们使用分区间胜率公式（37）–（39），并结合 $\mathrm{Reg}(\pi) = v(C) - R^{r^\star}(\pi)$。

区间 I（$\beta \le \varepsilon\delta$）。由（37），$R^{r^\star}(\pi_\beta^\chi) = \varepsilon/2$，因此

$$\mathrm{Reg}(\pi_\beta^\chi) = v(C) - \frac{\varepsilon}{2} = 1 - \frac{\alpha}{2}. \tag{43}$$

区间 II（$\varepsilon\delta < \beta \le \mu$）。由（38），$R^{r^\star}(\pi_\beta^\chi) = \frac{\varepsilon}{2} + \left(1 - \frac{\alpha}{2}\right)\pi_\beta^\chi(C)$，因此

$$\begin{aligned} \mathrm{Reg}(\pi_\beta^\chi) &= v(C) - R^{r^\star}(\pi_\beta^\chi) \\ &= \left(1 - \frac{\alpha}{2} + \frac{\varepsilon}{2}\right) - \left(\frac{\varepsilon}{2} + \left(1 - \frac{\alpha}{2}\right)\pi_\beta^\chi(C)\right) \\ &= \left(1 - \frac{\alpha}{2}\right)\left(1 - \pi_\beta^\chi(C)\right) \\ &= \left(1 - \frac{\alpha}{2}\right)\left[\frac{\varepsilon}{\alpha} + \frac{\varepsilon\delta}{\beta}\left(1 - \frac{\varepsilon}{\alpha}\right)\right], \end{aligned} \tag{44}$$

其中，在最后一行中我们使用了 $\pi_\beta^\chi(C) = (1-\varepsilon/\alpha)(1-\varepsilon\delta/\beta)$（由（34））。特别地，与（42）相比，

$$\mathrm{Reg}(\pi_\beta^\chi) = \underbrace{\left(1-\frac{\alpha}{2}\right)\frac{\varepsilon}{\alpha}}_{=\mathrm{Reg}(\pi_M)} + \left(1-\frac{\alpha}{2}\right)\cdot\frac{\varepsilon\delta}{\beta}\left(1-\frac{\varepsilon}{\alpha}\right).$$

上式中的第二项是由 χ^2 倾斜所导致的额外惩罚：尽管 π_M 只需包含投毒尖峰 P 的 $(1-\alpha/2)\cdot\varepsilon/\alpha$ 部分，但 χ^2 解仍会产生额外损失，其大小与分数差 δ 成正比、与正则化参数 β 成反比，这与我们在困难实例设定中的直觉一致。

区间 III（$\beta > \mu$）。尽管我们已在（39）中说明如何计算此区间内的胜率，但现在我们将介绍一种更简单、也更便于进行分类讨论的方法。在此区间内，我们有$\lambda = \mu - \beta < 0$，因此对所有 $y \in \{B, C, P\}$，

$$w_\beta(y) = \frac{\widehat{r}(y)-\lambda}{\beta} = \frac{\widehat{r}(y)-(\mu-\beta)}{\beta} = 1 + \frac{\widehat{r}(y)-\mu}{\beta},$$

因此

$$\pi_\beta^\chi(y) = \pi_{\mathrm{ref}}(y)\, w_\beta(y) = \pi_{\mathrm{ref}}(y)\left(1+\frac{\widehat{r}(y)-\mu}{\beta}\right).$$

将 $\pi = \pi_\beta^\chi$ 代入（36）中的胜率公式，得到

$$\begin{aligned} R^{r^\star}(\pi_\beta^\chi) &= \sum_t \pi_{\mathrm{ref}}(t)\left(1+\frac{\widehat{r}(t)-\mu}{\beta}\right)v(t) \\ &= \underbrace{\sum_t \pi_{\mathrm{ref}}(t)v(t)}_{=R^{r^\star}(\pi_{\mathrm{ref}})} + \frac{1}{\beta}\underbrace{\sum_t \pi_{\mathrm{ref}}(t)\left(\widehat{r}(t)-\mu\right)v(t)}_{=:S(\alpha,\varepsilon,\delta)}. \end{aligned}$$

其中常数 S（α，ε，δ）仅取决于实例（而与 β无关）。事实上，只需重排（39）即可轻松验证这一点。因此，对于 $\beta > \mu$，

$$R^{r^\star}(\pi_\beta^\chi) = R^{r^\star}(\pi_{\mathrm{ref}}) + \frac{S(\alpha,\varepsilon,\delta)}{\beta},$$

Con因此，根据遗憾定义（41），以基准 $v(C)$，对所有 $\beta > \mu$，

$$\mathrm{Reg}(\pi_\beta^\chi) = v(C) - R^{r^\star}(\pi_{\mathrm{ref}}) - \frac{S(\alpha,\varepsilon,\delta)}{\beta}, \tag{45}$$

它关于 $1/\beta$ 是仿射的。

比较不同情形下的遗憾。回顾（42）可知

$$\mathrm{Reg}(\pi_M) = \left(1-\frac{\alpha}{2}\right)\frac{\varepsilon}{\alpha}, \tag{46}$$

并由（41）可知，$\mathrm{Reg}(\pi) = v(C) - R^{r^\star}(\pi)$ 当 $v(C) = 1 - \frac{\alpha-\varepsilon}{2}$。现在，我们将通过分析每个 $\inf_{\beta>0}\mathrm{Reg}(\pi_\beta^\chi)$，给出 β-区间的下界。

区间 I ($\beta \le \varepsilon\delta$)。由（43），$\mathrm{Reg}(\pi_\beta^\chi) = 1 - \frac{\alpha}{2}$。利用 $\alpha = k\sqrt{\varepsilon}$ 和 $\varepsilon \le \frac{1}{4k^2}$，可得

$$\frac{\varepsilon}{\alpha} = \frac{\sqrt{\varepsilon}}{k} \le \frac{1}{2k^2}.$$

代入（46）可得

$$\mathrm{Reg}(\pi_M) = \left(1 - \frac{\alpha}{2}\right)\frac{\varepsilon}{\alpha} \le \left(1 - \frac{\alpha}{2}\right)\frac{1}{2k^2}.$$

因此,

$$\mathrm{Reg}(\pi_\beta^\chi) = 1 - \frac{\alpha}{2} \ge 2k^2\,\mathrm{Reg}(\pi_M) \ge c\,\mathrm{Reg}(\pi_M),$$

由于 $k^2 \ge 2c$。

区间 II（$\varepsilon\delta < \beta \le \mu$）。由（44），

$$\mathrm{Reg}(\pi_\beta^\chi) = \left(1 - \frac{\alpha}{2}\right)\left[\frac{\varepsilon}{\alpha} + \frac{\varepsilon\delta}{\beta}\left(1 - \frac{\varepsilon}{\alpha}\right)\right],$$

该式关于 β递减，因此在 $\beta = \mu$处取得最小值:

$$\inf_{\varepsilon\delta<\beta\le\mu} \mathrm{Reg}(\pi_\beta^\chi) = \mathrm{Reg}(\pi_\mu^\chi) = \left(1 - \frac{\alpha}{2}\right)\frac{\varepsilon}{\mu}.$$

除以（46）可得

$$\frac{\mathrm{Reg}(\pi_\mu^\chi)}{\mathrm{Reg}(\pi_M)} = \frac{\alpha}{\mu} = \frac{1}{(1-\delta) + \delta\,\varepsilon/\alpha}. \tag{47}$$

使用 $\varepsilon/\alpha = \sqrt{\varepsilon}/k \le \frac{1}{2k^2}$（来自 $\varepsilon \le \frac{1}{4k^2}$），

$$(1-\delta) + \delta\,\varepsilon/\alpha \le (1-\delta) + \frac{\delta}{2k^2} \le \frac{1}{2c} + \frac{1}{4c} = \frac{3}{4c} \le \frac{1}{c},$$

其中第二个不等式中用到了 $\delta \ge 1 - \frac{1}{2c}$ 和 $k^2 \ge 2c$。代入（47）得 $\mathrm{Reg}(\pi_\mu^\chi) \ge c\,\mathrm{Reg}(\pi_M)$。由于 $\mathrm{Reg}(\pi_\beta^\chi)$ 在区间 II 上关于 β单调递减，因此对所有 $\beta \in (\varepsilon\delta,\ \mu]$ 均有

$$\mathrm{Reg}(\pi_\beta^\chi) \ge \mathrm{Reg}(\pi_\mu^\chi) \ge c\,\mathrm{Reg}(\pi_M).$$

区间 III（$\beta > \mu$）。由（45），

$$\mathrm{Reg}(\pi_\beta^\chi) = v(C) - R^{r^\star}(\pi_{\mathrm{ref}}) - \frac{S(\alpha,\varepsilon,\delta)}{\beta},$$

因此，$\mathrm{Reg}(\pi_\beta^\chi)$ 关于 $1/\beta$ 在 $\beta > \mu$ 上是仿射的。因此，在 $\beta > \mu$ 上的最小值在端点处取得:

$$\inf_{\beta>\mu} \mathrm{Reg}(\pi_\beta^\chi) = \min\left\{\mathrm{Reg}(\pi_\mu^\chi),\ \mathrm{Reg}(\pi_{\mathrm{ref}})\right\}. \tag{48}$$

第一个端点 $\mathrm{Reg}(\pi_\mu^\chi)$ 已由区间 II 得到下界 $c\mathrm{Reg}(\pi_M)$。接下来只需为第二个端点取下界，即 $\mathrm{Reg}(\pi_{\mathrm{ref}})$。根据平局各半约定，利用独立同分布抽样的对称性，

$$R^{r^\star}(\pi_{\mathrm{ref}}) = \mathbb{P}_{y,y'\sim\pi_{\mathrm{ref}}}\left(r^\star(y) > r^\star(y')\right) + \frac{1}{2}\mathbb{P}_{y,y'\sim\pi_{\mathrm{ref}}}\left(r^\star(y) = r^\star(y')\right) = \frac{1}{2}.$$

因此，

$$\mathrm{Reg}(\pi_{\mathrm{ref}}) = v(C) - R^{r^\star}(\pi_{\mathrm{ref}}) = \left(1 - \frac{\alpha-\varepsilon}{2}\right) - \frac{1}{2} = \frac{1-\alpha+\varepsilon}{2} \geq \frac{1}{4},$$

其中最后一个不等式由$\alpha \leq \frac{1}{2}$ 根据构造得出。另一方面，再次使用 $\varepsilon/\alpha \leq \frac{1}{2k^2}$，

$$\mathrm{Reg}(\pi_M) = \left(1 - \frac{\alpha}{2}\right)\frac{\varepsilon}{\alpha} \leq \frac{\varepsilon}{\alpha} \leq \frac{1}{2k^2} \leq \frac{1}{4c},$$

其中最后一个不等式使用 $k^2 \geq 2c$。因此$\mathrm{Reg}(\pi_{\mathrm{ref}}) \geq \frac{1}{4} \geq c\mathrm{Reg}(\pi_M)$。代入（48）得到$\mathrm{Reg}(\pi_\beta^\chi) \geq c\mathrm{Reg}(\pi_M)$ 对所有 $\beta > \mu$。

完成分离。综合这三种情形可得

$$\inf_{\beta>0} R^{r^\star}(\pi^\star) - R^{r^\star}(\pi_\beta^\chi) \geq c \cdot \left(\inf_{M>0} R^{r^\star}(\pi^\star) - R^{r^\star}(\pi_M)\right),$$

如所声称。 □

期望证明：针对任意比较策略的胜率

本节将证明，当以任意测度 q 作为比较基准时，BoN 和 $\mathcal{E}_M$-正则化 BoN 性能的上界。本节安排如下：

1. 第 E.1 节 给出了附录 B 中引入的胜率迁移引理的一个修正版，该引理针对胜率比较对象 π_{ref}，并假设 q与 π_{ref}之间的密度比有界。
2. 第 E.2 节 基于第 2.3 节中引入的 $\mathcal{E}_M$-散度，给出了 BoN 遗憾的一般上界，并针对胜率引入了一种新的基线比较策略。
3. 在第 E.3 节，我们针对以任意 q为比较基准的胜率，给出了我们实现的 $\mathcal{E}_M$-正则化 BoN 策略遗憾的一般上界。

在开始证明之前，回顾一下第 6 节中我们对以 q 为比较基准的胜率的定义：

$$R_q^r(\pi) := \mathbb{E}_{y\sim\pi,\, y'\sim q}[\phi_r(y,y')] \quad \text{and} \quad \phi_r(y,y') := \mathbf{1}\{r(y) > r(y')\} + \tfrac{1}{2}\mathbf{1}\{r(y) = r(y')\}.$$

该定义推广了公式（1）中针对 π_{ref} 的胜率定义，在这里，我们现在抽取 $y' \sim q$，以便将其与来自我们策略的随机样本进行比较。本节中我们始终假设 $q \ll \pi_{\mathrm{ref}}$，并且

$$w_q(y) := \frac{dq}{d\pi_{\mathrm{ref}}}(y) \leq L \qquad \pi_{\mathrm{ref}}\text{-a.s.} \tag{49}$$

对于某个常数 $L \geq 1$。我们继续用相对于 π_{ref} 的成对奖励模型误差来衡量：

$$\varepsilon_{\mathrm{pw}}(\widehat{r}) = \mathbb{E}_{y,y'\sim\pi_{\mathrm{ref}}}\left[\left|\phi_{\widehat{r}}(y,y') - \phi_{r^\star}(y,y')\right|\right].$$

E.1 针对 q胜率的迁移引理

引理 12（针对 q 的胜率迁移）。 固定 π_{ref}， q 满足 (49)，一个真实奖励 $r^\star$以及一个学习得到的奖励模型 $\widehat{r}$。对于任意具有密度比 $\pi \ll \pi_{\text{ref}}$ 的策略 $w_\pi := d\pi/d\pi_{\text{ref}}$ 以及任意 $M \geq 1$,

$$\left|R_q^{r^\star}(\pi) - R_q^{\widehat{r}}(\pi)\right| \;\leq\; L\,(M \cdot \varepsilon_{\text{pw}}(\widehat{r}) + \mathcal{E}_M(\pi\|\pi_{\text{ref}}))\,, \tag{50}$$

其中 $\mathcal{E}_M(\pi\|\pi_{\text{ref}}) := \mathbb{E}_{y\sim\pi_{\text{ref}}}[(w_\pi(y) - M)_+]$。

证明。 证明采用与引理 4相同的策略。我们首先定义 $\ell(y,y') := |\phi_{r^\star}(y,y') - \phi_{\widehat{r}}(y,y')| \in [0,1]$。然后

$$\begin{aligned}\left|R_q^{r^\star}(\pi) - R_q^{\widehat{r}}(\pi)\right| &= |\mathbb{E}_{y\sim\pi,\,y'\sim q}[\phi_{r^\star}(y,y') - \phi_{\widehat{r}}(y,y')]| \\ &\leq \mathbb{E}_{y\sim\pi,\,y'\sim q}[\ell(y,y')]\,.\end{aligned}$$

接下来，我们可以使用 $w_\pi = d\pi/d\pi_{\text{ref}}$ 和 $w_q = dq/d\pi_{\text{ref}}$进行测度变换:

$$\mathbb{E}_{y\sim\pi,\,y'\sim q}[\ell(y,y')] = \mathbb{E}_{y,y'\sim\pi_{\text{ref}}}[w_\pi(y)\,w_q(y')\,\ell(y,y')] \leq L\cdot\mathbb{E}_{y,y'\sim\pi_{\text{ref}}}[w_\pi(y)\,\ell(y,y')]\,,$$

其中，该不等式利用了 $w_q \leq L\,\pi_{\text{ref}}$-几乎处处成立。现在将 $w_\pi(y) = (w_\pi(y)\wedge M) + (w_\pi(y) - M)_+$分解，并使用 $\ell \in [0,1]$:

$$\begin{aligned}\mathbb{E}_{y,y'\sim\pi_{\text{ref}}}[w_\pi(y)\,\ell(y,y')] &= \mathbb{E}_{y,y'\sim\pi_{\text{ref}}}[(w_\pi(y)\wedge M)\,\ell(y,y')] + \mathbb{E}_{y,y'\sim\pi_{\text{ref}}}[(w_\pi(y) - M)_+\,\ell(y,y')] \\ &\leq M\cdot\mathbb{E}_{y,y'\sim\pi_{\text{ref}}}[\ell(y,y')] + \mathbb{E}_{y\sim\pi_{\text{ref}}}[(w_\pi(y) - M)_+] \\ &= M\cdot\varepsilon_{\text{pw}}(\widehat{r}) + \mathcal{E}_M(\pi\|\pi_{\text{ref}}).\end{aligned}$$

两边乘以 L得到（50）。 □

E.2 定理 6 的证明

接下来，我们证明当胜率是相对于任意测度 q进行比较时，BoN 策略的遗憾上界。按照我们处理 π_{ref}-胜率情形的相同方式对下述上界进行优化，可得与定理 3 相同的结果（相差常数因子）。

定理 9。 在式 q （(49)）中密度比满足逐点上界的条件下，对于任意 $N \geq 1$、参考模型π_{ref}、比较策略 $\pi^\star \ll \pi_{\text{ref}}$、真实奖励 $r^\star$和学习得到的奖励模型 $\widehat{r}$，有

$$R_q^{r^\star}(\pi^\star) - R_q^{r^\star}(\widehat{\pi}_{\text{BoN}}) \lesssim L\cdot N\cdot\varepsilon_{\text{pw}}(\widehat{r})\cdot\log\left(1/\varepsilon_{\text{pw}}(\widehat{r})\right) + \mathcal{E}_{N/\log(1/\varepsilon_{\text{pw}}(\widehat{r}))}\left(\pi^\star\|\pi_{\text{ref}}\right).$$

证明。 我们沿用定理 3 在C.1 节中给出的证明，将 $R^{(\cdot)}$ 替换为 $R_q^{(\cdot)}$。具体而言，对于任意 $M > 0$, 同样的遗憾分解成立:

$$R_q^{r^\star}(\pi^\star) - R_q^{r^\star}(\widehat{\pi}_{\text{BoN}}) = \underbrace{R_q^{r^\star}(\pi^\star) - R_q^{\widehat{r}}(\pi_M^\star)}_{\text{(T1)}} + \underbrace{R_q^{\widehat{r}}(\pi_M^\star) - R_q^{\widehat{r}}(\widehat{\pi}_{\text{BoN}})}_{\text{(T2)}} + \underbrace{R_q^{\widehat{r}}(\widehat{\pi}_{\text{BoN}}) - R_q^{r^\star}(\widehat{\pi}_{\text{BoN}})}_{\text{(T3)}}, \tag{51}$$

其中 $\pi_M^\star$ 是相同的 M-截断全变差投影，定义见（13）。为界定 (T1)，我们按照C.1 节进行分解:

$$R_q^{r^\star}(\pi^\star) - R_q^{\widehat{r}}(\pi_M^\star) = R_q^{r^\star}(\pi^\star) - R_q^{r^\star}(\pi_M^\star) + R_q^{r^\star}(\pi_M^\star) - R_q^{\widehat{r}}(\pi_M^\star).$$

第一个差项的控制方法与之前完全相同：对于固定的 $(r,\ q)$，胜率泛函有界，因此关于全变差是 1-利普希茨的。因此，

$$R_q^{r^\star}(\pi^\star) - R_q^{r^\star}(\pi_M^\star) \le \mathrm{TV}(\pi^\star, \pi_M^\star) = \mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}}),$$

其中该等式使用了与 Block 和 Polyanskiy [2] 相同的论证，正如定理 3的证明中那样。对于第二个差值，应用引理 12 并取 $\pi = \pi_M^\star$；由于 $d\pi_M^\star / d\pi_{\mathrm{ref}} \le M$，我们有 $\mathcal{E}_M(\pi_M^\star \| \pi_{\mathrm{ref}}) = 0$，因此

$$R_q^{r^\star}(\pi_M^\star) - R_q^{\widehat{r}}(\pi_M^\star) \le L \cdot M\, \varepsilon_{\mathrm{pw}}(\widehat{r}).$$

因此，

$$(\mathrm{T1}) \le \mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}}) + L\, M\, \varepsilon_{\mathrm{pw}}(\widehat{r}). \tag{52}$$

接下来，为了界定 (T2)，我们引入相同的近似拒绝采样选择规则$\pi_{M,\ \mathrm{rej}}^\star$，与 C.1 节 中相同，并分解

$$R_q^{\widehat{r}}(\pi_M^\star) - R_q^{\widehat{r}}(\widehat{\pi}_{\mathrm{BoN}}) = R_q^{\widehat{r}}(\pi_M^\star) - R_q^{\widehat{r}}(\pi_{M,\mathrm{rej}}^\star) + R_q^{\widehat{r}}(\pi_{M,\mathrm{rej}}^\star) - R_q^{\widehat{r}}(\widehat{\pi}_{\mathrm{BoN}}).$$

第一个差值的控制方式与之前完全相同，即通过 Block 和 Polyanskiy [2], 的近似拒绝采样分析，仅使用胜率的有界性：

$$R_q^{\widehat{r}}(\pi_M^\star) - R_q^{\widehat{r}}(\pi_{M,\mathrm{rej}}^\star) \le \mathrm{TV}(\pi_M^\star, \pi_{M,\mathrm{rej}}^\star) \le \tfrac{1}{2} \exp(-N/M),$$

因为 $\mathcal{E}_M(\pi_M^\star \| \pi_{\mathrm{ref}}) = 0$。对于第二个差值，在任意固定批次 $\mathcal{Y}_N = (y_1, \ldots\ldots, y_N)$ 的条件下，BoN 输出一个达到 $\max_i \widehat{r}(y_i)$ 的索引。我们定义

$$g_{r,q}(r(y)) := \mathbb{P}_{y' \sim q}\left(r(y') < r(y)\right) + \tfrac{1}{2}\, \mathbb{P}_{y' \sim q}\left(r(y') = r(y)\right) \in [0, 1],$$

因此，对于任意策略 π, $R_q^{\widehat{r}}(\pi) = \mathbb{E}_{y \sim \pi}[g_{\widehat{r},q}(\widehat{r}(y))]$，通过对 y 条件化，完全如引理 5 中所述。从这一点可以清楚看出，$g_{\widehat{r},q}$ 关于 $\widehat{r}(y)$ 单调不减。因此，对于任意 $y \in \mathcal{Y}_N$，我们有 $g_{\widehat{r},q}(\widehat{r}(y)) \le \max_{i \in [N]} g_{\widehat{r},q}(\widehat{r}(y_i))$，并且定理 3 的证明中使用的相同耦合论证给出

$$R_q^{\widehat{r}}(\pi_{M,\mathrm{rej}}^\star) - R_q^{\widehat{r}}(\widehat{\pi}_{\mathrm{BoN}}) \le 0.$$

结合两个差值的界可得

$$(\mathrm{T2}) \le \tfrac{1}{2} \exp(-N/M). \tag{53}$$

为了界定 (T3)，回顾一下，在 C.1 节 中我们证明了 BoN 满足 $d\widehat{\pi}_{\mathrm{BoN}} / d\pi_{\mathrm{ref}} \le N$ 逐点成立，因此$\mathcal{E}_N(\widehat{\pi}_{\mathrm{BoN}} \| \pi_{\mathrm{ref}}) = 0$。应用 引理 12，取 $\pi = \widehat{\pi}_{\mathrm{BoN}}$并进行截断 $M = N$，可得

$$(\mathrm{T3}) = R_q^{\widehat{r}}(\widehat{\pi}_{\mathrm{BoN}}) - R_q^{r^\star}(\widehat{\pi}_{\mathrm{BoN}}) \le L \cdot N\, \varepsilon_{\mathrm{pw}}(\widehat{r}). \tag{54}$$

将 (52)、(53) 和 (54) 与 (51) 结合可得

$$R_q^{r^\star}(\pi^\star) - R_q^{r^\star}(\widehat{\pi}_{\mathrm{BoN}}) \lesssim \mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}}) + L(M+N)\, \varepsilon_{\mathrm{pw}}(\widehat{r}) + \exp(-N/M).$$

令 $M = N / \log\left(1/\varepsilon_{\mathrm{pw}}(\widehat{r})\right)$，并注意到对于这样的 M，有 $\exp(-N/M) = \varepsilon_{\mathrm{pw}}(\widehat{r})$，即可得到所需结果。 □

E.3 定理7的证明

最后，我们给出在胜率以任意测度 $\mathcal{E}_M$ 为比较对象时， q-正则化 BoN 策略的遗憾上界。我们仍假设$\widehat{\pi}_M$ 的密度比逐点有界（49）。请注意，尽管下述上界在定性上与定理 3的结果相似，但我们得到的 N 收敛速率较慢，同时仍保持极小极大最优。

定理 10. 对于任意 M、 $N \geq 1$、参考模型 π_{ref}、比较策略 $\pi^\star \ll \pi_{\text{ref}}$、真实奖励 $r^\star$以及学习得到的奖励模型 $\widehat{r}$，均有

$$R_q^{r^\star}(\pi^\star) - R_q^{r^\star}(\widehat{\pi}_M) \lesssim L \cdot \mathcal{E}_M(\pi^\star \| \pi_{\text{ref}}) + L \cdot M \cdot \varepsilon_{\text{pw}}(\widehat{r}) + \sqrt{\frac{M}{N}}.$$

证明。固定 $M \geq 1$ ，并令 $k := \lceil N/M \rceil$。回顾 $\widehat{\pi}_M$ 是随机输出y_{out} 的（无条件）分布，该输出由如下过程产生：采样 y_1，……， y_N 独立同分布$\sim \pi_{\text{ref}}$，以及独立同分布的 V_1，……， $V_N \sim$均匀分布$[0,1]$ 且相互独立；令 I为字典序（$\widehat{r}(y_i)$ V_i）下前k 个样本的索引；输出 y_J ，其中 J 在给定$\widehat{\ }(y_N, V_{1:N})$的条件分布下于 I 上服从均匀分布。与定理 5 关于 π_{ref}-胜率的证明一样，我们考虑相同的遗憾分解：

$$R_q^{r^\star}(\pi^\star) - R_q^{r^\star}(\widehat{\pi}_M) = \underbrace{\left(R_q^{r^\star}(\pi^\star) - R_q^{\widehat{r}}(\pi_M^\star)\right)}_{\text{(T1)}} + \underbrace{\left(R_q^{\widehat{r}}(\pi_M^\star) - R_q^{\widehat{r}}(\widehat{\pi}_M)\right)}_{\text{(T2)}} + \underbrace{\left(R_q^{\widehat{r}}(\widehat{\pi}_M) - R_q^{r^\star}(\widehat{\pi}_M)\right)}_{\text{(T3)}}.$$

这里， $\pi_M^\star$ 的定义与之前完全相同：它是在所有满足 $\frac{d\pi}{d\pi_{\text{ref}}} \leq M$ π_{ref}-几乎处处条件的 $\pi \ll \pi_{\text{ref}}$ 中，使其与 $\pi^\star$ 的全变差距离最小的对象。为了界定 (T1)，我们将其分解为

$$R_q^{r^\star}(\pi^\star) - R_q^{\widehat{r}}(\pi_M^\star) = R_q^{r^\star}(\pi^\star) - R_q^{\widehat{r}}(\pi^\star) + R_q^{\widehat{r}}(\pi^\star) - R_q^{\widehat{r}}(\pi_M^\star).$$

对于第一个差值，我们应用 引理 12得到

$$R_q^{r^\star}(\pi^\star) - R_q^{\widehat{r}}(\pi^\star) \;\leq\; L\left(M\,\varepsilon_{\text{pw}}(\widehat{r}) + \mathcal{E}_M(\pi^\star \| \pi_{\text{ref}})\right).$$

对于第二个差值，我们注意到，对于固定的 (q， $\widehat{r}$)，胜率泛函有界，因此关于全变差距离是 1-利普希茨的。因此，

$$R_q^{\widehat{r}}(\pi^\star) - R_q^{\widehat{r}}(\pi_M^\star) \leq \text{TV}(\pi^\star, \pi_M^\star) = \mathcal{E}_M(\pi^\star \| \pi_{\text{ref}}),$$

其中，最后一个等式使用了与之前相同的 Block 和 Polyan-skiy [2] 近似拒绝采样论证。综上，

$$\text{(T1)} \;\lesssim\; L\,\mathcal{E}_M(\pi^\star \| \pi_{\text{ref}}) \;+\; L\,M\,\varepsilon_{\text{pw}}(\widehat{r}). \tag{55}$$

对于 (T2)，与之前一样，我们引入理想的 top-$1/M$ 阈值策略

$$\pi_M(\cdot) := \pi_{\text{ref}}\left(\cdot \mid \widehat{r}(\cdot) \geq \lambda_M\right),$$

使得 $\mathbb{P}_{y\sim\pi_{\text{ref}}}(\widehat{r}(y) \geq \lambda_M) = \frac{1}{M}$，并将其分解为

$$R_q^{\widehat{r}}(\pi_M^\star) - R_q^{\widehat{r}}(\widehat{\pi}_M) = R_q^{\widehat{r}}(\pi_M^\star) - R_q^{\widehat{r}}(\pi_M) + R_q^{\widehat{r}}(\pi_M) - R_q^{\widehat{r}}(\widehat{\pi}_M).$$

对于第一个差值，与定理 9的证明中一样，我们定义

$$g_{r,q}(r(y)) := \mathbb{P}_{y'\sim q}\left(r(y) > r(y')\right) + \tfrac{1}{2}\,\mathbb{P}_{y'\sim q}\left(r(y) = r(y')\right) \in [0,1],$$

因此，对于任意策略 π, $R_q^{\widehat{r}}(\pi) = \mathbb{E}_{y\sim\pi}[g_{\widehat{r},q}(\widehat{r}(y))]$，通过对 y 条件化可得，这与引理 5 中完全相同。由此显然可知，$g_{\widehat{r},q}$ 关于 $\widehat{r}(y)$ 单调不减，因此 π_M（由命题 4 在 $s = \widehat{r}$ 条件下给出）在密度比受 M 约束的策略类中使 $R_q^{\widehat{r}}(\pi)$ 最大。回想一下，$\pi_M^\star$ 也满足这一约束，因此可行，由此可得 $R_q^{\widehat{r}}(\pi_M^\star) \le R_q^{\widehat{r}}(\pi_M)$，于是

$$R_q^{\widehat{r}}(\pi_M^\star) - R_q^{\widehat{r}}(\pi_M) \le 0.$$

还需对 $R_q^{\widehat{r}}(\pi_M) - R_q^{\widehat{r}}(\widehat{\pi}_M)$进行界定。如 D.2 节所示，我们令 $F_{\widehat{r}}$ 和 $F_{\widehat{r}}(\cdot^-)$ 表示 $\widehat{r}(y)$ 的累积分布函数，其中 $y \sim \pi_{\text{ref}}$ 如 引理 5中所示。给定一个与 y独立的辅助随机种子 $V \sim$ 均匀分布$[0,1]$，我们定义

$$\tilde{u}(y;V) := F_{\widehat{r}}(\widehat{r}(y)^-) + V\left(F_{\widehat{r}}(\widehat{r}(y)) - F_{\widehat{r}}(\widehat{r}(y)^-)\right) \in [0,1].$$

根据 引理 5中的相同均匀化论证， $\tilde{u}(y;V) \sim$均匀分布$[0,1]$,，而且

$$F_{\widehat{r}}^{-1}(\tilde{u}(y;V)) = \widehat{r}(y)$$

π_{ref}-几乎处处，其中 $F_{\widehat{r}}^{-1}$是广义逆。如定理 5的证明所示，按 $(\widehat{r}(y_i)\ V_i)$ 排序等价于按 $\tilde{u}_i := \tilde{u}(y_i;\ V_i)$ 排序。我们令 $u_1,\dots,u_N$独立同分布$\sim$均匀分布$[0,1]$表示这些随机化秩，并记为 $u_{(1)} \le \cdots \le u_{(N)}$。现在令 $t := 1 - \frac{1}{M}$，并定义经验截断值 $\tau := u_{(N-k)}$（因此，选出的集合对应于排名前 k 的项）。我们还定义有界单调函数

$$h(u) := g_{\widehat{r},q}\left(F_{\widehat{r}}^{-1}(u)\right) \in [0,1],$$

使得 $R_q^{\widehat{r}}(\pi) = \mathbb{E}_{y\sim\pi}[h(\tilde{u}(y;\ V))]$。在 π_M 下，我们有 $\tilde{u}(y;\ V) \sim$均匀分布$[t,1]$，因此

$$R_q^{\widehat{r}}(\pi_M) = \mu(t),$$

其中 $\mu(s) := \mathbb{E}[h(u) \mid u \ge s]$且 $u \sim$均匀分布$[0,1]$。在 $\widehat{\pi}_M$ 下，输出秩 u_{out}在前 k 个秩中均匀分布。按照与引理 11相同的逻辑，标准的次序统计量性质表明，给定 τ 后，高于 k 的 τ个秩构成的多重集是独立同分布的，服从均匀分布$[\tau,1]$；因此 $u_{\text{out}} \mid \tau \sim$均匀分布$[\tau,1]$且

$$R_q^{\widehat{r}}(\widehat{\pi}_M) = \mathbb{E}\left[\mu(\tau)\right],$$

其中期望 是对 τ 中随机性的期望。至此，我们已将差距 t o

$$R_q^{\widehat{r}}(\pi_M) - R_q^{\widehat{r}}(\widehat{\pi}_M) = \mu(t) - \mathbb{E}[\mu(\tau)].$$

为了给这个量求上界，我们注意到显然

$$\mu(t) - \mathbb{E}[\mu(\tau)] \le \mathbb{E}\left[|\mu(t) - \mu(\tau)|\right].$$

回顾 $\mu(s) = \mathbb{E}[h(U) \mid U \ge s]$，其中 $U \sim$均匀分布$[0,1]$且 $h \in [0,1]$。等价地，若记 $P_s :=$均匀分布$[s,1]$，则 $\mu(s) = \mathbb{E}_{P_s}[h]$。因此，对于任意 s， $s' \in [0,1)$,

$$|\mu(s) - \mu(s')| = \left|\mathbb{E}_{P_s}[h] - \mathbb{E}_{P_{s'}}[h]\right| \le \mathrm{TV}(P_s, P_{s'}).$$

接下来，我们将说明如何界定这一全变差距离。不失一般性，假设 $s \le s'$。相应的密度为

$$f_s(u) = \frac{1}{1-s}\,\mathbf{1}\{u \in [s,1]\}, \qquad f_{s'}(u) = \frac{1}{1-s'}\,\mathbf{1}\{u \in [s',1]\},$$

so

$$\begin{aligned}\mathrm{TV}(P_s, P_{s'}) &= \frac{1}{2}\int_0^1 |f_s(u) - f_{s'}(u)|\, du \\ &= \frac{1}{2}\left(\int_s^{s'} \frac{1}{1-s}\, du \;+\; \int_{s'}^1 \left(\frac{1}{1-s'} - \frac{1}{1-s}\right) du\right) \\ &= \frac{1}{2}\left(\frac{s'-s}{1-s} \;+\; (1-s')\left(\frac{1}{1-s'} - \frac{1}{1-s}\right)\right) = \frac{s'-s}{1-s}.\end{aligned}$$

因此，对于所有 s， $s' \in [0,1)$,

$$|\mu(s) - \mu(s')| \le \frac{|s-s'|}{1-\min\{s,s'\}}.$$

将其应用于 $s = t$和 $s' = \tau$，得到

$$|\mu(t) - \mu(\tau)| \le \frac{|t-\tau|}{1-\min\{t,\tau\}} \le \frac{|t-\tau|}{1-t} = M\,|t-\tau|,$$

因为 $t = 1 - \frac{1}{M}$且 $\min\{t,\tau\} \le t$。因此

$$R_q^{\widehat{r}}(\pi_M) - R_q^{\widehat{r}}(\hat{\pi}_M) = \mu(t) - \mathbb{E}[\mu(\tau)] \le \mathbb{E}\left[|\mu(t) - \mu(\tau)|\right] \le M\,\mathbb{E}[|t-\tau|].$$

因此

$$R_q^{\widehat{r}}(\pi_M) - R_q^{\widehat{r}}(\widehat{\pi}_M) \le M\mathbb{E}\left[|t-\tau|\right], \tag{56}$$

并且仍需对 $\mathbb{E}\left[|t-\tau|\right]$进行界定。由于 $u_1, \ldots, u_N$ 独立同分布$\sim$ 均匀分布 $[0,1]$,，截断值 $\tau = u_{(N-k)}$ 服从贝塔分布$\tau \sim \mathrm{Beta}(N-k,\ k+1)$。特别地,

$$\mathbb{E}[\tau] = \frac{N-k}{N+1} \qquad \text{and} \qquad \mathrm{Var}(\tau) = \frac{(N-k)(k+1)}{(N+1)^2(N+2)} \;\le\; \frac{k+1}{(N+1)^2}.$$

利用三角不等式，我们可以将其分解为

$$\mathbb{E}[|t-\tau|] \;\le\; |t - \mathbb{E}[\tau]| + \mathbb{E}[|\tau - \mathbb{E}[\tau]|],$$

对第二项应用柯西-施瓦茨不等式，可得

$$|t - \mathbb{E}[\tau]| + \mathbb{E}[|\tau - \mathbb{E}[\tau]|] \;\le\; |t - \mathbb{E}[\tau]| + \sqrt{\mathrm{Var}(\tau)}.$$

由于 $t = 1 - \frac{1}{M}$和 $k = \lceil N/M \rceil$，很容易验证 $|\mathbb{E}[\tau] - t| \lesssim \frac{1}{N}$（由于舍入误差）且$\sqrt{\text{方差}(\tau)} \lesssim \frac{1}{\sqrt{MN}}$，因此

$$\mathbb{E}[|t-\tau|] \;\lesssim\; \frac{1}{N} + \frac{1}{\sqrt{MN}}.$$

结合（56）可得

$$R_q^{\widehat{r}}(\pi_M) - R_q^{\widehat{r}}(\hat{\pi}_M) \;\lesssim\; M\left(\frac{1}{N} + \frac{1}{\sqrt{MN}}\right) \;=\; \frac{M}{N} + \sqrt{\frac{M}{N}},$$

因此我们可以得出结论:

$$(\mathrm{T2}) = R_q^{\widehat{r}}(\pi_M^\star) - R_q^{\widehat{r}}(\widehat{\pi}_M) \;\lesssim\; \frac{M}{N} + \sqrt{\frac{M}{N}}, \tag{57}$$

其中，当 N较大时，第二项占主导。对于 (T3)，与定理 5的证明类似，我们有 $\mathcal{E}_M(\widehat{\pi}_M \| \pi_{\mathrm{ref}}) = 0$ ，因为 $w_{\widehat{\pi}_M} \le N/k \le M$ π_{ref}-几乎处处成立。对 引理 12 应用于 $\pi = \widehat{\pi}_M$ 并利用（49）可得

$$(\mathrm{T3}) = R_q^{\widehat{r}}(\widehat{\pi}_M) - R_q^{r^\star}(\widehat{\pi}_M) \;\le\; L\,(M\,\varepsilon_{\mathrm{pw}}(\widehat{r}) + \mathcal{E}_M(\widehat{\pi}_M \| \pi_{\mathrm{ref}})) = L\,M\,\varepsilon_{\mathrm{pw}}(\widehat{r}). \tag{58}$$

最后，结合（55）、（57）和（58）可得

$$R_q^{r^\star}(\pi^\star) - R_q^{r^\star}(\widehat{\pi}_M) \lesssim L \cdot \mathcal{E}_M(\pi^\star \| \pi_{\mathrm{ref}}) + L \cdot M \cdot \varepsilon_{\mathrm{pw}}(\widehat{r}) + \sqrt{\frac{M}{N}}.$$

□

尽管我们没有给出计算上的下界来证明 $\sqrt{M/N}$ 估计项在完全一般的情形下是紧的，但这一较慢的收敛速率可以看作一般q 情形下的一个产物。关键障碍在于，对于一般的 q，胜率是随机 尾部泛函，作用于 $\widehat{r}$-秩，而非如 定理 3 在 C.1 节的证明中那样，是一个可显式计算的线性统计量。具体来说，我们写作

$$R_q^{\widehat{r}}(\pi) = \mathbb{E}_{y\sim\pi,\,V}\big[h(\tilde{u}(y;V))\big], \qquad h(u) := g_{\widehat{r},q}\big(F_{\widehat{r}}^{-1}(u)\big) \in [0,1],$$

其中，在（ h 49）下，它仅有界（且单调）。当 $q = \pi_{\mathrm{ref}}$时，相应的变换是线性的（事实上 $h(u) = u$），因此经验 top-$1/M$ 策略可以通过精确的次序统计量计算得到 $1/N$ 项。然而，对于一般的 q，在固定截断值 h 附近，$t = 1 - \frac{1}{M}$的形状不受控制： h 几乎可以将其所有变化都集中在 $O(1/M)$ 邻域内，围绕 t。在这类最坏情形下，经验截断值的小幅波动 τ 围绕 t可能转化为如下形式的敏感性界

$$|\mu(t) - \mu(\tau)| \;\lesssim\; M\,|t - \tau|,$$

这对所有有界的 h 都成立，并刻画了这样一个事实：将阈值改变 $|t-\tau|$ 会扰动质量为 $1/M$的尾部。由于 τ以 t为中心、在 $\mathbb{E}|t-\tau| \asymp 1/\sqrt{MN}$的尺度上波动（当 $k \approx N/M$时），因此会导致量级为 $\sqrt{M/N}$ 的估计误差，而无需对 q （等价地，对h）作出超出（49）的额外正则性假设。