arXiv:2302.04658v3 [stat.ML] 23 Feb 2024

近似拒绝抽样的样本复杂度及其在平滑在线学习中的应用

Adam Block、Yury Polyanskiy,
麻省理工学院

摘要

假设我们可以访问来自分布 μ 的 n 个独立样本，并希望从中输出一个样本，使输出的分布尽可能接近目标分布 ν。本文证明，最优全变差距离关于 n 的函数为 $\tilde{\Theta}\left(\frac{D}{f'(n)}\right)$，其适用范围为所有满足有界 f-散度$D_f(\nu\|\mu) \leq D$ 的分布对 ν、μ。此前，这一问题仅针对 ν 相对于 μ 的拉东-尼科迪姆导数有一致上界的情形进行过研究。接着，我们探讨一个看似截然不同的领域——平滑在线学习中的应用，并证明，即使对对手的约束有所放宽（要求其具有有界的 f-散度，而非有界的拉东-尼科迪姆导数），关于极小极大遗憾和预言机高效算法遗憾的最新结果仍然成立。最后，我们还研究了重要性采样在函数类上均匀估计均值方面的效果，并比较了重要性采样与拒绝采样。

1 引言

考虑以下问题：给定来自某个基准分布 μ的 n 个独立样本，学习者如何生成一个来自目标分布 ν的样本？这个简单的问题可以追溯到几十年前，其首个形式化解法——拒绝采样，早在 Von Neumann[1951] 时便已提出。由于方法简单，这一采样问题作为基本工具，出现在机器学习、理论计算机科学和密码学的众多应用中 [Lyubashevsky, 2012, Liu, 1996, Naesseth 等人, 2017, Ozols 等人, 2013]；因此，构造高效解法一直是许多研究的重点 [Grover 等人, 2018, Gilks and Wild, 1992, Martino and M´ıguez, 2011]。或许令人惊讶的是，拒绝采样最初的解法 [Von Neumann, 1951] 至今仍是一种广受欢迎的方法。

给定 $X_1, \ldots, X_n \sim \mu$，回顾一下，拒绝采样需要一个参数 M，它是拉东-尼科迪姆导数 $\frac{d\nu}{d\mu}$ 的一个统一上界；对于每个 $1 \leq i \leq n$，以概率 $\frac{1}{M} \cdot \frac{d\nu}{d\mu}(X_i)$ 接受X_i，并将任意一个被接受的 X_i 作为来自 ν 的样本返回。容易验证，若 $M \geq \left\|\frac{d\nu}{d\mu}\right\|_\infty$，那么任何被接受的样本都服从分布 ν。此外，不难看出，任何样本被接受的概率为 $\frac{1}{M}$，且与其他样本相互独立

因此，若希望以高概率至少获得一个被接受的样本，就需要满足 $n = \Theta(M)$。信息论界已投入大量工作来改进这一上界 [Liu and Verdu, 2018, Harsha 等，2007]；统计学界也致力于在强结构假设下提高采样效率 [Gilks and Wild, 1992, G¨or¨ur and Teh, 2011]，。然而，这些工作的适用范围大多受限于条件 $\left\|\frac{d\nu}{d\mu}\right\|_\infty < \infty$。在许多场景中，这一假设并不成立 [Block 等人，2023]；因此，我们关注一个类似的问题，但不再要求满足严格的统一上界假设。遗憾的是，不难看出，存在一些例子：若没有这一统一上界，我们根本无法从 ν 中精确恢复样本（示例见定理 31）。因此，我们放宽目标，转而考虑近似采样。具体而言，我们提出以下问题：

> 我们需要从源分布μ 中抽取多少个独立样本 X_1、……、X_n，才能选出某个 $j^* \in [n]$，使得 X_{j^*} 的分布律与 ν之间的全变差距离不超过 ε？

尽管这个问题很简单，但据我们所知，迄今为止文献中尚未有人研究过。我们在附录A的相关工作中强调了几个特殊情形。本文针对所有具有实际意义的超线性 f-散度，给出了这个问题的完整答案，并得到了基本匹配的上界和下界。上界通过一种改进的拒绝采样器得到，其分析基于经典研究，难度不大；而下界则需要采用更具技术性的途径。为了量化 ν与 μ之间的差异，我们使用信息论概念—— f-散度。对于定义在同一集合上的两个测度 $\nu \ll \mu$ 以及一个凸函数 f，我们定义

$$D_f(\nu||\mu) = \mathbb{E}_\mu\left[f\left(\frac{d\nu}{d\mu}(Z)\right)\right].$$

下面我们将给出更正式的定义，但这里先指出， f-散度这一概念推广了包括全变差、KL 散度、Rényi 散度和$\mathcal{E}_\gamma$ 散度[Polyanskiy and Wu, 2022+, Van Erven and Harremos, 2014, Asoodeh 等, 2021]在内的常见散度。我们将作出如下假设：对于某个凸的 f，源测度和目标测度满足$D_f(\nu||\mu) < \infty$，并考察在这一约束下， ε-近似拒绝抽样的样本复杂度。有趣的是，答案取决于 f的尾部行为；特别是，如果 sup $f'(x) < \infty$，那么仅凭这一约束，拒绝抽样就无法奏效（见命题4）。然而，如果我们有一个 f-散度约束，且 $f'(\infty) = \infty$，我们将看到

$$n = \widetilde{\Theta}\left((f')^{-1}\left(\frac{8 \cdot D_f(\nu||\mu)}{\varepsilon}\right)\right)$$

样本既是必要的也是充分的，以生成一个 X_{j^*} 满足 ε-全变差距离的样本。事实上，我们证明了 Von Neumann 最初提出的拒绝采样器对于这一问题基本上是最优的，并且不需要使用 Harsha 等 [2007], Liuand Verdu [2018]为精确抽样而引入的更复杂的采样器。如上所述，上界相对标准，而大部分技术工作涉及下界的构造。

尽管上述结果本身颇有意义，我们仍要强调这些结果在一个看似毫不相关的领域中的关键用途：平滑在线学习。下面简要回顾其设定。

对于一般的在线学习，我们固定一个上下文集合 $\mathcal{X}$、一个目标集合 $\mathcal{Y}$ 和一个函数类$\mathcal{F}$：$\mathcal{X} \to \mathcal{Y}$，以及一个损失函数 ℓ：$\mathcal{Y} \times \mathcal{Y} \to [0,1]$。对于某个时域 T，在线学习按轮次进行，在每个时刻 $1 \le t \le T$ 会发生以下情况：

1. 自然选择某个上下文 x_t 和标签 y_t。
2. 学习者选择某个预测 $\widehat{y}_t \in \mathcal{Y}$。
3. 学习者观察到 y_t，并承受损失 $\ell(\widehat{y}_t, y_t)$。

与 Block 等人 [2022], Haghtalab 等人 [2022a], 我们区分 适当 和 非适当 设定。在前一种设定中，学习者必须先选择某个函数 $f_t \in \mathcal{F}$，然后才能看到 x_t，接着预测 $\widehat{y}_t = f_t(x_t)$。在后一种设定中，学习者观察到 x_t，然后进行任意$\widehat{y}_t \in \mathcal{Y}_t$预测。两种情况下的目标都是最小化相对于事后最优函数的期望遗憾，其中

$$\mathrm{Reg}_T = \sum_{t=1}^{T} \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t).$$

如前所述，自然选择上下文和标签时不受任何限制，这称为对抗性设定。尽管这种机制因其稳健性且无需假设而广受欢迎，但完全对抗性设定存在两个主要问题：首先，像一维阈值函数这样简单的函数类，在数据独立出现时很容易学习，却会在对抗性机制下变得不可学习 [Rakhlin 等，2015年，Littlestone， 1988]；其次，即使函数类可学习，通常也无法高效地学习 [Hazan and Koren，2016]。为了解决第一个问题，平滑在线学习的概念近来逐渐受到关注 [Rakhlin 等，2011，Haghtalab 等，2022a、2022b，Block 等，2022，Block and Simchowitz，2022]。受算法的平滑分析启发，Rakhlin 等人 [2011], Haghtalab 等人 [2022b] 考虑以下设定。对于某个集合 $\mathcal{X}$上的固定基准测度 μ，若测度ν满足 $\left\| \frac{d\nu}{d\mu} \right\|_\infty \le \frac{1}{\sigma}$，则称其关于 μ是 σ-平滑的。若对于所有 t，在给定全部历史的条件下， x_t 的分布 p_t 都是 σ-平滑的，则称对手相对于某个固定的 μ 是 σ-平滑的。这一定义的一种动机是设想自然处于完全对抗状态，但受到少量噪声的扰动。例如，若 $\mathcal{X} = \mathbb{R}^d$，我们可以设想向对抗性输入中加入少量均匀噪声或高斯噪声 [Block 等， 2023]。在 Block 等 [2022],Haghtalab 等人 [2022b], 中，研究者推导出了平滑在线学习的极小极大最优速率，误差仅相差多对数因子。举例来说，若令 vc ($\mathcal{F}$) 表示某个二值函数类 $\mathcal{F}$的 Vapnik-Chervonenkis维数 [Blumer等人， 1989] ，那么存在某种算法，使其能够针对指示损失达到

$$\mathbb{E}\left[\mathrm{Reg}_T\right] = O\left(\sqrt{T \cdot \mathrm{vc}\left(\mathcal{F}\right) \cdot \log\left(\frac{T}{\sigma}\right)}\right).$$

遗憾的是，在许多常见情形下，可能无法获得关于 $\frac{dp_t}{d\mu}$ 的一致上界。例如，再次考虑在 $\mathbb{R}^d$ 中向对抗性输入添加少量高斯噪声的情形。一个自然的选择是令 μ 为某个固定的高斯分布，但我们无法

确保 $\left\|\frac{dp_t}{d\mu}\right\|_\infty$ 有限。即使拉东-尼科迪姆导数有一致上界，在许多高维情形下，这个上界也可能过大，导致由此得出的结论缺乏实际意义。因此，在第 4 节中，我们提出一种更一般的概念，即 f-平滑对手，其中上下文的分布 p_t 在给定历史 x_t 的条件下满足$D_f\left(p_t||\mu\right) \leq \frac{1}{\sigma}$。在这一更具挑战性的设定中，Block 等人 [2022], Haghtalab 等人 [2022a,b]的结果不再适用，因为这些结果的证明都依赖于一个关键技术步骤，而该步骤在此处失效。在第 4 节中，我们运用近似拒绝抽样样本复杂度的界，推广这些工作的思路，并为 f- 平滑在线学习的信息论速率给出上界；对于某些 f-散度，这些上界是紧的。

尽管Block 等人 [2022], Haghtalab 等人 [2022b]给出的信息论速率很重要，但实现这些速率的算法在计算上不可处理，因此还提出了两种预言机高效算法。在这些算法中，学习者可以访问一个经验风险最小化（ERM）预言机，该预言机返回在某个数据集上评估的加权累积损失函数在 $\mathcal{F}$ 上的极小化者（正式定义见定义36 ）。同样，这两种算法的分析无法推广到标准平滑设定之外；在第4节中，我们再次应用拒绝采样的样本复杂度界来证明，只需调整这两种算法的超参数，便仍可在更一般得多的 f平滑在线学习设定下保证无悔。

受篇幅所限，相关工作的讨论留待附录A中进行。现在，我们总结主要贡献：

- 在定理3中，我们给出了一个上界，刻画在能够访问来自 μ 的样本时，从某个目标测度 ν 近似采样的复杂度。具体而言，我们证明，通过改进经典拒绝采样，只需 $\widetilde{\Theta}\left((f')^{-1}\left(\frac{D_f(\nu||\mu}{\varepsilon}\right)\right)$ 个样本，即可获得一个与目标测度的全变差距离至多为 ε 的样本。
- 在命题4以及定理5和定理6中，我们证明了拒绝采样给出的上界在本质上是紧的。具体而言，我们证明，拒绝采样在某种意义上具有普适性：使用来自 μ 的样本生成来自 ν 的样本时，“最佳”方法就是上述方法。此外，我们证明，如果 f' 有上界，那么近似采样问题是不可能解决的；如果 f' 无界，我们在定理5和定理6中证明，定理 3推导出的样本复杂度在本质上是紧的，且 $\varepsilon \downarrow 0$。具体而言，定理 5表明，对于所有 n，都存在具有有界 f-散度的分布，使得生成一个来自目标测度的 ε-近似样本需要 $\Omega\left((f')^{-1}\left(\frac{D_f(\nu||\mu}{\varepsilon}\right)\right)$ 个样本；而在定理 6中，我们证明，对于满足温和增长条件的一个稍小的 f 类，存在一些分布，使得前述下界对 n一致成立。
- 在第 4 节中，我们将此前关于平滑在线学习的结果推广到更一般得多的 f-平滑在线学习设定。具体而言，我们推导出不考虑计算时间的极小极大上界，并证明 Block 等人 [2022] 提出的两种预言机高效算法（一种适当算法和一种不适当算法）即使在更一般的 f-平滑在线学习设定中仍然是无悔的。此外，在定理 12 中，我们通过证明 FTPL 的一个实例具有Block 等人 [2022] ，回答了他们提出的一个开放问题。

遗憾的尺度为 $\sigma^{-1/4}$，而非 $\sigma^{-1/2}$；其中 σ 是对手的平滑度参数。这将Haghtalab 等人 [2022a] 的一项结果推广到了任意上下文空间。

- 在附录 B中，我们证明了关于重要性采样质量的新界，用于相对于目标测度估计均值，并在函数类 ν 上统一成立 $\mathcal{F}$，前提是我们可以获取来自 μ 的样本。随后，我们将这些结果与假设 $D_f(\nu||\mu) < \infty$ 时采用拒绝采样所得的估计进行比较，针对 χ^2-散度这一特殊情形，并将这些结果与 Chatterjee 和 Diaconis [2018], Cortes 等人 [2010]此前给出的界进行比较。

记号 在下文中，我们始终用 μ 表示集合 $\mathcal{X}$ 上的一个基准测度，其对应的σ-代数为 $\mathscr{F}$。我们用 $X_{1:n} = (X_1, \ldots, X_n)$表示来自 μ的 n 个独立样本组成的向量，并将 j^* 记作一个选择规则。我们用 ν表示目标测度，并将字母 ε、δ、γ保留用于表示较小的正实数常量。此外，我们用 f 表示一个将正实数映射到正实数的凸函数，并满足 $f(1) = f'(1) = 0$。此外，对于这样的 f，我们令 $f^{-1}(u) = \inf\{t > 0 | f'(t) \geq u\}$，其中采用标准约定，将空集的下确界视为无穷大。对于给定的随机变量 Y，我们用 P_Y表示 Y的分布。我们用 $O(\cdot)$、$\Omega(\cdot)$表示渐近大 O 记号，并用波浪号省略多对数因子。

2 问题设置与符号说明

本节将正式定义必要的信息论量，并陈述问题。首先，我们定义 f-散度。关于信息论概念的更多信息，请参见Polyanskiy and Wu [2022+]

定义 1. 设 f: $[0, \infty] \to \mathbb{R}_{\geq 0} \cup \{\infty\}$是满足 $f(1) = f'(1) = 0$ 的凸函数。对于某个空间 $\mathcal{X}$ 上的两个概率测度 ν, μ，定义 f-散度为:

$$D_f(\nu||\mu) = \mathbb{E}_\mu \left[f\left(\frac{d\nu}{d\mu}(Z)\right) \mathbb{I}\left[\frac{d\nu}{d\mu}(Z) < \infty\right]\right] + f'(\infty)\mu\left(\frac{d\nu}{d\mu}(Z) = \infty\right).$$

注意，如果 $\nu \ll \mu$，那么我们可以忽略第二项。

备注2。作为一个技术性补充，本文中我们将使用 f'和 f''分别表示定义 f 中出现的1的函数的一阶和二阶导数。根据 Rademacher 定理[Rademacher, 1919], f 几乎处处可微，但对于 f 不可微的点，我们将 f'取为最大次梯度。由于 f是递增的，f'是非递减的，因此我们可以将 f''取为 f'的右导数，它总是有定义的。

我们将用一般 f下的 $D_f(\nu||\mu)$来表述我们的结果，但全文还会涉及几个重要的例子。在正式介绍该问题之前，我们先给出几个众所周知的 f-散度的例子:

例 1（全变差）。考虑 $f(x) = |x-1| - (x-1)$。在这种情况下，我们有

$$D_f(\nu||\mu) = \mathrm{TV}(\nu, \mu) = \sup_{A \in \mathscr{F}} |\nu(A) - \mu(A)|$$

全变差 距离，其中 $\mathscr{F}$是 $\mathcal{X}$上的公共 σ代数，ν、μ定义在其上。

例 2（KL 散度）。若令 $f(x) = x\log(x) - x + 1$，则得到 $D_f(\nu||\mu) = D_{KL}(\nu||\mu)$，即 KL 散度。

例 3（Rényi 散度）。若我们令 $f(x) = x^\lambda - \lambda x + \lambda - 1$，则可得

$$D_f(\nu||\mu) = e^{(\lambda-1)D_\lambda(\nu||\mu)}$$

其中，$D_\lambda(\nu||\mu)$ 是阶数为 λ 的 Rényi 散度。$D_\lambda(\nu||\mu)$ 的特殊情形包括 $\lambda \downarrow 1$ 的情况，此时我们再次得到 KL 散度；以及 $\lambda = 2$ 的情况，此时我们得到标准 χ^2 散度（经单调变换后）。

例 4（$\mathcal{E}_\gamma$ 散度）。若对于 $\gamma \geq 1$，令 $f_\gamma(x) = (x-\gamma)_+$，则用 $\mathcal{E}_\gamma(\nu||\mu)$ 表示与该 f 相关的散度。该散度最初由 [Polyanskiy 于 2010 年在（2.141）] 中定义，用于研究信道编码。此后，它在差分隐私研究中 [Asoodeh 等，2021] 以及窃听信道研究中 [Liu 等，2017] 都发挥了重要作用。它对于下文下界的证明也至关重要。

现在我们定义研究的主要对象。给定 $X_{1:n} = (X_1、\ldots\ldots、X_n)$ 为 $\mathcal{X}$ 中元素组成的元组，我们将选择规则 j^* 定义为：取值于 $[n]$ 且以任意方式依赖于 $X_{1:n}$ 的任意随机变量。现在我们可以正式陈述主要问题：

> 问题：设 $\mathcal{X}$ 为任意集合，其上配备 σ-代数 $\mathscr{F}$；设 μ、ν 为关于 $\mathscr{F}$ 的概率测度，并且对于某个固定的 f，满足 $D_f(\nu||\mu) < \infty$。对于固定的 $\varepsilon > 0$，n 需要多大，才能存在一个选择规则 j^*，使得 $\mathrm{TV}\left(P_{X_{j^*}}、\nu\right) < \varepsilon$?

作为一个例子，我们考虑传统的拒绝采样。我们通过以概率 $\frac{1}{M}\cdot\frac{d\nu}{d\mu}(X_j)$ 将 j 加入 $\mathcal{S}$，来构造随机集合 $\mathcal{S} \subset [n]$；根据 $M > \left|\left|\frac{d\nu}{d\mu}\right|\right|_\infty$ 这一假设，该概率至多为 1。如果 $\mathcal{S}$ 非空，我们令 j^* 为其中任意一个元素；否则，我们均匀随机地选择 j^*。正如我们将为完整起见所证明的（见附录 D 中的引理 26），$\mathcal{S}$ 为空的概率至多为 $e^{-\frac{n}{M}}$；如果 $\mathcal{S}$ 非空，则 X_{j^*} 服从 ν 的分布。因此，如果 $n = M\log\left(\frac{1}{\delta}\right)$，则以至少 $1-\delta$ 的概率，X_{j^*} 服从 ν 的分布。由于我们要求 $M > \left|\left|\frac{d\nu}{d\mu}\right|\right|_\infty$，可知 $\Theta\left(\left|\left|\frac{d\nu}{d\mu}\right|\right|_\infty \log\left(\frac{1}{\delta}\right)\right)$ 个样本足以以高概率从 ν 中进行精确采样。其必要性将在下一节的下界中作为一个非常特殊的情形加以说明。

3 拒绝采样的样本复杂度

在本节中，我们陈述并概述关于拒绝采样的主要结果的证明，并完整回答第 2 节提出的问题。我们将结果分为两个定理：一个使用拒绝采样的改进版本给出上界，另一个给出几乎匹配的下界。我们先介绍上界：

定理 3（上界）。假设 μ、ν 是集合 $\mathcal{X}$ 上的概率分布，并假设 $X_1, \ldots, X_n \sim \mu$ 相互独立。固定一个满足定义中条件的 f

1。对于 $\varepsilon > 0$，如果

$$n \geq \frac{2}{1-\varepsilon} \log\left(\frac{2}{\varepsilon}\right) (f')^{-1}\left(\frac{4 \cdot D_f(\nu||\mu)}{\varepsilon}\right) \vee 2$$

那么存在一个选择规则 j^*，满足 TV $\left(P_{X_{j^*}}, \nu\right) \leq \varepsilon$。

我们将讨论分为两种情形：超线性情形，即 $f'(t) \uparrow \infty$ 当 $t \uparrow \infty$；以及线性情形，即 $f'(t)$ 有上界。在前一种情形中，我们将看到，随着 $n \uparrow \infty$，我们总能使用拒绝采样，从 ν中得到越来越好的近似样本，因为 $(f')^{-1}$ 在整个正实数轴上有限。然而，在线性情形中，我们很快将证明，没有任何选择规则能够使结果在全变差意义下任意接近 ν。在概述定理 3的证明之前，我们先给出一些例子。

例 5 (全变差)。回顾一下，全变差是满足以下条件的 f-散度：$f(x) = |x-1| - x + 1$。注意，对于所有 $f'(x) = 0$，都有 $x > 1$，因此 $(f')^{-1}(M)$ 在 $M > 0$时为无穷大。因此，正如预期的那样，当我们只能控制全变差时，定理 3 便是空洞的。

例 6（KL 散度）。正如我们在例 2 中所看到的，KL 散度是将 $f(x) = x\log(x) - x + 1$ 设定为 f-散度的情形。在此情形下，我们看到 $f'(x) = \log(x)$，因此定理 3 告诉我们，为了在全变差意义下达到 ε-接近，只需 $\widetilde{O}\left(\exp\left(\frac{D_{KL}(\nu||\mu}{\varepsilon}\right)\right)$ 个样本即可。

例 7（Rényi 散度）。回顾例 3，$f(x) = x^\lambda - \lambda x + \lambda - 1$ for$\lambda > 1$ 定义了 Rényi 散度。在此情形下，我们看到 $\widetilde{O}\left(e^{D_\lambda(\nu||\mu)}\varepsilon^{-\frac{1}{\lambda-1}}\right)$ 个样本即可。由于 $\lambda \uparrow \infty$，我们令 $\varepsilon \downarrow 0$ 并注意到 $D_\infty(\nu||\mu) = \left|\left|\frac{d\nu}{d\mu}\right|\right|_\infty$，从而得到标准的拒绝采样界。在 $\lambda = 2$ 的特殊情形下，我们注意到 Rényi 散度恢复了 χ^2-散度，并注意到只需 $\widetilde{O}\left(\frac{\chi^2(\nu||\mu}{\varepsilon}\right)$ 个样本即可。

现在概述上界的证明，详细内容留待附录 D:

定理 *3* 的证明。令 ν_M 表示在事件 $\frac{d\nu}{d\mu} \leq M$ 发生的条件下的测度 ν，并令 $\widetilde{\nu}$ 表示从 ν_M 中进行拒绝采样并使用 n 个样本所产生样本的分布律。拒绝采样的标准分析告诉我们，如果 $n = \Omega\left(\log\left(\frac{1}{\varepsilon}\right) M\right)$，那么 $\mathrm{TV}(\widetilde{\nu}, \nu_M) \leq \varepsilon$。我们将在引理 27 中证明，如果 $M > 1$，则

$$\nu\left(\frac{d\nu}{d\mu} > M\right) \leq \frac{2 \cdot D_f(\nu||\mu)}{f'(M/2)}.$$

利用这一结果，我们证明 TV $(\nu_M, \nu) \leq \frac{2 \cdot D_f(\nu||\mu)}{f'(M/2)}$，最后应用三角不等式得出结论。
■

现在我们转而讨论下界。具体而言，我们证明，对于任意 f-散度，都存在满足 $D_f(\nu||\mu) < \infty$ 的分布 μ, ν，使得若要存在一个能保证 TV $\left(P_{X_{j^*}}, \nu\right) < \varepsilon$ 的选择规则，就要求 n 足够大。我们将再次分别讨论线性情形和超线性情形。对于线性情形，有如下下界:

命题4（下界，线性情形）。设 f 是定义 1 中的凸函数，满足对所有 $t>1$，均有 $f'(t) \leq C<\infty$。则存在分布 μ, ν，使得 $D_f(\nu||\mu)<\infty$，且 $\varepsilon=\varepsilon(f, D_f(\nu||\mu))>0$，从而对所有 n 和 $X_1, \ldots X_n \overset{iid}{\sim} \mu$。

$$\inf_{j^*} \mathrm{TV}\left(P_{X_{j^*}}, \nu\right) \geq \varepsilon$$

其中，下确界取遍所有选择规则 j^*。

注意，命题 4 与定理 f 3 中的线性情形上界相符，并反映了这样一个事实：对于 f 不超线性增长的情形，$D_f(\nu||\mu)<\infty$ 对 ν的控制非常弱。直观上这应当很清楚：注意，如果 f处于线性区域，那么即使 $D_f(\nu||\mu)$ 仍然有限，ν 也可能相对于 μ 是奇异的，因此从 μ 中抽取的样本永远无法指望以任意精度逼近ν。完整证明见附录D。

接下来讨论更有意思的超线性 f情形，我们给出一个下界，与定理 3 中针对所有超线性 f的上界相匹配。

定理 5 (下界，超线性情形)。设 f 是定义 *1* 中所述的凸函数，且其超线性增长。则对于所有 $0<\varepsilon \leq 1/4$ 以及 $\delta>2f(1/2)$，存在一对测度ν，μ使得 $D_f(\nu||\mu) \leq \delta$，且任何满足条件 j^* 的选择规则 $\mathrm{TV}(P_{X_{j^*}}, \nu) \leq \varepsilon$ 都需要

$$n \geq \frac{1}{2} \cdot (f')^{-1}\left(\frac{\delta}{2\varepsilon}\right). \tag{1}$$

我们在附录 D中给出完整细节，此处给出证明概要：

证明。引理 29 中的一个简单计算表明，如果 $\widetilde{\nu}$ 是 X_{j^*}的分布律，那么 $\widetilde{\nu}$ 相对于 μ 的拉东-尼科迪姆导数一致地有界于 n。引理 32 中的另一个计算表明，如果 $\widetilde{\nu}$ 的似然比有界于 n，那么我们可以将 TV $(\widetilde{\nu}, \nu)$ 下界为 $\mathcal{E}_n(\nu||\mu)$。结合这些事实可知，只需找出两个分布 μ, ν，使得对于所有不满足 (1) 的 n，都有 $D_f(\nu||\mu) \leq \delta$ 且 $\mathcal{E}_n(\nu||\mu) \geq \varepsilon$。因此，我们将证明归结为判断点 (ε, δ) 是否位于 μ 和 ν 的联合值域中的某个点之上，即集合 $\{(\mathcal{E}_n(\nu||\mu)\ D_f(\nu||\mu))\}$，其中 μ 和 ν 遍历所有分布。Harremo¨es 和 Vajda [2011], 指出，使联合值域达到极值的分布通常是成对的伯努利随机变量。因此，我们考虑 $\mu=$ 伯努利分布 $\left(\frac{\varepsilon}{n}\right)$ 和 $\nu=$ 伯努利分布(2ε)，并证明 $\mathcal{E}_n(\nu||\mu)=\varepsilon$，而 $D_f(\nu||\mu) \leq \delta$，除非 n 足够大，足以满足 (1)。结论得证。■

注意，定理 5 告诉我们，忽略对数因子后，定理 3 确定的样本复杂度是最优的。不过，上述结果有一个缺点：正如证明中所清楚表明的，分布 μ 和 ν 取决于 n，因此定理 5 中的量词顺序弱于定理 3。为了解决这一不足，我们在 f 上施加一个温和的增长条件，并证明一个稍弱的下界：

定理 6。设 f 是定义*1* 中所述的超线性增长凸函数。假设f 满足一个温和的增长条件（正式表述见定理 *34* ）。那么，对于任意 $\zeta>0$,

存在分布 μ,ν，使得 $D_f(\nu||\mu) < \infty$，并且对于所有充分大的 $n \in \mathbb{N}$，在X_1、……、X_n从 μ 中独立采样的情况下，有

$$\inf_{j^*} \mathrm{TV}\left(P_{X_{j^*}}, \nu\right) \geq \frac{\zeta^{1+\zeta}}{8} \cdot \left(\frac{D_f(\nu||\mu)}{f'(n)}\right)^{1+\zeta} \tag{2}$$

其中，下确界取遍所有选择规则。

我们注意到，定理 6 所要求的温和增长条件纯粹是技术性的，通过更精细的分析很可能可以去掉；另一方面，据我们所知，所有常用的超线性 f-散度都满足这一条件。由定理 3，可知如果

$$n = \widetilde{O}\left((f')^{-1}\left(\frac{D_f(\nu||\mu)}{\varepsilon}\right)\right),$$

那么拒绝采样足以从 ν生成一个 ε-近似样本。另一方面，令 $\zeta = o(1)$为 $\varepsilon \downarrow 0$，定理 6 告诉我们，在最坏情况下，我们需要

$$n = \widetilde{\Omega}\left((f')^{-1}\left(\frac{D_f(\nu||\mu)}{\varepsilon^{1-o(1)}}\right)\right)$$

使式（2）右侧低于ε所需的样本数。因此，当 $\varepsilon \downarrow 0$时，这些界基本吻合。特别地，由于例 f中的 6 和 7 中的散度满足温和增长条件，这些例子中推导出的样本复杂度上界，对于所有充分大的 n而言确实是紧的。

定理 6 的详细证明推迟至附录 D。该方法与定理 5 类似，因为我们将问题归约为对分布 $\mathcal{E}_n(\nu||\mu)$ 满足有界 f-散度时，给出 单个 对 (μ,ν)，该对取决于 f 但独立于 n，并且满足所需性质。

结合定理 3、5 和 6，我们证明了 $D_f(\nu||\mu)$ 个样本对于生成一个 ε-近似样本 ν既是必要的，也是充分的。根据这些结果，一个直接应用是：当给定来自 ν 的样本时，在某个函数类 $\mathcal{F}$ 上对 μ进行均值估计。附录B中，我们将使用定理3 的估计量与经典的重要性采样方法进行比较。由于篇幅所限，这部分内容放在附录中；现在我们继续讨论关于平滑在线学习的主要应用。

4 平滑在线学习

我们最重要的直接应用，是如引言所述，将平滑在线学习推广这一问题。在本节中，我们将针对平滑对手证明的结果 [Rakhlin 等人，2011年，Block 等人，2022年，Haghtalab 等人，2022b，a] 在引言中所述的结果进行扩展，以允许更强大的自然对手。为此，我们采用以下定义：

定义 7。固定某个集合 $\mathcal{X}$ 上的基准测度 μ。如果测度 ν满足$\leq \frac{1}{\sigma}$，我们就称其相对于 μ是 (f,σ)-平滑（或f-平滑）的。若对于所有 $1 \leq t \leq T$，在给定全部历史的条件下，x_t的分布 p_t都是 (f,σ)-平滑的，则称对手相对于 μ是 (f,σ)-平滑的。

定义 7 引出了一个显而易见的问题：即使我们只要求自然满足 f-平滑，能否在完全对抗的设定下实现改进？答案当然取决于我们选择的 f。例如，对于最终呈线性的 f，我们会看到，一般而言不可能实现改进：

命题 8。 假设 $\mathcal{F} = \{x \mapsto \mathbb{I}[x \geq \theta] | \theta \in [0,1]\}$ 是一维阈值函数类。设 f 是定义 *1* 中的凸函数，且最终呈线性，即f'有上界。对于所有 $0 < \sigma < 1$，存在一个 (f, σ)-平滑对手，使得任何学习者都会遭受 $\mathbb{E}[\text{遗憾}_T] = \Omega(T)$。

这一结果在附录 E中得到证明。鉴于完全对抗式在线学习 $\mathcal{F}$ 是不可能的，这一结果并不令人意外；如果 f 是线性的，那么自然可以将最坏情形对手与基准分布混合，同时仍以有限的 $D_f(\nu||\mu)$造成线性遗憾。更有趣的是更强的 f-散度情形。在介绍我们的结果之前，我们先陈述主要的技术工具，它推广了 Haghtalab 等人 [2022b] 提出并由 Block 等人[2022]扩展的一种技术。在这些工作中，作者构造了一个耦合，将平滑的自适应对手产生的上下文序列，与从基准测度中独立抽取的更大样本集联系起来。借助第 3节中发展的工具，我们将这一技术推广到拉东-尼科迪姆导数不再具有一致上界的情形：

引理 9。 设 $\mathcal{X}$为一个集合，且 μ为 $\mathcal{X}$上的某个测度。假设对手相对于 μ是 f, σ)-平滑的，其中 μ满足 f定义 *1* 的条件，且 $\sup f'(t) = \infty$。对于任意 T和任意 ε， $\delta > 0$若

$$n \geq \frac{1}{1-\varepsilon} \log\left(\frac{T}{\delta}\right) (f')^{-1}\left(\frac{1}{\varepsilon\sigma}\right)$$

则存在 $(x_1, \ldots, x_T)$ 与 $\{Z_{t,j} | 1 \leq t \leq T$和 $1 \leq j \leq n\}$ 之间的一个耦合，使得 $(x_1, \ldots, x_T)$ 按照对手进行分布，$Z_{t,j} \sim \mu$相互独立，并且以至少 $1 - \delta$ 的概率，存在选择规则 j_t^*，使得 TV $\left(P_{x_t}, P_{Z_{t,j_t^*}}\right) \leq \varepsilon$。

耦合的构造推迟至附录 E；现在我们先关注其含义。我们的第一个结果将 Block 等人 [2022, 定理 3] 和 Haghtalab 等人 [2022b，定理 31]推广到 f-平滑在线学习的情形。虽然我们在附录 E中针对一般的实值函数类陈述该结果，但为简化起见，我们在此仅关注二值 $\mathcal{F}$。

定理 10. 设 $\mathcal{F} \to \{\pm 1\}$ 是一个二值函数类，并令 $\text{vc}(\mathcal{F})$ 表示其 Vapnik-Chervonenkis 维数。设 (x_t, y_t) 由一个 (f, σ)-平滑对手按照定义 *7* 生成，且满足 $f'(\infty) = \infty$。那么存在一个算法，使得

$$\mathbb{E}[\text{Reg}_T] \lesssim \sqrt{T \log(T) \cdot \text{vc}(\mathcal{F})} + \inf_{0<\varepsilon<1} \varepsilon T + \sqrt{T \text{vc}(\mathcal{F}) \log\left(T (f')^{-1}\left(\frac{1}{\varepsilon\sigma}\right)\right)}.$$

我们指出，定理 10 是更一般的定理 35 的一个特例。后者适用于任意实值函数类，我们在附录E中陈述并证明。证明沿用了 Block 等人 [2022] 的方法，只是改为应用更一般的耦合

引理 9，证明推迟至附录。这里，我们考虑 f-散度的两种具体情形。首先，对于 Rényi 散度的情形（见例3 和 7），我们看到，对于 Rényi 平滑对手，可以达到 $\widetilde{O}\left(\left(1+\frac{1}{\lambda-1}\right)\sqrt{T\mathrm{vc}(\mathcal{F})\log\left(\frac{1}{\sigma}\right)}\right)$ 量级的遗憾。注意，当$\lambda \uparrow \infty$时，我们得到 Block 等人 [2022]的结果。另一方面，如果 λ 与 1 保持一定距离（这涵盖了对手的 χ^2 散度有界的情形），我们可以看到，仅假设 $D_\lambda(p_t||\mu) < \infty$ 所付出的代价，相比标准设定仅多出常数量级。若假设对手在 KL 散度意义下经过 f-平滑，情况则有所不同：此时我们只能得到按 $\widetilde{O}\left(T^{2/3}\left(\mathrm{vc}(\mathcal{F})/\sigma\right)^{1/3}\right)$缩放的遗憾。虽然 Rényi 散度下的结果在多对数因子范围内是最优的，但针对 KL 平滑对手的遗憾能否改进，仍是一个值得探讨的开放问题。

尽管定理 10 给出了 f-平滑在线学习的信息论速率，因而十分重要，但其中给出的算法在计算上不可处理。现在我们证明，Block 等人 [2022], 和 Haghtalab 等人 [2022为]平滑在线学习提出的两种算法，即使将假设放宽至包含(f, σ)-平滑对手，仍然是无悔的。这些算法是预言机高效的，也就是说，它们对函数类 $\mathcal{F}$的经验风险最小化（ERM）预言机调用次数很少。ERM 预言机的正式定义见附录E （参见定义36），它返回定义在函数类 $\mathcal{F}$上的加权累积损失函数的最小化器。为求简洁，我们再次仅陈述二值 $\mathcal{F}$ 情形下的结果，并将更一般的表述和证明放在附录中。

定理 11。假设 $\mathcal{F}\colon \mathcal{X} \to \{\pm 1\}$ 是一个 VC 维数为 vc $(\mathcal{F})$ 的函数类，并且$\ell\colon [-1,1]\times[-1,1]\to[0,1]$ 是一个关于第一个参数为凸且 1-Lipschitz 的损失函数。那么，存在一种非适当算法，每个时刻 t 需要调用 ERM 预言机 2 次；若对手是 (f,σ)-平滑的，则遗憾有如下界：

$$\mathbb{E}\left[\mathrm{Reg}_T\right] \lesssim \inf_{\alpha>0}\left\{\alpha T + \sqrt{\mathrm{vc}\left(\mathcal{F}\right)\cdot T\cdot\log(T)\cdot\left(f'\right)^{-1}\left(\frac{1}{\alpha\sigma}\right)}\right\}. \tag{3}$$

我们在附录 E 中证明定理 11，并在证明中应用 Block 等人[2022]的论证。我们分别在两种情况下实例化式 (3) 中的界：Rényi 散度（例 3）和 KL 散度（例 2）。如果我们假设对手在 Rényi 散度意义下经过平滑处理，那么通过优化 α，我们得到一个预言机高效算法，其遗憾的量级为 $\widetilde{O}\left(\mathrm{vc}\left(\mathcal{F}\right)^{\frac{\lambda-1}{2\lambda-1}}\cdot T^{\frac{\lambda}{2\lambda-1}}\cdot\sigma^{-\frac{1}{2\lambda-1}}\right)$。注意，如果 $\left\|\frac{dp_t}{d\mu}\right\|_\infty \le (\sigma')^{-1}$，那么我们可以取 $\sigma = (\sigma')^{\lambda-1}$。由此可见，当 $\lambda\uparrow\infty$ 时，在极限情况下，我们恢复了 Block 等人 [2022, 定理 7] 中的 $\widetilde{O}\left(\sqrt{\mathrm{vc}\left(\mathcal{F}\right)}\cdot T/\sigma'\right)$ 阶遗憾。在特殊情形 $\lambda = 2$ 下，遗憾的量级为 $\widetilde{O}\left(\left(\mathrm{vc}\left(\mathcal{F}\right)/\sigma\right)^{1/3}\cdot T^{2/3}\right)$。另一方面，如果我们采用较弱的假设，即对手仅在 KL 意义下经过平滑处理，那么定理 11 得到的遗憾量级仅为 $\sim O(\log(d)T/(\sigma\log(T)))$，它关于 T 是次线性的，但收敛速度非常慢。

现在转而讨论适当算法的情形。如 Block 等人 [2022], 所述，我们使用高斯过程扰动来实例化 Follow the Perturbed Leader（FTPL）；同样地，我们将引理 9 应用于 Block 等人 [2022, 附录 E]中采用的证明技术。为简便起见，我们将重点限制在具有线性损失的二值函数类上。

定理 12。 假设我们处于定理 *11*的情形，且损失函数 ℓ 是线性的，即 $\ell\left(\widehat{y},\ y\right)\ =\ \left(1-\widehat{y}\cdot y\right)/2$。进一步假设我们的对手是 $(f,\ \sigma)$-光滑的（依 Rényi 散度定义），即对某个 $\lambda\geq 2,\ D_{\lambda}\left(p_t||\mu\right)\ \leq 1/\sigma$ 对所有 p_t均成立。那么存在一个 适当算法，每轮仅需调用 1 次经验风险最小化预言机，且其遗憾有如下界：

$$\mathbb{E}\left[\mathrm{Reg}_T\right]=\widetilde{O}\left(\sqrt{\mathrm{vc}\left(\mathcal{F}\right)}\cdot T^{\frac{2\lambda+1}{4\lambda-1}}\cdot\sigma^{-\frac{1}{4\lambda-1}}\right).$$

请注意，在取12 时，我们定理中的遗憾实际上优于 [Block 等人，2022年定理10] 所得到的结果 $\lambda\uparrow\infty$。事实上，若处于强平滑情形，使得对手分布的拉东-尼科迪姆导数一致有界于 σ'^{-1}，那么在极限下，我们得到的期望遗憾具有如下标度：$\widetilde{O}\left(\sqrt{\mathrm{vc}(\mathcal{F})}\cdot T\cdot(\sigma')^{-\frac{1}{4}}\right)$，这与 Haghtalab 等人 [2022a] 针对离散 $\mathcal{X}$所给出的 FTPL 实例相匹配。因此，通过考察 f-平滑对手，我们回答了Block 等人 [2022] 关于如何改善高斯扰动 FTPL 期望遗憾对 σ' 的依赖性这一开放问题。

我们将定理11 和12中算法的遗憾是否紧致这一问题留作一个有趣的后续研究方向。正如Block 等人 [2022], Haghtalab 等人 [2022a],所示，即使在强平滑对手的情形下，也存在统计-计算差距：预言机高效算法的期望遗憾对 σ 的依赖必须是多项式级的，但定理 10 给出的统计速率对同一参数的依赖仅为多对数级。然而，即使在对抗性设定下，对于预言机高效的非适当算法[Hazan and Koren， 2016]是否存在这种指数级差距仍属未知。

最后，我们注意到，定理12 仅适用于在 Rényi 意义下的 f平滑对手 $\lambda\geq 2$。我们的证明采用测度变换论证，即将关于基准测度 μ 的期望替换为关于对手分布 p_t的期望；对于较弱的 f散度（如 KL 散度），类似的陈述需要对指数矩进行界定，这将需要显著更强的分析。关于 KL 平滑对手是否存在预言机高效的适当算法，我们将其留作另一个有趣的后续研究方向。

致谢

AB 感谢美国国家科学基金会研究生研究奖学金（拨款编号 1122374）的支持，以及美国海军研究办公室（拨款 N00014-20-1-2336）和美国能源部（拨款 DE-SC0022199）的支持。YP 部分得到 MIT-IBM Watson 人工智能实验室和美国国家科学基金会拨款 CCF-2131115 的支持。我们还感谢一位匿名审稿人指出一篇相关参考文献，并感谢 Gergely Flamich 和 William Lennie Wells 指出本文早期版本中引理27 证明中的一个错误。

参考文献

Shweta Agrawal、Damien Stehle 和 Anshu Yadav。《基于格的最优轮数阈值签名：再探》。密码学电子印刷档案，

2022/634，2022年。网址：https://eprint.iacr.org/2022/634。https://eprint.iacr.org/2022/634。Shahab Asoodeh、Maryam Aliakbarpour 和 Flavio P Calmon。局部差分隐私等价于 $\mathcal{E}_\gamma$-散度的收缩。arXiv预印本 arXiv:2102.01258，2021。Peter L Bartlett、Philip M Long 和 Robert C Williamson。脂肪分解维数与实值函数的可学习性。载于《第七届计算学习理论年度会议论文集》，第299–310页，1994年。Matthias Bauer 和 Andriy Mnih。用于变分自编码器的重采样先验。载于《第22届国际人工智能与统计学会议》，第66–75页。PMLR，2019年。Bernard Bercu、Elisabeth Gassiat 和 Emmanuel Rio。自归一化经验过程的集中不等式、大偏差与中等偏差。《概率年刊》，30(4)：1576–1604，2002年。Adam Block 和 Max Simchowitz。高效且近乎最优的广义线性函数平滑在线学习。arXiv预印本 arXiv:2205.13056，2022年。Adam Block、Yuval Dagan 和 Alexander Rakhlin。主控测度、序列复杂度与在线学习。载于《学习理论会议》，第587–590页。PMLR，2021年。Adam Block、Yuval Dagan、Noah Golowich 和 Alexander Rakhlin。平滑在线学习与统计学习一样容易。arXiv预印本 arXiv:2202.04690，2022年。Adam Block、Max Simchowitz 和 Russ Tedrake。分段仿射系统中的预测平滑在线学习，2023年。Anselm Blumer、Andrzej Ehrenfeucht、David Haussler 和 Manfred K Warmuth。可学习性与 Vapnik-Chervonenkis维数。《ACM学报》（JACM），36(4)：929–965，1989年。Nicolo Cesa-Bianchi 和 G´abor Lugosi。《预测、学习与博弈》。剑桥大学出版社，2006年。Sourav Chatterjee 和 Persi Diaconis。重要性采样所需的样本量。《应用概率年刊》，28(2)：1099–1135，2018年。Corinna Cortes、Yishay Mansour 和 Mehryar Mohri。重要性加权的学习界。《神经信息处理系统进展》，23，2010年。Julien Devevey、Omar Fawzi、Alain Passel`egue 和 Damien Stehl´e。关于 Lyubashevsky签名方案中的拒绝采样。密码学电子印刷档案，2022年。Bryan Eikema、Germ´an Kruszewski、Christopher R Dance、Hady Elsahar 和 Marc Dymetman。带散度诊断的基于能量的模型近似采样器。《机器学习研究汇刊》，2022年。Bernard D Flury。让接受-拒绝采样变得简单。《SIAM评论》，32(3)：474–476，1990年。

Walter R Gilks 和 Pascal Wild。用于吉布斯采样的自适应拒绝采样。《英国皇家统计学会杂志：C辑（应用统计）》，41(2):337–348，1992年。Evarist Gin´e 和 Richard Nickl。无限维统计模型的数学基础。剑桥大学出版社，2021。Dilan G¨or¨ur 和 Yee Whye Teh。凹凸自适应拒绝采样。《计算与图形统计学杂志》，20(3):670–691，2011年。Aditya Grover、Ramki Gummadi、Miguel Lazaro-Gredilla、Dale Schuurmans 和 Stefano Ermon。变分拒绝采样。载于《人工智能与统计国际会议》，第823–832页。PMLR，2018年。Nika Haghtalab、Tim Roughgarden 和 Abhishek Shetty。在线学习与差分隐私学习的平滑分析。《神经信息处理系统进展》，33:9203–9215，2020。Nika Haghtalab、Yanjun Han、Abhishek Shetty 和 Kunhe Yang。面向超越最坏情况对手的预言机高效在线学习。arXiv预印本 arXiv:2202.08549，2022a。Nika Haghtalab、Tim Roughgarden 和 Abhishek Shetty。针对自适应对手的平滑分析。载于《2021年IEEE第62届计算机科学基础年会（FOCS）》，第942–953页。IEEE，2022b。Peter Harremo¨es 和 Igor Vajda。关于 f散度对及其联合值域。《IEEE信息论汇刊》，57(6):3230–3235，2011年。Prahladh Harsha、Rahul Jain、David McAllester 和 Jaikumar Radhakrishnan。相关性的通信复杂度。载于《IEEE第二十二届计算复杂性年会（CCC'07）》，第10–23页。IEEE，2007年。Juha Harviainen、Antti R¨oysk¨o 和 Mikko Koivisto。利用深度拒绝采样近似永久。《神经信息处理系统进展》，34:213–224，2021年。Elad Hazan 和 Tomer Koren。优化在在线学习中的计算能力。载于《第四十八届ACM计算理论年会论文集》，第128–141页，2016年。Adam Kalai 和 Santosh Vempala。在线决策问题的高效算法。《计算机与系统科学杂志》，71(3):291–307，2005年。Michael J Kearns 和 Robert E Schapire。概率概念的高效分布无关学习。《计算机与系统科学杂志》，48(3):464–497，1994年。Tuen Kloek 和 Herman K Van Dijk。方程组参数的贝叶斯估计：蒙特卡罗积分的应用。《计量经济学》：计量经济学会杂志，第1–19页，1978年。Nick Littlestone。当无关属性大量存在时快速学习：一种新的线性阈值算法。《机器学习》，2(4):285–318，1988年。

Jingbo Liu 和 Sergio Verdu。矩约束下的拒绝采样与非因果采样。载于《2018年IEEE国际信息论研讨会（ISIT）》，第1565–1569页。IEEE，2018年。Jingbo Liu、Paul Cuff 和 Sergio Verd´u。 $\mathcal{E}_\gamma$ -可达性。《IEEE信息论汇刊》，63(5):2629–2658，2017年。doi: 10.1109/TIT.2016.2642111。Jun S Liu。与拒绝采样和重要性采样进行比较的Metropolis化独立采样。《统计与计算》，6(2):113–119，1996年。Jun S Liu 和 Jun S Liu。《科学计算中的蒙特卡罗策略》，第10卷。施普林格，2001年。Vadim Lyubashevsky。无陷门格签名。载于《密码技术理论与应用国际年会》，第738–755页。施普林格，2012年。Luca Martino 和 Joaqu´ın M´ıguez。自适应拒绝采样算法的推广。《统计与计算》，21(4):633–647，2011年。Christian Naesseth、Francisco Ruiz、Scott Linderman 和 David Blei。通过接受-拒绝采样算法进行重参数化梯度计算。载于《人工智能与统计》，第489–498页。PMLR，2017年。Maris Ozols、Martin Roetteler 和 J´er´emie Roland。量子拒绝采样。《ACM计算理论汇刊（TOCT）》，5(3):1–33，2013年。Y. Polyanskiy。信道编码：非渐近基本极限。博士论文，普林斯顿大学，美国新泽西州普林斯顿，2010年。URL http://people.lids.mit.edu/yp/homepage。Yury Polyanskiy 和 Yihong Wu。《信息论：从编码到学习》。剑桥大学出版社， 2022+.¨Hans Rademacher。论多变量函数的局部与整体可微性以及二重积分的变换。《数学年刊》，79(4):340–359，1919。Alexander Rakhlin、Karthik Sridharan 和 Ambuj Tewari。在线学习：随机、受约束和平滑对手。《神经信息处理系统进展》，24，2011年。Alexander Rakhlin、Karthik Sridharan 和 Ambuj Tewari。序列复杂度与一致鞅大数定律。《Probability Theory and Related Fields》，161(1):111–153，2015年。Sasha Rakhlin、Ohad Shamir 和 Karthik Sridharan。松弛与随机化：从价值到算法。《神经信息处理系统进展》，25，2012年。Mark Rudelson 和 Roman Vershynin。随机过程组合学与凸体截面。《数学年刊》，第603–648页，2006年。Rajan Srinivasan。重要性采样：在通信与检测中的应用。施普林格科学与商业媒体，2002年。

Surya T Tokdar 和 Robert E Kass。重要性采样：综述。Wiley Interdisciplinary Reviews: Computational Statistics，2(1):54–60，2010年。Tim Van Erven 和 Peter Harremos。R´enyi 散度与 Kullback-Leibler 散度。IEEE信息论汇刊，60(7):3797–3820，2014年。Ramon van Handel。普适Glivenko–Cantelli性质。Probability Theory and Related Fields，155(3-4):911–934，2013年。Ramon Van Handel。高维概率。技术报告，PRINCETON UNIV NJ，2014年。John Von Neumann。13. 随机数字相关的各种技术。Appl. Math Ser，12(36-38):3，1951年。Martin J Wainwright。高维统计学：非渐近视角，第48卷。剑桥大学出版社，2019年。Zhongxiang Zheng、Anyu Wang 和 Lingyue Qin。拒绝抽样再探：如何选择基于格的签名中的参数。Mathematical Problems in Engineering，2021，2021。

相关工作

拒绝采样 正如引言中所述，拒绝采样由Von Neumann [1951] 开创，因其简单且适用范围广而备受关注，相关文献不胜枚举 [Gilks and Wild，1992年，Liu，1996年，Flury，1990，Harviainen 等，2021，Bauer andMnih， 2019]。近年来，信息论界一直关注如何在假设存在额外结构（例如 Rényi散度界）的情况下，改进精确拒绝采样的样本复杂度界 [Liu and Verdu，2018 年，Harsha 等， 2007]。

或许令人意外的是，近似拒绝采样受到的关注相对较少。有几项研究提出了近似拒绝抽样的样本复杂度与生成样本准确度之间的权衡；其中一个例子是Grover 等人 [2018], , 他们展示了某一特定分布族中描述这种权衡的定性单调性，但没有提供任何定量保证。一些密码学研究[Lyubashevsky、2012年、Zheng 等人，2021，Agrawal 等人， 2022] 针对格上的离散高斯族这一特定情形，给出了样本复杂度的上界。更近期的一项同期研究 [Devevey 等人， 2022] 给出了精确拒绝采样的上界和下界，以及离散超立方体上概率分布的近似拒绝采样的一些上界。与这些先前研究不同，我们的结果适用于任意概率分布和更一般得多的 f-散度。此外，Eikema 等人 [2022] 研究了近似拒绝采样问题，并证明了关于近似拒绝采样的样本复杂度与生成样本准确度之间权衡的若干定性结果。我们的工作首次为一般情形下的近似拒绝采样提供了精确的定量保证。

平滑在线学习 在线学习的研究可追溯至数十年前，相关文献浩如烟海，无法一一列举。Cesa-Bianchi 和 Lugosi[2006] 对这一领域作了很好的入门介绍。近年来，人们对统计学习现象的序列类比兴趣日增 [Rakhlin 等人，2015、2012，Block 等人， 2021]。然而，由于这一研究领域面临统计和计算方面的挑战，一些研究提出了平滑在线学习设定 [Rakhlin 等，2011，Haghtalab 等人， 2020]，；Haghtalab 等人 [2022b],Block 等人 [2022] 给出了刻画平滑在线学习问题难度的统计速率，而 Block 等人 [2022], Haghtalab 等人 [2022a], Block 和 Simchowitz [2022] 则提出了预言机高效算法。在本文中，我们将 Block 等人 [2022],Haghtalab 等人 [2022a] 的结果推广到我们称为 f-平滑在线学习设定的情形，其中对手受到的约束较弱，只需满足较弱意义上的平滑性。

分布外学习 重要性采样由 Kloek 和 Van Dijk [1978]提出，随后因其广泛的适用性而得到深入研究。这个热门领域的参考文献实在太多，无法一一列举，但一些标准的论述见 Liu 和 Liu [2001],Srinivasan [2002], Tokdar 和 Kass [2010]。与我们工作最为相近的是 Chatterjee 和 Diaconis[2018], Cortes 等人 [2010]。在第一篇工作中，作者精确计算了重要性采样的样本复杂度，用于估计给定函数在某个目标测度 f 下的均值 ν。他们指出，$(\nu||\mu)$) 个样本对于完成这项任务既必要又充分，并给出了其主要上界的若干具体实例。遗憾的是，他们的上界过于宽松，无法用于在大型函数类上统一估计均值，而这正是学习理论所要求的。在 Cortes 等人 [2010], 的工作中，作者在假设函数类 $\mathcal{F}$ 具有有限伪维数的前提下，分别在有界似然和有界 Rényi 约束下证明了重要性采样样本复杂度的上界。他们还证明了一个下界。

B 一致均值估计中重要性采样的比较

在本附录中，我们将主要结果应用于一致均值估计。更具体地说，假设学习者可以获得 $X_1, \ldots, X_n \sim \mu$ 个独立样本，并且希望针对某个其他测度 ν，对某个函数类 $\mathcal{F}$ 中的所有函数 f 估计 $\mathbb{E}_{Y\sim\nu}[f(Y)]$。一个自然的问题是，为了使我们的估计以高概率与 $\mathbb{E}[f(Y)]$ 相差不超过 ε， n 必须有多大，以及 ν 必须与 μ 有多接近。本节将比较两种常见方法。第一种使用重要性采样，其中我们将估计量取为

$$I_n(f) = \frac{1}{n}\sum_{i=1}^{n} \frac{d\nu}{d\mu}(X_i) f(X_i),$$

第二种方法是使用拒绝采样，从 ν 中生成 n' 个独立样本，然后直接使用这些样本的经验均值来估计 $\mathbb{E}_{Y\sim\nu}[f(Y)]$。我们先考虑重要性采样。

在 $\mathcal{F} = \{f\}$ 是单元素集合的特殊情形下，以下 Chatterjee 和 Diaconis[2018] 的定理完整回答了样本复杂度问题：

定理 13 （Chatterjee 和 Diaconis [2018]）。设 X_1，.....，$X_n \sim \mu$ 相互独立，并且假设 $n \geq \exp(D_{KL}(\nu||\mu))$。若 f 是实值可测函数，则有

$$\mathbb{E}\left[|I_n(f) - \mathbb{E}_\nu\left[f(Y)\right]|\right] \leq ||f||_{L^2(\nu)} \left(\left(\frac{e^{D_{KL}(\nu||\mu)}}{n} \right)^{\frac{1}{4}} + 2\sqrt{\mathbb{P}_{Y\sim\nu}\left(\frac{d\nu}{d\mu}(Y) > \sqrt{e^{D_{KL}(\nu||\mu)} \cdot n} \right)} \right).$$

此外，如果 $n \leq e^{D_{KL}\,(\nu||\mu)}$，则以至少 $1-\delta$ 的概率，

$$I_n(1) \leq \sqrt{\frac{n}{e^{D_{KL}(\nu||\mu)}}} + \frac{\mathbb{P}\left(\frac{d\nu}{d\mu}(Y) \leq \sqrt{e^{D_{KL}(\nu||\mu)} \cdot n} \right)}{1-\delta}.$$

因此，定理 13 表明，需要 Θ（$\exp(D_{KL}$（$\nu||\mu)$）个样本，重要性采样才能生成一个在期望上接近真实均值的估计。不幸的是，上述保证过于薄弱，无法应用于大型函数类。虽然Cortes 等人 [2010] 给出了关于重要性采样的若干界，这些界在具有有界伪维数的函数类上是一致成立的，但我们在此给出其结果的一个推广，该推广适用于大多数 Donsker 函数类。在此之前，我们需要定义函数类的相关复杂度概念，即括号数。更多详情请参见Gin´e 和 Nickl [2021]。

定义 14. 设 $\mathcal{F}$：$\mathcal{X} \to [-1,1]$ 是一个实值函数类。给定实值函数f_-、f_+，我们将括号 $[f_-, f_+]$ 定义为满足 $f_- \leq f \leq f_+$逐点成立的函数 $f \in \mathcal{F}$所组成的集合。若 $[f_-, f_+]$ 是关于 μ 的一个 ε 括号，则称其为$||f_+ - f_-||_{L^2\,(\mu)} \leq \varepsilon$。我们将括号数 $N_{[]}(\mathcal{F}, \varepsilon)$ 定义为使得 $\mathcal{F}$包含于这些括号的并集中的 ε-括号的最小数量。最后，如果 $\mathcal{F}$ 具有有限括号积分，则

$$\int_0^2 \sqrt{\log N_{[]}\left(\mathcal{F}, \varepsilon\right)} d\varepsilon < \infty.$$

虽然在更一般的复杂度假设下，我们的结论也成立，但为简单起见，我们考虑括号积分。我们有如下结果：

定理 15. 假设 $\mathcal{F}$：$\mathcal{X} \to [-1,1]$ 是一个相对于 μ 具有有限括号积分的实值函数类。进一步假设 $\chi^2(\nu||\mu) < \infty$。则有如下不等式：

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |I_n(f) - \mathbb{E}_\nu\left[f(Y)\right]| \right] \lesssim \sqrt{1+\chi^2\left(\nu||\mu\right)} \cdot \max\left(n^{-1/3}, \sqrt{\frac{1+\chi^2\left(\nu||\mu\right)}{n}} \cdot \int_0^2 \sqrt{\log N_{[]}\left(\mathcal{F}, \alpha\right)} d\alpha \right).$$

定理 15 的证明基于以下结果。为便于后文引用，我们用 P_n表示从 n 中独立抽取的点所构成的经验测度，其中抽取点数为 μ。

定理 16 （Bercu 等人 [2002]中的定理 1.1）。设 $\mathcal{F}$ 是一个实值函数类，其括号数有限，并满足

$$E_{\mathcal{F}} = \sup_{n>0} \mathbb{E}\left[\sup_{f\in\mathcal{F}} \max\left(P_n\left(f - \mathbb{E}_\mu[f]\right), 0 \right) \right] < \infty.$$

对于任意 $\delta > 0$ 和 $\alpha > \sqrt{2}$，存在依赖于 δ 和 α 的常数 θ、n_0，使得

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{P_n(f-\mathbb{E}_\mu[f])}{\sqrt{P_n\left(f-\mathbb{E}_\mu[f]\right)^2}} \geq \frac{x+E_{\mathcal{F}}}{\sqrt{n}}\right) \leq e^{-\frac{x^2}{4\alpha^2(1+\delta)}}$$

对所有 $n \geq n_0$ 和 $x \leq \theta\sqrt{n}$。

我们注意到，Gin´e 和 Nickl [2021, 定理 3.5。13] 告诉我们，除一个常数因子外，E 可由以下函数类的括号积分界定：$\mathcal{F}$。现在我们可以证明重要性抽样界了：

定理 *15* 的证明。我们将把定理 16 应用于函数类

$$\widetilde{\mathcal{F}} = \frac{d\nu}{d\mu}\cdot\mathcal{F} = \left\{x \mapsto \frac{d\nu}{d\mu}(x)\cdot f(x) | f \in \mathcal{F}\right\}.$$

首先注意到，柯西–施瓦茨不等式告诉我们，若 $[f_-, f_+]$ 是 μ 的一个 ε-括号，则 $\left[\frac{d\nu}{d\mu}\cdot f_-, \frac{d\nu}{d\mu}\cdot f_+\right]$ 是 μ 的一个 $(\sqrt{1+\chi^2(\nu||\mu)}\cdot\varepsilon)$-括号。我们还注意到，

$$I_n(f) = P_n\left(\frac{d\nu}{d\mu}\cdot f\right) \qquad\qquad \mathbb{E}_\nu\left[f(Y)\right]\mathbb{E}_\mu\left[\frac{d\nu}{d\mu}\cdot f\right].$$

现在注意到，只需证明上尾界，而对称性也将推出下尾界。我们将应用定理 16，但首先必须界定相关量。注意到Gin´e 和 Nickl [2021, 定理 3.5。13] 推出 $E_{\mathcal{F}}$ 可由以下函数类的括号积分界定：$\mathcal{F}$；结合我们关于 ε 的 -括号 $\mathcal{F}$ 与 $\sim\mathcal{F}$ 的括号之间关系的观察，我们得到

$$E_{\widetilde{\mathcal{F}}} \lesssim \sqrt{1+\chi^2\left(\nu||\mu\right)}\cdot\int_0^2 \sqrt{\log N_{[]}\left(\mathcal{F},\alpha\right)}\,d\alpha.$$

进一步利用柯西-施瓦茨不等式可得

$$\begin{aligned}
\sqrt{P_n\left(f-\mathbb{E}_\mu[f]\right)^2} &\leq 2\cdot\sqrt{P_n\left(\frac{d\nu}{d\mu}\cdot f\right)^2 + \mathbb{E}_\mu\left[\left(\frac{d\nu}{d\mu}\cdot f\right)^2\right]} \\
&\leq 2\cdot\sqrt{P_n\left(\frac{d\nu}{d\mu}\right)^2\cdot P_n(f)^2 + \left(1+\chi^2\left(\nu||\mu\right)\right)\cdot\mathbb{E}_\mu\left[f^2\right]} \\
&\leq 2\cdot\sqrt{P_n\left(\frac{d\nu}{d\mu}^2\right) + 1 + \chi^2\left(\nu||\mu\right)},
\end{aligned}$$

其中我们使用了 由于 $\mathcal{F}$一致有界。我们进一步根据马尔可夫不等式可知

$$\mathbb{P}\left(P_n\left(\frac{d\nu}{d\mu}\right)^2 > u^2\right) \leq \frac{1+\chi^2\left(\nu||\mu\right)}{u^2}. \tag{4}$$

根据定理 16，有

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} I_n(f) - \mathbb{E}_\nu[f] > t \text{ and } P_n\left(\frac{d\nu}{d\mu}\right)^2 \le u^2\right) \le \exp\left(-C\left(\sqrt{\frac{n}{1+u^2+\chi^2(\nu||\mu)}}\cdot t - E_{\widetilde{\mathcal{F}}}\right)^2\right).$$

现在，令

$$u = n^{\frac{1}{6}}\cdot t^{\frac{2}{3}}\cdot\sqrt{1+\chi^2(\nu||\mu)}$$

我们看到，只要

$$t \ge C\sqrt{\frac{1+\chi^2(\nu||\mu)}{n}}\cdot E_{\widetilde{\mathcal{F}}},$$

就有

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} I_n(f) - \mathbb{E}_\nu[f] > t \text{ and } P_n\left(\frac{d\nu}{d\mu}\right)^2 \le u^2\right) \le C\exp\left(-C\frac{n^{2/3}t^{2/3}}{(1+\chi^2(\nu||\mu))}\right).$$

应用（4），我们看到

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} I_n(f) - \mathbb{E}_\nu[f] > t\right) \le \frac{1}{n^{1/3}\cdot t^{4/3}} + C\exp\left(-C\frac{n^{2/3}t^{2/3}}{(1+\chi^2(\nu||\mu))}\right).$$

对 $-\mathcal{F}$ 应用相同的论证，可得下尾也成立。最后，我们注意到

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in\mathcal{F}}|I_n(f)-\mathbb{E}_\nu[f]|\right] &= \int_0^\infty \mathbb{P}\left(\sup_{f\in\mathcal{F}}|I_n(f)-\mathbb{E}_\nu[f]|>t\right)dt\\ &\lesssim \sqrt{\frac{1+\chi^2(\nu||\mu)}{n}}\cdot E_{\widetilde{\mathcal{F}}} + \frac{\sqrt{1+\chi^2(\nu||\mu)}}{n^{\frac{1}{3}}}\\ &\lesssim \sqrt{1+\chi^2(\nu||\mu)}\cdot\max\left(n^{-1/3}, \sqrt{\frac{1+\chi^2(\nu||\mu)}{n}}\cdot\int_0^2\sqrt{\log N_{[]}(\mathcal{F},\alpha)}d\alpha\right)\end{aligned}$$

这正是所需的结果。 ■

注意，当 $n\uparrow\infty$时，定理 15 给出的速率按 $O\left(n^{-1/3}\right)$的量级缩放。Cortes 等人 [2010], 的研究带来了一项改进：他们假设函数类 $\mathcal{F}$ 具有有限伪维数，并得到按 $O\left(n^{-3/8}\right)$的量级缩放的速率。请注意，函数类具有有限括号积分这一性质，与具有有限伪维数并不相同；事实上，Van Handel [2013, 命题 17] 表明，即使对于 VC 维数有限的函数类，括号数也可能任意大；而 Gin´e 和 Nickl [2021] 则表明，像索伯列夫空间这样 VC 维数无限的函数类，其括号数仍可能很小。在 Cortes 等人 [2010, 的命题 2], 中，作者在假设 $\frac{d\nu}{d\mu}$ 一致有界且样本量相对于该一致界足够大的条件下，给出了重要性抽样的一个下界

界。本质上，这表明如果 $\chi^2(\nu||\mu)$ 是无限的，那么我们就不能期望得到一个重要性抽样估计量，使其以 $\Theta\left(n^{-\frac{1}{2}}\right)$ 速率收敛到总体均值。我们将弥合这两种速率之间差距的问题留作未来研究的一个有趣方向。

现在我们转向第二种估计量：使用拒绝采样从 ν 中生成独立样本，并计算样本均值。更准确地说，给定 $X_1, \ldots, X_n \sim \mu$ 相互独立，且对于某个 $m \in \mathbb{N}$ 将 n分成大小为 $[n]$ 的集合，假设我们将 m 划分成大小为 m 的集合，并在每个子集上执行定理3 中的拒绝采样过程，生成 n/m 个独立样本$X'_1, \ldots, X'_{n/m} \sim \nu_M$个独立样本，且其概率至少为 $1-n\delta$，其中某个 δ。对于给定的 m， n，令

$$J_{m,n}(f) = \frac{m}{n} \cdot \sum_{i=1}^{n/m} f(X'_i) \tag{5}$$

表示 $\mathbb{E}_\nu[f(Y)]$ 的拒绝采样估计量。我们有如下结果：

推论 17. 假设 $\mathcal{F}$: $\mathcal{X} \to [-1,1]$ 是一个关于 μ 具有有限括号积分的实值函数类，且 $D_f(\nu||\mu) < \infty$。进一步假设，对于某个 $\frac{1}{2} > \varepsilon > 0$,

$$m \geq 4 \log\left(\frac{1}{\varepsilon}\right) \cdot (f')^{-1}\left(\frac{D_f(\nu||\mu)}{\varepsilon}\right).$$

则有

$$\mathbb{E}\left[\sup_{g\in\mathcal{F}} |J_{m,n}(g) - \mathbb{E}_\nu[g(Y)]|\right] \lesssim \left(\int_0^2 \sqrt{\log N_{[]}(\mathcal{F},\alpha)}d\alpha\right) \cdot \sqrt{\frac{m}{n}} + 2\varepsilon$$

证明。 我们首先援引定理 3，并注意到我们可以选择 X'_1, ..., $X'_{n/m} \sim \widetilde{\nu}$相互独立，且存在某个 $\widetilde{\nu}$满足 TV（$\widetilde{\nu}$， ν） $\leq \varepsilon$。注意，根据 Gin´e 和 Nickl [2021, 定理 3.5， 13],有

$$\mathbb{E}\left[\sup_{g\in\mathcal{F}} \left|J_{m,n}(g) - \mathbb{E}_{\widetilde{\nu}}\left[g(\widetilde{Y})\right]\right|\right] \lesssim \sqrt{\frac{m}{n}} \cdot \left(\int_0^2 \sqrt{\log N_{[]}(\mathcal{F},\alpha)}d\alpha\right).$$

现在注意到 g 取值于 $[-1,1]$，我们得到

$$\mathbb{E}\left[\sup_{g\in\mathcal{F}} \left|\mathbb{E}_{\widetilde{\nu}}\left[g(\widetilde{Y})\right] - \mathbb{E}_\nu[g(Y)]\right|\right] \leq 2 \cdot \mathrm{TV}(\widetilde{\nu},\nu) \leq 2\varepsilon.$$

结论得证。 ■

重新缩放后，我们 看到推论 17 告诉我们，如果希望应用估计量（5），则需要

$$n = \Theta\left(\frac{\log\left(\frac{1}{\varepsilon}\right) \cdot (f')^{-1}\left(\frac{D_f(\nu||\mu)}{\varepsilon}\right)}{\varepsilon^2} \cdot \left(\int_0^2 \sqrt{\log N_{[]}(\mathcal{F},\alpha)}d\alpha\right)^2\right)$$

为了使我们的估计值在函数类 $\mathcal{F}$ 中一致地与真实总体均值相差不超过 ε，所需的样本数。在我们的 f-散度为 $\chi^2(\nu||\mu)$ 这一特殊情形下，我们看到 $\Theta(\chi^2(\nu||\mu) \cdot \varepsilon^{-3})$

个样本足以恢复均值的一致估计，误差不超过 ε。请注意，这与重要性采样给出的速率相符：根据定理 15， $O(\chi^2(\nu||\mu)\cdot\varepsilon^{-3})$ 个样本足以在相同情形下恢复期望意义下 ε-精度的一致估计。另一方面，推论 17 适用于任意 f-散度，因此具有显著更广的适用性；至于定理 15 能否推广到更一般的 f-散度，我们留待未来工作研究。不过我们注意到，拒绝抽样估计量的分析本质上是紧的，而重要性抽样估计量的分析则可能较为宽松。重要性采样优于拒绝采样的一种情形是 $\mathcal{F}$ 具有有限伪维数；在这种情况下，Cortes 等人 [2010, 定理 3] 告诉我们， $\widetilde{O}\left(\sqrt{\chi^2(\nu||\mu)}\cdot\varepsilon^{-8/3}\right)$ 个样本就足以进行重要性采样，其效果严格优于上文推导出的拒绝采样速率。

C 序列复杂度与极小极大遗憾

本节回顾在线学习问题的不同复杂度概念的一些基本定义，以及它们与遗憾的关系。这些结果将在整个附录 E 中使用。许多对抗性复杂度由 Rakhlin 等人 [2015] 提出，我们也紧密遵循该文以及 Block 等人 [2022] 的论述。我们首先回顾 Bartlett 等人 [1994] 提出的尺度敏感 VC 维数的定义：

定义18。设 $\mathcal{F}: \mathcal{X}\to\mathbb{R}$ 是一个函数类。对于任意 $\alpha>0$ 和点 $X=\{x_1,\ldots,x_m\}\subset\mathcal{X}$，如果点集 X 以尺度 α 、见证 s_1、 ..... 、 $s_m\in\mathbb{R}$ 打散 $\mathcal{F}$，那么对于所有 m 元符号 $\varepsilon=$ （ε_1，.....，ε_m），都存在某个 $f_\varepsilon\in\mathcal{F}$，使得对于所有 $1\le i\le m$:

$$\varepsilon_i(f_\varepsilon(x_i)-s_i)\ge\frac{\alpha}{2}.$$

我们令 $\mathrm{vc}(\mathcal{F},\alpha)$ 表示最大的 m，使得存在一个大小为 m 的集合 X，它能打散 $\mathcal{F}$；并令 $\mathrm{vc}(\mathcal{F})=\sup_\alpha \mathrm{vc}(\mathcal{F},\alpha)$。

众所周知，Kearns 和 Schapire [1994], Bartlett 等人 [1994] 证明了 $\mathrm{vc}(\mathcal{F},\alpha)$ 刻画了 $\mathcal{F}$ 在数据独立同分布时（即批量设定下）的可学习性。另一个相关量是 Rademacher复杂度，对某个集合x_1，...，$x_n\in\mathcal{X}$定义为

$$\mathfrak{R}_n(\mathcal{F})=\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^n\varepsilon_i f(x_i)\right], \tag{6}$$

其中， ε_i 是独立的 Rademacher 随机变量。以下结果可见于Rudelson 和 Vershynin [2006]:

命题 19。设 $\mathcal{F}: \mathcal{X}\to[-1,1]$ 为一个函数类。则有

$$\mathfrak{R}_n(\mathcal{F})\lesssim\inf_{\gamma>0}\gamma n+\sqrt{n}\int_\gamma^1\sqrt{\mathrm{vc}(\mathcal{F},\delta)}d\delta.$$

Block 等人 [2021]证明了一个适用于完全对抗性设定的类似结果。为了准确陈述其对应形式，我们首先回顾 Rakhlin 等人 [2011]提出的依赖分布的序列拉德马赫复杂度概念。对于给定深度 T、顶点标记为某个空间 $\mathcal{X}$ 中元素的完全二叉树 $\mathbf{x}$，以及一条路径 $\varepsilon \in \{\pm 1\}^T$，我们将 $\mathbf{x}_t(\varepsilon)$ 表示为该路径在步骤 t 处的顶点：从根节点出发，若 $\varepsilon_s = 1$，则在时间 s 取右子节点，否则取左子节点。对于给定的生成 x_1、……、x_T的对手，令 $P_{x_{1:T}}$表示 x_1、……、x_T的联合分布，并令 p_t 表示在给定历史条件下 x_t 的分布。在标签属于 $\mathcal{X}$、深度为 T 的二叉树有序对 $(\mathbf{x}, \mathbf{x}')$ 上递归定义测度 $\rho_{P_{x_{1:T}}}$ 如下。首先独立地抽取 $\mathbf{x}_0, \mathbf{x}_0' \sim p_0$。现在假设 $t > 0$，并且对于任意 $s < t$，令

$$\chi_s(\varepsilon) = \begin{cases} \mathbf{x}_s(\varepsilon) & \varepsilon_s = 1 \\ \mathbf{x}_s'(\varepsilon) & \varepsilon_s = -1 \end{cases}.$$

独立地采样 $\mathbf{x}_t(\varepsilon)$、$\mathbf{x}_t'(\varepsilon) \sim p_t(\cdot|\chi_1(\varepsilon) \ldots, \chi_{t-1}(\varepsilon))$，并继续采样，直到构造出两棵深度为 T 的二叉树。现在我们定义依赖分布的序列拉德马赫复杂度：

定义 20 (Rakhlin 等人 的定义 2 [2011])。令 $\mathcal{F}: \mathcal{X} \to [-1, 1]$ 为一个函数类，并固定一个定义在元组 $P_{x_{1:T}}$上的分布 $(x_1, \ldots, x_T)$。我们将依赖分布的序列拉德马赫复杂度定义如下：

$$\mathfrak{R}_T^{\text{seq}}(\mathcal{F}, P_{x_{1:T}}) = \mathbb{E}_{(\mathbf{x},\mathbf{x}') \sim \rho_{P_{x_{1:T}}}} \left[\mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{T} \varepsilon_t f(\mathbf{x}_t(\varepsilon)) \right] \right].$$

对于给定的分布类 $\mathscr{D} = \{P_{x_{1:T}}\}$，定义

$$\mathfrak{R}_T^{\text{seq}}(\mathcal{F}, \mathscr{D}) = \sup_{P_{x_{1:T}} \in \mathscr{D}} \mathfrak{R}_T^{\text{seq}}(\mathcal{F}, P_{x_{1:T}}).$$

请注意，如果 $\mathscr{D}$ 是 $\mathcal{X}$ 上所有分布的集合，那么就得到 Rakhlin 等人 [2015] 提出的标准序列拉德马赫复杂度概念。引入这一虽说较为技术性的复杂度概念，原因在于以下结果：

定理 21 (Rakhlin 等人的定理 3 和引理 20 [2011])。令 $\mathcal{F}: \mathcal{X} \to [-1, 1]$ 为一个函数类，且 $\ell: [-1, 1] \times [-1, 1] \to [0, 1]$ 为一个关于第一个参数满足利普希茨条件的损失函数。假设我们处于第 *4* 节所述的在线学习设定中，且对手只能根据 $x_{1:T}$ 从 $\mathscr{D}$中的某个分布进行选择。那么存在一个算法，使得

$$\mathbb{E}[\text{Reg}_T] \lesssim \mathfrak{R}_T^{\text{seq}}(\mathcal{F}, \mathscr{D}).$$

回到组合复杂度概念，Rakhlin 等人 [2015] 提出了以下尺度敏感 VC 维数的序列类比：

定义 22. 设 $\mathcal{F}: \mathcal{X} \to \mathbb{R}$ 是一个函数类。对于任意 $\alpha > 0$ 以及深度为 m的完全二叉树 $\mathbf{x}$，如果对所有

m元符号 $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$，都存在某个 $f_\varepsilon \in \mathcal{F}$，使得对所有 $1 \leq i \leq m$，均有

$$\varepsilon_i \left(f_\varepsilon(\mathbf{x}_i(\varepsilon)) - \mathbf{s}_i(\varepsilon) \right) \geq \frac{\alpha}{2}.$$

我们用 fat $(\mathcal{F}, \alpha)$ 表示序贯脂肪粉碎维数，即满足存在一棵深度为 m 的树 $\mathbf{x}$ 能够粉碎 $\mathcal{F}$ 的最大 m。

注意，一般而言，fat $(\mathcal{F}, \alpha) \gg$ vc $(\mathcal{F}, \alpha)$，且两者之差可能是无穷大，例如单位区间上的阈值函数就是如此。不过，对于有限定义域 $\mathcal{X}$，则可以得到反向界。我们将用到以下结果：

引理 23 (Block 等人 [2022]中的引理 21)。设 $\mathcal{F}: \mathcal{X} \to [-1, 1]$ 是一个函数类，并令 $\operatorname{fat}(\mathcal{F}, \alpha)$ 表示尺度为 α时的序贯脂肪粉碎维数。那么存在普适常数 C，c，使得对任意 $\beta > 0$，以下不等式成立：

$$\operatorname{fat}(\mathcal{F}, \alpha) \leq C \cdot \operatorname{vc}(\mathcal{F}, c\beta\alpha) \cdot \log^{1+\beta}\left(\frac{C|\mathcal{X}|}{\operatorname{vc}(\mathcal{F}, c\alpha)\,\alpha} \right).$$

为了将这些复杂度概念与当前问题联系起来，我们将用到以下结果：

命题 24 (Block 等人 [2021]中的推论 18)。设 $\mathcal{F}: \mathcal{X} \to [-1, 1]$ 是一个函数类。对于任意分布 $P_{x_{1:T}}$，有如下界：

$$\mathfrak{R}_T^{\text{seq}}(\mathcal{F}, P_{x_{1:T}}) \lesssim \inf_{\gamma > 0} \gamma T + \sqrt{T} \int_\gamma^1 \sqrt{\operatorname{fat}(\mathcal{F}, \delta)}\, d\delta.$$

结合命题24 和引理 23 可得 Block 等人 [2022]的以下结果：

引理 25. 假设我们处于命题 *24* 所述情形，并且 $|\mathcal{X}| < \infty$。那么对任意 $P_{x_{1:T}}$，以下结论成立：

$$\mathfrak{R}_T^{\text{seq}}(\mathcal{F}, P_{x_{1:T}}) \lesssim \inf_{\beta, \gamma > 0} \gamma T + \sqrt{T \cdot \log^{1+\beta}(|\mathcal{X}|)} \int_\gamma^1 \sqrt{\operatorname{vc}(\mathcal{F}, c\beta\delta) \cdot \log^{1+\beta}\left(\frac{1}{\operatorname{vc}(\mathcal{F}, c\delta)\,\delta} \right)}\, d\delta.$$

D 第3节的证明

D.1 上界

本节将证明主要的上界，即定理 3。为确保完整性，我们先陈述并证明拒绝采样的标准保证。

引理 26. 设 μ、ν 是 $(\mathcal{X}, \mathscr{F})$ 上的两个测度，并假设 $X \sim \mu$。设 μ、ν 满足以下条件：存在某个 $M < \infty$, $\left\| \frac{d\nu}{d\mu} \right\|_\infty \leq M$。设 ξ 是一个二值随机变量

，且在给定 X 的条件下，$\xi = 1$ 的概率为 $\frac{1}{M}\frac{d\nu}{d\mu}(X)$。那么 $\mathbb{P}(\xi = 1) = \frac{1}{M}$，并且对于任意 $A \in \mathscr{F}$，均有

$$\mathbb{P}(X \in A | \xi = 1) = \nu(A).$$

因此，如果从 μ 中独立抽取 $X_1, \ldots, X_n$，并按照上述方式构造 $\xi_1, \ldots, \xi_n$，那么至少有一个 ξ_j 等于 1 的概率至少为$1 - e^{-\frac{n}{M}}$。

证明。为证明第一个结论，根据条件期望的塔式性质，我们有

$$\mathbb{P}(\xi = 1) = \mathbb{E}\left[\frac{1}{M}\frac{d\nu}{d\mu}(X)\right] = \frac{1}{M}$$

根据拉东-尼科迪姆导数的定义。为证明第二个结论，我们有

$$\begin{aligned}
\mathbb{P}(X \in A | \xi = 1) &= \frac{\mathbb{P}(X \in A \text{ and } \xi = 1)}{\mathbb{P}(\xi = 1)} \\
&= M \cdot \mathbb{P}(X \in A \text{ and } \xi = 1) \\
&= M \cdot \mathbb{E}\left[\mathbb{I}[X \in A]\frac{1}{M}\frac{d\nu}{d\mu}(X)\right] \\
&= \nu(A)
\end{aligned}$$

如所需。最后，最后一个结论成立是因为

$$\begin{aligned}
\mathbb{P}(\xi_j = 1 \text{ for some } j) &= 1 - \mathbb{P}(\xi_j = 0 \text{ for all } j) \\
&= 1 - \prod_{j=1}^{n}(1 - \mathbb{P}(\xi_j = 1)) \\
&= 1 - \left(1 - \frac{1}{M}\right)^n \\
&\geq 1 - e^{-\frac{n}{M}}
\end{aligned}$$

如所需。 ■

遗憾的是，引理26 要求似然比具有一致上界，而这恰恰是我们希望避免的。为继续推进，我们证明一个关键的测度变换引理：

引理 27. 设 μ，ν 是某个集合 $\mathcal{X}$ 上的概率测度，并设 $Y \sim \nu$。对于任意满足定义 *1* 中条件的 f $M \geq 2$，下列不等式成立：

$$\mathbb{P}\left(\frac{d\nu}{d\mu}(Y) > M\right) \leq \frac{2 \cdot D_f(\nu||\mu)}{f'(M/2)}.$$

引理的证明 *27*。 注意，引理 27 并非由马尔可夫不等式直接推出，因为我们关注的是在 ν 下似然比较大时的概率，而 f-散度定义为在 μ下的期望。相反，我们注意到，对于任意 $s > M > 1$，由凸性可得

$$\frac{f(s)}{s} \geq \frac{f(M)}{M}.$$

事实上，令

$$\widetilde{f}(t)=\begin{cases}0 & t\le 1\\ f(t) & t\ge 1\end{cases},$$

~ ~根据我们对

f 的假设可知，f 是凸的。此外，对于 $s>M>1$，由 f 的凸性可得

$$\frac{f(s)}{s}=\frac{\widetilde{f}(s)-\widetilde{f}(0)}{s}\ge\frac{\widetilde{f}(M)-\widetilde{f}(0)}{M}=\frac{f(M)}{M}.$$

这意味着 因此，$s\le \frac{M}{f(M)}\cdot f(s)$。令 $s=\frac{d\nu}{d\mu}(X)$，并取期望可得 t

$$\begin{aligned}\nu\left(\frac{d\nu}{d\mu}(Y)>M\right)&=\mathbb{E}_\mu\left[\frac{d\nu}{d\mu}(X)\cdot\mathbb{I}\left[\frac{d\nu}{d\mu}(X)>M\right]\right]\\&\le\frac{M}{f(M)}\cdot\mathbb{E}\left[f\left(\frac{d\nu}{d\mu}(X)\right)\cdot\mathbb{I}\left[\frac{d\nu}{d\mu}(X)>M\right]\right]\\&\le\frac{M\cdot D_f(\nu||\mu)}{f(M)},\end{aligned}$$

其中，最后一个不等式由假设 $f\ge 0$ 得出。

现在注意到

$$f(M)=\int_1^M f'(t)dt\ge\int_{M/2}^M f'(t)dt\ge\frac{M}{2}\cdot f'(M/2),$$

其中，第一个不等式成立是因为对于 $t\ge 1$，$M\ge 2$ 和 $f'(t)\ge 0$；第二个不等式成立是因为 f' 单调不减。整理后可得

$$\frac{M}{f(M)}\le\frac{2}{f'(M/2)}.$$

由于 $f'(t)$ 是非递减的，因此 $f'^{1}{}_{(M)}\le f'^{2}{}_{(M/2)}$ 也成立。综上即可完成证明。■

注 28. 请注意，$M\ge 2$ 的要求是任意的。事实上，对于任意 $c>1$，相同的证明表明，对于 $M\ge c, \mathbb{P}\left(\frac{d\nu}{d\mu}(Y)>M\right)\le \frac{D_f(\nu||\mu)}{c-1\cdot f'()((1-1/c)M)}$。

现在我们可以证明主要上界了

定理 *3* 的证明。回顾一下，对于任意 $M\ge 1$，我们定义 ν_M 为 $\mathcal{X}$ 上的一个测度，使得

$$\frac{d\nu_M}{d\mu}(X)=\frac{1}{\mathbb{E}\left[\frac{d\nu}{d\mu}(X)\cdot\mathbb{I}\left[\frac{d\nu}{d\mu}(X)\le M\right]\right]}\cdot\frac{d\nu}{d\mu}(X)\cdot\mathbb{I}\left[\frac{d\nu}{d\mu}(X)\le M\right].$$

假设 $Y\sim\nu$，由构造及引理27可知，

$$\left|\left|\frac{d\nu_M}{d\mu}\right|\right|_\infty\le\frac{M}{\mathbb{P}\left(\frac{d\nu}{d\mu}(Y)\le M\right)}\le\frac{M}{1-\frac{2D_f(\nu||\mu)}{f'(M/2)}}=M'.$$

因此，根据引理 26，如果我们对 ν_M使用来自 μ的样本进行拒绝采样，共采样 $n \geq M' \log\left(\frac{1}{\delta}\right)$次，那么以至少 $1-\delta$ 的概率，我们会得到至少一个接受的样本，且该接受样本服从 ν_M。将全变差表示为密度之间 L^1距离的一半，经过显式计算可知，若 E 可测，且 ν^E 表示在事件 ν条件下的测度，则 E，全变差 $\left(\nu,\ \nu^E\right) = \nu(E^c)$。结合这一事实与引理 27 并化简，可得

$$\mathrm{TV}\left(\nu_M, \nu\right) = \mathbb{P}\left(\frac{d\nu}{d\mu}(Y) > M\right) \leq \frac{2 \cdot D_f\left(\nu||\mu\right)}{f'(M/2)}.$$

令

$$M = 2 \cdot (f')^{-1}\left(\frac{4 \cdot D_f\left(\nu||\mu\right)}{\varepsilon}\right)$$

则有 $\mathrm{TV}(\nu_M, \nu) \leq \frac{\varepsilon}{2}$。注意到

$$M' = \frac{M}{1-\varepsilon},$$

证明至此完成。 ■

我们在此指出，上述证明实际上证明了一个稍强的结论：也就是说，对于任意 $\delta > 0$，都存在一个“成功”事件 E，使得 $\mu^{\otimes n}(E) \geq 1-\delta$，并且在事件 E 条件下，X_{j^*} 的分布律与目标测度 ν 的全变差距离至多为 ε，只要 δ、ε、n 满足

$$n \geq \frac{2}{1-\varepsilon} \log\left(\frac{1}{\delta}\right) (f')^{-1}\left(\frac{4 \cdot D_f\left(\nu||\mu\right)}{\varepsilon}\right).$$

事实上，在我们的应用中，有用的正是上界的这种表述。

D.2 下界

在本节中，我们将证明主要下界，即定理 5，以及其替代版本，即定理 6。首先注意，任意选择规则 j^* 相对于 μ，其拉东-尼科迪姆导数都有一个上界：

引理 29. 假设 $X_1, \ldots, X_n \sim \mu$ 相互独立，并令 j^* 为满足 $P_{X_{j^*}} = \widetilde{\nu}$ 的选择规则。那么有

$$\left|\left|\frac{d\widetilde{\nu}}{d\mu}\right|\right|_{L^\infty(\mu)} \leq n.$$

证明. 令 $A \in \mathscr{F}$ 为可测集，并令 $\widetilde{\nu}$ 表示 X_{j^*} 的分布律。于是我们观察到

$$\begin{aligned}\widetilde{\nu}(A) &= \sum_{j=1}^{n} \mathbb{P}\left(X_j \in A \text{ and } j^* = j\right) \\ &= \sum_{j=1}^{n} \mu(A) \cdot \mathbb{P}\left(j^* = j | X_j \in A\right) \\ &= \mu(A) \cdot \sum_{j=1}^{n} \mathbb{P}\left(j^* = j | X_j \in A\right) \\ &\leq n \cdot \mu(A).\end{aligned}$$

上述计算对所有 $A \in \mathscr{F}$ 均成立，因此结论成立。 ■

引理 29 得证后，我们看到，只需将注意力转向相对于基准测度具有有界似然比的那些分布。为证明下界，我们将分析分为两种情况。较简单的情况是线性 f，即那些 f 具有有界f'的情形。我们有如下界：

引理 30. 假设 f 是定义*1* 中所述的凸函数，且满足 g

$$\lim_{t \to \infty} f'(t) = C < \infty.$$

对于任意 $0 < \Delta \leq C + f(0)$，存在某个仅依赖于 C、Δ 和 f 的 $\varepsilon > 0$，使得存在满足 $D_f(\nu||\mu) = \Delta$ 的分布，并且

$$\inf_{\widetilde{\nu} \text{ such that } ||\frac{d\widetilde{\nu}}{d\mu}|| < \infty} \mathrm{TV}\left(\widetilde{\nu}, \nu\right) > \varepsilon.$$

证明。固定某个无原子的 ν，并对某个待确定的 $\varepsilon < 1$，固定 $A \in \mathscr{F}$，使得 $\nu(A) = \varepsilon$。定义 μ，使得

$$\frac{d\mu}{d\nu}(X) = \mathbb{I}\left[Z \notin A\right] \frac{1}{1-\varepsilon}.$$

注意，这定义了一个有效的似然比，并且根据定义，

$$D_f\left(\nu||\mu\right) = f(1-\varepsilon) + \varepsilon f'(\infty) = f(1-\varepsilon) + \varepsilon C.$$

由于 f 关于 $\varepsilon \downarrow 0$ 连续，且 $f(1) = 0$，因此当 ε 从 0 变化到 1 时，$D_f(\nu||\mu)$ 会在 $C + f(0)$ 和 0 之间连续变化。因此，根据介值定理，存在某个 ε，使得 $D_f(\nu||\mu) = \Delta$。对于这个 ε，我们注意到，对于任意相对于 μ绝对连续的 $\widetilde{\nu}$，都有 $\widetilde{\nu}(A) = 0$。因此，

$$\mathrm{TV}\left(\widetilde{\nu}, \nu\right) \geq |\widetilde{\nu}(A) - \nu(A)| = \varepsilon > 0$$

结论由此得证。 ■

现在我们可以陈述并证明命题4的正式版本：

定理 31。 假设 f 是定义 *1* 中所述的凸函数，且满足

$$\lim_{t\to\infty} f'(t) = C < \infty.$$

对于任意 $0 < \Delta, \leq C + f(0)$，存在某个 $\varepsilon > 0$，它仅取决于 C, Δ 和 f，使得存在分布 μ，ν，满足 $D_f(\nu||\mu) = \Delta$，并具有以下性质：若 $X_1, \ldots, X_n \sim \mu$ 相互独立，则

$$\inf_{j^*} \mathrm{TV}\left(P_{X_{j^*}}, \nu\right) \geq \varepsilon$$

其中，下确界取遍所有选择规则 j^*。

证明。根据引理29 和30，有

$$\inf_{j^*} \mathrm{TV}\left(P_{X_{j^*}}, \nu\right) \geq \inf_{\widetilde{\nu} \text{ such that } \left|\left|\frac{d\widetilde{\nu}}{d\nu}\right|\right| < \infty} \mathrm{TV}\left(\widetilde{\nu}, \nu\right) \geq \varepsilon.$$

结论得证。 ■

注意，定理31 意味着，若 $D_f(\cdot||\cdot)$这一相似性概念过于粗糙，那么使用 μ 从 ν 进行近似采样可能是不可能的。

现在我们转向更复杂的情形，即超线性 f。我们首先证明，$\mathcal{E}_\gamma$散度（定义见例4）为选择规则的分布律与目标测度之间的全变差提供了一个下界：

引理 32. 设 μ, ν 为测度，且对于某个 $\gamma \geq 1$，令 $\mathcal{E}_\gamma$ 表示示例 *4*中给出的散度。则有

$$\inf_{\left|\left|\frac{d\widetilde{\nu}}{d\mu}\right|\right|_{L^\infty(\mu)} \leq \gamma} \mathrm{TV}\left(\widetilde{\nu}, \nu\right) \geq \mathcal{E}_\gamma(\nu||\mu).$$

证明。注意，对于任意满足 $\left|\left|\frac{d\widetilde{\nu}}{d\nu}\right|\right|_\infty \leq \gamma$ 的 $\widetilde{\nu}$，均有

$$\begin{aligned}
\mathrm{TV}\left(\widetilde{\nu}, \nu\right) &= \sup_{A \in \mathscr{F}} |\nu(A) - \widetilde{\nu}(A)| \\
&= \sup_{A \in \mathscr{F}} \left| \mathbb{E}\left[\mathbb{I}[X \in A] \left(\frac{d\nu}{d\mu}(X) - \frac{d\widetilde{\nu}}{d\mu}(X) \right) \right] \right| \\
&\geq \mathbb{E}\left[\mathbb{I}\left[\frac{d\nu}{d\mu}(X) > \gamma \right] \left(\frac{d\nu}{d\mu}(X) - \gamma \right) \right] \\
&= \mathbb{E}\left[\left(\frac{d\nu}{d\mu}(X) - \gamma \right)_+ \right] \\
&= \mathcal{E}_\gamma\left(\nu||\mu\right).
\end{aligned}$$

■

现在我们可以证明定理 5：

定理 5 的证明。 注意到，结合引理 29 和 32，只需给出两个测度 μ， ν 使得 $\mathcal{E}_n(\nu||\mu) > \varepsilon$ 且 $D_f(\nu||\mu)$ 有界。令 μ = 伯努利分布(q) 和 ν =伯努利分布(p)，对于

$$q = \frac{\varepsilon}{n}, \qquad\qquad p = 2\varepsilon.$$

我们注意到

$$\frac{1-p}{1-q} = \frac{1-2\varepsilon}{1-\frac{\varepsilon}{n}} \in \left[\frac{1}{2}, 1\right]$$

根据假设 $\varepsilon \leq \frac{1}{4}$。因此，对于 $n \geq 1$，我们有

$$\mathcal{E}_n(\nu||\mu) = q\left(\frac{p}{q} - n\right) = \varepsilon.$$

类似地，我们计算得

$$D_f(\nu||\mu) = q \cdot f\left(\frac{p}{q}\right) + (1-q) f\left(\frac{1-p}{1-q}\right) \leq \frac{\varepsilon}{n} \cdot f(2n) + f\left(\frac{1}{2}\right),$$

其中，不等式由 f 的凸性及上述计算可得。现在我们注意到，根据 f 的凸性，有

$$0 = f(1) \geq f(x) + (1-x) f'(x)$$

意味着对所有 $x > 1$， $f\left(\frac{x}{x-1}\right) \leq f'(x)$。因此，

$$D_f(\nu||\mu) \leq 2\varepsilon f'(2n) + f\left(\frac{1}{2}\right).$$

注意，如果

$$n < \frac{1}{2}(f')^{-1}\left(\frac{\delta}{2\varepsilon}\right),$$

则 $D_f(\nu||\mu) \leq \delta$，因此我们如所愿地构造出了一对 (ν, μ)。结论随之成立。 ■

现在我们转向定理 6 的证明。正如正文中所述，该方法与构造出具有较大 μ、 ν 以及 $\mathcal{E}_n(\nu||\mu)$ 的分布这一阶段之前的方法类似。我们有如下结果：

引理 33。 假设 f 是定义 1 中的一个凸函数，满足 $f(0) < \infty$，并且对于某个 $\zeta > 0$，函数

$$t \mapsto \frac{t f''(t)}{(f'(t))^{1+\zeta}}$$

对于足够大的 t，该函数单调递减，并在 $t \uparrow \infty$ 时趋于 0。于是，对于足够大的 n，存在分布 μ、 ν，使得 $D_f(\nu||\mu) < \infty$，并且

$$\mathcal{E}_n(\nu||\mu) \geq \frac{1}{8} \cdot \left(\frac{\zeta D_f(\nu||\mu)}{f'(n)}\right)^{1+\zeta}.$$

证明。我们先对这个结果给出一个粗略的直观解释。理想情况下，我们希望构造一对分布，使得 $D_f(\nu\|\mu) < \infty$，并且对所有 $\gamma \geq \gamma_0$ 都有

$$\mathcal{E}_\gamma(\nu\|\mu) = \frac{C}{f'(\gamma)}.$$

注意，Polyanskiy [2010,(2.144)] 表明 $\frac{d}{d\gamma}\mathcal{E}_\gamma(\nu\|\mu) = -\mathbb{P}\left(\frac{d\nu}{d\mu}(X) \geq \gamma\right)$，其中 $X \sim \mu$。因此，我们看到目标是构造一对分布，使得

$$\mathbb{P}\left(\frac{d\nu}{d\mu}(X) \geq \gamma\right) = C \cdot \frac{f''(\gamma)}{(f'(\gamma))^2}.$$

不过，通过简单的变量替换，我们可以看到右侧的表达式在 ∞ 处可积，但它不一定是单调的（因此还需要对 f 作额外假设）。然而，即使它是单调的，也可以看出 $D_f(\nu\|\mu)$ 并非有限。因此，我们在下面对这一构造稍作调整。

固定 δ、β、$\zeta > 0$，使其满足 $\delta < n$，并令

$$F(t) = \begin{cases} 0 & t < 0 \\ 1 - \frac{\beta f''\left((f')^{-1}(\delta)\right)}{\delta^{2+\zeta}} & 0 \leq t < (f')^{-1}(\delta) \\ 1 - \frac{\beta f''(t)}{(f'(t))^{2+\zeta}} & t \geq (f')^{-1}(\delta) \end{cases}.$$

我们声称，对于适当选取的 β、δ、$\zeta > 0$，F 是一个有效的累积分布函数。首先注意到，按构造，F 是右连续的。同样显然的是，当 $t \downarrow 0$ 时，$F(t) \downarrow 0$。如果 δ 足够大，使得

$$t \mapsto \frac{tf''(t)}{f'(t)^{2+\zeta}}$$

对所有 $t > \delta$ 都递减（根据陈述中的假设，这样的 δ 总是存在），由此可知 $F(t)$ 是非递减的。最后，要证明当 $t \uparrow \infty$ 时 $F(t) \uparrow 1$，注意到

$$\begin{aligned} \int_{(f')^{-1}(\delta)}^{\infty} \frac{\beta f''(t)}{f'(t)^{2+\zeta}} dt &= \lim_{N \uparrow \infty} \left(\frac{\beta}{(1+\zeta)\delta^{1+\zeta}} - \frac{\beta}{(1+\zeta)(f')^{-1}(N)^{1+\zeta}} \right) \\ &= \frac{\beta}{(1+\zeta)\delta^{1+\zeta}} \end{aligned}$$

根据假设，$f'(t)$ 当 $t \uparrow \infty$ 时 $\uparrow \infty$。特别地，有

$$\lim_{t \uparrow \infty} \frac{\beta f''(t)}{f'(t)^{2+\zeta}} = 0$$

因此，$F(t)$ 是一个累积分布函数。进一步注意，如果非负实数轴上的随机变量 Z 服从 F，那么根据 Fubini 定理，

$$\begin{aligned} \mathbb{E}[Z] &= \int_{(f')^{-1}(\delta)}^{\infty} \beta \frac{f''(t)}{(f'(t))^{2+\zeta}} dt \\ &= \frac{\beta}{(1+\zeta)\delta^{1+\zeta}}. \end{aligned}$$

因此，如果 $\beta = (1+\zeta)\delta^{1+\zeta}$，那么 $\mathbb{E}[Z] = 1$。于是，采用这一选择的 β，令 μ 在某个集合 $\mathcal{X}$上为无原子的，并令 $\frac{d\nu}{d\mu}(X)$ 服从 F，其中 $X \sim \mu$。我们首先利用 Fubini 定理计算 f-散度在 ν和 μ 之间的值：

$$\begin{aligned}
D_f(\nu||\mu) &= \mathbb{E}\left[f\left(\frac{d\nu}{d\mu}(X)\right)\right] \\
&= \mathbb{E}[f(Z)] \\
&= f(0)\mathbb{P}(Z=0) + \int_{(f')^{-1}(\delta)}^{\infty} f'(t)\mathbb{P}(Z>t)\,dt \\
&= f(0)\left(1 - \frac{\beta f''((f')^{-1}(\delta))}{\delta^{2+\zeta}}\right) + \int_{(f')^{-1}(\delta)}^{\infty} \frac{\beta f''(t)}{f'(t)^{1+\zeta}} dt \\
&= f(0)\left(1 - \frac{\beta f''((f')^{-1}(\delta))}{\delta^{2+\zeta}}\right) + \frac{\beta}{\zeta\delta^{\zeta}} \\
&\le f(0) + \frac{1+\zeta}{\zeta}\cdot\delta,
\end{aligned}$$

这里用到了凸函数的二阶导数非负这一事实，以及上文对 $\beta = (1+\zeta)\delta^{1+\zeta}$ 的计算。若取 δ、ζ 使得 $f(0) \le \frac{1+\zeta}{\zeta}\cdot\delta$，则有

$$D_f(\nu||\mu) \le 2\frac{1+\zeta}{\zeta}\delta.$$

再次根据 Fubini 定理，并利用 $n > \delta$ 这一事实，可得

$$\begin{aligned}
\mathbb{E}\left[\frac{d\nu}{d\mu}\mathbb{I}\left[\frac{d\nu}{d\mu} > n\right]\right] &= \int_n^{\infty} \frac{\beta f''(t)}{(f'(t))^{2+\zeta}} dt \\
&= \frac{\beta}{(1+\zeta)f'(n)^{1+\zeta}}.
\end{aligned}$$

最后，注意到

$$n\mathbb{P}\left(\frac{d\nu}{d\mu} > n\right) = \frac{n\beta f''(n)}{f'(n)^{2+\zeta}}.$$

综合以上各点，我们得到

$$\begin{aligned}
\mathcal{E}_n(\nu||\mu) &\ge \frac{\beta}{(1+\zeta)f'(n)^{1+\zeta}} - \frac{n\beta f''(n)}{f'(n)^{2+\zeta}} \\
&= \frac{\delta^{1+\zeta}}{f'(n)^{1+\zeta}} - \frac{n(1+\zeta)\delta^{1+\zeta}f''(n)}{f'(n)^{2+\zeta}} \\
&= \left(\frac{\delta}{f'(n)}\right)^{1+\zeta}\cdot\left(1 - \frac{n(1+\zeta)f''(n)}{f'(n)}\right) \\
&\ge \left(\frac{\zeta}{2(1+\zeta)}\cdot\frac{D_f(\nu||\mu)}{f'(n)}\right)^{1+\zeta}\cdot\left(1 - \frac{n(1+\zeta)f''(n)}{f'(n)}\right).
\end{aligned}$$

根据陈述中的假设，对于足够大的 n，我们有

$$\frac{nf''(n)}{f'(n)^{1+\zeta}} < \frac{1}{4}$$

因此该结果成立。 ■

我们注意到，要求 f（0） $< \infty$ 相对较弱，但对于某些重要的 f-散度并不成立。然而，通过在上述证明中将原子置于 1 2 而不是 0 处，可以轻松移除此要求；为简洁起见，此处不再展开，因为这样会导致证明更加复杂，而清晰度的提升却很有限。现在我们陈述并证明定理 6的正式版本：

定理 34。 设 f 是定义 *1* 中所述的凸函数，且 f（0） $< \infty$ ，并令 $\zeta > 0$ 为任意值。假设存在某个 $t_0 > 0$ 使得函数

$$t \mapsto \frac{tf''(t)}{(f'(t))^{1+\zeta}} \tag{7}$$

对于所有 $t > t_0$ 均为非递增函数，并且

$$\lim_{t\uparrow\infty} \frac{tf''(t)}{(f'(t))^{1+\zeta}} = 0.$$

那么存在分布 μ、 ν，满足 $D_f(\nu||\mu) < \infty$，使得如果 X_1、.....、 $X_n \sim \mu$ 相互独立，则

$$\inf_{j^*} \mathrm{TV}\left(P_{X_{j^*}}, \nu\right) \geq \frac{1}{8}\left(\frac{\zeta D_f\left(\nu||\mu\right)}{f'(n)}\right)^{1+\zeta}$$

其中下确界取遍所有选择规则。

证明。 该结果由结合引理 29 、32 和 33 得出。 ■

我们注意到，对于几乎所有常见的超线性 f-散度，例如 KL 散度和 Rényi 散度，在函数（7）上定义的假设均成立。

第 E 节 第4 节的证明

E.1 命题 8 的证明

令 μ 表示单位区间上的均匀测度，并对每个 t，令 $p_t = (1-\delta)\mu + \delta q_t$，其中某个 q_t 将在后文定义。假设 $f'(\infty) = C < \infty$。注意，无论 q_t 如何，均有

$$D_f\left(\nu||\mu\right) \leq \delta f'(\infty) = \delta C$$

因此，如果对手从 p_t 中采样 x_t，且 $\delta \leq \sigma/C$，那么该对手是 (f,σ)-平滑的。现在，对 $1 \leq t < \infty$ 定义 $\bar{x}_t$ 如下。令 $\bar{x}_1 = \frac{1}{2}$，并令 $\varepsilon_1, \varepsilon_2, \ldots$ 为相互独立的 Rademacher 随机变量。令

$$\bar{x}_t = \frac{1}{2} + \sum_{s=1}^{t-1} \varepsilon_s 2^{-s-1}$$

并注意到 $\bar{x}_t \in [0,1]$ 对所有 t 几乎处处成立。此外，注意到 $\bar{x}_t \to \bar{x}_\infty$ 几乎处处成立，并定义 $\theta^* = \bar{x}_\infty$。令 $y_t = \mathbb{I}[x_t \geq \theta^*]$ 对所有 t 成立，并注意到该对手是可实现的，即存在某个 $f \in \mathcal{F}$ 能够达到零遗憾。令 q_t 表示位于 $\bar{x}_t$ 处的一个原子，并假设该对手进行 $x_t \sim p_t$。如上所述，该对手是 f-平滑的。注意，当$x_t = \bar{x}_t$时，有 $y_t = \frac{1-2\varepsilon_t}{2}$成立。由 ε_t的独立性可知，对于任意 T，有

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}[y_t \neq \widehat{y}_t]\right] &\geq \sum_{t=1}^{T} \mathbb{P}\left(x_t = \bar{x}_t\right) \mathbb{P}\left(\widehat{y}_t \neq y_t | x_t = \bar{x}_t\right) \\
&= \sum_{t=1}^{T} \delta \cdot \frac{1}{2} \\
&= \frac{\delta T}{2}.
\end{aligned}$$

结论即得。

E.2 引理9的证明

我们按照 Haghtalab 等人 [2022b], Block 等人 [2022], 的方法进行，但使用我们的定理 3 而不是标准的拒绝采样界。我们首先对 $Z_{t,\,j}$ 进行采样，并对所有 $1 \leq t \leq T$和 $1 \leq j \leq n$分别独立采样。将定理 3 应用于 x_t 在给定历史条件下的分布，然后使用定义 7 对 $D_f\left(P_{x_t} || \mu\right)$ 进行界定，即可完成证明。

E.3 f-平滑在线学习的极小极大遗憾

本节将定理10 推广到任意函数类。我们的证明方法沿用了Block 等人 [2022] 的思路，不同之处在于使用了引理9中的新拒绝采样耦合。我们得到以下结果：

定理 35. 设 $\mathcal{F}$: $\mathcal{X} \to [-1,1]$ 是一个实值函数类，并令 vc $(\mathcal{F}, \alpha)$ 表示其尺度敏感 VC 维数。假设 ℓ: $[-1,1] \times [-1,1] \to [0,1]$ 是一个关于第一个参数满足利普希茨条件的损失函数。此外，设 f 是定义 1 中的凸函数，且 $f'(\infty) = \infty$。如果对手在第4节所述意义下是 (f,σ)-平滑的，那么存在普适常数 $c, C > 0$，使得存在一个算法，其 $\mathbb{E}\left[\mathrm{Reg}_T\right]$ 的上界为以下表达式：

$$C \cdot \inf_{\beta,\gamma,\varepsilon>0} (\varepsilon+\gamma)T + \sqrt{T \log^{1+\beta}\left(T(f')^{-1}\left(\frac{1}{\sigma\varepsilon}\right)\right)} \int_{\gamma}^{1} \sqrt{\mathrm{vc}\left(\mathcal{F}, c\beta\delta\right) \cdot \log^{1+\beta}\left(\frac{1}{\mathrm{vc}\left(\mathcal{F}, c\delta\right)\delta}\right)} d\delta.$$

证明。应用定理 21，我们看到，只需控制 $\mathfrak{R}_T^{\text{seq}}(\mathcal{F}, \mathscr{D})$，其中 $\mathscr{D}$是 (f, σ)-平滑对手类。固定某个 $\frac{1}{2} \geq \alpha, \delta > 0$，并令 Π 表示由引理 $x_1, \ldots, x_T$ 和 $\{Z_{t,j} | 1 \leq t \leq T, 1 \leq j \leq n\}$ 保证存在的耦合9，使得

$$n \geq 2 \log \left(\frac{T}{\delta} \right) (f')^{-1} \left(\frac{1}{\alpha \sigma} \right),$$

其中 $Z_{t,\ j} \sim \mu$ 相互独立，并且以至少 $1 - \delta$ 的概率，存在满足 TV $\left(P_{x_t}, P_{Z_{t,j_t^*}} \right) \leq \alpha$ 的选择规则 j_t^*。记这些选择规则存在这一事件为 $\mathcal{E}$，并注意到$\mathbb{P}(\mathcal{E}) \geq 1 - \delta$。现在固定某个 $P_{x_{1:T}}$ 并计算:

$$\begin{aligned}
\mathfrak{R}_T^{\text{seq}}(\mathcal{F}, P_{x_{1:T}}) &= \mathbb{E}_{\rho_{P_{x_{2:T}}}} \left[\sup_{g \in \mathcal{F}} \sum_{t=1}^{T} \varepsilon_t g(\mathbf{x}_t(\varepsilon)) \right] \\
&= \mathbb{E}_{\Pi, \varepsilon} \left[\sup_{g \in \mathcal{F}} \sum_{t=1}^{T} \varepsilon_t g(\mathbf{x}_t(\varepsilon)) \right] \\
&= \mathbb{E}_{\Pi, \varepsilon} \left[\mathbb{I}[\mathcal{E}] \sup_{g \in \mathcal{F}} \sum_{t=1}^{T} \varepsilon_t g(\mathbf{x}_t(\varepsilon)) \right] + \mathbb{E}_{\Pi, \varepsilon} \left[\mathbb{I}\left[\mathcal{E}^c\right] \sup_{g \in \mathcal{F}} \sum_{t=1}^{T} \varepsilon_t g(\mathbf{x}_t(\varepsilon)) \right].
\end{aligned}$$

对于第二项，注意到几乎处处，

$$\sup_{g \in \mathcal{F}} \sum_{t=1}^{T} \varepsilon_t g(\mathbf{x}_t(\varepsilon)) \leq T$$

因此

$$\mathbb{E}_{\Pi, \varepsilon} \left[\mathbb{I}\left[\mathcal{E}^c\right] \sup_{g \in \mathcal{F}} \sum_{t=1}^{T} \varepsilon_t g(\mathbf{x}_t(\varepsilon)) \right] \leq \delta T. \tag{8}$$

对于另一项，我们注意到

$$\begin{aligned}
\mathbb{E}_{\Pi, \varepsilon} \left[\mathbb{I}[\mathcal{E}] \sup_{g \in \mathcal{F}} \sum_{t=1}^{T} \varepsilon_t g(\mathbf{x}_t(\varepsilon)) \right] \leq\ & \mathbb{E}_{\Pi, \varepsilon} \left[\mathbb{I}[\mathcal{E}] \sup_{g \in \mathcal{F}} \sum_{t=1}^{T} \varepsilon_t g(\mathbf{x}_t(\varepsilon)) - \varepsilon_t g\left(Z_{t,j_t^*}\right) \right] \\
&+ \mathbb{E}_{\Pi, \varepsilon} \left[\mathbb{I}[\mathcal{E}] \sup_{g \in \mathcal{F}} \sum_{t=1}^{T} \varepsilon_t g\left(Z_{t,j_t^*}\right) \right].
\end{aligned}$$

对于第一项，我们看到

$$\begin{aligned}
\mathbb{E}_{\Pi, \varepsilon} \left[\mathbb{I}[\mathcal{E}] \sup_{g \in \mathcal{F}} \sum_{t=1}^{T} \varepsilon_t g(\mathbf{x}_t(\varepsilon)) - \varepsilon_t g\left(Z_{t,j_t^*}\right) \right] &\leq \sum_{t=1}^{T} \mathbb{E}_{\Pi, \varepsilon} \left[\sup_{g \in \mathcal{F}} \varepsilon_t g(\mathbf{x}_t(\varepsilon)) - \varepsilon_t g\left(Z_{t,j_t^*}\right) \right] \\
&\leq T \max_{1 \leq t \leq T} \text{TV} \left(P_{\mathbf{x}_t(\varepsilon)}, P_{Z_{t,j_t^*}} \right) \\
&\leq \alpha T
\end{aligned} \tag{9}$$

根据构造可知。对于第二项，我们使用詹森不等式和条件期望的塔式性质进行计算：

$$\begin{aligned}\mathbb{E}_{\Pi,\varepsilon}\left[\mathbb{I}[\mathcal{E}]\sup_{g\in\mathcal{F}}\sum_{t=1}^{T}\varepsilon_t g\left(Z_{t,j_t^*}\right)\right] &\leq \mathbb{E}_{\Pi,\varepsilon}\left[\sup_{g\in\mathcal{F}}\sum_{t=1}^{T}\varepsilon_t g\left(Z_{t,j_t^*}\right)\right] \\ &= \mathbb{E}_{Z_{t,j}}\left[\mathbb{E}_{\varepsilon}\left[\sup_{g\in\mathcal{F}}\sum_{t=1}^{T}\varepsilon_t g\left(Z_{t,j_t^*}\right)\mid\{Z_{t,j}\}\right]\right] \\ &= \mathbb{E}_{Z_{t,j}}\left[\mathbb{E}_{\varepsilon}\left[\sup_{g\in\mathcal{F}_{\{Z_{t,j}\}}}\sum_{t=1}^{T}\varepsilon_t g\left(Z_{t,j_t^*}\right)\mid\{Z_{t,j}\}\right]\right].\end{aligned}$$

注意到 $|\{Z_{t,j}\}| = Tn$，现在我们可以应用引理 25，得出最后这个表达式的上界为：

$$\inf_{\beta,\gamma>0}\gamma T + \sqrt{T\cdot\log^{1+\beta}(nT)}\int_{\gamma}^{1}\sqrt{\mathrm{vc}\left(\mathcal{F},c\beta\delta\right)\cdot\log^{1+\beta}\left(\frac{1}{\mathrm{vc}\left(\mathcal{F},c\delta\right)\delta}\right)}d\delta. \tag{10}$$

令 $\delta = \frac{1}{T}$，并结合式（8）、（9）和（10），即可完成证明。 ■

E.4 预言机高效算法

本节将转向计算上可处理的算法。具体而言，我们关注的是只需对下文定义的经验风险最小化（ERM）预言机进行多项式次调用的算法。值得注意的是，ERM 预言机是在线学习领域常见的计算访问模型 [Kalai and Vempala，2005年，Hazan and Koren，2016 年，Block 等人，2022年，Haghtalab 等人，2022a]，这既是因为它们足以支持统计设定下的学习（其中数据独立出现），也是因为许多我们感兴趣的问题都有广受欢迎的计算启发式方法，可用于实现这些预言机。我们将考虑两种算法：一种非适当算法，每轮需要两次预言机调用，其遗憾随 Rademacher复杂度增长（见式（6））；另一种适当算法，每轮需要一次预言机调用。这两种算法均由 Block 等人 [2022] 提出。我们采用类似的分析来界定它们的遗憾，只是将 Block 等人 [2022] 提出的耦合替换为我们更一般的版本，即引理 9。首先定义经验风险最小化预言机：

定义 36. 我们假设学习者可以访问经验风险最小化预言机。该预言机给定一组元组 $(x_1,y_1)\ldots,(x_m,y_m)\in\mathcal{X}\times[-1,1]$、权重列表 $w_1,\ldots,w_m\in\mathbb{R}$ 以及一个由 $[0,1]$- 值损失函数 $\ell_1,\ldots,\ell_m$ 构成的序列后，会返回某个满足以下条件的 $\widehat{g}\in\mathcal{F}$satisfying

$$\sum_{i=1}^{m} w_i\ell_i\left(\widehat{g}(x_i),y_i\right)\leq\inf_{g\in\mathcal{F}}\sum_{i=1}^{m} w_i\ell_i(f(x_i),y_i).$$

稍弱一些的假设允许存在一定程度的近似，即 ERMOracle 返回某个$\widehat{g}$，它与实际最小化器的距离不超过 ε。为简便起见，我们在此仅关注精确预言机，但所有结果都适用于更一般的情形，只需在证明中几乎不作修改，并允许加性 εT。

E.4.1 通过松弛实现的非适当算法

现在我们转向第一个预言机高效算法，该算法的灵感来自Rakhlin 等人 [2012]的松弛框架。我们基本遵循Block 等人 [2022]的阐述。首先，我们定义一个松弛：

定义 37. 对于固定的时域 T、函数类 $\mathcal{F}$、上下文空间 $\mathcal{X}$ 和测度 μ，如果任意松弛序列 $\mathbf{Rel}_T(\mathcal{F}|x_1, y_1 \ldots, x_t, y_t) : \mathcal{X}^{\times t} \times [-1,1]^{\times t} \to \mathbb{R}$ 满足：对于任意序列 x_1、...、x_T 以及任意 $1 \le t \le T$，以下两个性质成立，则称其为松弛：

$$-\inf_{g\in\mathcal{F}} \sum_{t=1}^{T} \ell(g(x_t), y_t) \le \mathbf{Rel}_T\left(\mathcal{F}|x_1, y_1 \ldots, x_T, y_T\right)$$

$$\text{上确界}_{p_t} \mathbb{E}_{x_t\sim p_t} \text{下确界}_{q_t\in\Delta([-1,1])} \text{上确界}_{y'_t\in[-1,1]} \{\mathbb{E}_{\widehat{y}_t\sim q_t}[\ell(\widehat{y}_t, y'_t)] + \mathbf{Rel}_T(\mathcal{F}|x_1, y_1 \ldots, x'_t, y'_t)\} \le \mathbf{Rel}_T(\mathcal{F}|x_1, y_1 \ldots, x_{t-1}, y_{t-1})$$

其中，第一个上确界取遍关于 μ 的（f，σ）-平滑分布，下确界取遍 $[-1,1]$ 上的分布。

松弛的关键性质，正如Rakhlin 等人 [2012, 命题 1], 所证明的，是任何满足 q_t 定义 37 中第二个不等式的策略，其遗憾都不超过$\mathbf{Rel}_T(\mathcal{F}|\emptyset)$。我们的第一个结果表明，只需稍作修改，Block 等人 [2022] 提出的松弛在 f-平滑情形下仍然是有效的松弛。

引理 38。假设对手是 (f,σ)-平滑的，且损失函数 ℓ 关于第一个参数是凸且利普希茨连续的。设 $0 < \alpha \le \frac{1}{2}$，并且假设

$$n \ge 8\log(T)(f')^{-1}\left(\frac{1}{\alpha\sigma}\right).$$

Then

$$\mathbf{Rel}_T\left(\mathcal{F}|x_1, y_1 \ldots, x_t, y_t\right) = \mathbb{E}_{\mu,\varepsilon}\left[2\sup_{g\in\mathcal{F}} \sum_{j=1}^{n} \sum_{s=t+1}^{T} \varepsilon_{s,j} g(Z_{s,j}) - \sum_{s=1}^{t} \ell(g(x_s), y_s)\right] + T\alpha + \frac{T-t}{nT}$$

是一个松弛，其中期望是相对于 $Z_{s,\ j} \sim \mu$ 以及独立的 Rademacher 随机变量而言的。

证明。我们沿用Block 等人 [2022, 命题 6], 中的证明技巧，将他们的Block 等人[2022, 引理 14] 替换为我们的引理9。我们引入同样方便的简写：

$$L_t(g) = \sum_{s=1}^{t} \ell(g(x_s), y_s)$$

对所有 $g \in \mathcal{F}$。现在注意到，定义 37中的第一个条件显然成立，因为 $\mathbf{Rel}_T(\mathcal{F}|x_1, y_1 \ldots, x_T, y_T)$实际上等于定义不等式左侧的下确界。因此，只需证明对于所有 t 以及所有取值 $x_1, y_1, \ldots, x_{t-1}, y_{t-1}$，定义 37 中的第二个不等式成立。我们固定某个 (f,σ)-平滑的 p_t，并针对这个任意的平滑分布进行论证。

由于损失函数 ℓ 关于第一个参数是凸的，我们可以用其期望替换对分布 q_t 的抽样来源，其中 $\widehat{y}$ 从该分布中抽取，即，

$$\begin{aligned}\inf_{q_t\in\Delta([-1,1])}\sup_{y_t'\in[-1,1]}&\left\{\mathbb{E}_{\widehat{y}_t\sim q_t}[\ell(\widehat{y}_t,y_t')]+\mathbf{Rel}_T\left(\mathcal{F}|x_1,y_1\ldots,x_t',y_t'\right)\right\}\\&=\inf_{\widehat{y}_t\in[-1,1]}\sup_{y_t'\in[-1,1]}\left\{\ell(\widehat{y}_t,y_t')+\mathbf{Rel}_T\left(\mathcal{F}|x_1,y_1\ldots,x_t',y_t'\right)\right\}.\end{aligned}$$

现在，按照Block 等人 [2022], 中的论证，我们可以看到，对于任意 $x_t'\in\mathcal{X}$,

$$\begin{aligned}&\inf_{\widehat{y}_t}\sup_{y_t'}\left\{\ell(\widehat{y}_t,y_t')+\mathbb{E}_{\mu,\varepsilon}\left[\sup_{g\in\mathcal{F}}2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j})-L_t(g)\right]\right\}\\&\quad=\inf_{\widehat{y}_t}\sup_{y_t'}\mathbb{E}_{\mu,\varepsilon}\left\{\sup_{g\in\mathcal{F}}2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j})-L_{t-1}(g)+\ell(\widehat{y}_t,y_t')-\ell(g(x_t'),y_t')\right\}\\&\quad\le\inf_{\widehat{y}_t}\sup_{y_t'}\mathbb{E}_{\mu,\varepsilon}\left\{\sup_{g\in\mathcal{F}}2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j})-L_{t-1}(g)+\partial\ell(\widehat{y}_t,y_t')\left(\widehat{y}_t-g(x_t')\right)\right\}\\&\quad\le\inf_{\widehat{y}_t}\max_{\gamma_t\in\{\pm1\}}\mathbb{E}_{\mu,\varepsilon}\left\{\sup_{g\in\mathcal{F}}2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j})-L_{t-1}(g)+\gamma_t(\widehat{y}_t-g(x_t'))\right\},\end{aligned}$$

其中，第一个不等式源于 ℓ 关于第一个参数是凸的，而第二个不等式源于 ℓ 关于同一参数是利普希茨的。现在我们应用极小极大定理；由于该集合不具有凸性，我们不得不对 $\{\pm1\}$ 上的分布取上确界。我们再次沿用 Block 等人 [2022], 的论证，令 d_t 表示 $\{\pm1\}$ 上的一个分布，并从中抽取 $\gamma_t\sim d_t$。我们计算：

$$\begin{aligned}&\inf_{\widehat{y}_t}\max_{\gamma_t\in\{\pm1\}}\mathbb{E}_{\mu,\varepsilon}\left\{\sup_{g\in\mathcal{F}}2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j})-L_{t-1}(g)+\gamma_t(\widehat{y}_t-g(x_t'))\right\}\\&\quad=\sup_{d_t}\inf_{\widehat{y}_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\sup_{g\in\mathcal{F}}2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j})-L_{t-1}(g)+\gamma_t(\widehat{y}_t-g(x_t'))\right\}\\&\quad\le\sup_{d_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\inf_{\widehat{y}_t}\mathbb{E}_{\gamma_t'\sim d_t}[\gamma_t'\widehat{y}_t]+\sup_{g\in\mathcal{F}}2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j})-L_{t-1}(g)-\gamma_t\cdot g(x_t')\right\}\\&\quad\le\sup_{d_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\sup_{g\in\mathcal{F}}2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j})-L_{t-1}(g)+\mathbb{E}_{\gamma_t'\sim d_t}[\gamma_t'g(x_t)]-\gamma_t\cdot g(x_t')\right\}\\&\quad\le\sup_{d_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\sup_{g\in\mathcal{F}}2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j})-L_{t-1}(g)+2\varepsilon_t\gamma_t\cdot g(x_t')\right\}\\&\quad\le\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\sup_{g\in\mathcal{F}}2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j})-L_{t-1}(g)+2\varepsilon_t\cdot g(x_t')\right\},\end{aligned}$$

其中，倒数第二个不等式源于对称化，最后一个不等式源于收缩。现在我们注意到，因为

$$n \geq 8\log(T)(f')^{-1}\left(\frac{1}{\alpha\sigma}\right),$$

我们有

$$n \geq 4\log(nT)(f')^{-1}\left(\frac{1}{\alpha\sigma}\right)$$

根据 $2\log(n) \leq n$ 对所有 $n > 0$。现在应用引理 9，并注意到根据 n 的定义，我们可以在 x_t 和 $Z_{t\text{ 之间建立耦合，}j}$ 对于 $1 \leq j \leq n$，以至少 $1-(nT)^{-2}$ 的概率，存在某个 j^*，使得 $\mathrm{TV}\left(p_t, P_{Z_{t,\,j^*}}\right) \leq \alpha$；令 $\mathcal{E}$ 表示存在这样的 j^* 的事件。现在利用 p_t 是平滑的这一事实，并对先前任意选取的 $x'_t \sim p_t$ 取期望。上述推导于是表明

$$\begin{aligned}
&\mathbb{E}_{x'_t\sim p_t}\inf_{q_t\in\Delta([-1,1])}\sup_{y'_t\in[-1,1]}\left\{\mathbb{E}_{\widehat{y}_t\sim q_t}[\ell(\widehat{y}_t,y'_t)]+\mathbf{Rel}_T\left(\mathcal{F}|x_1,y_1\ldots,x'_t,y'_t\right)\right\}\\
&\quad\leq \mathbb{E}_{x'_t\sim p_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\sup_{g\in\mathcal{F}}2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j})-L_{t-1}(g)+2\varepsilon_t\cdot g(x'_t)\right\}\\
&\quad= \mathbb{E}_{x'_t\sim p_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\mathbb{I}[\mathcal{E}]\sup_{g\in\mathcal{F}}2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j})-L_{t-1}(g)+2\varepsilon_t\cdot g(x'_t)\right\}\\
&\quad+\mathbb{E}_{x'_t\sim p_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\mathbb{I}[\mathcal{E}^c]\sup_{g\in\mathcal{F}}2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j})-L_{t-1}(g)+2\varepsilon_t\cdot g(x'_t)\right\}.
\end{aligned}$$

注意，对于上面的第二项，被积函数中的表达式至多为 $n(T-t+1)$，因此

$$\begin{aligned}
&\mathbb{E}_{x'_t\sim p_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\mathbb{I}[\mathcal{E}^c]\sup_{g\in\mathcal{F}}2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j})-L_{t-1}(g)+2\varepsilon_t\cdot g(x'_t)\right\}\\
&\qquad\leq \mathbb{P}\left(\mathcal{E}^c\right)\cdot n(T-t+1)\leq\frac{1}{nT}
\end{aligned}$$

对于第一项，我们有

$$\begin{aligned}
&\mathbb{E}_{x'_t\sim p_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\mathbb{I}[\mathcal{E}]\sup_{g\in\mathcal{F}}2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j})-L_{t-1}(g)+2\varepsilon_t\cdot g(x'_t)\right\}\\
&\quad= \mathbb{E}_{x'_t\sim p_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\mathbb{I}[\mathcal{E}]\sup_{g\in\mathcal{F}}2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j})-L_{t-1}(g)+2\varepsilon_t\cdot g(Z_{t,j^*})+2\varepsilon_t(g(x'_t)-g(Z_{t,j^*}))\right\}\\
&\quad\leq 2\,\mathrm{TV}\left(p_t,P_{Z_{t,j^*}}\right)+\mathbb{E}_{x'_t\sim p_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\sup_{g\in\mathcal{F}}2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j})-L_{t-1}(g)+2\varepsilon_t\cdot g(Z_{t,j^*})\right\}.
\end{aligned}$$

上面的第一项至多为 2α。对于第二项，我们应用詹森不等式，得到

$$\begin{aligned}
&\mathbb{E}_{x_t'\sim p_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\sup_{g\in\mathcal{F}} 2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j}) - L_{t-1}(g) + 2\varepsilon_t\cdot g(Z_{t,j^*})\right\} \\
&\quad\le \mathbb{E}_{x_t'\sim p_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\sup_{g\in\mathcal{F}} 2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j}) - L_{t-1}(g) + 2\varepsilon_t\cdot g(Z_{t,j^*}) + \sum_{j\neq j^*}2\mathbb{E}[\varepsilon_{t,j}g(Z_{t,j})]\right\} \\
&\quad\le \mathbb{E}_{x_t'\sim p_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\sup_{g\in\mathcal{F}} 2\sum_{j=1}^{n}\sum_{s=t}^{T}\varepsilon_{s,j}g(Z_{s,j}) - L_{t-1}(g)\right\}.
\end{aligned}$$

由此可见，

$$\begin{aligned}
&\mathbb{E}_{x_t'\sim p_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\sup_{g\in\mathcal{F}} 2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j}) - L_{t-1}(g) + 2\varepsilon_t\cdot g(x_t')\right\} \\
&\quad\le \mathbb{E}_{x_t'\sim p_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\sup_{g\in\mathcal{F}} 2\sum_{j=1}^{n}\sum_{s=t}^{T}\varepsilon_{s,j}g(Z_{s,j}) - L_{t-1}(g)\right\} + 2\alpha + \frac{n(T-t+1)}{n^2T^2}.
\end{aligned}$$

代入引理陈述中给出的松弛定义，证明即告完成。

■

虽然引理 38 提供了一种实现低遗憾的算法，但由于必须计算期望，目前尚不清楚该算法是否具有预言机效率。因此，正如 Block 等人 [2022],所采用的方法，我们使用 Rakhlin 等人 [2012] 提出的随机模拟思想来给出一种具有预言机效率的算法。在继续之前，我们回顾一个经典观察：由于损失函数 ℓ关于第一个参数具有凸性，ℓ 只需假设其为线性的；事实上，我们只需在每个时间步将损失替换为损失的梯度。由于凸性，采用后一种反馈的算法的遗憾上界了采用原损失函数时的遗憾。关于这一经典论证的更多细节，请参见例如 Rakhlin 等人 [2012, 第 5] 节，或 Haghtalab 等人 [2022的附录 G。2]因此，我们将重点限制在线性损失上，并得到以下结果：

定理 39. 设 ℓ 是关于第一个参数凸且满足利普希茨条件的损失函数，并令 $\mathcal{F}$：$\mathcal{X}\to[-1,1]$ 表示一个函数类。考虑这样一种算法：在每个时刻 t，对 $t+1\le s\le T$ 和 $1\le j\le n$ 进行采样 $Z_{s,j}\sim\mu$，并选择

$$\widehat{y}_t = \operatorname*{argmin}_{\widehat{y}\in[-1,1]}\sup_{y_t\in[-1,1]}\left\{\ell(\widehat{y},y_t) + \sup_{g\in\mathcal{G}}\left[12\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j}) - \sum_{s=1}^{t}\partial\ell(g(x_s),y_s)\cdot g(x_s)\right]\right\}.$$

假设

$$n\ge 8\log(T)\cdot(f')^{-1}\left(\frac{1}{\varepsilon\sigma}\right)$$

且对手经过 (f,σ)-平滑处理。那么学习者所经历的

$$\mathbb{E}\left[\mathrm{Reg}_T\right]\le 2\mathbb{E}_\mu\left[\mathfrak{R}_{nT}(\mathcal{F})\right] + \varepsilon T + 1. \tag{11}$$

此外，每轮 t 中，$\widehat{y}_t$ 可通过调用 经验风险最小化预言机 2 次来计算。

证明。我们首先指出，只需考虑线性损失 $\ell(\widehat{y}, y) = \frac{1-\widehat{y}\cdot y}{2}$。事实上，这是文献中常见的从在线凸优化到在线线性优化的标准归约。关于这一经典论证的更多细节，可参见例如 Rakhlin 等人 [2012,第 5] 节，或 Haghtalab 等人 [2022a，附录 G.2]。因此，我们将重点限于线性损失，并假设 $\partial\ell(g(x_s)\ y_s) = -y_s \cdot g(x_s) = \ell(g(x_s))$。现在我们注意到，正如 Rakhlin 等人 [2012] 或 Block 等人 [2022, 引理 26]所指出的，只需调用预言机两次即可计算 $\widehat{y}_t$。因此，只需证明对于所有 (f, σ)-光滑的 p_t，都有

$$\mathbb{E}_{x_t\sim p_t}\left[\sup_{y_t\in[-1,1]}\left\{\ell(\widehat{y}, y_t) + \sup_{g\in\mathcal{G}}\left[6\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j}) - \sum_{s=1}^{t}\partial\ell(g(x_s), y_s)\cdot g(x_s)\right]\right\}\right]$$
$$\leq \mathbf{Rel}_T\left(\mathcal{F}|x_1, y_1\ldots, x_{t-1}, y_{t-1}\right)$$

对于 $\mathbf{Rel}_T(\mathcal{F}|x_1,\ y_1\ldots,\ x_{t-1},\ y_{t-1})$ 的定义同引理 38。事实上，若此条件成立，则 Rakhlin 等人 [2012,的命题 1] 保证最终遗憾有界于 $\mathbf{Rel}_T(\mathcal{F}|\emptyset)$，这恰好就是式（11）给出的表达式。不过，上述显示式中的界是通过应用 Block 等人 [2022, 的定理 7] 和引理38得到的。结论得证。 ■

现在我们可以证明定理 11 是一个特例：

定理 *11*的证明。 注意，这是一个经典结论 [Wainwright, 2019] 即，若 $\mathcal{F}$ 是一个二值类，则

$$\mathbb{E}_\mu\left[\mathfrak{R}_T(\mathcal{F})\right] \lesssim \sqrt{\mathrm{vc}\left(\mathcal{F}\right)\cdot T}.$$

应用定理 39即可得出结论。 ■

E.4.2 通过 FTPL 实现适当算法

现在我们转向一种适当算法，即 Block 等人 [2022]提出的跟随扰动领导者（FTPL）的建议实例。鉴于证明存在技术上的困难，我们将重点限定在具有线性损失 $\ell(\widehat{y}, y) = \frac{1-\widehat{y}\cdot y}{2}$ 的二值函数类 $\mathcal{F}$ 上。回顾一下，我们用$L_t(g)$表示函数 $g\in\mathcal{F}$ 在数据 $x_1, y_1, \ldots, x_t, y_t$上的累积损失。Block 等人 [2022] 提出的算法在每一轮进行如下操作：抽取 $\gamma_{t,\ 1}, \ldots, \gamma_{t,\ m}$ 个相互独立的标准高斯随机变量，以及 $Z_{t,\ 1}, \ldots, Z_{t,\ m} \sim \mu$，并调用预言机进行评估

$$g_t \in \operatorname*{argmin}_{g\in\mathcal{F}} L_{t-1}(g) + \eta\omega_{t,m}(g) \tag{12}$$

其中

$$\omega_{t,m}(g) = \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\gamma_{t,i}g(Z_{t,i}).$$

注意，该过程是适当的，因为它不依赖于 x_t。随后，玩家执行 $g_t(x_t)$。我们将证明，该算法针对 Rényi 平滑对手可以实现零遗憾。

定理 40。 设 $\mathcal{F}$： $\mathcal{X} \to \{\pm 1\}$ 是一个二值函数类，并且 ℓ 是线性损失函数。设对于某个 $\lambda \geq 2$，对手是 $(f,\ \sigma)$ 关于阶数为 λ 的 Rényi 散度平滑的，即 $e^{(\lambda-1)D_\lambda(p_t||\mu)} \leq \frac{1}{\sigma}$。如果学习者执行非适当算法 (12)，则

$$\mathbb{E}\left[\mathrm{Reg}_T\right] = \widetilde{O}\left(\sqrt{\mathrm{vc}\left(\mathcal{F}\right)} \cdot T^{\frac{2\lambda+1}{4\lambda-1}} \cdot \sigma^{-\frac{1}{4\lambda-1}}\right).$$

在证明主要结果之前，我们需要回顾几个中间结论。与非适当算法的情形一样，我们将修改 Block 等人 [2022] 的方法，使其适用于我们的设定，并应用引理 9。我们需要的第一个结果是 Kalai and Vempala[2005]提出的经典 Be-the-Leader 引理；我们将以下述形式陈述它：

引理 41（Block 等人的引理 32 [2022]）。假设我们处于定理 40 的情形，并令 $(x'_1, y'_1)\ldots$, (x'_T, y'_T) 为若干元组，使得在给定历史的条件下，(x_t, y_t) 和 (x'_t, y'_t) 独立同分布（换言之，x'_t, y'_t 构成一个切向序列）。那么，按照 (12) 进行决策的学习者的期望遗憾上界为

$$2\eta\mathbb{E}\left[\sup_{g\in\mathcal{F}} \omega_{1,m}(g)\right] + \sum_{t=1}^{T} \mathbb{E}\left[\ell(g_t(x'_t), y'_t) - \ell(g_{t+1}(x'_t), y'_t)\right] + \sum_{t=1}^{T} \mathbb{E}\left[\ell(g_{t+1}(x'_t), y'_t) - \ell(g_{t+1}(x_t), y_t)\right].$$

第一项控制扰动的大小，第二项在 Block 等人 [2022], 的文献中称为稳定性项，最后一项则称为泛化误差。第一项很容易控制：

引理 42。 假设我们处于定理 *40*的情形。那么

$$\mathbb{E}\left[\sup_{g\in\mathcal{F}} \omega_{1,m}(g)\right] \lesssim \sqrt{\mathrm{vc}\left(\mathcal{F}\right)}.$$

证明。 我们要控制以 $\mathcal{F}$中的元素为索引的高斯过程的上确界。例如， Wainwright [2019], VanHandel [2014] 中的一个初等链式论证可立即得出该结论。 ■

对于第二项，我们需要修改 Block 等人 [2022] 的一个论证，以考虑我们较弱的假设。首先使用以下事实：

引理 43（Block 等人 [2022]中的引理 33）。假设我们处于定理 *40*的情形，并令 $\widehat{\mu}_m$ 表示 $Z_{t,1}, \ldots, Z_{t,m}$的经验分布。那么对于任意 $\alpha > 0$，有

$$\mathbb{P}\left(\sup_{x_t, y_t} ||g_t - g_{t+1}||_{L^2(\widehat{\mu}_m)} > \alpha\right) \lesssim \frac{1}{\alpha^4\eta^2} + \frac{1}{\alpha^2\eta}\mathbb{E}\left[\sup_{g\in\mathcal{F}} \omega_{t,m}(g)\right].$$

现在应用引理 43 来证明 Block 等人 [2022, 引理 34], , 允许 f-平滑对手。

引理 44. 假设我们处于定理 *40*的情形。此外，假设对于某个$\Delta > 0$，有

$$\sup_{g,g'\in\mathcal{F}} \left| ||g - g'||^2_{L^2(\mu)} - ||g - g'||^2_{L^2(\widehat{\mu}_m)} \right| \leq \Delta.$$

Then

$$\mathbb{E}\left[\ell(g_t(x_t), y_t) - \ell(g_{t+1}(x'_t), y'_t)\right] \lesssim \mathbb{E}\left[1 + \sup_{g\in\mathcal{F}} \omega_{t,m}(g)\right] \log(\eta)\sigma^{-\frac{1}{\lambda-1}} \cdot \frac{1}{\eta^{1-\frac{1}{\lambda}}} + \sigma^{-\frac{1}{\lambda-1}} \cdot \Delta^{\frac{\lambda-1}{\lambda}}.$$

证明。 我们采用 Block 等人 [2022]的技巧。由于 (x'_t, y'_t) 与 (x_t, y_t)同分布， g_t与 (x_t, y_t)独立，且期望具有线性性，因此只需将 (x_t, y_t) 替换为 (x'_t, y'_t)来证明该结果。对于任意 $0 < \beta < \alpha$，我们计算

$$
\begin{aligned}
&\mathbb{E}_{x'_t \sim p_t}\left[\ell(g_t(x'_t), y'_t) - \ell(g_{t+1}(x'_t), y'_t)\mathbb{I}\left[\beta < ||g_t - g_{t+1}||_{L^2(\widehat{\mu}_m)} \leq \alpha\right]\right] \\
&\quad \leq \mathbb{E}_{x'_t \sim p_t}\left[|g_t(x'_t) - g_{t+1}(x'_t)|\,\mathbb{I}\left[\beta < ||g_t - g_{t+1}||_{L^2(\widehat{\mu}_m)} \leq \alpha\right]\right] \\
&\quad = \mathbb{E}_{Z_t \sim \mu}\left[\frac{dp_t}{d\mu}(Z_t)\,|g_t(Z_t) - g_{t+1}(Z_t)|\,\mathbb{I}\left[\beta < ||g_t - g_{t+1}||_{L^2(\widehat{\mu}_m)} \leq \alpha\right]\right] \\
&\quad \leq e^{D_\lambda(p_t||\mu)} \cdot \left(\mathbb{E}_{Z_t \sim \mu}\left[|g_t(Z_t) - g_{t+1}(Z_t)|^{\frac{\lambda}{\lambda-1}}\,\mathbb{I}\left[\beta < ||g_t - g_{t+1}||_{L^2(\widehat{\mu}_m)} \leq \alpha\right]\right]\right)^{\frac{\lambda-1}{\lambda}} \\
&\quad = \sigma^{-\frac{1}{\lambda-1}} \cdot \left(\mathbb{E}_{Z_t \sim \mu}\left[|g_t(Z_t) - g_{t+1}(Z_t)|^2\,\mathbb{I}\left[\beta < ||g_t - g_{t+1}||_{L^2(\widehat{\mu}_m)} \leq \alpha\right]\right]\right)^{\frac{\lambda-1}{\lambda}} \\
&\quad \leq \sigma^{-\frac{1}{\lambda-1}} \cdot \mathbb{P}\left(||g_t - g_{t+1}||_{L^2(\widehat{\mu}_m)} > \beta\right) \cdot \left(\alpha^2 + \Delta\right)^{\frac{\lambda-1}{\lambda}},
\end{aligned}
$$

其中，第一个不等式由利普希茨性得出，第二个不等式由霍尔德不等式得出，最后一个不等式由我们的假设得出；第二个等式成立是因为 g_t， g_{t+1}取值于 $\{0, 1\}$。现在我们采用 Block 等人 [2022] 提出的求和论证，并令 $S = \left\lceil \log \min\left(\sqrt{\eta}, \frac{1}{\sqrt{\Delta}}\right)\right\rceil$和 $\alpha_i = 2^{1-i}$:

$$
\begin{aligned}
&\mathbb{E}\left[\ell(g_t(x'_t), y'_t) - \ell(g_{t+1}(x'_t), y'_t)\right] \\
&\quad \leq \mathbb{E}\left[\left(\ell(g_t(x'_t), y'_t) - \ell(g_{t+1}(x'_t), y'_t)\right)\mathbb{I}\left[||g_t - g_{t+1}||_{L^2(\widehat{\mu}_m)} \leq \alpha_S\right]\right] \\
&\quad + \sum_{i=0}^{S-1} \mathbb{E}\left[\left(\ell(g_t(x'_t), y'_t) - \ell(g_{t+1}(x'_t), y'_t)\right)\mathbb{I}\left[\alpha_{i+1} < ||g_t - g_{t+1}|| \leq \alpha_i\right]\right] \\
&\quad \leq \sigma^{-\frac{1}{\lambda-1}}\left(\alpha_S^2 + \Delta\right)^{\frac{\lambda-1}{\lambda}} + \sum_{i=0}^{S-1} \sigma^{-\frac{1}{\lambda-1}} \cdot \mathbb{P}\left(||g_t - g_{t+1}||_{L^2(\widehat{\mu}_m)} > \alpha_{i+1}\right) \cdot \left(\alpha_i^2 + \Delta\right)^{\frac{\lambda-1}{\lambda}} \\
&\quad \leq \sigma^{-\frac{1}{\lambda-1}}\left(\alpha_S^2 + \Delta\right)^{\frac{\lambda-1}{\lambda}} + C\sum_{i=0}^{S-1} \sigma^{-\frac{1}{\lambda-1}} \cdot \left(\frac{1}{\alpha_{i+1}^4 \eta^2} + \frac{1}{\alpha_{i+1}^2 \eta}\mathbb{E}\left[\sup_{g \in \mathcal{F}} \omega_{t,m}(g)\right]\right) \cdot \left(\alpha_i^2 + \Delta\right)^{\frac{\lambda-1}{\lambda}} \\
&\quad \leq \sigma^{-\frac{1}{\lambda-1}}\left(\alpha_S^2 + \Delta\right)^{\frac{\lambda-1}{\lambda}} + 2C\sum_{i=0}^{S-1} \sigma^{-\frac{1}{\lambda-1}} \cdot \left(\frac{1}{\alpha_{i+1}^{2+\frac{2}{\lambda}} \eta^2} + \frac{1}{\alpha_{i+1}^{\frac{2}{\lambda}} \eta}\mathbb{E}\left[\sup_{g \in \mathcal{F}} \omega_{t,m}(g)\right]\right) \\
&\quad \leq C'\mathbb{E}\left[1 + \sup_{g \in \mathcal{F}} \omega_{t,m}(g)\right] \log(\eta)\sigma^{-\frac{1}{\lambda-1}} \cdot \frac{1}{\eta^{1-\frac{1}{\lambda}}} + C'\sigma^{-\frac{1}{\lambda-1}} \cdot \Delta^{\frac{\lambda-1}{\lambda}}.
\end{aligned}
$$

这里我们用到了以下事实： $\alpha_i^2 \geq \Delta$ 和 $\frac{1}{\alpha_i \eta} \leq \sqrt{\eta}$。结论由此得出。 ■

一种标准的经验过程理论方法可以帮助我们界定 Δ；我们将直接引用Block 等人 [2022, 引理 36]。最后，我们需要界定泛化误差。下面的引理给出了证明:

引理 45。 假设我们处于定理 *40* 和 $\eta \geq \sqrt{m}$所述的情形。进一步假设存在某个 $k \in \mathbb{N}$ 满足

$$m \geq 4k \log(T) \cdot (\varepsilon\sigma)^{-\frac{1}{\lambda-1}}.$$

Then

$$\mathbb{E}\left[\ell(g_{t+1}(x_t'), y_t') - \ell(g_{t+1}(x_t), y_t)\right] \leq 2k\varepsilon + \frac{2}{k}\mathfrak{R}_k\left(\mathcal{F}\right) + \frac{2}{T}$$

证明。 首先注意到，根据关于 m 的假设以及引理 9，以至少$1 - (mT)^{-2}$的概率，存在 k 个指标 $i_1, \ldots, i_k$，使得 $i_j \in \left\{1 + (j-1) \cdot \frac{m}{k}, \ldots, j \cdot \frac{m}{k}\right\}$并且TV $\left(P_{Z_{t,i_j}}, p_t\right) \leq \varepsilon$对所有 $1 \leq j \leq k$均成立。为此，注意到若

$$m \geq 4k \log(T) \cdot (\varepsilon\sigma)^{-\frac{1}{\lambda-1}}$$

then

$$m \geq 2k \log(mT) \cdot (\varepsilon\sigma)^{-\frac{1}{\lambda-1}}.$$

根据 $2\log(m) \leq m$ 这一事实。令 $\widetilde{\omega}_{t,m}$ 表示高斯过程 $\omega_{t,m}$，并对其作如下修改：将 Z_{t,i_j}替换为 Z'_{t,i_j}，其中 $Z_{t,i'_j} \sim p_t$；现在，令

$$\widetilde{g}_{t+1} = \operatorname*{argmin}_{g \in \mathcal{F}} L_t(g) + \widetilde{\omega}_{t,m}.$$

注意，并集界告诉我们

$$\mathrm{TV}\left(P_{g_{t+1}}, P_{\widetilde{g}_{t+1}}\right) \leq k\varepsilon.$$

根据 Block 等人 [2022, 的引理， 38], 可知，若 $\eta \geq \sqrt{m}$，则有

$$\mathbb{E}\left[\ell(\widetilde{g}_{t+1}(x_t'), y_t') - \ell(\widetilde{g}(x_t), y_t)\right] \leq \frac{2}{k}\mathfrak{R}_k\left(\mathcal{F}\right) + \frac{2m+T}{(mT)^2}.$$

综上所述，我们得到

$$\begin{aligned}
\mathbb{E}\left[\ell(g_{t+1}(x_t'), y_t') - \ell(g_{t+1}(x_t), y_t)\right] &= \mathbb{E}\left[\ell(g_{t+1}(x_t'), y_t') - \ell(\widetilde{g}_{t+1}(x_t'), y_t')\right] \\
&\quad + \mathbb{E}\left[\ell(\widetilde{g}_{t+1}(x_t'), y_t') - \ell(\widetilde{g}(x_t), y_t)\right] + \mathbb{E}\left[\ell(\widetilde{g}_{t+1}(x_t), y_t) - \ell(g_{t+1}(x_t), y_t)\right] \\
&\leq 2k\varepsilon + \frac{2}{k}\mathfrak{R}_k\left(\mathcal{F}\right) + \frac{2}{T}.
\end{aligned}$$

结论由此得证。 ■

现在我们可以结合这些引理，证明主要结果：

定理 *40* 的证明。 我们首先应用引理 41，该引理告诉我们，期望遗憾上界为

$$2\eta\mathbb{E}\left[\sup_{g \in \mathcal{F}} \omega_{1,m}(g)\right] + \sum_{t=1}^{T} \mathbb{E}\left[\ell(g_t(x_t'), y_t') - \ell(g_{t+1}(x_t'), y_t')\right] + \sum_{t=1}^{T} \mathbb{E}\left[\ell(g_{t+1}(x_t'), y_t') - \ell(g_{t+1}(x_t), y_t)\right].$$

根据引理 42，第一项为 $O\left(\eta\sqrt{\mathrm{vc}\left(\mathcal{F}\right)}\right)$。对于第二项，我们首先注意到，根据Block 等人 [2022, 引理 36], ，我们可以以至少 $1-T^{-1}$ 的概率取,

$$\Delta \lesssim \frac{1}{\sqrt{m}}\left(\frac{1}{m}\mathfrak{R}_m(\mathcal{F})+\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{m}}\right) \lesssim \frac{\sqrt{\mathrm{vc}\left(\mathcal{F}\right)+\log\left(T\right)}}{m}$$

引理 44中，后一个不等式由命题 19推出。因此，应用引理45，可知期望遗憾满足

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Reg}_T\right] &\lesssim \eta\sqrt{\mathrm{vc}\left(\mathcal{F}\right)}+\sqrt{\mathrm{vc}\left(\mathcal{F}\right)}\cdot\log(\eta)\sigma^{-\frac{1}{\lambda-1}}\cdot\frac{T}{\eta^{1-\frac{1}{\lambda}}}+T\cdot\sigma^{-\frac{1}{\lambda-1}}\cdot\left(\frac{\sqrt{\mathrm{vc}\left(\mathcal{F}\right)+\log\left(T\right)}}{m}\right)^{\frac{\lambda-1}{\lambda}} \\
&\quad +Tk\varepsilon+\frac{T}{k}\mathfrak{R}_k\left(\mathcal{F}\right)+1 \\
&\lesssim \eta\sqrt{\mathrm{vc}\left(\mathcal{F}\right)}+\sqrt{\mathrm{vc}\left(\mathcal{F}\right)}\cdot\log(\eta)\sigma^{-\frac{1}{\lambda-1}}\cdot\frac{T}{\eta^{1-\frac{1}{\lambda}}} \\
&\quad +T\cdot\sigma^{-\frac{1}{\lambda-1}}\cdot\left(\frac{\sqrt{\mathrm{vc}\left(\mathcal{F}\right)+\log\left(\frac{1}{\delta}\right)}}{k\log(T)\cdot(\varepsilon\sigma)^{-\frac{1}{\lambda-1}}}\right)^{\frac{\lambda-1}{\lambda}}+Tk\varepsilon+T\cdot\sqrt{\frac{\mathrm{vc}\left(\mathcal{F}\right)}{k}},
\end{aligned}$$

最后一步同样由命题19推出。令

$$k=\varepsilon^{-\frac{2}{3}} \qquad \eta=\sqrt{m} \qquad m=k\log(T)\cdot(\varepsilon\sigma)^{-\frac{1}{\lambda-1}} \qquad \varepsilon=T^{-\frac{6\lambda-6}{4\lambda-1}}\cdot\sigma^{-\frac{3}{4\lambda-1}}$$

代入后即可完成证明。 ■