有界 f-散度下的配分函数估计

Adam Block*[1,2] 和 Abhishek Shetty†[3]

[1]哥伦比亚大学计算机科学系[2]哥伦比亚大学电气工程系[3]麻省理工学院

arXiv:2602.23535v1 [stat.ML] 26 Feb 2026

摘要

我们研究在能够获取提议分布的样本以及目标分布的未归一化密度比的条件下，估计配分函数所需的统计复杂度。尽管配分函数估计是一项经典问题，但现有保证通常依赖于关于定义域或模型几何结构的结构假设。我们则给出一种一般性的信息论刻画，它仅取决于提议分布与目标分布之间的关系。我们的分析引入了积分覆盖剖面这一泛函，用以量化密度比大的区域中包含多少目标质量。我们证明，积分覆盖能够紧确刻画乘法型配分函数估计的样本复杂度，并给出了匹配的下界。我们进一步用 f-散度表示这些界，揭示了根据 f 的增长率而出现的尖锐相变；该结果既将经典结论作为特例纳入其中，也推广到了重尾情形。匹配下界证实了所有情形下的紧确性。作为应用，我们推导出重要性采样和自归一化重要性采样更优的有限样本保证，并证明在相同的散度约束下，近似采样与计数的复杂度存在严格分离。我们的结果统一并推广了重要性采样、拒绝采样和重尾均值估计的既有分析，在最少假设下建立了配分函数估计理论。在此过程中，我们还引入了新的技术工具，包括覆盖与 f-散度之间的新联系，以及对经典 Paley-Zygmund不等式的推广。

1 引言

在许多特定情形下，配分函数估计问题都受到了广泛关注，研究对象包括统计力学和组合学中的经典模型，例如 Ising模型 [22, 41] ，也包括科学应用中施加了物理结构的 λ [11]更一般的情形。其目标是估计未归一化密度的归一化常数，这在统计学、机器学习和计算机科学等领域有诸多应用，包括贝叶斯推断[16, 23], 、图模型 [43], 、基于能量的模型 [26], 、统计物理学和化学 [11, 15],, 以及用于语言模型后训练的强化学习 [5, 10], 等。事实上，由于这一问题备受关注，许多先前研究都致力于设计和分析用于完成这项重要任务的算法。尽管这些研究在所探讨的问题领域中提供了大量理论和实践方面的

*adam.block@columbia.edu
†shetty@麻省理工学院.edu

见解，但它们往往依赖于关于 λ 结构的结构假设（例如，图上学习背景下出现的假设），或关于 $\mathcal{X}$ 几何性质的假设（例如，在欧几里得空间中要求光滑或满足其他正则性条件），从而限制了它们在更一般情形下的适用性。事实上，当前文献中存在一个令人意外的空白：尽管配分函数估计问题至关重要，却鲜有一般性结果能够根据底层分布 μ 和 ν的自然属性及信息论属性，刻画配分函数估计的统计复杂度。鉴于语言建模等现代应用中定义域缺乏结构，且 λ 往往对应复杂的学习模型和奖励函数 [37, 45]，这一点尤为重要。

在这项工作中，我们旨在通过提出以下问题来填补这一空白：为了以期望的精度估计目标分布 ν 的配分函数，需要从基分布 μ 中抽取多少个样本？这取决于 μ 与 ν之间哪些自然的信息论量？我们通过给出所需样本复杂度 n 的紧界，完整回答了这一问题；该复杂度以目标分布 ν 与提议分布 μ之间的覆盖率曲线（定义 1）来刻画，这一曲线由 [8, 10]考察。覆盖率曲线定量衡量 ν 在密度比 $d\nu/d\mu$ 较大的区域上分配了多少质量，因此能够反映密度比的尾部行为。具体而言，我们通过 ν 与 μ 之间的一种差异度量来刻画这种衰减，并将其称为积分覆盖，记作 $\mathrm{ICov}_M(\nu\|\mu)$（定义 1）。我们考虑这样一种设定：可以访问 μ 的样本，并能计算未归一化密度比 λ，其中 $\lambda = Z \cdot d\nu/d\mu$ 对应某个未知的归一化常数$Z = \int \lambda(x)d\mu(x)$。我们证明，积分覆盖能够精确刻画配分函数估计的样本复杂度，具体如下：

定理 1（定理 4 和7的非正式表述）。设 μ，ν是 $\mathcal{X}$上的两个概率测度。令 M 满足 $M^{-1} \cdot \mathrm{ICov}_M(\nu\|\mu)$，$\leq \varepsilon$则需要且只需 $n = \Theta(M \cdot \varepsilon^{-1})$ 个样本，即可将 Z 估计到乘法精度 $(1 \pm \varepsilon)$。

这一结果不仅用自然的信息论量精确刻画了配分函数估计的样本复杂度，还统一、推广并应用了重要性采样和均值估计方面的若干先前结果 [8, 13]。

我们界限的精确性还可以通过 ν与 μ之间更熟悉的差异度量来进一步理解；f-散度 [12, 35]。回顾一下，f-散度（定义 2）定义为密度比的凸函数 f 的期望，并推广了全变差、KL 散度和 Rényi 散度等概念。事实上，以下是用f-散度表述的主要结果的非正式版本：

定理 2（定理 5 和8的非正式陈述）。给定两个概率测度 μ，ν 定义在 $\mathcal{X}$ 上，且它们之间的f-散度$\mathrm{D}_f(\nu\|\mu)$，则存在一个函数 γ_f，它依赖于 f，使得

$$n = \Theta\left(\left[\gamma_f\left(\Theta(1) \cdot \varepsilon^{-1} \cdot \mathrm{D}_f(\nu\|\mu)\right)\right] \vee \left[\mathrm{D}_{\chi^2}(\nu\|\mu) \cdot \varepsilon^{-2}\right]\right)$$

估计 Z 至乘法精度 $(1 \pm \varepsilon)$ 且具有常数概率，需要且只需要这么多个样本，其中$\mathrm{D}_{\chi^2}(\cdot\|\cdot)$ 是 χ^2-散度。

作为这些结果的一项应用，我们根据目标函数以及分布，对重要性采样和自归一化重要性采样（SNIS）估计量进行了更精确的有限样本分析 μ和 ν （定理 9和10），并为已有结果提供了统一的视角 [8, 34]。特别是，我们的结果有助于指导重要性采样提议分布的设计，使其针对给定的一组目标函数，将所需样本复杂度降至最低；与经典的方差最小化提议分布相比，这种设计更加灵活 [29, 34]。

我们还发现，这与以下情形下采样的样本复杂度存在有趣的联系：从 ν 给定来自 μ 的样本，并且可以获取未归一化密度比 λ。在 [3, 14], 的结果基础上加以推广，我们证明了可以根据覆盖和 f-散度，对采样的样本复杂度作出紧确刻画，具体如下：

定理 3（命题 3 的非正式表述）。设 μ, ν 是 $\mathcal{X}$ 上的概率测度。设 M 满足 ~，使得 $Cov_M(\nu\|\mu)\varepsilon$。那么，$n = \Theta(M \cdot log\,(1/\varepsilon))$ 是从 μ 的样本以及未归一化密度比的访问权限出发，生成与 ν 相距全变差距离至多为 ε 的近似样本所必需且充分的。特别地，当 $n \gtrsim log\,(1/\varepsilon) \cdot \gamma_f(\Theta(1) \cdot D_f(\nu\|\mu)/\varepsilon)$ 时，这一结论成立。

这一结果表明，采样比估计更容易。与一些设置（例如自归约性）形成鲜明对比的是，在那些设置中，采样的复杂度通常与估计大致相同[22, 41]。我们将在第 3.3 节详细比较采样与估计的样本复杂度。

下一节将正式提出我们关注的问题，并介绍和定义全文中使用的关键量。随后，我们将在第 3 节陈述主要结果，包括配分函数估计样本复杂度的上界和下界，以及近似采样的界。接着，我们将在第 4 节 讨论这些结果在重要性采样和自归一化重要性采样中的应用。然后，我们将在第 5 节概述用于证明主要上界的证明技巧，并将所有结果的完整证明放在附录中。最后，我们将在第 6 节 讨论开放问题和未来方向。我们的证明引入了若干具有独立研究价值的新技术工具，包括f-散度与积分覆盖之间的联系、截断密度比的方差界（引理 5），以及 Paley-Zygmund 不等式的一个强推广（引理 1）。

2 问题设置与预备知识

我们考虑可测空间 $\mathcal{X}$ 和其上的基准概率测度 μ。我们还将考虑定义在 $\mathcal{X}$ 上的目标分布 ν，我们希望从中生成样本，或在能够获取来自 μ 的样本以及未归一化密度比 λ 的情况下，估计其归一化常数；其中 $\mathcal{X} \to \mathbb{R}_{\geq 0}$ 根据 ν 相对于 μ 的 Radon-Nikodym 导数定义为 $\lambda(x) = Z \cdot d\nu/d\mu(x)$，其中 Z 是某个未知的归一化常数。我们主要关注获取归一化常数的一个估计值 Z 的难度；更确切地说，我们提出以下问题：给定独立同分布样本 $X_1, \ldots, X_n \sim \mu$，并且能够在每个样本上计算未归一化密度比 λ，为了确保存在一个估计量 $\hat{Z} = \hat{Z}(X_1, \ldots, X_n, \lambda(X_1) \ldots, \lambda(X_n))$，使得 $(1-\varepsilon)Z \leq \hat{Z} \leq (1+\varepsilon)Z$ 以至少 $1-\delta$ 的概率成立，n 必须关于 μ、ν、ε 和 δ 取多大?

获得此类估计值的速度取决于两个分布 μ 和 ν之间的相似程度。例如，如果 μ 和 ν 的支撑集不相交，那么无论从 μ中抽取多少样本，都不足以估计 Z。本文的重点是根据 μ 和 ν之间相似程度的定量测度，对样本复杂度进行精确刻画。

衡量 μ 与 ν 之间差异的关键概念是 [8, 10]所考虑的覆盖。从概念上讲，覆盖衡量的是 ν 在密度比 $\frac{d\nu}{d\mu}$ 较大的区域所放置的质量，因此，良好的覆盖意味着 μ 在 ν的高密度区域放置了足够的质量。形式上，覆盖定义如下。

定义 1（覆盖与积分覆盖）。 对于 $\mathcal{X}$ 和 $M > 0$上的两个概率测度 μ, ν，在 M 处的覆盖 函数定义为

$$\mathrm{Cov}_M(\nu\|\mu) = \nu\left(\{x \in \mathcal{X} : {}^{d\nu}/_{d\mu}(x) \geq M\}\right),$$

其中， ${}^{d\nu}/_{d\mu}$ 是相对于 μ的 ν Radon-Nikodym 导数。进一步地，我们引入一个称为积分覆盖剖面的概念，其定义为

$$\mathrm{ICov}_M(\nu\|\mu) = \int_0^M \mathrm{Cov}_t(\nu\|\mu)\, dt.$$

我们首先指出覆盖和积分覆盖的若干基本性质。首先，显然，对于所有 $M > 0$，$\mathrm{Cov}_M(\nu\|\mu) \leq 1$，因此 $\mathrm{ICov}_M(\nu\|\mu) \leq M$；此外，$\mathrm{Cov}_M(\nu\|\mu)$ 显然关于 M 是非递增的。进一步地，经过简单计算可知， $M \mapsto {}^{\mathrm{ICov}_M(\nu\|\mu)}/_M$关于 M 同样是非递增的，并且当 $\nu \ll \mu$时，该量随着 $M \to \infty$趋于零。我们将看到，后一个映射对于刻画配分函数估计的样本复杂度至关重要。

覆盖这一概念此前已被用于分析重要性采样的样本复杂度 [8], 拒绝采样的样本复杂度 [3]以及有趣的是，用于理解预训练语言模型的有效性 [10, 19, 20]。本文引入的积分覆盖概念，是衡量 μ 与 ν 之间关系的更精细测度，能够对样本复杂度给出更精确的刻画。

两种测度之间更为标准的差异度量是 f-散度 [12, 35], ，其定义如下。

定义 2. 设 f: $\mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$是一个凸函数，满足 $f(1) = f'(1) = 0$。对于定义在 $\mathcal{X}$ 上的两个概率测度 μ, ν, ν与 μ之间的 f-散度定义为

$$\mathrm{D}_f(\nu\|\mu) = \mathbb{E}_{X\sim\mu}\left[f\left(\frac{d\nu}{d\mu}(X)\right)\right] + \nu\left(\frac{d\nu}{d\mu} = \infty\right) \cdot f'(\infty). \tag{1}$$

f-散度概括了概率测度之间许多著名的距离概念，包括全变差距离（其中 $f(t) = {}^1/_2|t-1|$）、Kullback-Leibler（KL）散度（其中$f(t) = t\log t$），以及 α-Rényi 散度（的单调变换）（其中 $f(t) = t^\alpha - \alpha t$，当$\alpha > 0$, $\alpha \neq 1$时）。特别值得注意的是参数为 $\alpha = 2$ 的 α-Rényi 散度，它对应于 χ^2散度，定义为 $\chi^2(\nu\|\mu) = \mathbb{E}_{X\sim\mu}\left[({}^{d\nu}/_{d\mu}(X) - 1)^2\right]$。

从概念上说，有界的 f-散度能够控制密度比 ${}^{d\nu}/_{d\mu}$ 的尾部行为； f 增长得越快，控制就越强。正如我们在附录C中所示，这一直觉可通过多种方式精确表述，其中最关键的是观察到，对于任意 f-散度，$\mathrm{Cov}_M(\nu\|\mu) \leq {}^{M\cdot \mathrm{D}f(\nu\|\mu)}/_M$；因此，有界的 f-散度意味着覆盖会迅速衰减。正如我们将看到的，涉及 f-散度的样本复杂度结果取决于 f 的增长率，具体通过我们称为 γ_f函数的量体现。该函数定义为映射 $t \mapsto {}^{f(t)}/_t$ 在 $[1, \infty)$ 上的逆函数，即，

$$\gamma_f(M) = \inf\left\{t \geq 1 : {}^{f(t)}/_t \geq M\right\}. \tag{2}$$

f 的增长率决定了 γ_f的行为；例如，当 f 是超线性的，即 $\lim_{t\to\infty} {}^{f(t)}/_t = \infty$ 时，函数 γ_f 在整个 $\mathbb{R}_{\geq 0}$上都有定义；超线性 f 的例子包括对应 KL 散度的情形（此时 $\gamma_f(M) \asymp \exp(M)$）以及满足 $\alpha > 1$的 α-Rényi 散度（此时 $\gamma_f(M) \asymp M^{\frac{1}{\alpha-1}}$）。我们还需要区分超二次的 f（其中 $\lim_{t\to\infty} {}^{\overline{f(t)}}/_{t^2} = \infty$，例如满足 $\alpha > 2$的 α-Rényi 散度）与非超二次的 f（例如 KL 散度和满足$1 < \alpha \leq 2$的 α-Rényi 散度，如 χ^2）。当 f为线性时，对于 $M > \sup_t {}^{f(t)}/_t$，我们令 $\gamma_f(M) = \infty$；请注意，根据凸性假设， f 不可能是次线性的。这些区别决定了样本复杂度结果中的具体增长率。

3 主要结果

现在，我们介绍关于估计配分函数的主要结果：已知基分布的样本以及未归一化密度比时，估计配分函数所需的样本复杂度。

3.1 估计的上界

我们首先陈述关于配分函数估计样本复杂度的主要上界，该上界以定义 1中定义的积分覆盖剖面表示。

定理 4。 设 μ、ν 是 $\mathcal{X}$ 上的概率测度。假设我们可以获得独立同分布样本 X_1、...、$X_n \sim \mu$ 以及一个未归一化密度比 λ。令 M_ε 满足 $M^{-1} \cdot \mathrm{ICov}_M(\nu\|\mu) \le \varepsilon/4$。^那么存在一个估计量 Z，使得只要

$$n \gtrsim {M_\varepsilon \cdot \log(1/\delta)}/{\varepsilon}, \tag{3}$$

^则以至少 $1-\delta$ 的概率有 $(1-\varepsilon) \cdot Z \le Z \le (1+\varepsilon)Z$。

如上文所述，只要 $\nu \ll \mu$，对于任意正的 ε，这样的 M_ε 总是存在；因此，在绝对连续情形下，上述定理总能为以 $(1\pm\varepsilon)$ 精度估计配分函数提供有限的样本复杂度界。

为了理解 (3)，考虑 $\chi^2(\nu||\mu)$ 有界这一简单情形。此时，一个简单的尾界表明，对所有 M 均有 $\mathrm{ICov}_M(\nu\|\mu) \le \chi^2(\nu||\mu)$。因此，我们可以在定理 4 中取 $M_\varepsilon \asymp \chi^2(\nu||\mu)/\varepsilon$，并得到 $n \gtrsim \chi^2(\nu||\mu)/\varepsilon^2$ 个样本足以以高概率将配分函数估计到 $(1\pm\varepsilon)$ 精度。这恢复了通过计算方差 [34] 来界定重要性采样样本复杂度的标准结果。于是自然要问： M_ε 作为 ε的函数能增长得多慢？或许有人希望，借助改进的矩界，可以通过M_ε获得更快的估计速率，但事实并非如此。事实上，这一点可以直观地看出：由于对任意两个分布 ν, μ 都有 $\mathrm{Cov}_0(\nu\|\mu) = 1$，故当 $\varepsilon \downarrow 0$ 时必有 $M_\varepsilon \gtrsim 1/\varepsilon$，而定理 4 永远无法实现优于 $\Omega(\varepsilon^{-2})$ 的样本复杂度缩放，这与中心极限定理所预期的、将均值估计到 $(1\pm\varepsilon)$ 精度时的情形一致。因此，我们更应将上述定理视为将该界推广到更一般的重尾情形，其中 χ^2 散度可能为无穷。

因此，我们可以进一步探讨：能否用更熟悉的 f-散度来给出更明确的样本复杂度界？这相当于控制密度比的尾部。这正是下一个定理要阐述的内容。

定理 5。 设 μ, ν 是 $\mathcal{X}$ 上的概率测度，并设 f 是一个满足定义 2 中条件的凸函数，且存在[1] $c \ge 1$，使得对于 $t \ge c$，函数 $t \mapsto f(t)/t^2$ 单调不增。假设我们可以访问独立同分布样本 $X_1, \ldots, X_n \sim \mu$ 以及一个未归一化密度比 λ。那么存在一个^估计量 Z，只要

$$n \gtrsim \frac{\gamma_f\left({6 \cdot \mathrm{D}_f(\nu\|\mu)}/{\varepsilon}\right) \cdot \log\left({1}/{\delta}\right)}{\varepsilon} \vee \frac{c^2}{\varepsilon^2}, \tag{4}$$

^那么，以至少 $1-\delta$ 的概率，有$(1-\varepsilon) \cdot Z \le Z \le (1+\varepsilon)Z$。

[1]定理 5 中的实际常数 $c \ge 1$ Theorem 5是一个技术性产物，通常应认为它等于 1。事实上，这个条件只是为了确保 $f(t)$ 最终次二次，详见下文。

f-散度的视角为定理 4中的基于覆盖率的界提供了一个互补的观点。我们可以根据 图1中总结的 f 的增长率，将估计配分函数的样本复杂度划分为三个区域： 线性、超线性 但 次二次，以及超二次。实际上，当 f 为线性，即 $\lim_{t\to\infty} f(t)/t < \infty$ 时，函数 γ_f 仅在有界定义域上有定义，因此对于足够小的 ε，该定理是空洞的；例如，总变差距离或海林格距离就是这种情况。在超线性但次二次的情形下，即 $\lim_{t\to\infty} f(t)/t^2 = C < \infty$，我们看到 γ_f 在整个 $\mathbb{R}_{\geq 0}$ 上都有定义，且增长速度至少与 ε^{-1}一样快，这意味着式（4）中的第一项占主导地位。KL 散度就属于这种情况，其中

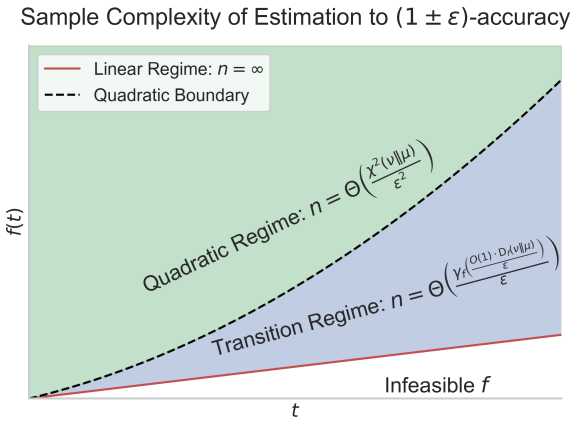

图1：根据定义 $D_f(\nu\|\mu)$ 的 f 增长率，估计 Z to 达到 $(1\pm\varepsilon)$ 精度所需样本复杂度（n）的不同情形。

$n \gtrsim \exp(c \cdot D_{KL}(\nu\|\mu)/\varepsilon)/\varepsilon$ 个样本也足够；对于 α-Rényi 散度，在 $1 < \alpha \leq 2$ 时，$n \gtrsim \varepsilon^{-\alpha} \cdot D_f(\nu\|\mu)^{1/\alpha-1}$ 个样本足够。最后，在 $\lim_{t\to\infty} f(t)/t^2 = \infty$ 的超二次情形下，我们看到 γ_f 增长得太慢，无法在较小的 ε 时主导样本复杂度，因此式 (4) 的第二项开始起作用；对于满足 $\alpha > 2$ 的 α-Rényi 散度而言，$n \gtrsim 1/\varepsilon^2$ 个样本足够。

请注意，如上所述，对于较小的 ε，我们最紧的上界也无法优于 ε^{-2} 缩放。因此，我们以上界讨论的最后部分采用一种非对称方法，它能够改善 ε 下尾的样本复杂度缩放，但代价是上尾界显著变差。

定理 6。 设 μ、ν 为 $\mathcal{X}$ 上的概率测度，并假设 $X_1 \ldots X_n \sim \mu$ λ、均与一个未归一化密度比独立。则存在一个估计量 Z，使得对于任意 $0 < \varepsilon < 1$，若

$$n \gtrsim \log(1/\delta) \cdot M/\varepsilon \quad \textit{where } \mathrm{Cov}_M(\nu\|\mu) \lesssim \varepsilon, \tag{5}$$

则以至少 $1-\delta$ 的概率有 $(1-\varepsilon) \cdot Z \leq Z \leq M \cdot Z$。特别地，如果 $D_f(\nu\|\mu) < \infty$，那么只需取 $M = \gamma_f(6 \cdot D_f(\nu\|\mu)/\varepsilon)$。

注意，在 M （5）中的值不大于 M_ε 定理 4中的值，因为对于所有 M，根据覆盖递减这一事实，有 $\mathrm{ICov}_M(\nu\|\mu) \leq M \cdot \mathrm{Cov}_M(\nu\|\mu)$。因此，在 f 为超二次的情形下，我们可以实现从下方估计 Z 的样本复杂度，其值显著小于 ε^{-2} 定理 4和 5的缩放。

3.2 下界

为了补充我们的上界，我们针对配分函数估计的样本复杂度给出若干下界，分别基于积分覆盖以及 f-散度，以此证明我们的界是紧的。下界构造将在附录 D中给出，部分灵感来自 Block 和 Polyanskiy [3]，尤其是 Harremoës 和 Vajda 的结果 [18]，该结果表明，成对的伯努利随机变量能够使 f-散度的联合值域达到极值。我们首先给出积分覆盖下界。

定理 7。 对于任意非原子 μ 和任意 $\varepsilon \leq 1/2$，都存在一个分布族 ν，使得任何估计量 Z 若要以至少 $2/3$ 的概率达到 $(1\pm\varepsilon)$ 乘法精度，就至少需要$\varepsilon^{-1} \cdot M_\varepsilon$ 个样本，其中 M_ε 满足

$$\mathrm{ICov}_{M_\varepsilon}(\nu\|\mu) \leq M_\varepsilon \cdot \varepsilon.$$

特别是，定理 7 表明，定理 4 的样本复杂度界在对积分覆盖和 ε的依赖关系上是紧的。不幸的是，定理 7 不能直接转化为关于 f-散度的下界；虽然确实可以通过 f-散度对积分覆盖进行紧确控制，但尚不清楚这种紧确控制是否能由 定理 7的下界构造实现。因此，我们通过针对上述三种情形分别给出下界来补充定理 5 ，首先从线性情形开始。

命题 *1.* 设 f 是定义 *2* 中的凸函数，且 $t \mapsto f(t)/t$ 有界。对于任意 $0 < C < f'(\infty)$，存在测度 μ、ν，使得 $D_f(\nu\|\mu) \le C$，从而对于任意 $\varepsilon \le f'(\infty)/(f'(\infty)-C)$，任何仅能访问来自 μ 的有限个样本的估计量，都无法以至少 $3/4$ 的概率给出 Z 的 $1\pm\varepsilon$ 近似。此外，如果 $C \ge f'(\infty)$，则存在测度 μ、ν，使得 $D_f(\nu\|\mu) \le C$，且任何此类估计量都无法以至少 $3/4$ 的概率给出 Z 的 $1\pm\varepsilon$ 近似，对任意 $\varepsilon > 0$ 均如此。

命题 1 表明，当 f 为线性函数时，无论样本数量多少，都不足以将配分函数估计到任何非平凡的乘法精度范围内。鉴于具有线性 f-散度（例如全变差或海灵格） f 无法以任何有意义的方式控制密度比 $\frac{d\nu}{d\mu}$ 的尾部，且实际上允许出现以下情况： ν 相对于 μ是奇异的，此时显然不可能估计 Z，因此这一事实并不令人意外。现在我们转向第二种情况，即 f 为超线性但次二次。

定理 8. 对于任意 $C > 2$ 与 $\varepsilon \in (0, 1-4)$，存在 $\mathcal{X}$ 上的概率测度 μ 以及 $\mathcal{X}$ 上的概率测度类 $\mathcal{V}$，使得对所有 $\nu \in \mathcal{V}$，均有 $D_f(\nu\|\mu) \le C + f(1-\varepsilon)$，并且任何满足 $(1-\varepsilon)Z \le \widehat{Z} \le (1+\varepsilon)Z$、对所有 $\nu \in \mathcal{V}$ 以至少 $2\,3$ 的概率成立的估计量 $\widehat{Z}$，都必须至少使用 $n \gtrsim \gamma_f(\frac{C}{2\cdot\varepsilon})/\varepsilon$个样本。

虽然该下界对任意 f-散度均成立，但它仅在超线性/次二次情形下是紧的，其中 $\gamma_f(1/\varepsilon)\cdot\varepsilon^{-1} \gg \varepsilon^{-2}$；这适用于 $1 < \alpha \le 2$ 时的 KL 散度与 Rényi 散度。对于超二次情形，我们有如下下界。

命题 2。 固定 $0 < \varepsilon < 1/8$。那么，存在一族定义在 $\mathcal{X}$ 上的概率测度 $\mathcal{U}$，以及定义在 $\mathcal{X}$ 上的概率测度 ν，使得对所有 $\mu \in \mathcal{U}$，都有 $\|d\nu/d\mu\|_\infty \le 4$；并且，任何估计量 Z 若对所有 $\nu \in \mathcal{V}$以至少 $2/3$ 的概率给出 Z 的 $1\pm\varepsilon$ 近似，则至少需要 $n \gtrsim \frac{1}{\varepsilon^2}$ 个样本。

综上，本节给出的下界表明，我们的每个上界在各自适用的范围内都是紧的，无论是从积分覆盖还是从 f-散度的角度来看。

3.3 与采样的比较

虽然我们的主要关注点是配分函数估计，但将我们的结果与相关问题的结果进行比较也很有启发性：从目标分布 ν 中进行采样，其中可访问来自基分布 μ 的样本以及未归一化密度比 λ。在 Block 和 Polyanskiy [3], 中，作者研究了近似拒绝采样问题：学习者可以访问来自 μ的 n 个独立同分布样本以及归一化 密度比 $\frac{d\nu}{d\mu}(\cdot)$，并且必须生成一个来自分布 ν_n的样本；在 ν与 μ之间的 f-散度有界的假设下，该分布与 ν在全变差距离意义下接近。我们将他们的结果以及 Flamich 和 Wells [14]的结果推广到未归一化 密度比和任意 f-散度的情形。

命题 3。 设 μ，ν 是定义在 $\mathcal{X}$ 上的概率测度，并设 f 是定义 *2* 中的凸函数定义 *2*。假设我们可以访问独立同分布样本 X_1，……， $X_n \sim \mu$ 以及未归一化密度比 λ。那么

存在一种算法，可以从分布 ν_n 中生成样本 $X_{\hat{j}}$，使得$\mathrm{TV}(\nu_n, \nu) \leq \varepsilon$，只要

$$n \gtrsim M \cdot \log(1/\varepsilon) \quad \textit{where } \mathrm{Cov}_M(\nu\|\mu) \lesssim \varepsilon. \tag{6}$$

特别地，这对

$$n \gtrsim \log(1/\varepsilon) \cdot \gamma_f\left(\Theta(1) \cdot \mathrm{D}_f(\nu\|\mu)/\varepsilon\right). \tag{7}$$

我们注意到，Block 和 Polyanskiy 关于近似拒绝采样的下界 [3] 已经表明，（7）在对数因子范围内是紧的，因为该工作中的区域严格强于我们考虑的区域（即，可以访问归一化密度比，而非未归一化密度比）。我们将命题 3 的证明留待 A^*附录 F [14], ，届时我们将推广 Flamich 和 Wells [27] 中近似拒绝采样算法的分析。

值得注意的是，采样的样本复杂度严格低于配分函数估计的样本复杂度。为说明这一点，注意到对于所有 M，$\mathrm{Cov}_M(\nu\|\mu) \leq M^{-1} \cdot \mathrm{ICov}_M(\nu\|\mu)$，因为 $\mathrm{Cov}_M(\nu\|\mu)$ 是 M 的递减函数。因此，直接可得，与定理 4 中的 (6) 相比 (4)，样本复杂度至少改善了 ε^{-1} 倍。然而，视覆盖率曲线的变化情况而定，这种改善甚至可能更加显著。事实上，我们已经看到，随着 M 增大，$M^{-1} \cdot \mathrm{ICov}_M(\nu\|\mu)$ 的下降速度不会快于 M^{-1}；因此，在超二次区域中，当 $\mathrm{D}_f(\nu\|\mu)$ 对快速增长的 f 有界时，采样与配分函数估计的样本复杂度之间几乎可以相差二次。在极端情况下，若密度比一致有界，即 $\|d\nu/d\mu\|_\infty \leq M$，采样只需 $n \gtrsim \log(1/\varepsilon)$ 个样本，就能从 ν 中生成一个在全变差距离下误差为 ε 的近似样本；而配分函数估计则需要 $n \gtrsim \varepsilon^{-2}$ 个样本。一种理解方式是，估计取决于整个覆盖率曲线，而采样只取决于 M 的单个取值处的覆盖率。

由此我们看到，与众所周知的“自归约”问题类截然不同，“计数”（即配分函数估计）问题在一般 f-散度或覆盖约束下严格难于“采样”。

4 应用：重要性采样的改进有限样本界

重要性采样是统计学和机器学习中的一项基本技术，并在包括因果推断、强化学习 [36, 39], 变分推断[6], 以及概率编程 [44]在内的许多领域中充当基本原语。它是一种工具，用于利用来自提议分布 μ的样本，估计函数 g（我们关注有界正值情形）在目标分布 ν 下的期望。给定独立同分布样本 $X_1, \ldots, X_n \sim \mu$，$\nu_g = \mathbb{E}_\nu[g(X)]$的重要性采样估计量为

$$\hat{g}_{\mathrm{IS}} = \frac{1}{n}\sum_{i=1}^{n} \frac{d\nu}{d\mu}(X_i) g(X_i).$$

该估计量是无偏的，即 $\mathbb{E}[\hat{g}_{\mathrm{IS}}] = \nu_g$，而关于重要性采样估计量的标准结果给出了$\hat{g}_{\mathrm{IS}}$ 的方差上界；该方差可借助目标分布与提议分布之间的 χ^2-散度加以界定，因为方差以 $\chi^2(\nu\|\mu)$为界。

为获得更精确的界，我们考虑由函数 g 加权的目标分布，即测度 $\nu \cdot g$，其相对于 μ 的密度为 $g(x)/\nu_g \cdot d\nu/d\mu(x)$。由此可定义

加权目标分布 $\nu \cdot g$ 与提议分布 μ 之间的覆盖，在此语境下可表示为

$$\mathrm{Cov}_M\left(\nu \cdot g \| \mu\right) = \frac{1}{\nu_g} \int g(x) \frac{d\nu}{d\mu}(x) \mathbb{I}\left[g(x)\frac{d\nu}{d\mu}(x) \geq M \cdot \nu_g\right] d\mu(x)$$

此外，这使我们能够以自然的方式定义加权目标分布 $\nu \cdot g$ 与提议分布 μ 之间的积分覆盖和 f-散度 $\mathrm{D}_f(\nu \cdot g \| \mu)$。

应用定理 4，我们可以得到以一般 f-散度表示的、适用于重要性采样估计量的有限样本界。

定理9（重要性采样）。设 μ、ν 为测度，g 为有界正函数。则重要性采样估计量 $\hat{}g_{\mathrm{IS}}$ 满足 $|\hat{}g_{\mathrm{IS}} - \nu_g| \leq \varepsilon \cdot \nu_g$，其概率至少为 $1-\delta$，只要$n \gtrsim {}^{M_{\varepsilon,\,\delta}}/_{\varepsilon}$。其中$M_{\varepsilon,\,\delta}$ 满足 积分覆盖$M_{\varepsilon,\,\delta}\left(g\nu \| \mu\right) \leq {}^{\varepsilon\delta}/_{6}$。

这推广了重要性采样的标准方差界，该界对应于χ^2-散度的情形。概率参数 δ 对结果的影响可以通过$\log(1/\delta)$ 来减弱，方法与定理 4的证明中使用的均值中位数法相同；但我们将其表述为这种形式，以保留实践中常用的估计量形式。

更重要的是，我们的结果将误差界定为 g、ν 和 μ 的联合函数，而非分别用 $D_f(\nu\|\mu)$ 和 $\|g\|$ [8]等量来界定。当我们可以根据目标函数 g 和分布 ν 来设计提议分布 μ 时，这一点尤其有用：上述界可用于优化提议分布 μ，从而针对给定的一组目标函数 g，使所需的样本复杂度最小化。特别是，重要性采样中的最优提议分布是一个研究充分的问题（综述请见 [29, 34] ），通常选择使重要性采样估计量的方差最小的分布。我们的结果表明，更精细的目标是最小化加权目标分布之间的积分覆盖。

我们的方法还可推广到自归一化重要性采样（SNIS） [34], ，当我们只能从提议分布 μ 中采样，并且只能计算未归一化密度比 $\lambda(x) = \frac{d\nu}{d\mu}(x) \cdot Z$（其中归一化常数 $Z = \int \lambda(x) d\mu(x)$ 未知）时，它可用于估计期望值 $\mathbb{E}_\nu[g(X)]$ 。给定独立同分布样本 $X_1, \ldots, X_n \sim \mu$，SNIS估计量为 $\hat{}g_{\mathrm{SNIS}} = \sum_{i=1}^n \lambda(X_i) g(X_i) / \sum_{i=1}^n \lambda(X_i)$。在假设 μ 关于 ν 绝对连续的条件下，SNIS估计量是渐近无偏的 [34, 定理 9.2]。其渐近方差为 $\mathbb{E}_{X\sim\mu}\left[(g(X) - \mathbb{E}_{X\sim\nu}[g(X)])^2 \cdot \left(\frac{d\nu}{d\mu}(X)\right)^2\right]$；对于有界函数 g，该方差至多为 $\|g\|_\infty^2 \cdot \chi^2(\nu\|\mu)$。这些渐近界给出了有限样本下的样本复杂度 $n \geq {}^{4\|g\|_\infty^2 \chi^2(\nu\|\mu)}/_{\delta\varepsilon^2}$ [32,命题 9]。

我们的结果针对一般的 f-散度，对 SNIS估计量给出了改进的有限样本分析，即使 χ^2-散度为无穷大时也适用。

定理 10（自归一化重要性采样）。设 μ, ν 为测度，且 g 为有界正函数。那么，自归一化重要性采样估计量 $\hat{}g_{\mathrm{SNIS}}$ 以至少 $1-\delta$的概率满足 $|\hat{}g_{\mathrm{SNIS}} - \nu_g| \leq \varepsilon \cdot \nu_g$，只要 $n \gtrsim {}^{M_{\varepsilon,\delta}}/_{\varepsilon}$，其中 $M_{\varepsilon,\delta}$ 满足 $\mathrm{ICov}_{M_{\varepsilon,\delta}}(g\nu\|\mu) \leq {}^{\varepsilon\delta}/_{6}$ 以及$\mathrm{ICov}_{M_{\varepsilon,\delta}}(\nu\|\mu) \leq {}^{\varepsilon\delta}/_{6}$。

注意，对加权目标分布 $g\nu$ 和目标分布 ν 的积分覆盖提出要求是必要的，因为 SNIS估计量是两个重要性采样估计量的比值，其中分母用于估计归一化常数 Z。

上界的证明技巧

本节概述了用于建立主要上界的一些关键证明技巧，并重点介绍可能具有独立价值的技术性结果。所有结果的完整证明将在附录中给出。

5.1 样本复杂度的上界

我们的上界依赖于相同的一般证明策略和相同的估计量，即均值的中位数 [1, 30] 形式化地说，给定 n 个来自 μ的样本，我们将其划分为 k 组，每组大小为 $m = \lfloor n/k \rfloor$，并计算每组内的样本均值，即对于每个组 $j \in [k]$，令

$$\widehat{Z}_j = \frac{1}{m}\sum_{i=1}^{m} \lambda(X_i^{(j)}) \quad \text{and} \quad \widehat{Z} = \mathrm{Med}\left(\widehat{Z}_1, \ldots, \widehat{Z}_k\right), \tag{8}$$

其中$\mathrm{Med}(\cdot)$ 表示中位数运算符。根据标准分析（参见引理 4），只需证明$\widehat{}$每组均值 Z_j 以常数概率处于真实配分函数 Z 的常数倍范围内，$\widehat{}$例如 $2/3$。因此，主要的技术挑战在于分析样本均值 Z_j 围绕Z的集中性。完整的定理 4和5 证明将留待附录B和C.2节，下面我们概述主要思路。

$\widehat{}$**证明概要：定理 4。** 分析 Z_j 的集中性时，最自然的第一种方法是应用标准集中不等式，例如切尔诺夫界。不幸的是，如前所述，该方法会失效，因为密度比 $d\nu/d\mu$ 可能是重尾的，因而求和项的方差可能为无穷。为了克服这一挑战，我们考察$\widehat{}$这一密度比的截断版本的集中性；请注意，这种截断纯粹是一种分析工具，而估计量 Z_j 并不涉及任何截断。因此，我们分别分析密度比的“主体”和“尾部”。为了对$\sim Z_j$给出下界，我们可以完全忽略尾部；而为了给出估计值的上界，我们对覆盖的定义应用马尔可夫不等式。

因此，大部分工作在于控制截断密度比的平均值，其期望可借助覆盖的定义得到适当控制。人们起初可能会希望应用标准切尔诺夫界，因为截断到 M 后求和项是有界的，但这充其量只能得到按 M/ε^2 缩放的样本复杂度，其中 M 满足$\mathrm{Cov}_M(\nu\|\mu) \leq \varepsilon$，这较为松弛。相反，我们应用伯恩斯坦不等式，并对截断密度比的方差加以界定；我们证明，它可由积分覆盖本身加以控制（引理 5）。这种有趣的自归一化型性质将截断密度比的方差与截断所引入的偏差联系起来，是使我们能够改进样本复杂度中对 ε 的依赖关系的关键技术工具。■

接下来，我们可以利用这一结果，通过将积分覆盖与定理 5 联系起来，证明 f-散度，具体思路如下。

定理 5 的证明概要。为了获得依赖于 f-散度的样本复杂度，我们需要将积分覆盖与 f-散度联系起来。乍看之下，这似乎很简单，因为应用马尔可夫不等式便能直接联系这两个量（引理 7）。遗憾的是，这种直接关系不足以得到我们所需的积分覆盖界。为此，我们需要利用函数 f的增长性质。具体而言，由于我们主要关注次二次区域，因此利用假设 $t^{-2} \cdot f(t)$ 在 $t \geq c$ 上递减，从而将积分覆盖与 f-散度联系起来，如引理 9所示。■

5.2 基于覆盖的非对称估计

为了获得对 ε更优的依赖关系，在定理 6 中，我们放宽了对称估计的要求：只需严格控制下尾，同时允许上尾具有较宽松的常数因子。该技术方法基于以下引理；它可视为帕利–齐格蒙德不等式对 f-散度的推广，其证明见E.1 节。

引理 1. 设 μ, ν 是定义在 $\mathcal{X}$ 上的概率测度，且设 f 是一个凸函数，如 定义 2所述。那么对于任意 $0 < \varepsilon, u < 1$，均有

$$\mathbb{P}_{X\sim\mu}\left(\frac{d\nu}{d\mu}(X) \geq 1-\varepsilon\right) \geq {}^{(1-u)\varepsilon}\!/_{M} \quad \textit{where } \mathrm{Cov}_M\left(\nu\|\mu\right) \leq u\cdot\varepsilon.$$

特别地，$\mathbb{P}_{X\sim\mu}\left({}^{d\nu}\!/_{d\mu}(X) \geq 1-\varepsilon\right) \geq \sup_{0<u<1} {}^{(1-u)\cdot\varepsilon}\!/_{\gamma_f\left({}^{\mathrm{D}_f(\nu\|\mu)}\!/_{u\cdot\varepsilon}\right)}$。

尽管引理 1 是用 f-散度和似然比来表述的，但也可以将其表述为适用于任意非负随机变量的形式，从而给出此类随机变量超过其均值的 $(1-\varepsilon)$ 倍的概率下界。事实上，取 $f(t) = (t-1)^2$ 并令 $Y \geq 0$，可知我们得到的 Paley-Zygmund 不等式与其相差至多一个因子 4。类似地，取 $f(t) = t^p - 1$，其中 $p > 1$，即可得到 Paley-Zygmund 不等式关于高阶矩的标准推广。如果我们仅假设存在 $|X|\log(|X|)$ 阶矩，则可导出一个据我们所知尚属新颖的界，表明非负随机变量以至少$\varepsilon \cdot \exp\left(-O(1)\cdot {}^{\mathbb{E}[X\log(X)]}\!/_{\varepsilon\cdot\mathbb{E}[X]}\right)$ 的概率超过其均值的 $(1-\varepsilon)$ 倍。现在，我们利用引理 1 概述定理 6 的证明；完整细节见附录 E。

定理 6 的证明概要。与之前的上界不同，这里我们不使用中位数均值估计量，而是直接取样本的$(1-\alpha)$分位数，其中 α经过适当选择。核心思路是利用引理 1 对单个样本超过 $(1-\varepsilon)Z$的概率给出下界，再通过切尔诺夫界控制分位数估计量的下尾。随后，我们根据覆盖的定义控制估计量的上尾，从而以高概率确保该分位数不超过 $M\cdot Z$。■

6 讨论与相关工作

本文完整刻画了配分函数估计的样本复杂度，该复杂度以积分覆盖以及目标分布与提议分布之间的 f-散度来衡量。现在我们讨论相关工作。

配分函数估计。配分函数估计是一项经典问题，其渊源可追溯至 Gibbs 和玻尔兹曼在统计力学中的研究；在此后的岁月里，人们发展并分析了许多经典技术 [15, 22, 33, 34]。理论计算机科学和统计力学领域的一条常见研究路线假设问题具有某种离散结构，并构造复杂的退火调度和采样器，以估计复杂模型（例如 Ising模型）的配分函数，或解决矩阵永久式等计数问题 [2, 21, 22, 25, 38]。另一条研究路线，尤其是在科学和贝叶斯应用中，则聚焦于欧几里得空间中的连续分布，并借助强正则性假设来确保结果质量 [9, 17, 28]。与这些研究不同，我们关注的是一个假设最少的设定，在不作任何结构性或正则性假设的情况下，依据纯粹的信息论量来刻画配分函数估计的难度。

采样与重要性采样。 与配分函数估计相对应，一个密切相关的问题是：在能够获取目标分布的样本 ν 和提议分布的样本 μ 并能够计算密度比 $d\nu/d\mu$ 的条件下，构造目标分布下的均值估计；这一问题也已得到广泛研究 [7, 24, 34]。与我们工作最相关的是 Chatterjee 和 Diaconis 的研究 [8], Devroye 等人的研究 [13]。前者虽未明确研究配分函数估计，却考察了多项式矩约束下的重尾均值估计；当考虑 Rényi 散度时，我们的 f-散度结果可得出他们的结论。另一方面，Chatterjee 和 Diaconis [8]研究了重要性采样，并考察 KL约束下的样本复杂度，证明要使重要性采样成功，所需样本数关于 KL 散度呈指数级增长，且这一数量级既必要又充分；但他们未考虑更一般的 f-散度或覆盖问题。我们的结果对他们的工作进行了推广和加强，将他们的界作为特例重新得到，并统一了这些研究范式。近期，Block 和 Polyanskiy[3] 研究了近似拒绝采样问题，这与我们的结果密切相关，但他们假设可以获取归一化的密度比。另一方面，Li 和 El Gamal [27] 研究了使用未归一化密度比进行采样，但假设密度比有界；Flamich 和 Wells [14]则将其推广到有界 KL约束。我们采用 Maddison 等人提出的同一算法 [31]，并将这些已有结果推广到任意 f-散度和覆盖约束。

参考文献

[1] Noga Alon、Yossi Matias 和 Mario Szegedy。《近似频率矩的空间复杂度》。载于第 *28* 届 *ACM*计算理论研讨会论文集，第 20–29 页，1996 年。[2] Ivona Bezáková、Daniel Štefankovi˘c、Vijay V Vazirani 和 Eric Vigoda。《加速矩阵永久式及组合计数问题的模拟退火》。*SIAM*计算杂志，37 (5):1429–1454，2008年。[3] Adam Block 和 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。[4] Stéphane Boucheron、Gábor Lugosi 和 Pascal Massart。《集中不等式：独立性的非渐近理论》。牛津大学出版社，2013 年。[5] Kianté Brantley、Mingyu Chen、Zhaolin Gao、Jason D Lee、Wen Sun、Wenhao Zhan 和 Xuezhou Zhang。《通过最优优势回归加速大语言模型推理的强化学习》。*arXiv* 预印本 *arXiv:2505.20686*，2025 年。[6] Yuri Burda、Roger Grosse 和 Ruslan Salakhutdinov。《重要性加权自编码器》。载于国际学习表征会议，2016 年。[7] Olivier Cappé、Eric Moulines 和 Tobias Rydén。《隐马尔可夫模型中的推断》。施普林格，2005 年。[8] 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。[9] Omar Chehab、Aapo Hyvarinen 和 Andrej Risteski。《退火在估计归一化常数方面的可证明优势：重要性采样、噪声对比估计及其他方法》。神经信息处理系统进展，36:45945–45970，2023 年。

[10] Fan Chen、Audrey Huang、Noah Golowich、Sadhika Malladi、Adam Block、Jordan T Ash、Akshay Krishnamurthy 和 Dylan J Foster。《覆盖原则：预训练如何助力后训练》。*arXiv* 预印本 *arXiv:2510.15020*，2025 年。[11] Christophe Chipot 和 Andrew Pohorille。《自由能计算：化学与生物学中的理论与应用》。*Springer*化学物理丛书，86，2007 年。[12] Imre Csiszár 和 Paul C Shields。《信息论与统计学：教程》。Now Publishers Inc，2004 年。[13] Luc Devroye、Matthieu Lerasle、Gabor Lugosi 和 Roberto I Oliveira。《次高斯均值估计量》。统计年刊，44(6):2695–2725，2016 年。[14] Gergely Flamich 和 Lennie Wells。《关于近似信道模拟样本复杂度的一些说明》。载于 *2024*年*IEEE*信息论国际研讨会（*ISIT-W*），第 1–6 页。IEEE，2024 年。[15] Andrew Gelman 和 Xiao-Li Meng。《模拟归一化常数：从重要性采样到桥式抽样，再到路径抽样》。统计科学，13(2):163–185，1998 年。[16] John Geweke。《利用蒙特卡罗积分对计量经济模型进行贝叶斯推断》。《计量经济学》，57(6):1317–1339，1989 年。[17] Roger B Grosse、Chris J Maddison 和 Russ R Salakhutdinov。《通过对矩取平均在分布之间进行退火》。神经信息处理系统进展，26，2013 年。[18] Peter Harremoës 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。[19] 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》。载于第十三届国际学习表征会议，2025 年。[20] Audrey Huang、Adam Block、Qinghua Liu、Nan Jiang、Akshay Krishnamurthy 和 Dylan J Foster。《n选优是否是最优之选？推理时对齐中的覆盖、缩放与最优性》。载于第四十二届国际机器学习会议，2025 年。[21] Mark Huber。《吉布斯分布归一化常数的近似算法》。2015 年。[22] Mark Jerrum。《计数、采样与积分：算法与复杂性》。Birkhäuser，巴塞尔，2003 年。[23] Robert E. Kass 和 Adrian E. Raftery。《贝叶斯因子》。美国统计学会会刊，90(430):773–795，1995 年。[24] Donald E Knuth。《数学与计算机科学：应对有限性——计算能力的进步使我们大幅接近最终极限》。《科学》，194(4271): 1235–1242，1976 年。[25] Vladimir Kolmogorov。《吉布斯配分函数的更快近似算法》。载于学习理论会议，第 228–249 页。PMLR，2018 年。

[26] Yann LeCun、Sumit Chopra、Raia Hadsell、Marc'Aurelio Ranzato 和 Fu Jie Huang。基于能量的学习教程。《预测结构化数据》，1(0)，2006年。[27] Cheuk Ting Li 和 Abbas El Gamal。强函数表示引理及其在编码定理中的应用。《IEEE信息论汇刊》，64(11):6967–6978，2018年。[28] Qiang Liu、Jian Peng、Alexander Ihler 和 John Fisher III。通过判别采样估计配分函数。载于《第三十一届人工智能不确定性会议论文集》，第514–522页，2015年。[29] Fernando Llorente 和 Luca Martino。重要性采样的最优性：简明综述，2025年。网址：https://arxiv.org/abs/2502.07396。[30] Gábor Lugosi 和 Shahar Mendelson。重尾分布下的均值估计与回归：综述。《计算数学基础》，19(5):1145–1190，2019年。[31] Chris J Maddison、Daniel Tarlow 和 Tom Minka。A*采样。《神经信息处理系统进展》，27，2014年。[32] Alberto Maria Metelli、Matteo Papini、Nico Montali 和 Marcello Restelli。策略优化中的重要性采样技术。《机器学习研究杂志》，21(141):1–75，2020年。[33] Radford M. Neal。退火重要性采样。《统计与计算》，11(2):125–139，2001年。[34] Art B Owen。《蒙特卡罗理论、方法与实例》。作者自出版，2013年。[35]Yury Polyanskiy 和 Yihong Wu。《信息论：从编码到学习》。https: // people.lids.mit.edu/yp/homepage/data/itbook.pdf，2022年。草稿。[36] Doina Precup、Richard S. Sutton 和 Satinder Singh。用于离策略策略评估的资格迹。载于《第十七届国际机器学习会议论文集》，第759–766页。摩根·考夫曼，2000年。[37] 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。[38] Daniel Štefankovi˘c、Santosh Vempala 和 Eric Vigoda。自适应模拟退火：采样与计数之间的近最优联系。《ACM学报》，56(3):1–36，2009年。[39] Philip S. Thomas、Georgios Theocharous 和 Mohammad Ghavamzadeh。高置信度离策略评估。载于《第二十九届AAAI人工智能会议论文集》，第3000–3006页。AAAI出版社，2015年。[40] Roman Vershynin。《高维概率：兼论其在数据科学中的应用》。剑桥大学出版社，2018年。[41] Eric Vigoda。采样与计数。载于《近似算法与元启发式手册》。CRC出版社，第2版，2024年。另见2007年早期版本。[42] 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。

[43] Martin J. Wainwright 和 Michael I. Jordan。《图模型、指数族与变分推断》，第1卷。Now Publishers，2008年。[44] David Wingate、Andreas Stuhlmüller 和 Noah D. Goodman。通过变换式编译实现概率编程语言的轻量级实现。载于第十四届人工智能与统计国际会议论文集，第770–778页。JMLR研讨会与会议论文集，2011年。[45] Tengyang Xie、Dylan J Foster、Akshay Krishnamurthy、Corby Rosset、Ahmed Awadallah 和 Alexander Rakhlin。探索式偏好优化：利用隐式 q* 近似实现样本高效的基于人类反馈的强化学习。*arXiv*预印本 *arXiv:2405.21046*，2024年。

目录

1 引言 **1**
2 问题设定与预备知识 **3**
3 主要结果3.1 估计上界 .3.2 下界 .3.3 与采样的比较 .**4 应用：改进重要性采样的有限样本界** **5**
5 上界证明技巧5.1 样本复杂度上界 .5.2 通过覆盖进行非对称估计 .**6 讨论与相关工作** **A 技术引理**A.1 集中不等式 .**B 定理 4 的证明**B.1 估计量的选择 .B.2 样本均值分析 .**C 覆盖与 f-散度之间的关系及定理 5 的证明**C.1 覆盖与 f-散度 .C.2 定理 5 的证明 .**D 下界证明**D.1 定理 8 的证明 .D.2 定理 7 的证明 .D.3 命题 1 的证明 .D.4 命题 2 的证明 .

3
5
5
6
7
8
10
10
11
11
16
16
17
17
17
20
20
22
22
22
24
25
25

E 定理 6 的证明 **26**E.1 广义帕利-齐格蒙德不等式 .26E.2 完成定理 6 的证明 .27

F 命题 3 的证明 **29**

A 技术引理

本节汇集了证明主要结果所用的一些技术引理。我们先介绍两个集中不等式，然后证明 引理 7 ，它将覆盖与 f-散度联系起来。

A.1 集中不等式

我们首先陈述标准的切尔诺夫界（例如参见 [4]）。

引理 2 （切尔诺夫界）。设 $X_1, \dots, X_m$ 为独立随机变量，使得对每个 $i \in [m]$，几乎必然有 $X_i \in [0, M]$，且 $\mu = \mathbb{E}[X_i]$。那么对任意 $1 \geq t > 0$，如果成立，则有

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m} X_i \leq (1-t)\mu\right) \leq e^{-\frac{m\mu t^2}{2M}}.$$

此外，对于任意 $t > 0$，有

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m} X_i \geq (1+t)\mu\right) \leq \exp\left(-m\mu/M \cdot ((1+t)\log(1+t) - t)\right).$$

由于我们通常考虑 $t \lesssim \varepsilon \ll 1$，因此常利用 $(1+t)\log(1+t) - t \geq t^2/3$ for $t \in (0,1)$ 这一事实来简化上尾界，从而得到

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right| \geq t\mu\right) \leq 2\exp\left(-m\mu t^2/3M\right)$$

对于 $t \in (0,1)$。

我们还会用到经典的伯恩斯坦不等式（例如参见 [4]），为完整起见，现将其陈述如下。

引理 3 （伯恩斯坦不等式）。设 $\xi_1, \dots, \xi_m$ 为独立随机变量，且对每个 $i \in [m]$，以下性质均成立：

1. 这些随机变量均以零为中心，即 $\mathbb{E}[\xi_i] = 0$。
2. 这些随机变量几乎必然有界，即 $|\xi_i| \leq M$。
3. 这些随机变量的方差至多为 σ^2，即 $\mathbb{E}[\xi_i^2] - \mathbb{E}[\xi_i]^2 \leq \sigma^2$。

则对任意 $0 < \delta < 1$，以至少 $1 - \delta$的概率成立，

$$\left|\frac{1}{m}\sum_{i=1}^{m} \xi_i\right| \leq \sqrt{\frac{2\sigma^2 \log(2/\delta)}{m}} + \frac{2M\log(2/\delta)}{3m}.$$

B 证明定理 4

B.1 估计量的选择

在本节中，我们定义将用于证明上界的估计量。回顾 k个样本 $Y_1, \ldots, Y_k$ 的中位数定义为 $y = \text{Med}(\{Y_1, \ldots, Y_k\})$，使得 $y \in \{Y_1, \ldots, Y_k\}$ 满足

$$|\{j \in [k] | Y_j \leq y\}| \geq \frac{k}{2} \quad \text{and} \quad |\{j \in [k] | Y_j \geq y\}| \geq \frac{k}{2}.$$

我们将使用均值中位数估计量来证明上界。令 $k = \lceil 8\log(1/\delta) \rceil$，并将$n$ 个样本划分为 k 组，每组大小为 $m = \lfloor n/k \rfloor$。回顾一下，我们定义估计量的方式是：对于每个组$j \in [k]$，令

$$\widehat{Z}_j = \frac{1}{m} \sum_{i=1}^{m} \lambda(X_i^{(j)}) \quad \text{and} \quad \widehat{Z} = \text{Med}\left(\widehat{Z}_1, \ldots, \widehat{Z}_k\right).$$

我们将使用的关于中位数的关键事实是以下标准引理，它允许我们使用均值中位数法将常数概率界转换为高概率界。该引理通常用于具有有界方差的估计问题中，但为完整起见，我们在此以更一般的形式陈述它。

引理 4。令 Y_1、……、Y_k 为独立随机变量，使得对于每个 $i \in [k]$，都有$\mathbb{P}(Y_i \in (a, b)) \geq 1/2 + c$，其中某个 $c > 0$成立。则有

$$\mathbb{P}(\text{Med}(Y_1, \ldots, Y_k) \in (a, b)) \geq 1 - \exp\left(-2kc^2\right).$$

证明。令 $y = \text{Med}(\{Y_1, \ldots, Y_k\})$，并令 $X_i = \mathbb{I}[Y_i \in (a, b)]$。注意，对所有 $i \in [k]$，$\mathbb{E}[X_i] \geq 1/2 + c$，并令 $S = \sum_{i=1}^{k} X_i$。注意，若 $y \notin (a, b)$，则必有 $S \leq k/2$。因此我们得到

$$\mathbb{P}(y \notin (a, b)) \leq \mathbb{P}(S \leq k/2) \leq \mathbb{P}(\text{Bin}(k/2 + c, k) \leq k/2) \leq e^{-2kc^2}$$

其中最后一个不等式由标准的切尔诺夫界（引理 2）得出。 □

B.2 样本均值分析

我们的上界证明将通过证明 Z_j 以 ε-接近于 Z，且对每个$j \in [k]$都以常数概率成立，然后应用引理 4得出结论。此外，由于我们的估计量是尺度不变的，且保证是乘法形式，因此不失一般性，我们可以分析 $Z = 1$ 的情形。

我们需要的第一个关键引理涉及用积分覆盖来控制截断版本的密度比$\frac{d\nu}{d\mu}$ 的方差。

—

引理 5. 对于任意两个概率测度 μ、ν，有

$$\mathbb{E}_{X \sim \mu}\left[\left(\frac{d\nu}{d\mu}(X)\right)^2 \mathbb{I}\left[\frac{d\nu}{d\mu}(X) \leq M\right]\right] \leq \text{ICov}_M(\nu \| \mu)$$

证明. 我们采用标准的尾部积分论证。事实上，有

$$\begin{aligned}
\mathbb{E}_{X\sim\mu}\left[\left(\frac{d\nu}{d\mu}(X)\right)^2 \mathbb{I}\left[\frac{d\nu}{d\mu}(X)\le M\right]\right] &= \mathbb{E}_{X\sim\nu}\left[\frac{d\nu}{d\mu}(Y)\mathbb{I}\left[\frac{d\nu}{d\mu}(Y)\le M\right]\right] \\
&= \int_0^\infty \mathbb{P}_{Y\sim\nu}\left(\frac{d\nu}{d\mu}(Y)\cdot\mathbb{I}\left[\frac{d\nu}{d\mu}(Y)\le M\right]\ge t\right)dt \\
&= \int_0^M \mathbb{P}_{Y\sim\nu}\left(t\le\frac{d\nu}{d\mu}(Y)\le M\right)dt \\
&= \int_0^M \left(\mathrm{Cov}_t\left(\nu\|\mu\right)-\mathrm{Cov}_M\left(\nu\|\mu\right)\right)dt \\
&= \int_0^M \mathrm{Cov}_t\left(\nu\|\mu\right)dt - M\cdot\mathrm{Cov}_M\left(\nu\|\mu\right).
\end{aligned}$$

该结果由此得出。

□

接下来，我们将基于积分覆盖证明关于密度比经验平均的集中性结果。关键思路是注意到，尽管密度比的方差可能无界，我们仍可以分析密度比的截断版本来控制方差。

引理6 (基于积分覆盖的样本均值集中)。 对于任意两个概率测度 μ,ν，令 $X_1,\ldots,X_m$ 为来自 μ的独立同分布样本。假设$M_{\varepsilon,\delta}>0$ 定义为满足

$$\mathrm{ICov}_{M_{\varepsilon,\delta}}\left(\nu\|\mu\right)\le\varepsilon\delta M_{\varepsilon,\delta},$$

按照约定，若不存在这样的有限值，则$M_{\varepsilon,\delta}=\infty$ 。那么对于任意 $0<\varepsilon,\delta<1$，只要

$$m\ge\frac{M_{\varepsilon,\delta}\cdot\log(1/\delta)}{\varepsilon},$$

那么，以至少 $1-\delta$ 的概率，有

$$1-\varepsilon\le\frac{1}{m}\sum_{i=1}^m\frac{d\nu}{d\mu}(X_i)\le 1+\varepsilon.$$

证明。 固定一个 M，其取值稍后确定，并令 $\eta_i=\frac{d\nu}{d\mu}(X_i)\cdot\mathbb{I}\left[\frac{d\nu}{d\mu}(X_i)\le M\right]$，其中 $i\in[m]$。显然，$\mathbb{E}\left[\eta_i\right]\le 1$。此外，根据定义1，我们有

$$\mathbb{E}\left[\eta_i\right]\ge 1-\mathrm{Cov}_M\left(\nu\|\mu\right) \tag{9}$$

根据引理 5，可得

$$\mathbb{E}\left[\eta_i^2\right]\le\mathrm{ICov}_M\left(\nu\|\mu\right)$$

且显然 $0\le\eta_i\le M$ 几乎必然。因此，根据引理 3，以至少 $1-\delta$，

$$\left|\frac{1}{m}\sum_{i=1}^m\eta_i-\mathbb{E}\left[\eta_i\right]\right|\le\sqrt{\frac{2\mathrm{ICov}_M\left(\nu\|\mu\right)\log\left(2/\delta\right)}{m}}+\frac{2M\log\left(2/\delta\right)}{3m}. \tag{10}$$

现在我们分别分析上下尾。

下尾。 结合（10）和（9），我们有至少以 $1-\delta$ 的概率，

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\frac{d\nu}{d\mu}(X_i) &\geq \frac{1}{m}\sum_{i=1}^{m}\eta_i \\ &\geq 1-\mathrm{Cov}_M(\nu\|\mu)-\sqrt{\frac{2\mathrm{ICov}_M(\nu\|\mu)\log(2/\delta)}{m}}-\frac{2M\log(2/\delta)}{3m}.\end{aligned}$$

应用杨氏不等式（即对于 $a,b,\lambda\geq 0$，有 $\sqrt{ab}\leq a/\lambda+\lambda b$），可得

$$\sqrt{\frac{2\mathrm{ICov}_M(\nu\|\mu)\log(2/\delta)}{m}}\leq\frac{\mathrm{ICov}_M(\nu\|\mu)}{M}+\frac{M\log(2/\delta)}{2m}.$$

这表明，以至少 $1-\delta$ 的概率，

$$\frac{1}{m}\sum_{i=1}^{m}\frac{d\nu}{d\mu}(X_i)\geq 1-\mathrm{Cov}_M(\nu\|\mu)-\frac{\mathrm{ICov}_M(\nu\|\mu)}{M}-\frac{2M\log(2/\delta)}{3m}.$$

令 M，使得 $\mathrm{Cov}_M(\nu\|\mu)\leq\varepsilon/3$、$\mathrm{ICov}_M(\nu\|\mu)\leq\varepsilon M/3$，并令 m，使得 $\frac{2M\log(2/\delta)}{3m}\leq\varepsilon/3$，则以至少 $1-\delta$ 的概率，

$$\frac{1}{m}\sum_{i=1}^{m}\frac{d\nu}{d\mu}(X_i)\geq 1-\varepsilon.$$

上尾。 我们将样本均值分解为

$$\frac{1}{m}\sum_{i=1}^{m}\frac{d\nu}{d\mu}(X_i)=\frac{1}{m}\sum_{i=1}^{m}\eta_i+\frac{1}{m}\sum_{i=1}^{m}\frac{d\nu}{d\mu}(X_i)\cdot\mathbb{I}\left[\frac{d\nu}{d\mu}(X_i)>M\right]$$

并分别分析每一项。根据马尔可夫不等式，有

$$\begin{aligned}\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m}\frac{d\nu}{d\mu}(X_i)\cdot\mathbb{I}\left[\frac{d\nu}{d\mu}(X_i)>M\right]>t\right)&\leq\frac{\mathbb{E}\left[\frac{d\nu}{d\mu}(X_i)\cdot\mathbb{I}\left[\frac{d\nu}{d\mu}(X_i)>M\right]\right]}{t}\\ &\leq\frac{\mathrm{Cov}_M(\nu\|\mu)}{t}.\end{aligned}$$

因此，只要 $\mathrm{Cov}_M(\nu\|\mu)\leq\frac{\varepsilon\delta}{4}$，就有以至少 $1-\delta/2$ 的概率，

$$\frac{1}{m}\sum_{i=1}^{m}\frac{d\nu}{d\mu}(X_i)\cdot\mathbb{I}\left[\frac{d\nu}{d\mu}(X_i)>M\right]\leq\frac{\varepsilon}{2}.$$

我们应用引理 3 再次t o控制第一项。事实上，以下结论以至少 $1-\delta/3$ 的概率成立，即

$$\frac{1}{m}\sum_{i=1}^{m}\xi_i\leq 1-\mathrm{Cov}_M(\nu\|\mu)+\sqrt{\frac{2\mathrm{ICov}_M(\nu\|\mu)\log(2/\delta)}{m}}+\frac{2M\log(2/\delta)}{3m}$$

和之前一样，应用杨氏不等式可得，以至少 $1-\delta/3$ 的概率，

$$\frac{1}{m}\sum_{i=1}^{m}\xi_i\leq 1-\mathrm{Cov}_M(\nu\|\mu)+\frac{\mathrm{ICov}_M(\nu\|\mu)}{M}+\frac{7M\log(3/\delta)}{6m}.$$

令 M，使得 $\mathrm{Cov}_M(\nu\|\mu) \le \frac{\varepsilon}{2}$，$\mathrm{ICov}_M(\nu\|\mu) \le \frac{\varepsilon M}{2}$，并令 m，使得 $\frac{7M\log(3/\delta)}{6m} \le \frac{\varepsilon}{2}$，则以至少 $1-\delta/3$ 的概率，

$$\frac{1}{m}\sum_{i=1}^{m}\xi_i \le 1+\varepsilon.$$

结合上尾的两部分并应用并集界，即可得到所需结果。 □

定理 4 的证明。由前一个引理可知，对于每个 $j \in [k]$，只要

$$m \ge \frac{M_{\varepsilon,1/4}\cdot\log(4)}{\varepsilon/2},$$

以至少 $3/4$ 的概率，

$$1-\varepsilon \le \widehat{Z}_j \le 1+\varepsilon.$$

应用引理 4，其中 $c=1/4$ 和 $k=\lceil 8\log(1/\delta)\rceil$ 可得，以至少 $1-\delta$ 的概率，

$$1-\varepsilon \le \widehat{Z} \le 1+\varepsilon$$

符合要求。 □

C 覆盖与 f-散度的关系及定理5的证明

在本附录中，我们建立覆盖与 f-散度之间的若干关系，并据此证明定理 5 （见C.2节）。我们从这些关系入手，最终得到一个关键关系（引理 9），该关系用于引理 10的证明，而后者又是证明我们的上界所必需的。

C.1 覆盖与 f-散度

我们首先陈述覆盖与 f-散度之间的一个基本关系，该关系将有助于证明以 f-散度表示的样本复杂度上界。

引理 7. 对于任意两个概率测度 μ、ν （定义在 $\mathcal{X}$ 上）以及任意凸函数 f （如 定义 2所述），对所有 $M>1$ 均成立：

$$\mathrm{Cov}_M(\nu\|\mu) \le \frac{M\cdot \mathrm{D}_f(\nu\|\mu)}{f(M)}.$$

因此，若 $M \ge \gamma_f\left(\mathrm{D}_{f(\nu\|\mu)}/\varepsilon\right)$ 对某个 $\varepsilon\in(0,1)$成立，则$\mathrm{Cov}_M(\nu\|\mu) \le \varepsilon$。

证明。 根据覆盖的定义，我们有

$$\begin{aligned}
\mathrm{Cov}_M(\nu\|\mu) &= \mathbb{P}_{X\sim\nu}\left(\frac{d\nu}{d\mu}(X)\ge M\right) = \mathbb{E}_{X\sim\mu}\left[\frac{d\nu}{d\mu}(X)\cdot\mathbb{I}\left[\frac{d\nu}{d\mu}(X)\ge M\right]\right] \\
&\le \mathbb{E}\left[\left(1+\frac{M-1}{f(M)}\cdot f\left(\frac{d\nu}{d\mu}(X)\right)\right)\cdot\mathbb{I}\left[\frac{d\nu}{d\mu}(X)\ge M\right]\right] \\
&\le \mathbb{P}_{X\sim\mu}\left(\frac{d\nu}{d\mu}(X)\ge M\right) + \frac{M-1}{f(M)}\cdot\mathrm{D}_f(\nu\|\mu),
\end{aligned}$$

其中，第一个不等式由 引理 8 在 $x = \frac{d\nu}{d\mu}(X)$对 $X \sim \mu$成立得出，第二个不等式则由 (f-散度在 (1) 中的定义以及 $f \geq 0$这一事实得出。最后，我们注意到，对于 $M \geq 1$，有 f 单调递增，因此可以应用马尔可夫不等式，得到

$$\mathbb{P}_{X\sim\mu}\left(\frac{d\nu}{d\mu}(X) \geq M\right) \leq \mathbb{P}_{X\sim\mu}\left(f\left(\frac{d\nu}{d\mu}(X)\right) \geq f(M)\right) \leq \frac{\mathbb{E}_{X\sim\mu}\left[f\left(\frac{d\nu}{d\mu}(X)\right)\right]}{f(M)} \leq \frac{\mathrm{D}_f\left(\nu\|\mu\right)}{f(M)}.$$

将此结果与前一个界相加，即可得到所需结论。 □

引理 8。 设 f： $\mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$是一个凸函数，使得 $f(1) = f'(1) = 0$，并设 $M \geq 1$。则对于任意 $x \geq M$，都有

$$x \leq 1 + \frac{M-1}{f(M)} \cdot f(x).$$

证明。 设 $t = \frac{M-1}{x-1}$；根据假设 $1 < M \leq x$，我们有 $t \in [0, 1)$。注意到 $M = t \cdot 1 + (1-t) \cdot x$，因此由 f 的凸性可得

$$f(M) \leq t \cdot f(1) + (1-t) \cdot f(x) = (1-t) \cdot f(x) = \frac{M-1}{x-1} \cdot f(x).$$

对上述不等式重新排列即可得到该结论。 □

现在我们证明截断密度比的二阶矩与两个测度之间的 f-散度之间的一个关键关系。

引理 9。 设 f 是一个凸函数，如 定义 2 中所述，且存在某个 $c > 1$，使得对于所有$M \geq t \geq c$均有

$$\frac{f(M)}{M^2} \leq \frac{f(t)}{t^2}.$$

那么，对于 $\mathcal{X}$ 上的任意概率测度 μ、ν 以及任意 $M \geq c$，有

$$\frac{\mathrm{ICov}_M\left(\nu\|\mu\right)}{M} \leq \frac{c^2}{M} + \frac{M \cdot \mathrm{D}_f\left(\nu\|\mu\right)}{f(M)}.$$

证明。 首先注意，如果 $f(M)/M^2 \leq f(t)/t^2$，那么

$$t^2 \leq \frac{M^2 \cdot f(t)}{f(M)}. \tag{11}$$

我们计算：

$$\begin{aligned}
\mathbb{E}_{X\sim\mu}\left[\left(\frac{d\nu}{d\mu}(X) \cdot \mathbb{I}\left[\frac{d\nu}{d\mu}(X) \leq M\right]\right)^2\right] &= \mathbb{E}_{X\sim\mu}\left[\left(\frac{d\nu}{d\mu}(X) \cdot \mathbb{I}\left[\frac{d\nu}{d\mu}(X) \leq M\right]\right)^2 \mathbb{I}\left[\frac{d\nu}{d\mu}(X) \leq c\right]\right] \\
&\quad + \mathbb{E}_{X\sim\mu}\left[\left(\frac{d\nu}{d\mu}(X) \cdot \mathbb{I}\left[\frac{d\nu}{d\mu}(X) \leq M\right]\right)^2 \mathbb{I}\left[M \geq \frac{d\nu}{d\mu}(X) > c\right]\right] \\
&\leq c^2 + \mathbb{E}_{X\sim\mu}\left[\frac{M^2 \cdot f\left(\frac{d\nu}{d\mu}(X)\right)}{f(M)} \mathbb{I}\left[M \geq \frac{d\nu}{d\mu}(X) > c\right]\right] \\
&\leq c^2 + \frac{2M^2 \cdot \mathrm{D}_f\left(\nu\|\mu\right)}{f(M)},
\end{aligned}$$

其中第一个不等式由（11）得出，第二个不等式则由被积函数的非负性得出。

□

注意，该引理将我们上界中使用的关键量$\mathrm{ICov}_M(\nu\|\mu)/M$与两个测度之间的$f$-散度联系起来。特别地，它表明，若选取足够大的 M，使得 $M\cdot \mathrm{D}_f(\nu\|\mu)/f(M)$很小，则$\mathrm{ICov}_M(\nu\|\mu)/M$也很小。现在，我们已准备好证明以 f-散度表示的主要上界。

C.2 定理5的证明

与前文一样，我们将使用式（8）中定义的中位数均值估计量，因此只需理解样本均值估计量在 f-散度下的行为。

引理 10（基于 f-散度的样本均值集中）。对于任意两个概率测度 μ，ν，令X_1、……、X_m为来自 μ的独立同分布样本。设 f为定义 *2* 中所述的凸函数，且存在某个 $c>1$ 使得对于所有 $M\geq t\geq c$，均有

$$\frac{f(M)}{M^2}\leq\frac{f(t)}{t^2}.$$

那么，对于任意 $0<\varepsilon$、$\delta<1$，只要

$$m\geq\frac{\gamma_f\left(\mathrm{D}_f(\nu\|\mu)/\varepsilon\delta\right)\cdot\log(1/\delta)}{\varepsilon}\vee\frac{c^2\cdot\log(1/\delta)}{\varepsilon^2\delta},$$

那么，以至少 $1-\delta$ 的概率，有

$$1-\varepsilon\leq\frac{1}{m}\sum_{i=1}^{m}\frac{d\nu}{d\mu}(X_i)\leq 1+\varepsilon.$$

证明。第一步是以下引理，它将覆盖与 f-散度联系起来；其证明推迟到附录C。

要应用引理6，我们需要选择 M，使得积分覆盖 $\mathrm{ICov}_M(\nu\|\mu)\leq\varepsilon\delta M$。由引理 9 可知，我们需要 $c^2/M+M\cdot\mathrm{D}_f(\nu\|\mu)/f(M)\leq\varepsilon\delta$。回顾式 (2) 中 γ_f 的定义，若设 $M=\gamma_f(\mathrm{D}_f(\nu\|\mu)/\varepsilon\delta)$，则根据 γ_f 的定义，有 $M\cdot\mathrm{D}_f(\nu\|\mu)/f(M)\leq\varepsilon\delta$。因此，令 $M\geq\max\left\{c^2/\varepsilon\delta,\gamma_f\left(\mathrm{D}_f(\nu\|\mu)/\varepsilon\delta\right)\right\}$，再应用引理6，即可完成证明。 □

D 下界结果的证明

在本附录中，我们给出下界结果的完整证明。我们首先证明最有趣的情形，即定理 8 （D.1节）中针对超线性但次二次 f-散度的一般下界。接着，我们给出关于积分覆盖的下界（定理 7），见D.2节，然后证明命题1 （D.3节）中的线性下界；最后，我们证明命题2（D.4节）中的超二次下界。

D.1 定理8

我们首先概述构造过程，然后证明该命题。令 μ 为任意测度，并考虑由 $p\in[0,1]$ 和 $\varepsilon\in(0,1/4)$ 参数化的一族分布 $\nu_{p,\varepsilon}$，其定义满足

$$\frac{d\nu_{p,\varepsilon}}{d\mu}(X)=\begin{cases}1-\varepsilon & \text{w.p. } 1-p,\\ 1+\varepsilon\left(\frac{1}{p}-1\right) & \text{w.p. } p.\end{cases}\tag{12}$$

我们令 $\mathcal{V} \subset \{\nu_{p,\varepsilon}\}$ 满足：对于所有 $\nu \in \mathcal{V}$，均有 $\mathrm{D}_f(\nu\|\mu) \leq C + f(1-\varepsilon)$。为此，我们首先计算 $\nu_{p,\varepsilon}$ 与 μ之间的 f-散度。

引理 11. 假设

$$\gamma_f\left(\frac{C}{2\varepsilon}\right) \geq 2.$$

那么，当.....时，有 $\mathrm{D}_f(\nu_{p,\varepsilon}\|\mu) \leq C + f(1-\varepsilon)$

$$p \geq \frac{2\cdot\varepsilon}{\gamma_f\left(\frac{C}{2\cdot\varepsilon}\right)}.$$

证明. 根据 f-散度的定义以及（12），我们有

$$\begin{aligned}\mathrm{D}_f(\nu\|\mu) &= (1-p)\cdot f(1-\varepsilon) + p\cdot f\left(1+\varepsilon\left(\frac{1}{p}-1\right)\right)\\ &\leq f(1-\varepsilon) + p\cdot f\left(1-\varepsilon+\frac{\varepsilon}{p}\right).\end{aligned}$$

因此，只要

$$p\cdot f\left(1-\varepsilon+\frac{\varepsilon}{p}\right) \leq C, \tag{13}$$

就满足 f-散度条件。

我们首先断言，若（13）对 p成立，则对所有 $p < q \leq 1$也成立。为此，注意到对于这样的q,

$$\begin{aligned}q\cdot f\left(1-\varepsilon+\frac{\varepsilon}{q}\right) &= (q(1-\varepsilon)+\varepsilon)\cdot\frac{f\left(\frac{q(1-\varepsilon)+\varepsilon}{q}\right)}{\frac{q(1-\varepsilon)+\varepsilon}{q}}\\ &\leq (p(1-\varepsilon)+\varepsilon)\cdot\frac{f\left(\frac{p(1-\varepsilon)+\varepsilon}{p}\right)}{\frac{p(1-\varepsilon)+\varepsilon}{p}}\\ &= p\cdot f\left(1-\varepsilon+\frac{\varepsilon}{p}\right) \leq C,\end{aligned}$$

其中，第一个不等式是由以下事实得出的： $t \mapsto {}^{f(t)}\!/_{t}$关于 $t \geq 1$ 单调不减。现在注意到，

(13) 当且仅当.....时成立 y if

$$\frac{f\left(\frac{p(1-\varepsilon)+\varepsilon}{p}\right)}{\frac{p(1-\varepsilon)+\varepsilon}{p}} \leq \frac{C}{p(1-\varepsilon)+\varepsilon}.$$

根据 γ_f 函数在（2）中的定义，只要

$$\frac{p(1-\varepsilon)+\varepsilon}{p} \leq \gamma_f\left(\frac{C}{p(1-\varepsilon)+\varepsilon}\right),$$

重新整理后，只要满足以下条件即可

$$p \geq \frac{\varepsilon}{\gamma_f\left(\frac{C}{p(1-\varepsilon)+\varepsilon}\right)-1}.$$

根据假设 $\gamma_f\left(\frac{C}{2\varepsilon}\right) \geq 2$，最后这一点可由 $p \geq \gamma_f^{2\cdot\varepsilon}\left(\frac{C}{p\ 1-\varepsilon()+\varepsilon}\right)$ 推出。注意，对于 $p \leq \varepsilon$，成立

$$\frac{2\cdot\varepsilon}{\gamma_f\left(\frac{C}{p(1-\varepsilon)+\varepsilon}\right)} \leq \frac{2\cdot\varepsilon}{\gamma_f\left(\frac{C}{2\cdot\varepsilon}\right)},$$

从而结论成立。 □

因此，对于固定的 ε，令

$$\mathcal{V} = \left\{ \nu_{p,\varepsilon'} \middle| \frac{2\varepsilon'}{\gamma_f\left(\frac{C}{2\varepsilon'}\right)} \leq p \leq 1,\ 0 \leq \varepsilon' \leq \frac{1}{4} \right\} \tag{14}$$

现在我们可以完成该命题的证明。

证明定理 8。 设 μ 为任意概率测度，并设 $\mathcal{V}$ 如 (14) 中所定义。由 引理 11，对所有 $\nu \in \mathcal{V}$，我们有 $\mathrm{D}_f(\nu\|\mu) \leq C + f(1-\varepsilon)$。

注意，对于任意固定的 p，为了区分分布 $\nu_0 = \mu$ 与 $\nu_{p,\ \varepsilon}$，并以至少 $2/3$ 的概率做到这一点，必须至少观测到一个来自 $\nu_{p,\ \varepsilon}$ 的高密度区域的样本，该事件以概率 p 在 μ 下发生。另一方面，在 n 个样本之后仅观测到低密度样本的概率为

$$(1-p)^n \geq e^{-2np}.$$

因此，对于 $n \leq \log(3/2)/2p$，任何估计量都无法以至少 $2/3$ 的概率区分 ν_0 和 $\nu_{p,\ \varepsilon}$。另一方面，在前一种情况下，归一化常数满足 $Z = \lambda(X_1)$；而在后一种情况下，归一化常数满足 $Z = \lambda(X_1)/1-\varepsilon$，因此，任何能以至少 $2/3$ 的概率达到 $(1-\varepsilon)Z \leq Z \leq (1+\varepsilon)Z$ 的估计量，都必须能够区分这两个分布。令 $p = \gamma_f^{2\cdot\varepsilon}\left(\frac{C}{2\cdot\varepsilon}\right)$ 即得结果。 □

D.2 定理7的证明

我们采用与定理 8的证明相同的构造。事实上，令$\nu_{p,\ \varepsilon}$ 如（12）中所定义。注意，对于任意 $\nu_{p,\ \varepsilon}$，都有

$$\mathrm{Cov}_M\left(\nu_{p,\varepsilon}\|\mu\right) = \begin{cases} 1 & M < 1-\varepsilon \\ p & 1-\varepsilon \leq M < 1+\varepsilon\left(\frac{1}{p}-1\right) \\ 0 & M \geq 1+\varepsilon\left(\frac{1}{p}-1\right) \end{cases}$$

因此有

$$\mathrm{ICov}_M\left(\nu_{p,\varepsilon}\|\mu\right) = \begin{cases} M & M < 1-\varepsilon \\ (1-\varepsilon)+p\cdot(M-(1-\varepsilon)) & 1-\varepsilon \leq M < 1+\varepsilon\left(\frac{1}{p}-1\right) \\ 1+\varepsilon\left(\frac{1}{p}-1\right) & M \geq 1+\varepsilon\left(\frac{1}{p}-1\right) \end{cases}.$$

按照定理 8的证明中相同的论证，要以至少 的概率达到 () 乘法精度，就必须能够以至少 的概率区分 和 ，ε，这至少需要 $\log(3/2)/2p$ 个样本。令 $p = \varepsilon^2/1+\varepsilon$，我们声称，对于 $M = \varepsilon/p$，有 $\mathrm{ICov}_M(\nu_{p,\varepsilon}\|\mu) \le M \cdot \varepsilon$。事实上，注意到对于这样的 M，显然处于上述中间区域，因此

$$\frac{\mathrm{ICov}_M(\nu\|\mu)}{M} = \frac{(1-\varepsilon)(1-p)}{M} + p \le \frac{p}{4\cdot\varepsilon} + p \le \varepsilon.$$

另一方面，我们已经证明，要达到所需的精度，至少需要$\log(3/2)/2p = \Theta(M\varepsilon/\varepsilon)$个样本。由此得证。■

D.3 命题1的证明

我们证明命题 1 的一个略作重述的版本。

命题 4. 设 f 是定义 2 中的凸函数，且映射 $t \mapsto f(t)/t$ 一致有界。对于任意 $0 < C < f'(\infty)$，存在测度 μ、ν，使得 $D_f(\nu\|\mu) \le C$，并且对于任意 $\varepsilon \le f'(\infty)/(f'(\infty)-C)$，任何算法即使获得来自 μ 的任意有限个独立样本，并可预言机访问 $\frac{d\nu}{d\mu}(\cdot)\cdot Z$，也无法产生满足对 Z 的 $1\pm\varepsilon$ 近似的 Z 估计值，且其成功概率至少为 $3/4$。

此外，对于任意 $C \ge f'(\infty)$，存在测度 μ、ν，使得 $\mathrm{D}_f(\nu\|\mu) \le C$，并且对于任意 $\varepsilon > 0$，任何算法即使获得来自 μ 的任意有限个独立样本，并可预言机访问 $\frac{d\nu}{d\mu}(\cdot)\cdot Z$，也无法产生满足对 Z 的 $1\pm\varepsilon$ 近似的估计值 Z，且其成功概率至少为 $3/4$。

证明。 令 $\mu = \delta(0)$ 和 $\nu = \mathrm{Ber}(1-q)$。则有

$$\mathrm{D}_f(\nu\|\mu) = qf'(\infty).$$

因此，对于任意有限的 $C > 0$，选择 $q \le C/f'(\infty)$ 可确保 $\mathrm{D}_f(\nu\|\mu) \le C$。然而，由于 μ 是位于 0 的点质量，从 μ中抽取任意有限数量的样本，都无法揭示有关 q的任何信息。

在 $C \le f'(\infty)$ 的情况下，我们有

$$1 - \frac{C}{f'(\infty)} \le q \le 1.$$

因此， Z 的最紧常数因子近似至少为

$$\frac{f'(\infty)}{f'(\infty)-C}.$$

在 $C > f'(\infty)$ 的情况下，我们可以任意选择 $q \in [0,1]$，因此从 μ中抽取任意有限数量的样本，都无法得到 Z的任何常数因子近似。

□

D.4 命题2的证明

令 $\mathcal{X} = \{0,1\}$ 满足 $\mu_p(0) = 1-p$ 且 $\mu_p(1) = p$；令 $\mathcal{U} = \{\mu_p : 1/4 \le p \le 1/2\}$。令 $\nu = \delta_1$，于是

$$\frac{d\nu}{d\mu}(X) = \begin{cases} 0 & X = 0 \\ \frac{1}{p} & X = 1 \end{cases}.$$

因此，根据 p的下界，对所有 $\mu \in \mathcal{U}$都有 $\left\|\frac{d\nu}{d\mu}\right\|_\infty \le 4$。令 $\lambda(X) = \mathbb{I}[X=1]$为相对于 μ_p的 ν 的未归一化密度比，并注意到，估计 Z等价于估计 p。令 $p = 1/2$和 $p' = 1/2 - 2\varepsilon$，并考虑两个测度 $\mu = \mu_p$和 $\mu' = \mu_{p'}$。为了以至少 $2/3$的概率，在 μ和 μ'下都将 Z估计到因子 $(1 \pm \varepsilon)$以内，必须以至少 $1/3$的概率区分这两个分布。根据标准的假设检验下界，例如 Le Cam 方法（参见 Wainwright [42]），错误概率的下界为 $1/2\left(1 - \mathrm{TV}\left(\mu^{\otimes n}, \mu'^{\otimes n}\right)\right)$。因此，我们要求 $\mathrm{TV}\left(\mu^{\otimes n}, \mu'^{\otimes n}\right) \ge 1/3$。另一方面，根据 Pinsker 不等式和 KL 散度的链式法则，有

$$\mathrm{TV}\left(\mu^{\otimes n}, \mu'^{\otimes n}\right) \le \sqrt{\frac{n}{2} \cdot \mathrm{D}_{\mathrm{KL}}\left(\mu \| \mu'\right)}.$$

因此，若我们能证明以下结论，则该结果成立。

$$\mathrm{D}_{\mathrm{KL}}\left(\mu \| \mu'\right) \lesssim \varepsilon^2.$$

我们计算得到

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}\left(\mu \| \mu'\right) &= \frac{1}{2} \cdot \left(\log\left(\frac{1/2}{1/2 - 2\varepsilon}\right) + \log\left(\frac{1/2}{1/2 + 2\varepsilon}\right)\right) \\
&= \frac{1}{2} \cdot \log\left(\frac{1}{1 - 16\varepsilon^2}\right) \\
&\le 8\varepsilon^2,
\end{aligned}$$

其中最后一个不等式源于如下事实：对所有 $x \in (0, 1/2)$，均有 $\log\left(1/1-x\right) \le 2x$。至此证明完成。■

E 定理 6的证明

在本附录中，我们证明定理 6。我们首先证明广义 Paley-Zygmund 不等式，然后应用该不等式来表明，对于适当选取的 α，样本的经验$(1-\alpha)$-分位数以高概率具有所述性能保证。

E.1 广义帕利-齐格蒙德不等式

我们先给出一个比 引理 1 中的广义 Paley-Zygmund 不等式稍强的结果。

引理 12. 设 μ、ν 是 $\mathcal{X}$ 上的概率测度。那么对于任意 $0 < \varepsilon < 1$，对任意 $0 < u < 1$，都有

$$\mathbb{P}_{X \sim \mu}\left(\frac{d\nu}{d\mu}(X) \ge 1 - \varepsilon\right) \ge \frac{(1-u)\cdot\varepsilon}{M} \quad \textit{where } \mathrm{Cov}_M\left(\nu \| \mu\right) \le u \cdot \varepsilon.$$

证明。 令 $Y = \frac{d\nu}{d\mu}(X)$ 对于 $X \sim \mu$。根据定义，我们有 $\mathbb{E}[Y] = 1$，因此

$$1 = \mathbb{E}[Y] \le 1 - \varepsilon + \mathbb{E}\left[Y \cdot \mathbb{I}[Y \ge 1 - \varepsilon]\right].$$

整理后，对于任意 M，我们有

$$\begin{aligned}\varepsilon &\leq \mathbb{E}\left[Y \cdot \mathbb{I}\left[Y \geq 1-\varepsilon\right]\right] \\ &= \mathbb{E}\left[Y \cdot \mathbb{I}\left[1-\varepsilon \leq Y < M\right]\right] + \mathbb{E}\left[Y \cdot \mathbb{I}\left[Y \geq M\right]\right] \\ &\leq M \cdot \mathbb{P}\left(Y \geq 1-\varepsilon\right) + \mathbb{E}\left[Y \cdot \mathbb{I}\left[Y \geq M\right]\right] \\ &= M \cdot \mathbb{P}\left(Y \geq 1-\varepsilon\right) + \mathrm{Cov}_M(\nu\|\mu).\end{aligned}$$

通过整理并利用假设 $\mathrm{Cov}_M(\nu\|\mu) \leq u \cdot \varepsilon$，即可得到该结果。 □

现在我们通过应用引理 1，从引理 12 推导出引理 7。

引理 *1* 的证明。陈述的第一部分就是引理 12。对于第二部分，根据引理 7，对

$$M \geq \gamma_f\left(\frac{\mathrm{D}_f\left(\nu\|\mu\right)}{u \cdot \varepsilon}\right),$$

我们有协方差$_M(\nu\|\mu) \leq u \cdot \varepsilon$。将其代入引理 12 并对$u$进行优化，即可得到该结果。

□

E.2 补全定理6的证明

我们将结果连同常数项一起重述如下。

定理 11。 设 μ、ν 是 $\mathcal{X}$ 上的概率测度，并假设 X_1、……、$X_n \sim \mu$ 相互独立，且 $\curlywedge\lambda$ 是未归一化密度比。那么存在一个估计量 Z，使得对于任意 $0 < \varepsilon < 1$，若

$$n \geq \frac{18M}{\varepsilon} \cdot \log\left(2/\delta\right) \quad \textit{where } \mathrm{Cov}_M\left(\nu\|\mu\right) \leq \varepsilon/4,$$

^成立，则以至少 $1-\delta$ 的概率有 $(1-\varepsilon) \cdot Z \leq Z \leq M \cdot Z$。

我们首先定义该估计量，其构造受到引理 12的启发。事实上，对于任意 u，t，我们将选取某个 M，使得 $\mathrm{Cov}_M(\nu\|\mu) < u \cdot \varepsilon$，并令 $\alpha = (1-t)(1-u \cdot \varepsilon)$。~~接着，~~我们将样本 $\lambda(X_1)\ldots,\lambda(X_n)$ 的 $(1-\alpha)$ 分位数定义为该估计量；也就是说，令 $\lambda_{(1)} \leq \lambda_{(2)} \overset{M}{\leq} \cdots \leq \lambda_{(n)}$ 表示这些样本的次序统计量^，定义 $Z = \lambda_{(\lceil(1-\alpha)n\rceil)}$。现在我们分析 $t = u = 1/2$ 时该估计量的质量。

定理 *11* 的证明。 我们分别分析两侧尾部，然后应用并集界得出结论。

下尾。 设 $\eta_i = \mathbb{I}\left[\frac{d\nu}{d\mu}(X_i) \geq 1-\varepsilon\right]$ 对于 $i \in [n]$。根据 引理 12，对于 M，存在 $\mathrm{Cov}_M(\nu\|\mu) \leq u \cdot \varepsilon$ 使得

$$\mathbb{E}[\eta_i] = \mathbb{P}_{X\sim\mu}\left(\frac{d\nu}{d\mu}(X) \geq 1-\varepsilon\right) \geq \frac{(1-u)\cdot\varepsilon}{M}.$$

应用标准的切尔诺夫界（引理 2），我们得到

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\eta_i \leq \frac{(1-t)(1-u)\cdot\varepsilon}{M}\right) \leq \exp\left(-\frac{nt^2(1-u)\cdot\varepsilon}{2M}\right).$$

现在注意，根据分位数的定义，

$$\mathbb{P}\left(\widehat{Z} \leq (1-\varepsilon)Z\right) \leq \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\eta_i \leq \alpha\right) \leq \exp\left(-\frac{nt^2(1-u)\cdot\varepsilon}{2M}\right).$$

因此，通过设定

$$n \geq \frac{2M}{t^2(1-u)\cdot\varepsilon}\cdot\log\left(2/\delta\right), \tag{15}$$

我们以至少 $1-\delta/2$ 的概率有 $Z \geq (1-\varepsilon)Z$。

上尾。 固定 $M' \geq 1$ 并令 $\xi_i = \mathbb{I}\left[\frac{d\nu}{d\mu}(X_i) \geq M'\right]$ 对于 $i \in [n]$。根据 引理13，可得

$$\mathbb{E}[\xi_i] = \mathbb{P}_{X\sim\mu}\left(\frac{d\nu}{d\mu}(X) \geq M'\right) \leq \frac{\mathrm{Cov}_{M'}(\nu\|\mu)}{M'}.$$

因此，根据标准的切尔诺夫界（引理 2），对于 $0 < s \leq 1$，有

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i \geq (1+s)\frac{\mathrm{Cov}_{M'}(\nu\|\mu)}{M'}\right) \leq \exp\left(-\frac{ns^2\cdot\mathrm{Cov}_{M'}(\nu\|\mu)}{3M'}\right).$$

进而可知，只要

$$n \geq \frac{3M'}{s^2\cdot\mathrm{Cov}_{M'}(\nu\|\mu)}\cdot\log\left(2/\delta\right) \quad \text{and} \quad (1+s)\frac{\mathrm{Cov}_{M'}(\nu\|\mu)}{M'} \leq \alpha, \tag{16}$$

我们就以至少 $1-\delta/2$ 的概率有 $Z \leq M'Z$。

证明总结。 现在令 $t = u = 1/2$ 使得 $\alpha = \frac{\varepsilon}{4M}$。因此，(15) 蕴含：只要 $n \geq 16M/\varepsilon\cdot\log\left(2/\delta\right)$ 且 $\mathrm{Cov}_M(\nu\|\mu) \leq \varepsilon/2$，我们就以至少 $1-\delta/2$ 的概率有 $Z \geq (1-\varepsilon)Z$。此外，令 $s = 1/2$，由 (16) 可知，只要

$$n \geq \frac{12}{2\alpha/3}\cdot\log\left(2/\delta\right) = \frac{18M}{\varepsilon}\cdot\log\left(2/\delta\right)$$

以至少 $1-\delta/2$ 的概率，对于满足 $\mathrm{Cov}_{M'}(\nu\|\mu) \leq \varepsilon/4$ 的任意 M'，都有 $Z \leq M'\cdot Z$。应用并集界即可得到结果。 □

We 现在证明上述 定理 11 的证明中所使用的必要技术引理 e.

Lemma 13. 设 μ、ν 是 $\mathcal{X}$ 上的概率测度，且 $M \geq 1$。则有

$$\mathbb{P}_{X\sim\mu}\left(\frac{d\nu}{d\mu}(X) \geq M\right) \leq \frac{\mathrm{Cov}_M(\nu\|\mu)}{M}.$$

证明。 我们计算

$$\begin{aligned}\mathbb{P}_{X\sim\mu}\left(\frac{d\nu}{d\mu}(X)\geq M\right) &= \mathbb{E}_{X\sim\mu}\left[\mathbb{I}\left[\frac{d\nu}{d\mu}(X)\geq M\right]\right]\\ &= \mathbb{E}_{Y\sim\nu}\left[\frac{1}{\frac{d\nu}{d\mu}(Y)}\cdot\mathbb{I}\left[\frac{d\nu}{d\mu}(Y)\geq M\right]\right]\\ &\leq \frac{1}{M}\cdot\mathbb{E}_{Y\sim\nu}\left[\frac{d\nu}{d\mu}(Y)\cdot\mathbb{I}\left[\frac{d\nu}{d\mu}(Y)\geq M\right]\right]\\ &= \frac{\mathrm{Cov}_M(\nu\|\mu)}{M}.\end{aligned}$$

□

我们现在通过从定理 11推导出定理 6 来结束本节。

定理 6的证明。 由引理 7可知，对于

$$M\geq\gamma_f\left(\frac{4\cdot\mathrm{D}_f(\nu\|\mu)}{\varepsilon}\right),$$

我们有 即 $\mathrm{Cov}_M(\nu\|\mu)\leq\varepsilon/4$。 将此代入定理 11即可得到该结果 . □

F 命题 3的证明

我们使用经典的 A^* 采样算法 [27] ， 已知在 Radon-Nikodym 导数一致有界的情形下，或更一般地在 $\mathrm{D}_{\mathrm{KL}}(\nu\|\mu)$ [14] 有限的条件下，该算法能够产生来自目标分布的近似样本。该算法首先独立采样 X_1, ……， $X_n\sim\mu$ 和E_1，……， $E_n\sim$ Exp(1)。我们令 $N_i=\sum_{j=1}^{i}E_j$ 为泊松到达时间，并令

$$S_i=\frac{N_i}{d\nu/d\mu(X_i)}.$$

为了构造一个样本，我们令 $j=\mathrm{argmin}_{1\leq i\leq n}S_i$ 并返回 $X_{\widehat{j}}$。现在我们分析 $X_{\widehat{j}}$ 与 ν的分布之间的全变差距离。

引理 14。 </style}令 、为定义在 上的概率测度，并令 、…、相互独立。令 按上述方式定义。记 的分布为 ，则有 TV $(\nu_n,\nu)\leq\varepsilon$ ， 只要

$$n\geq 2M\cdot\log(3/\varepsilon)\quad \textit{where}\quad \mathrm{Cov}_M(\nu\|\mu)\leq\varepsilon/3.$$

证明。 我们采用 Li 和 El Gamal [27, 定理 1]中的证明策略。我们按如下方式构造 $Z_{\widehat{j}}$ 和 ν 的分布之间的耦合。将序列 X_i, E_i 延拓为无限序列，并令 $K=\mathrm{argmin}_{i\geq1}\frac{N_i}{d\nu/d\mu}(X_i)$。根据 Li 和 El Gamal [27], 中的论证，有 $Z_K\sim\nu$。因此有

$$\mathrm{TV}(\nu_n,\nu)\leq\mathbb{P}\left(K\neq\widehat{j}\right)=\mathbb{P}(K>n).$$

令 $\Theta = \min^{i \geq 1} {}^{N_i}_{d\nu/d\mu}(X_i)$，根据 Li 和 El Gamal [27]，以下结论成立：(i) $\Theta \sim \mathrm{Exp}(1)$ 并且 (ii) 在给定 Θ 和 X_K 的条件下，$K-1$ 是一个速率至多为 $\Theta \frac{d\nu}{d\mu}(X_K)$ 的泊松随机变量。因此有

$$\mathbb{P}(K > n) \leq \mathbb{E}_{Y\sim\nu}\left[\int_0^\infty e^{-\theta}\mathbb{P}\left(\mathrm{Pois}\left(\theta\frac{d\nu}{d\mu}(Y)\right) > n\right)d\theta\right].$$

注意到，速率较大的泊松随机变量在随机序上支配速率较小的泊松随机变量；因此，对任意 $M > 1$，都有

$$\begin{aligned}\mathbb{E}_{Y\sim\nu}\left[\int_0^\infty e^{-\theta}\mathbb{P}\left(\mathrm{Pois}\left(\theta\frac{d\nu}{d\mu}(Y)\right) > n\right)d\theta\right] &\leq \mathrm{Cov}_M(\nu\|\mu) + \int_0^\infty e^{-\theta}\mathbb{P}(\mathrm{Pois}(\theta M) > n)\,d\theta \\ &\leq \mathrm{Cov}_M(\nu\|\mu) + \inf_{t>0} e^{-t} + \mathbb{P}(\mathrm{Pois}(tM) > n) \\ &\leq 2\varepsilon/3 + \mathbb{P}(\mathrm{Pois}(M\cdot\log(3/\varepsilon)) > n).\end{aligned}$$

利用标准泊松尾界 [40], ，对于 λ，$t > 0$，有

$$\mathbb{P}(\mathrm{Pois}(\lambda) > (1+t)\lambda) \leq e^{-\frac{t^2\lambda}{2(1+t)}}.$$

因此，对于 $n \geq 2M\log(3/\varepsilon)$，有 $\mathbb{P}(\mathrm{Pois}(M\cdot\log(3/\varepsilon)) > n) \leq \varepsilon/3$。结论得证。 □

现在，我们可以借助上述引理和 引理 7 来证明该命题。

证明 命题 3。 根据 引理 7，有

$$\mathrm{Cov}_M(\nu\|\mu) \leq \frac{M\cdot \mathrm{D}_f(\nu\|\mu)}{f(M)}.$$

因此，令 M 使得 $\frac{M\cdot \mathrm{D}_f(\nu\|\mu)}{f(M)} \leq \varepsilon/3$，则有

$$n \gtrsim \log(3/\varepsilon)\cdot\gamma_f\left(\frac{3\cdot\mathrm{D}_f(\nu\|\mu)}{\varepsilon}\right)$$

样本足以确保 $\mathrm{TV}(\nu_n, \nu) \leq \varepsilon$。 □