arXiv:2602.05744v1 [cs.IT] 5 Feb 2026

负 Tsallis 熵的 Bregman 散度的广义 Pinsker 不等式

Guglielmo Beretta[(1), (2)]、Tommaso Cesari[(3)]、Roberto Colomboni[(4), (5)]

(1)DAIS，Università Ca' Foscari Venezia，意大利(2)DAUIN，Politecnico di Torino，意大利(3)电气工程与计算机科学学院，渥太华大学，加拿大(4)DEIB，米兰理工大学，意大利(5)计算机科学系，米兰大学，意大利

{guglielmo.beretta}@unive.it
{tommaso.cesari}@uottawa.ca
{roberto.colomboni}@polimi.it

2026年2月6日

摘要

Pinsker 不等式用全变差 D_{KL} 给出了 Kullback–Leibler 散度的下界，并提供了一种将 D_{KL} 控制转化为 $\|\cdot\|_1$-控制的规范方法。受 Tsallis 损失在概率预测和在线学习中的应用启发，我们针对由负 α-Tsallis 熵生成的 Bregman 散度 D_α（也称为 β-散度）建立了广义 Pinsker 不等式。具体而言，对于概率单纯形 Δ^K相对内部中的任意 p、q，我们证明了如下紧确界

$$D_\alpha(p\|q) \geq \frac{C_{\alpha,K}}{2} \cdot \|p-q\|_1^2 \,,$$

并且我们针对每一种 (α, K) 组合，明确确定最优常数 $C_{\alpha,K}$。

关键词： Pinsker 不等式、信息论、适当评分规则、适当损失、强适当性、Tsallis 熵、Burg 熵、Shannon 熵、Bregman 散度、 β-散度、板仓-斋藤散度、Kullback-Leibler散度

1 引言与相关工作

适当损失[1] 是概率预测的典范框架：给定可能结果的集合 $[K] := \{1, \ldots, K\}$（其中 $K \geq 2$ 为某个整数），学习器预测 $[K]$ 上的概率分布 q，并在结果 $y \in [K]$ 实现时承受损失 $\ell(q, y)$。若对每一个（真实）分布 p，期望损失 $q \mapsto \mathbb{E}_{Y\sim p}\big[\ell(q,Y)\big]$ 都在 $q = p$ [处被唯一最小化，则称损失 ℓ 为严格适当的（Gneiting 和 Raftery， 2007], ）；并且若该损失还可微，则超额风险 $\mathbb{E}_{Y\sim p}\big[\ell(q,Y)\big] - \mathbb{E}_{Y\sim p}\big[\ell(p,Y)\big]$ 具有经典的散度表示：它等于由凸函数 $-H$ 生成的 Bregman 散度，其中 H 是相应的贝叶斯风险（或广义熵） $H\colon\ p \mapsto \mathbb{E}_{Y\sim p}\big[\ell(p,Y)\big]$ [Reid 和 Williamson，2009b 年，Abernethy 和 Frongillo， 2012]。

[1]或者等价地，适当评分规则（至多差一个符号） .

由于学习保证通常控制超额风险，一个常见的下一步是将超额风险上的界转化为更易解释的分布接近性概念，例如全变差距离 [Reid 和 Williamson，2011年，Haghtalab 等人，2019]。一个典范例子是对数损失 $\ell_1(q,k) := -\ln q_k$。其贝叶斯风险是香农熵 S_1，其超额风险与概率单纯形 Δ^K 上的 Kullback–Leibler（KL）散度 D_{KL} 一致。从超额风险到全变差的精确转换由信息论中的经典 Pinsker 不等式给出 [Tsybakov，2008]

$$\|p-q\|_{\mathrm{TV}} \leq \sqrt{\tfrac{1}{2}\cdot D_{\mathrm{KL}}(p\|q)}\,,$$

其中 $\|\cdot\|_{\mathrm{TV}}$ 是全变差，[2] 或等价地（见 A 节），

$$D_{\mathrm{KL}}(p\|q) \geq \tfrac{1}{2}\cdot\|p-q\|_1^2\ .$$

在许多情形下，人们不再采用对数损失和香农熵，而其他严格适当损失则变得有用 [Savage, 1971, Lambert et al., 2008, Giummolè et al.，2019]。典型例子包括$Tsallis$ （或幂）损失

$$\ell_\alpha(q,k) := \frac{q_k^{\alpha-1}}{1-\alpha} + \frac{1}{\alpha}\sum_{i\in[K]} q_i^\alpha\,, \tag{1}$$

（其“极限” $\alpha=0$ 情形为 $\ell_0(q,k) := q_k^{-1} - K + \sum_{i\in[K]}\ln q_i$），这类损失已在评分规则文献中得到广泛研究，且存在多种无实质区别的加法或乘法归一化方式 [Dawid, 2007, Cichocki and Amari, 2010, Dawid and Musio, 2014, Dawid et al., 2016, Basu et al., 1998, Kanamori and Fujisawa, 2014, Mameli and Ventura，2015]。在应用文献中，这些损失主要通过诱导差异出现，即超额风险的期望 $\mathbb{E}_{Y\sim p}[\ell_\alpha(q,Y)] - \mathbb{E}_{Y\sim p}[\ell_\alpha(p,Y)]$,；经过重新参数化并忽略与 q无关的常数后，它与稳健统计中称为密度幂散度的目标函数相对应[Basu et al.，1998]，也与信号处理和矩阵分解中称为 β-散度的目标函数相对应 [Févotte 和 Idier，2011]；例如，Dawid et al. [2016]和 Jewson et al. [2018] 明确阐述了 Tsallis 评分与 β-散度／密度幂散度之间的联系。具体应用包括：通过最小密度幂散度对健康科学数据进行稳健推断 [Nandy et al.，2022],、遥感中的高光谱解混 [Févotte and Dobigeon，2015], 、动态 PET 因子分析 [Cavalcanti et al.，2019], ，以及利用非负矩阵分解进行音频源分离 [Ozerov and Févotte，2009]。

ℓ_α 的贝叶斯风险是 α-Tsallis 熵 S_α，而相应的超额风险是 $-S_\alpha$ 的 Bregman 散度 $D_\alpha(p\|q)$（见第 B 节）。因此，任何关于 $-S_\alpha$ 的 Bregman 散度的 Pinsker 型不等式，都能将关于 ℓ_α 的超额风险界转化为更易解释的预测分布 $\|\cdot\|_1$ （或全变差）控制，进而通过标准的中间不等式 $\mathbb{P}_{Y\sim p}\left[Y\neq k_q^\star\right] - \mathbb{P}_{Y\sim p}\left[Y\neq k_p^\star\right] \leq \|p-q\|_1$，给出相应代入规则的 0–1 损失的超额风险界；其中，对于任意 $q\in\Delta^K$，索引 $k_q^\star$ 都是在 $\operatorname{argmax}^{k\in[K]} q_k$ 中确定性地选出的（见第 C 节）。

Tsallis 熵也会作为在线学习/凸优化和多臂老虎机问题中的正则项出现。它们所诱导的几何结构和数据依赖行为介于香农熵（$\alpha=1$）与对数障碍/Burg 熵（$\alpha=0$）正则化之间，而相关的 Bregman 散度在分析相应算法时发挥着关键作用（例如，参见 Abernethy et al. 2015, Zimmert and Seldin 2019, 2021, Cesa-Bianchi et al.

[2]正如 Reid 和 Williamson [2011], 所述，比较不同作者的记号时应谨慎。例如，Reid 和 Williamson [2011] 所说的“变分散度”表示范数 $\|\cdot\|_1$。

表1：在 Δ^K 的相对内部上，负 α-Tsallis 熵的 Bregman 散度的尖锐 Pinsker 不等式常数 $C_{\alpha,K}$ 于 $D_\alpha(p\|q) \geq \frac{C_{\alpha,K}}{2}\cdot\|p-q\|_1^2$。

情形	常数 $C_{\alpha,K}$	注释
$\alpha \in (-\infty, 1]$	$2^{1-\alpha}$	无维数依赖常数
$\alpha \in (1, 2]$, K 偶数	$K^{1-\alpha}$	关于 K 的多项式依赖
$\alpha \in (1,2]$, K 奇数	$K^{1-\alpha}\cdot\sigma_{\alpha,K}$	$1 \leq \sigma_{\alpha,K} \leq 1+\frac{7}{6}\frac{(\alpha-1)(3-\alpha)}{K^2} \leq 1+\frac{7}{6}\frac{1}{K^2}$
$\alpha > 2$, $K = 2$	$2^{1-\max\{\alpha,3\}}$	1 4 </style}针对 $2 < \alpha \leq 3$，在 $\alpha = 3$ 处发生相变
$\alpha > 2$, $K \geq 3$	0	无 Pinsker：$\inf_{p\neq q} D_\alpha(p\|q)/\|p-q\|_1^2 = 0$

2022, Masoudian and Seldin 2021, Kulis and Bartlett 2010）。在这些情形下，关于 $-S_\alpha$ 的 Bregman 散度的 Pinsker 型不等式等价于其 $\|\cdot\|_1$-强凸性，因此确定了标准的跟随正则化领导者和镜像下降遗憾分析中所涉及的曲率参数（例如，参见 Shalev-Shwartz 2012, 定理2.11）。

因此，我们提出了以下问题：

> 由负 α-*Tsallis* 熵生成的 *Bregman* 散度，*Pinsker* 不等式的对应形式是什么？

我们通过建立 Pinsker 不等式的一个尖锐推广（定理 8）来回答这一问题。记 $D_\alpha(\cdot\|\cdot)$ 为由负 α-Tsallis 熵 $-S_\alpha$生成的 Bregman 散度（等价地，为满足 $\beta = \alpha$ 的 β-散度），我们给出了一个显式的 尖锐常数 $C_{\alpha,K}$ 使得对于任意 p，q 在 Δ^K 的相对内部[3] 中，

$$D_\alpha(p\|q) \geq \frac{C_{\alpha,K}}{2}\cdot\|p-q\|_1^2 \ ,$$

为此，我们根据 S_α 的海森矩阵 HS_α，推导出$C_{\alpha,K}$ 的变分刻画，从而将尖锐常数的计算归结为在切空间 $\|\cdot\|_1$-单位方向上优化参数化二次型。作为一个特例，由于负香农熵 $-S_1$ 的 Bregman 散度 D_1 与 Kullback-Leibler散度 D_{KL} 重合，我们的变分框架恢复了具有 $\alpha = 1$ 最优常数的 Pinsker 不等式。

常数 $C_{\alpha,K}$ 呈现出几种定性上截然不同的情形，这反映了随着 (α, K) 变化， $-S_\alpha$ 的 Bregman 散度几何性质也会改变（参见表1、图1，以及第3节开头的讨论）。当 $\alpha \in (-\infty, 1]$ 时，该界与维度无关：$C_{\alpha,K} = 2^{1-\alpha}$ 不依赖于 K，因此 Tsallis 超额风险能够以一个仅取决于 α 的常数，在整个单纯形上统一控制 $\|\cdot\|_1$ 度量。相比之下，当 $\alpha \in (1, 2]$，时，该常数会随维度呈幂律衰减，因为 $C_{\alpha,K} \approx K^{1-\alpha}$，所以从超额风险到$\|\cdot\|_1$ 的转换会随着 K而变差。此外，在这种情形下还存在一种特殊的奇偶效应：当 K 为偶数时，有精确恒等式 $C_{\alpha,K} = K^{1-\alpha}$；而当 K 为奇数时，最佳常数为 $C_{\alpha,K} = K^{1-\alpha}\cdot\sigma_{\alpha,K}$，其中 $1 \leq \sigma_{\alpha,K} = 1 + O(1/K^2)$。因此，偶数和奇数情形在主导阶上相同，但奇数情形会多出一个小的二阶修正项，该项在 $K \to \infty$时趋于零。最后，当 $\alpha > 2$时，情况会骤然改变。对于多分类问题，即 $K \geq 3$，不存在形如 $D_\alpha(p\|q) \geq c\,\|p-q\|_1^2$ 且满足 $c > 0$ 的统一 Pinsker 型不等式，因此需要对 C_α和/或 K 施加额外约束

[3]对于固定的、位于 Δ^K的相对内部中的 p，散度 $D_\alpha(p\|q)$可通过取极限，对 Δ^K 中的任意 q 给出一个定义良好的（可能为无穷的）延拓，因此该不等式平凡地推广到 Δ^K中的所有 q。下文为简洁起见，我们仅对同属于 Δ^K的相对内部的 p 与 q 陈述我们的结果。

才能得到类似的不等式（参见 G 节）。相比之下，在二元情形下，即 $K = 2$，这或许是最重要的特殊情形，Pinsker 型不等式确实对所有 $\alpha > 2$都成立，因此即使超过阈值 $\alpha = 2$，从超额风险到$\|\cdot\|_1$ 的转换仍然适用（例如，对于二分类）。

其他广义 Pinsker 不等式。 针对 Csiszár f-散度的 Pinsker 型不等式已有研究，因此 Tsallis 相对熵 （即 f-散度）也在研究之列 [Reid 和 Williamson, 2009a, Gilardoni, 2010, Rioul, 2023]。我们指出，这些结果与我们的研究相互独立：对于 $\alpha \neq 1$，由负 α-Tsallis 熵生成的 Bregman 散度并非Csiszár f-散度，而且它们仅 在香农极限 $\alpha \to 1$下与 f-散度一致；此时诱导出的 Bregman 散度退化为 Kullback-Leibler散度 D_{KL}。事实上，负 α-Tsallis 熵的 Bregman 散度是可分解的；在概率单纯形上，同时也是可分解 Bregman 散度的唯一 Csiszár f-散度（相差一个正的乘法常数）是 D_{KL} 及其反向形式（$p,\ q$） $\mapsto D_{\mathrm{KL}}(q\|p)$ [Amari, 2009]。

关于 Csiszár f-散度和 α-Tsallis 相对熵的背景，以及 α-Tsallis 相对熵的广义 Pinsker 不等式与我们结果之间关系的讨论，请参见附录 F。

2 Tsallis 熵的 Bregman 散度的 Pinsker 不等式

在本文余下部分中，我们假设 K 是任意一个大于或等于 2 的整数。对于任意 $\alpha > 0$ 和任意 $p \in \mathbb{R}^K$，我们用 $\|p\|_\alpha \coloneqq \left(\sum_{k\in[K]} |p_k|^\alpha\right)^{1/\alpha}$ 表示向量 p 的通常 Minkowski（拟）范数。我们用 $\Delta^K \coloneqq \left\{p \in [0,1]^K \colon \|p\|_1 = 1\right\}$ 表示（$K-1$）维概率单纯形，其相对内部记为 $\operatorname{relint}\left(\Delta^K\right) \coloneqq \Delta^K \cap (0,1)^K$。

我们首先回顾作为幂/Tsallis 损失族的贝叶斯风险而出现的熵；我们所研究的 Bregman 散度以这些熵的相反数作为凸生成元。

De定义 1 (Tsallis 熵)。 对于任意 $\alpha \in \mathbb{R}$，定义 $S_\alpha \colon [0,+\infty)^K \to [-\infty,+\infty)$ 为

$$S_\alpha(p) \coloneqq \begin{cases} \dfrac{\sum_{k\in[K]} p_k^\alpha}{\alpha(1-\alpha)}\,, & \text{if } \alpha \notin \{0,1\}\,, \\ \displaystyle\sum_{k\in[K]} \ln p_k\,, & \text{if } \alpha = 0\,, \\ \displaystyle -\sum_{k\in[K]} p_k \ln p_k\,, & \text{if } \alpha = 1\,, \end{cases} \tag{2}$$

约定当 α 为负数时， $0^\alpha = +\infty$，且 $\ln(0) = -\infty$ 但 $0 \cdot \ln(0) = 0$。对于任意 $p \in \Delta^K$，我们称量 $S_\alpha(p)$ 为 p 的 α-*Tsallis* 熵[4]。

在文献中， S_0 也称为 Burg（或对数）熵，而 S_1 是香农熵。将 S_0 和 S_1 纳入 Tsallis 熵族的原因，是一项将在备注 7 中显现的连续性论证。下一条备注给出了所有这些熵的一个共同性质，并据此为后续所有微分计算提供依据。

[4]文献中， α-Tsallis 熵的定义可能相差一个乘法因子，也可能相差 p 的仿射函数（尤其是加法常数）。对于 Bregman 散度，对生成元进行缩放会相应缩放散度，而加上任意仿射函数则不会改变散度：对于 $c > 0$、 $a \in \mathbb{R}$ 和 $u \in \mathbb{R}^K$，

$$D_{c\cdot f + a + \langle u,\cdot\rangle}(p,q) = c \cdot D_f(p,q)\,, \qquad \text{for any } p,q \in (0,+\infty)^K\,.$$

因此，更改 S_α 定义中的乘法因子，会使其诱导的散度按相同的比例缩放，因此定理 8 中的常数 $C_{\alpha,K}$ 也会按相同的因子缩放。

注2。 对于任意 $\alpha \in \mathbb{R}$，注意到 $-S_\alpha$ 在限制到 $(0,+\infty)^K$ 后是严格凸的，且属于 C^∞ 类。严格凸性直接由 $-S_\alpha$ 的海森矩阵正定得出（见注16）。

我们现在可以引入主要研究对象：由 $-S_\alpha$ 生成的 Bregman 散度。从评分函数/损失函数的角度看，它们表示贝叶斯风险为 S_α 的 Tsallis 评分函数/损失函数的超额风险（见附录B）。

定义3 （Bregman 散度）. 设 $f\colon (0,+\infty)^K \to \mathbb{R}$ 为可微函数。 f 的 Bregman 散度是函数

$$D_f\colon (0,+\infty)^K \times (0,+\infty)^K \to \mathbb{R}\,, \qquad (p,q) \mapsto f(p) - f(q) - \langle \nabla f(q), p-q\rangle\,,$$

其中 $\nabla f(q)$ 是 f 在 q 处的梯度。下文中，对任意 $\alpha \in \mathbb{R}$ 以及任意 $p,q \in (0,+\infty)^K$，我们用更简洁的记号 $D_\alpha(p\|q)$ 表示 $D_{-S_\alpha}(p,q)$。

读者 若不熟悉 Bregman 散度，可能会觉得下一条注记有用。

注记 4. 对每一个 $\alpha \in \mathbb{R}$， $-S_\alpha$ 在 $(0,+\infty)^K$ 上的严格凸性意味着 $D_\alpha(p\|q) \geq 0$ 对每一个 $p, q \in (0,+\infty)^K$ 成立，并且 $D_\alpha(p\|q) = 0$ 当且仅当 $q = p$。

从损失的角度来看，注4可以理解为：Tsallis 损失是严格适当的，即相应的超额风险散度 D_α 非负，且仅在 $q = p$处为零。

情形 $\alpha = 1$ 与 Kullback-Leibler散度之间存在一种有趣且已知的关系。注 5。 对于所有 p， $q \in \operatorname{relint}\left(\Delta^K\right)$, $D_1(p\|q) = D_{\mathrm{KL}}(p\|q)$。事实上：

$$D_1(p\|q) = \sum_{k\in[K]} \left(p_k \cdot \ln \frac{p_k}{q_k} - p_k + q_k \right) = \sum_{k\in[K]} p_k \cdot \ln \frac{p_k}{q_k} = D_{\mathrm{KL}}(p\|q)\,.$$

此外，定义3与统计学文献中的以下一参数 β-散度族有关 [Hennequin 等人，2010年, Févotte 和 Idier， 2011]。

定义6 （β-散度）。 对于任意 $\beta \in \mathbb{R}$， p的 β-散度， $q \in (0,\ +\infty)^K$ 定义为

$$d_\beta(p\|q) \coloneqq \begin{cases} \dfrac{1}{\beta} \cdot \displaystyle\sum_{k\in[K]} \frac{p_k^\beta + (\beta-1) q_k^\beta - \beta p_k q_k^{\beta-1}}{\beta-1}\,, & \text{if } \beta \notin \{0,1\}\,, \\ \displaystyle\sum_{k=1}^K \left(\frac{p_k}{q_k} - \ln \frac{p_k}{q_k} - 1 \right)\,, & \text{if } \beta = 0\,, \\ \displaystyle\sum_{k\in[K]} \left(p_k \cdot \ln \frac{p_k}{q_k} - p_k + q_k \right)\,, & \text{if } \beta = 1\,, \end{cases}$$

散度 d_0 也称为板仓-斋藤散度，而 d_1 称为 I-散度。定义3与定义6之间的联系将在下一个注释中说明；Hennequin 等人 [2010]已经指出了这一点。

注记7。对于任意 $\alpha \in \mathbb{R}$， $-S_\alpha$ 的 Bregman 散度 D_α 与 β-散度d_β一致，其中 $\beta = \alpha$，即，对于每个 p， $q \in (0,\ +\infty)^K$，以及 $\alpha \in \mathbb{R}$

$$D_\alpha(p\|q) = d_\alpha(p\|q)\,.$$

此外，对于每个 $p, q \in (0,+\infty)^K$ 以及每个 $\alpha \in \mathbb{R}$，均有

$$D_\alpha(p\|q) = \lim_{\alpha' \to \alpha} D_{\alpha'}(p\|q)\,.$$

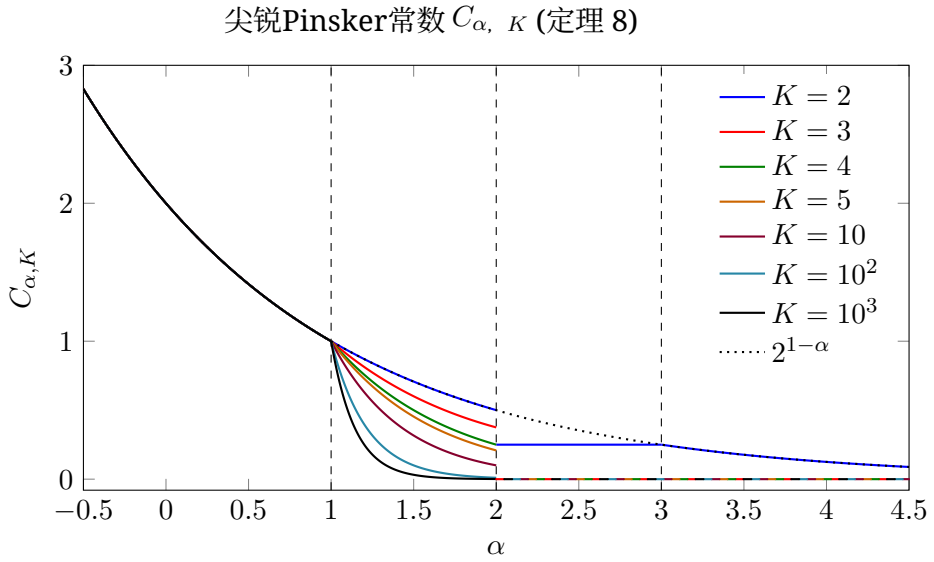


图1：所选 K 对应的尖锐Pinsker常数 $C_{\alpha,\ K}$。虚线标出发生相变的 $\alpha = 1$、2、3。

现在，我们已具备陈述主要结果所需的全部要素。我们的目标是建立 $D_\alpha(p\|q)$ 关于 $\|p-q\|_1^2$ 在 relint $\left(\Delta^K\right)$ 上的 Pinsker 型下界。下一个定理将确定最佳常数$C_{\alpha,K}$ 并给出其在所有情形下的显式表达式。

定理 8（Tsallis 熵的 Bregman 散度的 Pinsker 不等式）。设 $\alpha \in \mathbb{R}$ 并记 $C_{\alpha,\ K}$ 为最大的 $C \geq 0$，使得对每个 $p,\ q \in$ relint $\left(\Delta^K\right)$

$$D_\alpha(p\|q) \geq \frac{C}{2} \cdot \|p-q\|_1^2 \ . \tag{3}$$

于是，

$$C_{\alpha,K} = \begin{cases} 2^{1-\alpha} & \text{if } \alpha \in (-\infty, 1], \\ K^{1-\alpha} & \text{if } \alpha \in (1,2] \text{ and } K \text{ is even}, \\ K^{1-\alpha} \cdot \sigma_{\alpha,K} & \text{if } \alpha \in (1,2] \text{ and } K \text{ is odd}, \\ 2^{1-\max\{\alpha,3\}} & \text{if } \alpha \in (2,+\infty) \text{ and } K = 2, \\ 0 & \text{if } \alpha \in (2,+\infty) \text{ and } K \geq 3, \end{cases} \tag{4}$$

where $\sigma_{\alpha,K} := \left(\frac{(1-\frac{1}{K})^{\frac{1-\alpha}{3-\alpha}} + (1+\frac{1}{K})^{\frac{1-\alpha}{3-\alpha}}}{2} \right)^{3-\alpha}$ *(see Remark 11 for more about this term).*

注意，根据注5，且据我们所知，我们还首次为 $\beta \neq 1$ 的所有 β-散度（包括板仓-斋藤散度）建立了一个尖锐的 Pinsker 型不等式，并为 $\beta = 1$ 得到了经典的（尖锐）Pinsker 不等式

$$D_{\mathrm{KL}}(p\|q) = D_1(p\|q) \geq \frac{1}{2} \cdot \|p-q\|_1^2 \ , \qquad \text{for all } p, q \in \text{relint}\left(\Delta^K\right) .$$

作为定理 8 的直接推论，我们还得到了 $\|\cdot\|_1$相对于$-S_\alpha$的强凸常数，即参数 μ 。该曲率参数出现在采用 Tsallis正则化的标准跟随正则化领导者和镜像下降分析中。

推论 9. 负 *Tsallis* 熵 $-S_\alpha$ 相对于 $\|\cdot\|_1$是 $C_{\alpha,\ K}$-强凸的。此外，这个值 $C_{\alpha,\ K}$ 是最优的，也就是说，它是使该不等式成立的最大强凸参数。

现在，我们就 $C_{\alpha,K}$ 的相变行为提出一系列备注。除了其本身具有的信息论意义外，这些情形还阐明了维度、所选的 α以及二元与多类别设定 如何决定将 Tsallis 分数超额风险（等价地，散度 D_α）转换为 $\|\cdot\|_1$ 控制的强度。

备注 10. 对于 $\alpha \le 1$ $C_{\alpha,K}$ 的值不依赖于维度 K。

这意味着，对于 $\alpha \le 1$，定理 8 中从超额风险到$\|\cdot\|_1$ 的转换与维度无关；特别是，在大规模多分类情形下， $\alpha \le 1$ 在此转换中避免了维数灾难。

下一条备注量化了 $\sigma_{\alpha,K}$ 对 $\alpha \in (1,2]$ 且 K 为奇数）的奇偶校正，并明确指出 $C_{\alpha,K} = K^{1-\alpha}$ 除去低阶因子 $\sigma_{\alpha,K} = 1 + O(K^{-2})$ 。

注记 11. 对于 $\alpha \in (1,2]$ 且 K 为奇数），常数 $C_{\alpha,K}$ 等于 $K^{1-\alpha}$ 乘以缩放项 $\sigma_{\alpha,K}$，它满足当 $\alpha \to 1^+$ 时 $\sigma_{\alpha,K} > 1$ 以及 $\sigma_{\alpha,K} \to 1$ 。我们还有如下估计

$$1 \le 1 + \frac{\alpha-1}{3-\alpha}\cdot\frac{1}{K^2} \le \sigma_{\alpha,K} \le 1 + \frac{7(\alpha-1)}{6(3-\alpha)}\cdot\frac{1}{K^2} \le 1 + \frac{7}{6}\cdot\frac{1}{K^2}\ .$$

其证明见附录 D.1。特别地，当 $\alpha \in (1,2]$ 且 K为奇数）时，$K^{1-\alpha} < C_{\alpha,K} \approx K^{1-\alpha}$ 。

在情形 $\alpha = 2$ 中还可以注意到另一个有趣的事实，该情形尤其值得关注，因为 D_2与欧几里得距离平方的一半重合。[5]

注记 12. 在定理 8 中令 $\alpha = 2$ ，可得

$$C_{2,K} = \begin{cases} \frac{1}{K} & \text{for } K \text{ even}\ , \\ \frac{1}{K}\cdot\frac{K^2}{K^2-1} & \text{for } K \text{ odd}\ . \end{cases}$$

再回顾 $D_2(p\|q)$ 与 p 和 q之间的欧几里得距离的关系，我们便得到熟知的不等式 $\|p-q\|_2^2 \ge \frac{1}{K}\cdot\|p-q\|_1^2$，但另有一个观察：当 K为奇数且 p, q 被限制在 Δ^K中时，常数 $\frac{1}{K}$并非最优。

对于 $\alpha > 2$，情况发生了质的变化，二元情形 $K = 2$ 与多类别情形 $K \ge 3$之间出现了鲜明的区别。在二元情形下，非平凡的 Pinsker 型不等式仍然成立，因此较小的 Tsallis 损失超额风险仍能保证预测分布之间的 $\|\cdot\|_1$接近。具体而言，我们注意到以下几点：

注 13。 对于 $K = 2$ ，映射 $\alpha \mapsto C_{\alpha,K}$ 是 α 在 $[0,+\infty)$ 上的分段光滑函数，其唯一的不连续点是 $\alpha = 2$处的跳跃。

相比之下，在多分类情形 $K \ge 3$下， $\alpha > 2$ 区域本质上无法将超额风险界转化为 $\|\cdot\|_1$ 界。

注 14。 定理 8 指出，对于 $\alpha > 2$ 和 $K \ge 3$ ，无法找到一个正常数 C，使得对于每个 p, $q \in \text{relint}\ (\Delta^K)$，都有 $D_\alpha(p\|q) \ge C\cdot\|p-q\|_1^2$。

另请参见附录 E，其中明确构造了满足 $p(t)$ $q(t)$ $\in$relint(Δ^K) 的单参数族 $D_\alpha\big(p(t)\|q(t)\big) = o\big(\|p(t) - q(t)\|_1^2\big)$，因为 $t \to 0$。尽管有注14,

[5]事实上，根据定义1，我们有 $-S_2(p) = \frac{1}{2}\|p\|_2^2$，因此，对于任意 p, $q \in \text{relint}\ (\Delta^K)$，我们有 $D_2(p\|q) = \frac{1}{2}\|p\|_2^2 - \frac{1}{2}\|q\|_2^2 - \langle q,\ p-q\rangle = \frac{1}{2}\|p\|_2^2 + \frac{1}{2}\|q\|_2^2 - \langle p,\ q\rangle = \frac{1}{2}\|p - q\|_2^2$.

当 p 和 q 的所有坐标都远离零时，仍有可能证明 Pinsker 型不等式成立。感兴趣的读者可在 G 节找到相关证明。

我们通过指出：如果寻求 Pinsker 型不等式的范围不仅是 relint $\left(\Delta^K\right)$，而是整个 $(0, +\infty)^K$，将会发生什么，来结束本节。

备注 15（在 $(0, +\infty)^K$ 上不存在统一的 Pinsker 不等式，$\alpha = 2$ 除外）。如果 $\alpha \neq 2$，则不存在常数 $C > 0$ 使得对于所有 $p, q \in (0, +\infty)^K$，都有 $D_\alpha(p\|q) \geq C \cdot \|p - q\|_1^2$。如果 $\alpha = 2$，那么对于所有 $p, q \in (0, +\infty)^K$，都有 $D_2(p\|q) \geq \frac{1}{2K} \cdot \|p - q\|_1^2$，且常数 $\frac{1}{2K}$ 是最优的。 —

该备注的证明见 D 节。

3 定理 8 证明背后的几何直觉

现在，我们对用于推导最佳常数 $C_{\alpha,\,K}$ 显式表达式的方法做一些说明。首先，我们注意到，$C_{\alpha,\,K}$ 具有一种变分表述，其中涉及 S_α 的海森矩阵以及自变量 $(\gamma,\ v)$；其中，γ 位于 $\left(\Delta^K\right)$ 的相对内部，而 v 是一个 $\|\cdot\|_1$-单位向量（即 $\|v\|_1 = 1$），且与 $\left(\Delta^K\right)$ 的相对内部相切（即 $\sum_{k=1}^K v_k = 0$）。事实上，$C_{\alpha,\,K}$ 可以表示为关于 $(\gamma,\ v)$ 的下确界，其中被取下确界的是关于变量 v 的参数二次型，其系数（即 $\gamma_k^{\alpha-2}$）取决于 γ 和 α。这些系数的定性行为至关重要，它们在 $\alpha = 2$ 处表现出相变。具体而言，在 $\alpha \leq 2$ 的情形下，这些系数存在一个正的下界（因为 $\gamma_k < 1$ 且 $\alpha - 2 \leq 0$）；但在 $\alpha > 2$ 的情形下则不再如此（因为 $\alpha - 2 > 0$，相应的 γ_k 趋近于 0）。

为处理情形 $\alpha \leq 2$，我们先固定参数二次型中的 v，并对 γ的所有可能取值取下确界；这相当于在单纯形的相对内部 relint $\left(\Delta^K\right)$上，求 $-S_\alpha$ 沿方向 v 的最小二阶方向导数。结果表明，该下确界等于 $\|v\|_\beta$的平方，其中β 是一个正数（具体为$\beta = \frac{2}{3-\alpha}$），取决于 α。因此，$C_{\alpha,K}$ 与在切空间的 $\|\cdot\|_1$-单位向量中最小化 $\|\cdot\|_\beta$ 有关。有趣的是，$C_{\alpha,K}$的表达式体现了达到这一约束最小值的 v 的局部化。事实上，情形 $\alpha \in (-\infty, 1)$给出 $0 < \beta < 1$，而通过将 v 的质量集中在两个坐标上，可以使 $\|v\|_\beta$ 达到最小，这使得$C_{\alpha,K}$ 与 K无关。若 $\alpha = 1$，则 $\beta = 1$也成立，并且显然有$\|v\|_\beta = 1$，从而得到$C_{1,K} = 1$。这就得到了经典 Pinsker 不等式。情形$\alpha \in (1, 2]$ 则给出 $\beta > 1$，而将质量尽可能均匀地分散到各个坐标上，可以使 $\|v\|_\beta$ 达到最小，得到关于$C_{\alpha,K}$ 的表达式；该表达式随 K 递减，且 K的奇偶性也起着微妙作用。

对于 $\alpha > 2$，我们将看到，具体情况取决于 K，而且存在显著差异。事实上，如果 $K \geq 3$，则可以证明 $C_{\alpha,\,K} = 0$ 这归功于存在某个具有若干零分量的切向量 v。对于 $K = 2$，不存在类似的 v，并且实际上 $C_{\alpha,\,2} > 0$。这一次，相变发生在 $\alpha = 3$，此时映射 $\gamma_k \mapsto \gamma_k^{\alpha-2}$ 由凹变凸。

4 迈向证明：有用的结果

对于任意有限集 A，我们用记号 $|A|$ 表示 A中的元素个数。注 16。对于任意 $\alpha \in \mathbb{R}$和 $p \in (0, +\infty)^K$，我们有

$$\partial_{ij}^2 S_\alpha(p) = -p_i^{\alpha-2} \cdot \mathbb{I}\{i = j\}\,,$$

因此，$S_\alpha(p)$ 的海森矩阵是 $K \times K$矩阵 $HS_\alpha(p) = -\mathrm{diag}\left(p_1^{\alpha-2},\ \ldots\ldots,\ p_K^{\alpha-2}\right)$。关于此注释的证明可见 D 节。

现在我们定义定理 8 的表述中出现的关键量 $C_{\alpha,\ K}$。

定义 17. 对于每个 $\alpha \in \mathbb{R}$，我们定义

$$C_{\alpha,K} \coloneqq 2 \cdot \inf \left\{ \frac{D_\alpha(p\|q)}{\|p-q\|_1^2} : p,q \in \operatorname{relint}\left(\Delta^K\right), p \neq q \right\} . \tag{5}$$

注 18. 根据构造，定义 17 中给出的常数 $C_{\alpha,K}$ 是满足以下条件的最大 $C \geq 0$：对于每个 $p,\ q \in \operatorname{relint}\left(\Delta^K\right)$，都有 $D_\alpha(p\|q) \geq \frac{C}{2} \cdot \|p-q\|_1^2$。这一事实的证明见 D 节。定理 8 的证明归结为计算 $C_{\alpha,K}$ 的显式表达式。

定义 19. 我们设

$$\mathrm{B}^K \coloneqq \left\{ v \in \mathbb{R}^K : \|v\|_1 = 1, \sum_{k\in[K]} v_k = 0 \right\} . \tag{6}$$

注 20. 对于每个 $\gamma \in \operatorname{relint}\left(\Delta^K\right)$，向量空间 $\{v \in \mathbb{R}^K : \sum_{k\in[K]} v_k = 0\}$ 可视为 γ 中 relint $\left(\Delta^K\right)$ 的切空间，而 B^K 恰好是该空间在范数 $\|.\|_1$ 下的单位球面。

命题 21（通过二次型重新表述）。对任意 $\alpha \in \mathbb{R}$，我们有

$$C_{\alpha,K} = \inf \left\{ \sum_{k\in[K]} v_k^2 \gamma_k^{\alpha-2} : v \in \mathrm{B}^K, \gamma \in \operatorname{relint}\left(\Delta^K\right) \right\} . \tag{7}$$

证明。令 $C'_{\alpha,K}$ 为（7）的右端。给定两个不同的 $p,\ q \in \operatorname{relint}\left(\Delta^K\right)$，我们考虑 $S_\alpha(p)$ 的如下带有拉格朗日余项的一阶泰勒展开：

$$S_\alpha(p) = S_\alpha(q) + \langle \nabla S_\alpha(q), p-q \rangle + \frac{1}{2}(p-q)^\top H S_\alpha(\gamma)(p-q) , \tag{8}$$

其中 $\gamma \coloneqq xp + (1-x)q$，对某个适当的 $x \in (0,1)$ 成立，且 HS_α 是 S_α 的海森矩阵。特别地，γ 是 p 和 q 的凸组合，因此 $\gamma \in \operatorname{relint}\left(\Delta^K\right)$。定义 $v \coloneqq \frac{p-q}{\|p-q\|_1}$，注意到 $v \in \mathrm{B}^K$，并由（8）、定义3和注16可知

$$2 \cdot \frac{D_\alpha(p\|q)}{\|p-q\|_1^2} = -\left(\frac{p-q}{\|p-q\|_1}\right)^\top H S_\alpha(\gamma) \frac{p-q}{\|p-q\|_1} = \sum_{k\in[K]} v_k^2 \gamma_k^{\alpha-2} , \tag{9}$$

也就是说，$2 \cdot D_\alpha(p\|q)\, \|p-q\|_1^{-2} \geq C'_{\alpha,K}$。由于 p 和 q 是任意选取的，因此得到$C_{\alpha,K} \geq C'_{\alpha,K}$。为证明反向不等式，令 $\zeta \in \operatorname{relint}\left(\Delta^K\right)$，并令 $u \in \mathrm{B}^K$。对$t > 0$，定义点 $\tilde{p}(t) \coloneqq \zeta + t\frac{1}{2}u$ 和 $\tilde{q}(t) \coloneqq \zeta - t\frac{1}{2}u$；若 t 足够小，则它们属于 $\operatorname{relint}\left(\Delta^K\right)$，并满足 $u = \frac{\tilde{p}(t)-\tilde{q}(t)}{\|\tilde{p}(t)-\tilde{q}(t)\|_1}$。然后，像之前一样推导，可知存在某个 $r(t)$，它是 $\tilde{p}(t)$ 和 $\tilde{q}(t)$ 的凸组合，并且（9）对 $(p,q,v,\gamma) = (\tilde{p}(t)\ \tilde{q}(t)\ u, r(t))$ 成立。特别地，$C_{\alpha,K} \leq \sum_{k\in[K]} u_k^2 \left(r_k(t)\right)^{\alpha-2}$，因此 $C_{\alpha,K} \leq \sum_{k\in[K]} u_k^2 \zeta_k^{\alpha-2}$，因为 $r(t) \to \zeta$ 如 $t \to 0^+$。由 ζ 和 u 的任意性可知，$C_{\alpha,K} \leq C'_{\alpha,K}$，因而$C_{\alpha,K} = C'_{\alpha,K}$。

□

在下文中，我们将需要两个技术引理。第一个引理讨论某类齐次泛函在 $\operatorname{relint}\left(\Delta^K\right)$ 上的最小化。第二个引理量化 B^K 相对于所有闵可夫斯基（伪）范数与原点的接近程度。

引理 22（单纯形相对内部上的下确界）. 设 $\nu \geq 0$ 并考虑 $\lambda \in \mathbb{R}^K$使得 $\lambda \geq 0$。则

$$\inf\left\{\sum_{k\in[K]} \lambda_k \gamma_k^{-\nu} : \gamma \in \operatorname{relint}\left(\Delta^K\right)\right\} = \left(\sum_{k\in[K]} \lambda_k^{\frac{1}{\nu+1}}\right)^{\nu+1} .$$

证明。 定义 $S := \{k \in [K] : \lambda_k > 0\}$。我们假设 $S \neq \emptyset$ 和 $\nu > 0$，否则证明显然成立。在 $(0,+\infty)^K$上定义函数 $f(\gamma) := \sum_{k\in[K]} \lambda_k \gamma_k^{-\nu}$。对于 $S = [K]$的情况，注意到f在 $\operatorname{relint}\left(\Delta^K\right)$上严格凸，并且对任意 k， $f(\gamma) \to +\infty$as $\gamma_k \to 0$，因此，将 f限制在 $\operatorname{relint}\left(\Delta^K\right)$上后，它会在某个 $\hat{\gamma} \in \operatorname{relint}\left(\Delta^K\right)$处取得最小值。根据一阶最优性条件，存在某个 $\mu \in \mathbb{R}$，使得对每个 $k \in [K]$都有 $\partial_k f(\hat{\gamma}) = -\nu\lambda_k \hat{\gamma}_k^{-\nu-1} = \mu$。但此时， $\lambda_k^{\frac{1}{\nu+1}} \hat{\gamma}_k$ 的值与 k无关，因此 $\hat{\gamma}_k = \left(\sum_{i\in[K]} \lambda_i^{\frac{1}{\nu+1}}\right) \lambda_k^{\frac{1}{\nu+1}}$，并且

$$f(\hat{\gamma}) = \sum_{k\in[K]} \lambda_k \left(\sum_{i\in[K]} \lambda_i^{\frac{1}{\nu+1}}\right)^{\nu} \lambda_k^{-\frac{\nu}{\nu+1}} = \sum_{k\in[K]} \lambda_k^{\frac{1}{\nu+1}} \left(\sum_{i\in[K]} \lambda_i^{\frac{1}{\nu+1}}\right)^{\nu} = \left(\sum_{k\in[K]} \lambda_k^{\frac{1}{\nu+1}}\right)^{\nu+1} .$$

对于情形 $S \neq [K]$，令 $S = \{k_1, \ldots k_{|S|}\}$，并定义映射 $\pi\colon \operatorname{relint}\left(\Delta^K\right) \to \mathbb{R}^{|S|}$，如下： $\pi_i(\gamma) := \gamma_{k_i}/(\sum_{k\in S} \gamma_k)$，从而

$$f(\gamma) = \sum_{k\in S} \lambda_k \gamma_k^{-\nu} = \left(\sum_{k\in S} \gamma_k\right)^{-\nu} \sum_{i=1}^{|S|} \lambda_{k_i} \left(\pi_i(\gamma)\right)^{-\nu} . \tag{10}$$

对每个 $W \in (0,1)$，定义 $E_W := \left\{\gamma \in \operatorname{relint}\left(\Delta^K\right) : \sum_{k\in S} \gamma_k = W\right\}$。注意到 π将E_W 映射到 $\operatorname{relint}\left(\Delta^{|S|}\right)$，因此由 (10) 得到

$$\inf_{\gamma\in E_W} f(\gamma) = W^{-\nu} \inf_{\tilde{\gamma}\in\operatorname{relint}\left(\Delta^{|S|}\right)} \sum_{i=1}^{|S|} \lambda_{k_i} \tilde{\gamma}_i^{-\nu} = W^{-\nu} \left(\sum_{i=1}^{|S|} \lambda_{k_i}^{\frac{1}{\nu+1}}\right)^{\nu+1} ,$$

其中，最后一个等式由情形 $S = [K]$ 得出。特别地，这意味着

$$\inf_{\gamma\in\operatorname{relint}(\Delta^K)} f(\gamma) = \inf_{W\in(0,1)} \inf_{\gamma\in E_W} f(\gamma) = \inf_{W\in(0,1)} W^{-\nu} \left(\sum_{i=1}^{|S|} \lambda_{k_i}^{\frac{1}{\nu+1}}\right)^{\nu+1} = \left(\sum_{k\in[K]} \lambda_k^{\frac{1}{\nu+1}}\right)^{\nu+1} ,$$

其中我们用到了 $\inf_{W\in(0,1)} W^{-\nu} = \lim_{W\to 1^-} W^{-\nu} = 1$。 □

引理 23 (改变切空间上的范数)。 *1.* 对于 $0 < \beta < 1$， 我们有最小值$_{v\in \mathrm{B}^K} \|v\|_\beta = 2^{\frac{1}{\beta}-1}$，且该最小值在 $^-$

$$v_k = \begin{cases} (-1)^k \cdot \frac{1}{2} & \text{for } k = 1, 2 \\ 0 & \text{otherwise;} \end{cases} \tag{11}$$

2. 对于 $\beta > 1$， 我们有 最小值$_{v\in \mathrm{B}^K} \|v\|_\beta = 2^{-1}\left(\lfloor K/2\rfloor^{1-\beta} + \lceil K/2\rceil^{1-\beta}\right)^{1/\beta}$，且最小值在以下位置取得

$$v_k = \begin{cases} \frac{1}{2\lfloor \frac{K}{2}\rfloor} & \text{for } 1 \leq k \leq \lfloor \frac{K}{2}\rfloor \\ -\frac{1}{2\lceil \frac{K}{2}\rceil} & \text{for } \lfloor \frac{K}{2}\rfloor < k \leq K. \end{cases} \tag{12}$$

证明。令 $v \in \mathrm{B}^K$，并设 $P := \{k \in [K] : v_k > 0\}$ 和 $N := \{k \in [K] : v_k < 0\}$。根据 B^K 的定义，我们有 $\sum_{k\in P} |v_k| + \sum_{k\in N} |v_k| = 1$ 和 $\sum_{k\in P} |v_k| - \sum_{k\in N} |v_k| = 0$，因此 $\sum_{k\in P} |v_k| = \sum_{k\in N} |v_k| = \frac{1}{2}$。

1：对于 $0 < \beta < 1$，函数 $x \mapsto x^\beta$ 在 $[0, +\infty)$ 上严格凹。此外，对于每个 x、$y \geq 0$，我们有 $x^\beta + y^\beta \geq (x + y)^\beta$，因此

$$\sum_{k\in[K]} |v_k|^\beta = \sum_{k\in P} |v_k|^\beta + \sum_{k\in N} |v_k|^\beta \geq \left(\sum_{k\in P} |v_k|\right)^\beta + \left(\sum_{k\in N} |v_k|\right)^\beta = 2 \cdot \frac{1}{2^\beta} = 2^{1-\beta} ,$$

即 $\|v\|_\beta \geq 2^{\frac{1}{\beta}-1}$。最后请注意，当且仅当 $|P| = |N| = 1$时，该不等式取等号；对于按（11）定义的 v，这一条件成立。

2：对于 $\beta > 1$ 函数 $x \mapsto x^\beta$ 在 $[0, +\infty)$ 上严格凸，由詹森不等式可得

$$\sum_{k\in P} |v_k|^\beta = |P| \cdot \sum_{k\in P} \frac{|v_k|^\beta}{|P|} \geq |P| \cdot \left(\sum_{k\in P} \frac{|v_k|}{|P|}\right)^\beta = |P| \cdot \left(\frac{1}{2|P|}\right)^\beta = 2^{-\beta} \cdot |P|^{1-\beta} , \tag{13}$$

其中，当且仅当 v_k 对每个 $k \in P$都取相同的值时取等号，即当且仅当对于每个 $k \in P$，$v_k = 1/\big(2|P|\big)$。类似地，$\sum_{k\in N} |v_k|^\beta \geq 2^{-\beta}|N|^{1-\beta}$，其中当且仅当对于每个 $k \in N$，$v_k = -1/\big(2|N|\big)$ 时取等号。由此可得

$$\sum_{k\in[K]} |v_k|^\beta \geq 2^{-\beta} \cdot \Big(|P|^{1-\beta} + |N|^{1-\beta}\Big) . \tag{14}$$

现在我们来确定使 (14) 右端最小化的整数 $|P|$ 与 $|N|$ 的可行取值。不失一般性，我们假设 $|P| \leq |N|$，并观察到使 $|P|$、$|N|$ 最小化的取值满足约束 $|P| + |N| = K$，即 $\{P, N\}$ 是 $[K]$ 的一个划分，这是 $|N|^{1-\beta} \geq (K - |P|)^{1-\beta}$ 的推论。接下来，我们观察到函数 $\phi(x) := x^{1-\beta} + (K-x)^{1-\beta}$ 在区间 $(0, K/2]$ 上严格递减，因此 (14) 右端的最小值在 $\big(|P|, |N|\big) = \big(\lfloor \frac{K}{2} \rfloor, \lceil \frac{K}{2} \rceil\big)$ 处取得，且等于 $2^{-\beta} \cdot \Big(\lfloor \frac{K}{2} \rfloor^{1-\beta} + \lceil \frac{K}{2} \rceil^{1-\beta}\Big)$。这恰好是按 (12) 所定义的 v 对应的 $\|v\|_\beta^\beta$ 的值。 □

5 证明

现在，我们已经具备完成主要定理证明所需的全部工具。

定理 *8* 的证明。根据命题 21，我们知道

$$C_{\alpha,K} = \inf \left\{ \sum_{k\in[K]} v_k^2 \gamma_k^{\alpha-2} : v \in \mathrm{B}^K, \gamma \in \operatorname{relint}\left(\Delta^K\right) \right\} .$$

计算这个下确界时，针对 α 和 K的不同取值需要采用不同的策略。**情形 $\alpha \leq 2$：** 注意到 $2 - \alpha \geq 0$，因此根据引理 22

$$C_{\alpha,K} = \inf_{v\in\mathrm{B}^K} \inf_{\gamma\in\operatorname{relint}(\Delta^K)} \sum_{k\in[K]} v_k^2 \gamma_k^{\alpha-2} = \inf_{v\in\mathrm{B}^K} \left(\sum_{k\in[K]} |v_k|^{\frac{2}{3-\alpha}}\right)^{3-\alpha} = \inf_{v\in\mathrm{B}^K} \|v\|_\beta^2 ,$$

其中我们设定 $\beta := \frac{2}{3-\alpha}$。如果 $\alpha < 1$，那么 $0 < \beta < 1$，因此由引理 23 可得

$$C_{\alpha,K} = \left(2^{\frac{1}{\beta}-1}\right)^2 = 2^{\frac{2}{\beta}-2} = 2^{1-\alpha} .$$

若 $\alpha = 1$，那么 $\beta = 1$，因此 $C_{1,K} = 1$。如果 $\alpha \in (1,2]$，那么 $\beta > 1$，因此由引理 23 可得

$$C_{\alpha,K} = 2^{-2}\left(\lfloor K/2\rfloor^{1-\beta} + \lceil K/2\rceil^{1-\beta}\right)^{2/\beta},$$

且对于 K 偶数，这意味着

$$C_{\alpha,K} = 2^{-2}\left(\left(\frac{K}{2}\right)^{1-\beta} + \left(\frac{K}{2}\right)^{1-\beta}\right)^{2/\beta} = 2^{-2}\left(2\cdot\frac{K^{1-\beta}}{2^{1-\beta}}\right)^{2/\beta} = \left(K^{1-\beta}\right)^{2/\beta} = K^{\frac{2}{\beta}-2} = K^{1-\alpha}$$

而对于 K 奇数，我们有

$$C_{\alpha,K} = 2^{-2}\left(\left(\frac{K-1}{2}\right)^{1-\beta} + \left(\frac{K+1}{2}\right)^{1-\beta}\right)^{2/\beta} = \left(2^{-\beta}\cdot\frac{(K-1)^{1-\beta}+(K+1)^{1-\beta}}{2^{1-\beta}}\right)^{2/\beta}$$
$$= \left(\frac{(K-1)^{\frac{1-\alpha}{3-\alpha}}+(K+1)^{\frac{1-\alpha}{3-\alpha}}}{2}\right)^{3-\alpha} = K^{1-\alpha}\left(\frac{(1-\frac{1}{K})^{\frac{1-\alpha}{3-\alpha}}+(1+\frac{1}{K})^{\frac{1-\alpha}{3-\alpha}}}{2}\right)^{3-\alpha}.$$

情形 $\alpha > 2$，$K = 2$： 注意，$\mathrm{B}^2 = \{\pm(1/2, -1/2)\}$，因此

$$C_{\alpha,2} = \inf\left\{\sum_{k\in[2]} v_k^2\gamma_k^{\alpha-2} : v\in\mathrm{B}^2, \gamma\in\mathrm{relint}\left(\Delta^2\right)\right\} = \inf\left\{\sum_{k\in[2]}\frac{1}{4}\gamma_k^{\alpha-2} : \gamma\in\mathrm{relint}\left(\Delta^2\right)\right\}.$$

考虑函数 $\gamma \mapsto f(\gamma) = \frac{1}{4}(\gamma_1^{\alpha-2} + \gamma_2^{\alpha-2})$，它在 Δ^2 上连续。由稠密性可得 $C_{\alpha,2} = \min\left\{f(\gamma) : \gamma\in\Delta^2\right\}$。若 $\alpha\in(2,3)$，则函数 f 在 Δ^2 上是凹函数，因此最小值在 Δ^2 的某个极点处取得，从而得到

$$C_{\alpha,2} = f\big((1,0)\big) = f\big((0,1)\big) = \frac{1}{4}(1^{\alpha-2}+0^{\alpha-2}) = 2^{-2}.$$

若 $\alpha = 3$，则对每个 $\gamma\in\Delta^2$，$f(\gamma) = \frac{1}{4}$，因此 $C_{\alpha,2} = \frac{1}{4} = 2^{-2}$。若 $\alpha\in(3,+\infty)$，则函数 f 在 Δ^2 上是凸函数，且其关于 $\gamma = (1/2, 1/2)$ 的对称性意味着

$$C_{\alpha,2} = f\big((1/2,1/2)\big) = \frac{1}{4}\left(\left(\frac{1}{2}\right)^{\alpha-2} + \left(\frac{1}{2}\right)^{\alpha-2}\right) = 2^{1-\alpha}.$$

情形 $\alpha > 2$，$K \geq 3$： 根据关于 K 的假设，存在某个 $\bar{v}\in\mathrm{B}^K$，使得 $\bar{v}_1 = 0$；我们可以据此定义函数 $\gamma\mapsto g(\gamma) := \sum_{k\in[K]}\bar{v}_k^2\gamma_k^{\alpha-2}$。特别地，对每个 $\gamma\in\mathrm{relint}\left(\Delta^K\right)$，都有 $C_{\alpha,K}\leq g(\gamma)$；由 g 在 Δ^K 上的连续性可知，该不等式对每个 $\gamma\in\Delta^K$ 都成立。我们很容易看出，若 $\bar{\gamma}\in\Delta^K$ 是单纯形的顶点，且当且仅当 $k = 1$ 时 $\bar{\gamma}_k = 1$，则 $g(\bar{\gamma}) = 0$，这意味着 $C_{\alpha,K} = 0$。 □

6 结论

我们为负 Tsallis 熵所诱导的 Bregman 散度建立了一个紧确的 Pinsker 型不等式，其最优常数为显式形式 $C_{\alpha,K}$ 对每个 (α, K) 均成立，并对相变给出了几何解释（包括 $\alpha > 2$ 时多分类问题的紧确失效，以及 $\alpha\in(1,2]$ 时精确的维数／奇偶效应）。除了具有信息论意义之外，该不等式还带来了直接的学习理论后果。首先，它提供了一种紧确的方法，可将 Tsallis（幂）损失的超额风险控制转化为预测分布在单纯形上的全变差控制，且该控制具有统一性。其次，通过标准的代入论证，

它为从 Tsallis 替代性能推导多分类 0–1 分类遗憾界提供了严谨路径，明确说明这种转化何时与维度无关，何时会随 K 而减弱。第三，在在线学习中，该结果确定了 Tsallis 正则项最优的 $\|\cdot\|_1$- 强凸性（曲率），从而改进了 Follow-the-Regularized-Leader ／镜像下降分析中涉及的常数，并明确揭示了 α 的选择如何决定算法背后的几何结构。

致谢

TC 衷心感谢渥太华大学通过拨款 GR002837（启动资金）提供的支持，以及加拿大自然科学与工程研究委员会（NSERC）通过拨款 RGPIN-2023-03688（探索研究计划）和 DGECR2023- 00208（探索研究计划，DGECR——探索启动补助）提供的支持。

RC 部分受意大利大学与研究部PRIN计划拨款 2022EKNE5K（市场与社会中的学习）资助。该拨款由国家复苏与韧性计划框架下的下一代欧盟计划提供资金；此外，RC 还获得欧盟地平线计划CL4-2021-HUMAN-01研究与创新行动的支持，资助协议编号为 101070617，项目为 ELSA（欧洲安全与可信人工智能灯塔）。

参考文献

Jacob D. Abernethy 和 Rafael M. Frongillo。线性性质评分规则的刻画。载于第*25*届年度学习理论会议论文集，第23卷机器学习研究论文集，第27.1–27.13页，爱丁堡，苏格兰，2012年6月25–27日。PMLR。

Jacob D. Abernethy、Chansoo Lee 和 Ambuj Tewari。用一种新型平滑性对抗多臂老虎机。载于神经信息处理系统进展，第28卷，第2197–2205页。Curran Associates，2015年。

Shun’ichi Amari。 α-散度是唯一同时属于 f-散度和 Bregman 散度类别的散度。*IEEE*信息论汇刊，55(11):4925–4931，2009年。

Ayanendranath Basu、Ian R. Harris、Nils L. Hjort 和 Michael C. Jones。通过最小化密度幂散度进行稳健且高效的估计。*Biometrika*，85(3):549–559，1998年。

Yanna Cruz Cavalcanti、Thomas Oberlin、Nicolas Dobigeon、Cédric Févotte、Simon Stute、Maria-Joao Ribeiro 和 Clovis Tauber。动态PET图像的因子分析：超越高斯噪声。*IEEE*医学影像汇刊，38(9):2231–2241，2019年。

Nicolò Cesa-Bianchi 和 Gábor Lugosi。预测器对数损失的最坏情况界。机器学习，43(3):247–264，2001年。

Nicolò Cesa-Bianchi 和 Gábor Lugosi。《预测、学习与博弈》。剑桥大学出版社，2006年。ISBN 9780521841085。Nicolò Cesa-Bianchi、Tommaso Cesari、Roberto Colomboni、Claudio Gentile 和 Yishay Mansour。具有复合匿名反馈的非随机多臂老虎机。机器学习研究杂志，23(277):1–24，2022年。Andrzej Cichocki 和 Shun-ichi Amari。α、β 和 γ 散度族：灵活且稳健的相似性度量。*Entropy*，12(6):1532–1568，2010年。

Alexander Philip Dawid。适当评分规则的几何。统计数学研究所年刊，59(1):77–93，2007年。Alexander Philip Dawid 和 Monica Musio。适当评分规则的理论与应用。*Metron*，72(2):169–183，2014年。Alexander Philip Dawid、Monica Musio 和 Laura Ventura。最小评分规则推断。斯堪的纳维亚统计学杂志，43(1):123–138，2016年。Cédric Févotte 和 Nicolas Dobigeon。采用稳健非负矩阵分解进行非线性高光谱解混。*IEEE*图像处理汇刊，24(12):4810–4819，2015年。Cédric Févotte 和 Jérôme Idier。基于β-散度的非负矩阵分解算法。神经计算，23(9):2421–2456，2011年。Gustavo L. Gilardoni。关于 Csiszár 的 f-散度的 Pinsker 型和 Vajda 型不等式。*IEEE*信息论汇刊，56(11):5377–5386，2010年。Federica Giummolè、Valentina Mameli、Erlis Ruli 和 Laura Ventura。基于适当评分规则的客观贝叶斯推断。*Test*，28(3):728–755，2019年。Tilmann Gneiting 和 Adrian E. Raftery。严格适当评分规则、预测与估计。美国统计学会会刊，102(477):359–378，2007年。Nika Haghtalab、Cameron Musco 和 Bo Waggoner。迈向分布学习损失函数的特征刻画。载于神经信息处理系统进展，第32卷，第7237–7246页。Curran Associates，2019年。Romain Hennequin、Bertrand David 和 Roland Badeau。Beta散度是 Bregman 散度的一个子类。*IEEE*信号处理快报，18(2):83–86，2010年。Jack Jewson、Jim Q. Smith 和 Chris Holmes。使用一般散度准则进行贝叶斯推断的原则。*Entropy*，20(6):442，2018年。Takafumi Kanamori 和 Hironori Fujisawa。与适当复合评分规则相关的仿射不变散度及其应用。*Bernoulli*，20(4):2278–2304，2014年。Brian Kulis 和 Peter L. Bartlett。隐式在线学习。载于第*27*届国际机器学习会议（*ICML-10*）论文集，第575–582页，2010年。Nicolas S. Lambert、David M. Pennock 和 Yoav Shoham。概率分布属性的引出。载于第*9*届*ACM*电子商务会议论文集，EC '08，第129–138页，纽约，2008年。计算机协会。Valentina Mameli 和 Laura Ventura。评分规则的高阶渐近理论。统计规划与推断杂志，165:13–26，2015年。Saeed Masoudian 和 Yevgeny Seldin。对随机约束的对抗性多臂老虎机以及受对抗性污染的随机多臂老虎机中的 Tsallis-INF算法进行改进分析。载于第三十四届学习理论会议论文集，第134卷机器学习研究论文集，第3330–3350页。PMLR，2021年8月15日至19日。Amarnath Nandy、Ayanendranath Basu 和 Abhik Ghosh。健康科学中偏态数据的稳健推断。应用统计学杂志，49(8):2093–2123，2022年。

Alexey Ozerov 和 CédricFévotte。用于音频源分离的卷积混合中的多通道非负矩阵分解。*IEEE* 音频、语音与语言处理汇刊，18(3)：550–563，2009年。

Mark D. Reid 和 Robert C. Williamson。广义 Pinsker 不等式。载于 *COLT 2009 - 第 22 届学习理论会议，加拿大魁北克省蒙特利尔，2009年 6 月 18–21 日*。计算机协会，2009a年。

Mark D. Reid 和 Robert C. Williamson。适当损失的替代遗憾界。载于第 *26* 届年度国际机器学习会议论文集，ICML '09，第 897–904 页，纽约，2009b年。计算机协会。

Mark D. Reid 和 Robert C. Williamson。二元实验的信息、散度与风险。《机器学习研究杂志》，12(22)：731–817，2011年。

Olivier Rioul。《Schützenberger-Pinsker 不等式》的历史视角。载于 Frank Nielsen 和 Frédéric Barbaresco 编，《信息几何科学：第六届国际会议 GSI 2023，法国圣马洛，2023年8月30日至9月1日，会议论文集，第一部分》，第291–306页，Cham，2023年。Springer。

Leonard J. Savage。个人概率与期望的 elicitation。《美国统计学会期刊》，66(336)：783–801，1971年。

Shai Shalev-Shwartz。在线学习与在线凸优化。《机器学习基础与趋势》®，4(2)：107–194，2012年。

Alexandre B. Tsybakov。《非参数估计导论》。Springer，纽约，200 8.

Thijs van Ommen、Wouter M. Koolen、Thijs E. Feenstra 和 Peter D. Grünwald。稳健概率更新。《近似推理国际期刊》，74：30–57，2016年。

Julian Zimmert 和 Yevgeny Seldin。随机与对抗性多臂老虎机的最优算法。载于第二十二届国际人工智能与统计会议论文集，机器学习研究论文集第89卷，第467–475页。PMLR，2019年4月16–18日。

Julian Zimmert 和 Yevgeny Seldin。Tsallis-INF：随机与对抗性多臂老虎机的最优算法。《机器学习研究杂志》，22(28):1–49，2021年。

Pinsker 不等式：全变差与 $\|\cdot\|_1$

引理24（全变差与 $\|\cdot\|_1$）。设 $K \in \mathbb{N}$ 满足 $K \geq 2$，并设 p、$q \in \Delta^K$ 为 $[K]$ 上的概率分布。其中

$$\|p-q\|_{TV} \coloneqq \sup_{C \subset [K]} \left| \sum_{k \in C} p(k) - \sum_{k \in C} q(k) \right|, \qquad \|p-q\|_1 \coloneqq \sum_{k=1}^{K} |p(k)-q(k)| \ ,$$

则有

$$\|p-q\|_{TV} = \frac{1}{2} \cdot \|p-q\|_1 \ .$$

证明。对 $k \in [K]$，令 $d(k) \coloneqq p(k) - q(k)$。由于 $p,\ q \in \Delta^K$，

$$\sum_{k=1}^{K} d(k) = \sum_{k=1}^{K} p(k) - \sum_{k=1}^{K} q(k) = 1 - 1 = 0 \,.$$

定义指标集

$$P \coloneqq \big\{k \in [K] : d(k) \geq 0\big\} \,, \qquad N \coloneqq \big\{k \in [K] : d(k) < 0\big\} \,.$$

对任意 $C \subset [K]$，我们有

$$\sum_{k \in C} d(k) \leq \sum_{k \in C \cap P} d(k) \leq \sum_{k \in P} d(k) \,,$$

so s$\mathrm{up}_{C \subset [K]} \sum_{k \in C} d(k) \leq \sum_{k \in P} d(k)$，并且通过选取 $C = P$达到等式。因此 e

$$\sup_{C \subset [K]} \sum_{k \in C} d(k) = \sum_{k \in P} d(k) \,.$$

类似地，将同一论证应用于 $-d$ 可得

$$\sup_{C \subset [K]} \sum_{k \in C} \big(-d(k)\big) = \sum_{k \in N} \big(-d(k)\big) \,.$$

但由于 $\sum_{k=1}^{K} d(k) = 0$，我们有

$$\sum_{k \in P} d(k) = -\sum_{k \in N} d(k) = \sum_{k \in N} \big(-d(k)\big) \,.$$

因此，

$$\|p - q\|_{\mathrm{TV}} = \sup_{C \subset [K]} \left| \sum_{k \in C} d(k) \right| = \max \left(\sup_{C \subset [K]} \sum_{k \in C} d(k), \sup_{C \subset [K]} \sum_{k \in C} \big(-d(k)\big) \right) = \sum_{k \in P} d(k) \,.$$

另一方面，

$$\|p - q\|_1 = \sum_{k=1}^{K} |d(k)| = \sum_{k \in P} d(k) + \sum_{k \in N} \big(-d(k)\big) = 2 \cdot \sum_{k \in P} d(k) \,.$$

结合最后两个公式可得 $\|p - q\|_{\mathrm{TV}} = \frac{1}{2} \cdot \|p - q\|_1$。 – □

B Tsallis 损失及其 Bayes 风险与超额风险

本节证明，对于任意 $\alpha \in \mathbb{R}$，式（1）中的 Tsallis（或幂）损失 ℓ_α（其中 $\ell_0(q,\ k) \coloneqq q_k^{-1} - K + \sum_{i \in [K]} \ln\ q_i$，且 $\ell_1(q,\ k) \coloneqq -\ln\ q_k$）所对应的贝叶斯风险和超额风险，分别是 Tsallis 熵 S_α（定义1）及其 Bregman 散度 D_α（定义3）。具体而言，若 $\alpha \in \mathbb{R} \setminus \{0, 1\}$，则对于任意 $p,\ q \in \mathrm{relint}\ (\Delta^K)$，有

$$\begin{aligned} \mathbb{E}_{Y \sim p} \left[\ell_\alpha(q, Y)\right] &= \sum_{k \in [K]} \ell_\alpha(q, k) \cdot p_k = \sum_{k \in [K]} \left(\frac{q_k^{\alpha-1}}{1-\alpha} + \frac{1}{\alpha} \sum_{i \in [K]} q_i^\alpha \right) \cdot p_k \\ &= \sum_{k \in [K]} \frac{\alpha \cdot p_k \cdot q_k^{\alpha-1} + (1-\alpha) q_k^\alpha}{\alpha(1-\alpha)} \end{aligned}$$

在贝叶斯风险的情形 $p=q$ 下，上式变为

$$\mathbb{E}_{Y\sim p}\left[\ell_\alpha(p,Y)\right]=\sum_{k\in[K]}\frac{\alpha\cdot p_k\cdot p_k^{\alpha-1}+(1-\alpha)p_k^\alpha}{\alpha(1-\alpha)}=\sum_{k\in[K]}\frac{p_k^\alpha}{\alpha(1-\alpha)}=S_\alpha(p)\ ,$$

而超额风险变为

$$\begin{aligned}\mathbb{E}_{Y\sim p}\left[\ell_\alpha(q,Y)\right]-\mathbb{E}_{Y\sim p}\left[\ell_\alpha(p,Y)\right]&=\sum_{k\in[K]}\frac{\alpha\cdot p_k\cdot q_k^{\alpha-1}+(1-\alpha)q_k^\alpha}{\alpha(1-\alpha)}-\sum_{k\in[K]}\frac{p_k^\alpha}{\alpha(1-\alpha)}\\&=\frac{1}{\alpha}\cdot\sum_{k\in[K]}\frac{p_k^\alpha+(\alpha-1)q_k^\alpha-\alpha p_kq_k^{\alpha-1}}{\alpha-1}=d_\alpha(p\|q)=D_\alpha(p\|q)\ ,\end{aligned}$$

其中， d_α 是根据定义6为 $\beta=\alpha$ 定义的 β-散度，最后一个不等式由备注 7得出。

对于 $\alpha=0$的情形，如果 $p,\ q\in\mathsf{relint}\ (\Delta^K)$，则有

$$\mathbb{E}_{Y\sim p}\left[\ell_0(q,Y)\right]=\sum_{k\in[K]}\ell_0(q,k)\cdot p_k=\sum_{k\in[K]}\left(q_k^{-1}-K+\sum_{i\in[K]}\ln q_i\right)\cdot p_k=\sum_{k\in[K]}\left(\frac{p_k}{q_k}-1+\ln q_k\right)$$

对于贝叶斯风险情形 $p=q$，这变为

$$\mathbb{E}_{Y\sim p}\left[\ell_0(p,Y)\right]=\sum_{k\in[K]}\left(\frac{p_k}{p_k}-1+\ln p_k\right)=\sum_{k\in[K]}\ln p_k=S_0(p)$$

而超额风险变为

$$\begin{aligned}\mathbb{E}_{Y\sim p}\left[\ell_0(q,Y)\right]-\mathbb{E}_{Y\sim p}\left[\ell_0(p,Y)\right]&=\sum_{k\in[K]}\left(\frac{p_k}{q_k}-1+\ln q_k\right)-\sum_{k\in[K]}\ln p_k\\&=\sum_{k=1}^K\left(\frac{p_k}{q_k}-\ln\frac{p_k}{q_k}-1\right)=d_0(p\|q)=D_0(p\|q)\ ,\end{aligned}$$

其中， d_0 是根据定义6为 $\beta=0$ （或板仓-斋藤散度）定义的 β-散度，最后一个不等式由备注 7得出。

对于情形 $\alpha=1$，若 $p,\ q\in\mathsf{relint}\ (\Delta^K)$，则有

$$\mathbb{E}_{Y\sim p}\left[\ell_1(q,Y)\right]=\sum_{k\in[K]}\ell_1(q,k)\cdot p_k=\sum_{k\in[K]}(-\ln q_k)\cdot p_k=-\sum_{k\in[K]}p_k\cdot\ln q_k\ ,$$

对于贝叶斯风险情形 $p=q$，上式变为

$$\mathbb{E}_{Y\sim p}\left[\ell_1(p,Y)\right]=-\sum_{k\in[K]}p_k\cdot\ln p_k=S_1(p)$$

而超额风险变为

$$\begin{aligned}\mathbb{E}_{Y\sim p}\left[\ell_1(q,Y)\right]-\mathbb{E}_{Y\sim p}\left[\ell_1(p,Y)\right]&=-\sum_{k\in[K]}p_k\cdot\ln q_k+\sum_{k\in[K]}p_k\cdot\ln p_k\\&=\sum_{k\in[K]}p_k\cdot\ln\frac{p_k}{q_k}=D_{\mathrm{KL}}(p\|q)=D_1(p\|q)\ ,\end{aligned}$$

其中 $D_{\mathrm{K\,L}}$ 是 Kullback-Leibler散度，最后一个不等式由注5得出 .

C 0–1 损失遗憾由 $\|\cdot\|_1$ 距离控制

引理 25 (0-1 损失遗憾受 $\|\cdot\|_1$ 距离控制)。 对于任意 $p, q \in \Delta^K$，确定性地选取 $k_p^\star \in \operatorname{argmax}_{k\in[K]} p_k$ 和 $k_q^\star \in \operatorname{argmax}_k q_k$。于是

$$\mathbb{P}_{Y\sim p}\left[Y \neq k_q^\star\right] - \mathbb{P}_{Y\sim p}\left[Y \neq k_p^\star\right] \leq \|p-q\|_1 \ .$$

证明。 如果 $k_q^\star = k_p^\star$，则无需证明。因此假设 $k_q^\star \neq k_p^\star$。于是，我们有

$$\begin{aligned}
\mathbb{P}_{Y\sim p}\left[Y \neq k_q^\star\right] - \mathbb{P}_{Y\sim p}\left[Y \neq k_p^\star\right] &= 1 - p_{k_q^\star} - (1 - p_{k_p^\star}) = p_{k_p^\star} - p_{k_q^\star} \\
&= p_{k_p^\star} - q_{k_p^\star} + \underbrace{q_{k_p^\star} - q_{k_q^\star}}_{\leq 0} + q_{k_q^\star} - p_{k_q^\star} \\
&\leq p_{k_p^\star} - q_{k_p^\star} + q_{k_q^\star} - p_{k_q^\star} \\
&= \left|p_{k_p^\star} - q_{k_p^\star}\right| + \left|p_{k_q^\star} - q_{k_q^\star}\right| \leq \sum_{k=1}^K |p_k - q_k| = \|p-q\|_1 \ .
\end{aligned}$$

□

D 备注的证明

备注 7 的证明。 对于任意 $\alpha \in \mathbb{R}$， $-S_\alpha$ 的 Bregman 散度 D_α 与β-散度 d_β一致，其中 $\beta = \alpha$。事实上，对于任意 p， $q \in (0, +\infty)^K$，我们有

$$\begin{aligned}
D_\alpha(p\|q) &= -\frac{\sum_{k\in[K]} p_k^\alpha}{\alpha(1-\alpha)} + \frac{\sum_{k\in[K]} q_k^\alpha}{\alpha(1-\alpha)} - \sum_{k\in[K]} \frac{1}{\alpha-1} q_k^{\alpha-1}(p_k - q_k) \\
&= \frac{1}{\alpha} \cdot \sum_{k\in[K]} \frac{p_k^\alpha + (\alpha-1)q_k^\alpha - \alpha p_k q_k^{\alpha-1}}{\alpha-1} = d_\alpha(p\|q) \ , \qquad \alpha \notin \{0,1\}
\end{aligned}$$

而对于 $\alpha = 0$，我们有

$$D_0(p\|q) = -\sum_{k\in[K]} \ln p_k + \sum_{k\in[K]} \ln q_k + \sum_{k\in[K]} \frac{p_k - q_k}{q_k} = \sum_{k=1}^K \left(\frac{p_k}{q_k} - \ln\frac{p_k}{q_k} - 1\right) = d_0(p\|q) \ ,$$

这也称为板仓-斋藤散度；对于 $\alpha = 1$，我们有

$$\begin{aligned}
D_1(p\|q) &= \sum_{k\in[K]} p_k \ln p_k - \sum_{k\in[K]} q_k \ln q_k - \sum_{k\in[K]} (1 + \ln q_k)(p_k - q_k) \\
&= \sum_{k\in[K]} \left(p_k \cdot \ln\frac{p_k}{q_k} - p_k + q_k\right) = d_1(p\|q) \ ,
\end{aligned}$$

现在，我们证明连续性结论。仅有的两个非平凡情形是 $\alpha \in \{0,1\}$。对于情形 $\alpha = 0$，任取 p， $q \in (0, +\infty)^K$，注意到该极限是一个未定型 $0/0$，

我们可以应用洛必达法则 (H)，得到

$$\begin{aligned}
\lim_{\alpha\to 0} D_\alpha(p\|q) &= \lim_{\alpha\to 0}\frac{1}{\alpha}\cdot\sum_{k\in[K]}\frac{p_k^\alpha+(\alpha-1)q_k^\alpha-\alpha p_k q_k^{\alpha-1}}{\alpha-1}\\
&= \sum_{k\in[K]}\lim_{\alpha\to 0}\frac{-p_k^\alpha+(1-\alpha)q_k^\alpha+\alpha p_k q_k^{\alpha-1}}{\alpha}\\
&\overset{\text{(H)}}{=} \sum_{k\in[K]}\lim_{\alpha\to 0}\left(-p_k^\alpha\ln p_k-q_k^\alpha+(1-\alpha)q_k^\alpha\ln q_k+p_k q_k^{\alpha-1}+\alpha p_k q_k^{\alpha-1}\ln q_k\right)\\
&= \sum_{k\in[K]}\left(-\ln p_k-1+\ln q_k+p_k q_k^{-1}\right)\\
&= \sum_{k\in[K]}\left(\frac{p_k}{q_k}-\ln\frac{p_k}{q_k}-1\right)=D_0(p\|q)\ .
\end{aligned}$$

对于 $\alpha=1$ 的情形，对于任意 $p,\ q\in(0,\ +\infty)^K$，注意到该极限仍是一个未定式$0/0$，我们可以应用洛必达法则 (H) 得到

$$\begin{aligned}
\lim_{\alpha\to 1} D_\alpha(p\|q) &= \lim_{\alpha\to 1}\frac{1}{\alpha}\cdot\sum_{k\in[K]}\frac{p_k^\alpha+(\alpha-1)q_k^\alpha-\alpha p_k q_k^{\alpha-1}}{\alpha-1}\\
&= \sum_{k\in[K]}\lim_{\alpha\to 1}\frac{p_k^\alpha+(\alpha-1)q_k^\alpha-\alpha p_k q_k^{\alpha-1}}{\alpha-1}\\
&\overset{\text{(H)}}{=} \sum_{k\in[K]}\lim_{\alpha\to 1}\left(p_k^\alpha\ln p_k+q_k^\alpha+(\alpha-1)q_k^\alpha\ln q_k-p_k q_k^{\alpha-1}-\alpha p_k q_k^{\alpha-1}\ln q_k\right)\\
&= \sum_{k\in[K]}\left(p_k\ln p_k+q_k-p_k-p_k\ln q_k\right)\\
&= \sum_{k\in[K]}\left(p_k\cdot\ln\frac{p_k}{q_k}-p_k+q_k\right)=D_1(p\|q)\ .
\end{aligned}$$

□

注 *15* 的证明。对于 $\alpha\neq 2$，该注等价于证明

$$\inf_{\substack{p,q\in(0,+\infty)^K\\ p\neq q}}\frac{2D_\alpha(p\|q)}{\|p-q\|_1^2}=0\ .$$

令 $\mathbf{1}\in\mathbb{R}^K$ 为全一向量，e_1 为第一标准基向量。利用 $-S_\alpha$ 在 q 处的二阶泰勒展开，以及海森矩阵表达式 $H(-S_\alpha)(q)=\mathrm{diag}(q_1^{\alpha-2},\ \ldots,\ q_K^{\alpha-2})$（注16），对于 $q\in(0,\ +\infty)^K$ 和 $\varepsilon\to 0^+$，我们得到

$$D_\alpha(q+\varepsilon\cdot e_1\|q)=\frac{\varepsilon^2}{2}\cdot q_1^{\alpha-2}+o(\varepsilon^2)\ .$$

现在取 $q(t)\coloneqq t\cdot\mathbf{1}$，于是 $q_1(t)=t$，并注意到 $\left\|\big(q(t)+\varepsilon\cdot e_1\big)-q(t)\right\|_1=\varepsilon$。因此

$$\varphi_\alpha(\varepsilon,t)\coloneqq\frac{2D_\alpha\big(q(t)+\varepsilon\cdot e_1\|q(t)\big)}{\left\|\big(q(t)+\varepsilon\cdot e_1\big)-q(t)\right\|_1^2}=t^{\alpha-2}+o(1)\ ,\qquad\text{as }\varepsilon\to 0^+\ .$$

若 $\alpha<2$，令 $\varepsilon:=\frac{1}{t}$ 并发送 $t\to+\infty$，使得 $\varphi_\alpha\left(\frac{1}{t},\ t\right)\to 0$。若 $\alpha>2$，令 $\varepsilon:=t$ 并发送$t\to 0^+$，使得 $\varphi_\alpha(t,\ t)\to 0$。这就证明了该结论。

In c相反，当 $\alpha=2$ 时，正如我们在备注 12 的脚注中已经指出的，我们有 tha t

$$D_2(p\|q)=\frac{1}{2}\cdot\|p-q\|_2^2\qquad\text{for all } p,q\in(0,+\infty)^K\ .$$

因此，由詹森不等式可知，对任意 $x\in\mathbb{R}^K$，我们有

$$\|x\|_2=\left(\sum_{k\in[K]}x_k^2\right)^{1/2}=\sqrt{K}\cdot\left(\frac{1}{K}\sum_{k\in[K]}x_k^2\right)^{1/2}\geq\frac{1}{\sqrt{K}}\sum_{k=1}^K|x_k|=\frac{1}{\sqrt{K}}\cdot\|x\|_1\ ,$$

我们也可得到

$$D_2(p\|q)=\frac{1}{2}\cdot\|p-q\|_2^2\geq\frac{1}{2K}\cdot\|p-q\|_1^2,\qquad p,q\in(0,+\infty)^K\ ,$$

等号在 $q:=p+c\cdot\mathbf{1}$ 时成立；对任意 $c>0$ 以及任意 $p\in(0,\ +\infty)^K$。 □

备注 *16* 的证明。若 $\alpha\in(-\infty,\ 0)\cup(0,\ 1)\cup(1,\ +\infty)$，则对任意 $i,\ j\in[K]$，我们有 tha t

$$\partial_i S_\alpha(p)=-\frac{1}{\alpha-1}p_i^{\alpha-1},$$

由此立即可得，当且仅当 $j\neq i$ 时，$\partial_j\partial_i S_\alpha=0$，而

$$\partial_{ii}^2 S_\alpha(p)=\partial_i\left(-\frac{1}{\alpha-1}p_i^{\alpha-1}\right)=-p_i^{\alpha-2}\ .$$

若 $\alpha=0$，则有 $S_0(p)=\sum_{k\in[K]}\ln p_k$，因此对任意 $i\in[K]$,

$$\partial_i S_0(p)=\frac{1}{p_i},\qquad\partial_{ii}^2 S_0(p)=-\frac{1}{p_i^2},\qquad\partial_{ij}^2 S_0(p)=0\ \ (i\neq j).$$

由于当 $\alpha=0$ 时 $p_i^{-2}=p_i^{\alpha-2}$，这与所声称的公式一致。

若 $\alpha=1$，则有 $S_1(p)=-\sum_{k\in[K]}p_k\ln p_k$，因此对任意 $i\in[K]$,

$$\partial_i S_1(p)=-(1+\ln p_i),\qquad\partial_{ii}^2 S_1(p)=-\frac{1}{p_i},\qquad\partial_{ij}^2 S_1(p)=0\ \ (i\neq j).$$

由于当 $\alpha=1$ 时 $p_i^{-1}=p_i^{\alpha-2}$，这再次与所声称的公式一致。

综合以上各种情况，对每个 $\alpha\in\mathbb{R}$ 和 $p\in(0,+\infty)^K$,

$$\partial_{ij}^2 S_\alpha(p)=-\,p_i^{\alpha-2}\,\mathbb{I}\{i=j\},$$

因此 $HS_\alpha(p)=-\mathrm{diag}(p_1^{\alpha-2},\ldots,p_K^{\alpha-2})$。 □

注 *18* 的证明。给定 $C\geq 0$ 和 $p,\ q\in\mathrm{relint}\ (\Delta^K)$，不等式

$$D_\alpha(p\|q)\geq\frac{C}{2}\cdot\|p-q\|_1^2\tag{15}$$

当 $p=q$ 时显然成立，而对于 $p\neq q$，当且仅当以下条件成立时，该不等式才成立

$$2\cdot\frac{D_\alpha(p\|q)}{\|p-q\|_1^2}\geq C\ .$$

由此可见，$C\geq 0$ 满足 (15) 对任意 $p,\ q\in\mathrm{relint}\ (\Delta^K)$ 成立，当且仅当 if

$$2\cdot\inf\left\{\frac{D_\alpha(p\|q)}{\|p-q\|_1^2}:p,q\in\mathrm{relint}\left(\Delta^K\right),p\neq q\right\}\geq C\ ,\tag{16}$$

而最大的此类 C 等于（16）式的左侧。 □

D.1 备注 11 的证明：估计常数 $\sigma_{\alpha,K}$

命题 26。设 $\alpha \in (1, 2)$ 以及 $K \geq 3$。则

$$\sigma_{\alpha,K} \coloneqq \left(\frac{(1-\frac{1}{K})^{\frac{1-\alpha}{3-\alpha}} + (1+\frac{1}{K})^{\frac{1-\alpha}{3-\alpha}}}{2} \right)^{3-\alpha}$$

满足

$$1 + \frac{\alpha-1}{3-\alpha} \cdot \frac{1}{K^2} \leq \sigma_{\alpha,K} \leq 1 + \frac{7(\alpha-1)}{6(3-\alpha)} \cdot \frac{1}{K^2} \,. \tag{17}$$

证明。定义 $\nu \coloneqq (1-\alpha)/(3-\alpha)$，并注意到 $\nu \in (-1, 0)$。接下来，定义

$$z(\alpha, K) \coloneqq \frac{(1-\frac{1}{K})^{\nu} + (1+\frac{1}{K})^{\nu}}{2} - 1 \,,$$

通过考察 $(1+x)^{\nu}$ 和 $x = \pm 1/K$ 的二项级数，可将其展开为

$$z(\alpha, K) = \sum_{n=1}^{\infty} \binom{\nu}{2n} \frac{1}{K^{2n}} \,, \tag{18}$$

根据递推关系可知，系数均为非负数

$$\binom{\nu}{2n+2} = \binom{\nu}{2n} \cdot \frac{(\nu-2n)(\nu-2n-1)}{(2n+1)(2n+2)} \,. \tag{19}$$

特别地，函数 $K \mapsto z(\alpha, K)$ 是 $1/K^2$ 的增函数，并且

$$z(\alpha, K) \geq \binom{\nu}{2} \cdot \frac{1}{K^2} = \frac{\alpha-1}{(3-\alpha)^2} \cdot \frac{1}{K^2} \,.$$

该不等式与伯努利不等式结合可得

$$1 + \frac{\alpha-1}{(3-\alpha)} \cdot \frac{1}{K^2} \leq 1 + (3-\alpha) \cdot z(\alpha, K) \leq (1 + z(\alpha, K))^{3-\alpha} = \sigma_{\alpha,K} \,,$$

这就证明了（17）中的下界。对于上界，首先注意到 $K \geq 3$，并且根据（19），$\binom{\nu}{2n}$ 关于 n 单调递减，因此

$$z(\alpha, K) \leq \binom{\nu}{2} \cdot \sum_{n=1}^{\infty} \frac{1}{K^{2n}} \leq \binom{\nu}{2} \cdot \frac{1}{K^2} \cdot \sum_{n=0}^{\infty} \frac{1}{3^{2n}} = \frac{9(\alpha-1)}{8(3-\alpha)^2} \cdot \frac{1}{K^2} \,.$$

根据伯努利不等式，

$$\begin{aligned} \sigma_{\alpha,K} &= (1 + z(\alpha, K)) \cdot (1 + z(\alpha, K))^{2-\alpha} \\ &\leq (1 + z(\alpha, K)) \cdot (1 + (2-\alpha) \cdot z(\alpha, K)) \\ &= 1 + (3-\alpha) \cdot z(\alpha, K) \cdot \left(1 + \frac{2-\alpha}{3-\alpha} \cdot z(\alpha, K) \right) \,, \end{aligned}$$

因此

$$\begin{aligned}\sigma_{\alpha,K} &\le 1+(3-\alpha)\cdot z(\alpha,K)\cdot\left(1+\frac{2-\alpha}{3-\alpha}\cdot z(\alpha,3)\right)\\ &\le 1+\frac{9(\alpha-1)}{8(3-\alpha)}\cdot\frac{1}{K^2}\cdot\left(1+(2-\alpha)\frac{9(\alpha-1)}{8(3-\alpha)^3}\frac{1}{3^2}\right)\\ &= 1+\frac{9(\alpha-1)}{8(3-\alpha)}\cdot\left(1+\frac{(2-\alpha)(\alpha-1)}{8(3-\alpha)^3}\right)\cdot\frac{1}{K^2}\\ &\le 1+\frac{9(\alpha-1)}{8(3-\alpha)}\cdot\left(1+\frac{1}{32}\right)\cdot\frac{1}{K^2}\\ &= 1+\frac{297(\alpha-1)}{256(3-\alpha)}\cdot\frac{1}{K^2}\,,\end{aligned}$$

利用 $1/32$ 是 $(2-\alpha)(\alpha-1)/\big(8(3-\alpha)^3\big)$ 对 $\alpha\in(1,2)$ 的上界。注意到我们有 $\frac{297}{256}\le\frac{7}{6}$，即可得出结论。 □

备注 11 中给出的不等式基于命题 26，以及以下观察：$0\le\frac{\alpha-1}{3-\alpha}\le 1$ 对于 $\alpha\in(1,2]$。

E No当 $\alpha>2$ 且 $K\ge 3$ 时，不存在 Pinsker 型不等式

可以直接证明，对于 $\alpha>2$和 $K\ge 3$，散度 D_α 不满足 Pinsker 型不等式，也就是说，不可能找到一个正 常数 C，使得对于每个 $p,q\in\mathrm{relint}\,(\Delta^K)$，都有$D_\alpha(p\|q)\ge C\cdot\|p-q\|_1^2$。

定理 27. 设 $K\ge 3$ 为整数。对于每个 $\alpha\in(2,+\infty)$ 以及每个 $\varepsilon>0$，存在$p,q\in$相对内部 (Δ^K)，使得

$$D_\alpha(p\|q)<\varepsilon\cdot\|p-q\|_1^2\ .$$

证明。定义

$$p_k(t)\coloneqq\begin{cases}1-(K-1)\cdot t & \text{for } k=1\\ \frac{3}{4}t & \text{for } k=2\\ \frac{5}{4}t & \text{for } k=3\\ t & \text{for } 3<k\le K\end{cases}\qquad q_k(t)\coloneqq\begin{cases}1-(K-1)\cdot t & \text{for } k=1\\ \frac{5}{4}t & \text{for } k=2\\ \frac{3}{4}t & \text{for } k=3\\ t & \text{for } 3<k\le K.\end{cases}$$

对于 $0<t\le\frac{1}{K-1}$，$p\coloneqq p(t)$ 和 $q\coloneqq q(t)$ 均属于 $\mathrm{relint}\,(\Delta^K)$。根据构造，$S_\alpha\big(p(t)\big)=S_\alpha\big(q(t)\big)$，因此

$$\begin{aligned}D_\alpha\big(p\|q\big) &= \langle\nabla S_\alpha(q),p-q\rangle\\ &= -\sum_{k\in[K]}\frac{1}{\alpha-1}\cdot q_k^{\alpha-1}\,(p_k-q_k)\\ &= -\frac{1}{\alpha-1}\left(\left(\frac{5}{4}\cdot t\right)^{\alpha-1}\cdot\left(-\frac{1}{2}\cdot t\right)+\left(\frac{3}{4}\cdot t\right)^{\alpha-1}\cdot\left(\frac{1}{2}\cdot t\right)\right)\\ &= t^\alpha\cdot\frac{1}{2(\alpha-1)}\cdot\frac{5^{\alpha-1}-3^{\alpha-1}}{4^{\alpha-1}}\,,\end{aligned}$$

并注意到 $\|p-q\|_1 = t$，因此

$$\frac{D_\alpha(p\|q)}{\|p-q\|_1^2} = t^{\alpha-2}\cdot\frac{1}{2(\alpha-1)}\cdot\frac{5^{\alpha-1}-3^{\alpha-1}}{4^{\alpha-1}}\,.$$

根据关于 α 的假设，我们有 $t^{\alpha-2}\to 0$ ，因为 $t\to 0^+$，因此，对每个 $\varepsilon>0$取一个足够小的 $t=t(\varepsilon)$，即可得出结论。 □

我们强调，定理 27 的证明中给出的构造要求 $K\geq 3$。事实上，同样的论证对于 $K=2$并不成立。

Csiszár f-散度的 F Pinsker 不等式

本节回顾 Csiszár f-散度（特别是α-Tsallis相对熵）的关键定义与结果，并比较经典的、针对 f-散度的广义 Pinsker 不等式，以及我们针对负 α-Tsallis 熵所诱导的 Bregman 散度的广义 Pinsker 不等式。

定义 28. 如果 f 满足以下条件，则称其为可容许生成函数： $f\colon[0,+\infty)\to(-\infty,+\infty]$ 是一个凸函数，且对每个 $x>0$， f均为有限值；$(1)=0$，并且 $f(0)=\lim_{x\to 0^+}f(x)$。如果 f是可容许生成函数，且 $\mathbb{P}$ 和 $\mathbb{Q}$ 是定义在某个可测空间 $(\Omega,\mathcal{F})$ 上的概率测度，并满足 $\mathbb{P}\ll\mathbb{Q}$（这意味着 $\mathbb{P}$ 关于 $\mathbb{Q}$绝对连续，即对每个事件 $A\in\mathcal{F}$，若 $\mathbb{Q}[A]=0$则 $\mathbb{P}[A]=0$），那么$\mathbb{P}$ 相对于 $\mathbb{Q}$ 的 Csiszár f-散度定义为

$$D_f^{\mathrm{C}}(\mathbb{P},\mathbb{Q}) \coloneqq \int_\Omega f\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)\mathrm{d}\mathbb{Q}\,,$$

其中 $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}$ 是 $\mathbb{P}$ 关于 $\mathbb{Q}$的 Radon-Nikodym 导数。

对于任意在 1 处三次可微且满足 $f''(1)>0$的可容许生成函数 f，以及某个可测空间 $(\Omega,\mathcal{F})$ 上满足 $\mathbb{P}\ll\mathbb{Q}$的任意概率测度 $\mathbb{P}$与 $\mathbb{Q}$，Gilardoni [2010] 已证明如下形式的广义 Pinsker 不等式成立（且是尖锐的）

$$D_f^{\mathrm{C}}(\mathbb{P},\mathbb{Q})\geq 2f''(1)\cdot\|\mathbb{P}-\mathbb{Q}\|_{\mathrm{TV}}^2\,, \tag{20}$$

其中 $\|\mathbb{P}-\mathbb{Q}\|_{\mathrm{TV}}$是 $\mathbb{P}$与 $\mathbb{Q}$之间的全变差距离，形式上定义为 a s

$$\|\mathbb{P}-\mathbb{Q}\|_{\mathrm{TV}} \coloneqq \sup_{A\in\mathcal{F}}\left|\mathbb{P}[A]-\mathbb{Q}[A]\right|\,.$$

两个概率测度 $\mathbb{P}$ 与 $\mathbb{Q}$ 之间的 KL 散度 $D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q})$（它们定义在某个可测空间 $(\Omega,\mathcal{F})$ 上且满足 $\mathbb{P}\ll\mathbb{Q}$）确实是可容许生成函数f所对应的 Csiszár f-散度： $[0,+\infty)\to(-\infty,+\infty]$ 由 $x\mapsto f(x)\coloneqq x\ln x$给出（约定 $0\cdot\ln 0=0$），因为通过变量替换，我们有

$$D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q})\coloneqq\int_\Omega\ln\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)\mathrm{d}\mathbb{P}=\int_\Omega\ln\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\mathrm{d}\mathbb{Q}=\int_\Omega f\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)\mathrm{d}\mathbb{Q}=D_f^{\mathrm{C}}(\mathbb{P},\mathbb{Q})\,,$$

因此，对于公式（20），我们重新得到概率测度的（紧）Pinsker 不等式

$$D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q})=D_f^{\mathrm{C}}(\mathbb{P},\mathbb{Q})\geq 2f''(1)\cdot\|\mathbb{P}-\mathbb{Q}\|_{\mathrm{TV}}^2=2\cdot\|\mathbb{P}-\mathbb{Q}\|_{\mathrm{TV}}^2\,.$$

对于 $K\in\{2,3,\dots\}$，这最后一个不等式蕴含了 $(K-1)$ 维概率单纯形 Δ^K 上的经典 Pinsker 不等式。具体而言，若 $p,\ q\in\Delta^K$ 满足：对任意 $k\in[K]$,

每当 $q_k = 0$ 时亦有 $p_k = 0$，则 p 与 q 是离散空间 Ω 上两个概率测度 $\mathbb{P}$ 与 $\mathbb{Q}$ 的概率质量函数: $= [K]$，其中 $\mathbb{P} \ll \mathbb{Q}$，并且

$$\begin{aligned}\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} &= \sup_{A \in 2^K} \left|\mathbb{P}[A] - \mathbb{Q}[A]\right| = \sup_{A \in 2^K} \left|\sum_{k \in A} p_k - \sum_{k \in A} q_k\right| \\ &= \frac{1}{2} \cdot \sum_{k \in [K]} |p_k - q_k| = \frac{1}{2} \cdot \|p - q\|_1 = \|p - q\|_{\mathrm{TV}} \ ,\end{aligned}$$

（最后一个等式在 A 节中得到证明），由此可知（约定 $0/0 \coloneqq 0$ and$0 \cdot \ln(0) \coloneqq 0$）

$$D_{\mathrm{KL}}(p\|q) \coloneqq \sum_{k \in [K]} \ln\left(\frac{p_k}{q_k}\right) p_k = \int_\Omega \ln\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right) \mathrm{d}\mathbb{P} = D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q}) \geq 2 \cdot \|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}}^2 = 2 \cdot \|p - q\|_{\mathrm{TV}}^2 \ .$$

一类有趣的可容许生成函数具有如下形式

$$f_\alpha \colon [0, +\infty) \to (-\infty, +\infty] \ , \qquad x \mapsto \frac{x^\alpha - 1}{\alpha(\alpha - 1)} \ ,$$

对于 $\alpha \in (0, 1) \cup (1, +\infty)$，其对应的 Csiszár f_α-散度称为 α- Tsallis 相对熵（可能相差不同的归一化常数），此时公式（20）变为

$$D_\alpha^{\mathrm{TRE}}(\mathbb{P}\|\mathbb{Q}) \coloneqq D_{f_\alpha}^{\mathrm{C}}(\mathbb{P}, \mathbb{Q}) \geq 2 f_\alpha''(1) \cdot \|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}}^2 = 2 \cdot \|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}}^2 \ .$$

同样，若将这个最后的不等式特化到离散情形，对于任意 $p, q \in \mathrm{relint}\left(\Delta^K\right)$，若 $\mathbb{P}$ and$\mathbb{Q}$是概率测度，其相应的概率质量函数分别为 p 和 q，则可得

$$\begin{aligned}D_\alpha^{\mathrm{TRE}}(p\|q) \coloneqq \sum_{k \in [K]} f_\alpha\left(\frac{p_k}{q_k}\right) q_k = \int_\Omega f\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right) \mathrm{d}\mathbb{Q} = D_\alpha^{\mathrm{TRE}}(\mathbb{P}\|\mathbb{Q}) \qquad (21) \\ \geq 2 \cdot \|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}}^2 = 2 \cdot \|p - q\|_{\mathrm{TV}}^2 = \frac{1}{2} \cdot \|p - q\|_1^2 \ .\end{aligned}$$

探究 α-Tsallis 相对熵与负 α-Tsallis 熵的 Bregman 散度之间的关系、它们各自的广义 Pinsker 不等式，以及二者能否相互推导，是很有意义的。这里我们证明，对于 $\alpha \in (-\infty, 0) \cup (0, 1)$，Gilardoni 的 α-Tsallis 相对熵广义 Pinsker 不等式（公式 21）可推出我们关于负 α-Tsallis 熵的 Bregman 散度的 Pinsker 不等式（定理 8）的一个较宽松的版本；而当 $\alpha \in (1, +\infty)$ 时，反过来也成立。

具体而言，对于任意 $\alpha \in (-\infty, 0) \cup (0, 1) \cup (1, +\infty)$ 以及任意 $p, q \in \mathrm{relint}(\Delta^K)$，我们有 ve

$$\begin{aligned}D_\alpha(p\|q) &= \frac{1}{\alpha} \cdot \sum_{k \in [K]} \frac{p_k^\alpha + (\alpha - 1) q_k^\alpha - \alpha p_k q_k^{\alpha - 1}}{\alpha - 1} \\ &= \sum_{k \in [K]} \frac{\left(\frac{p_k}{q_k}\right)^\alpha + (\alpha - 1) - \alpha \frac{p_k}{q_k}}{\alpha(\alpha - 1)} \cdot q_k^\alpha =: \sum_{k \in [K]} g_\alpha\left(\frac{p_k}{q_k}\right) \cdot q_k^\alpha =: (\square) \ ,\end{aligned}$$

其中

$$g_\alpha \colon [0, +\infty) \to \mathbb{R} \ , \qquad x \mapsto \frac{x^\alpha + \alpha - 1 - \alpha x}{\alpha(\alpha - 1)} \ .$$

现在，注意 注意， g_α 是一个非负函数（其最小值在 1 处取得，且函数值为 0），因此 e

$$(\square) \geq \sum_{k\in[K]} g_\alpha\left(\frac{p_k}{q_k}\right)\cdot q_k\ , \qquad \textit{if } \alpha\in(-\infty,0)\cup(0,1)\ ,$$
$$(\square) \leq \sum_{k\in[K]} g_\alpha\left(\frac{p_k}{q_k}\right)\cdot q_k\ , \qquad \textit{if } \alpha\in(1,\infty)\ .$$

现在，我们有

$$\begin{aligned}\sum_{k\in[K]} g_\alpha\left(\frac{p_k}{q_k}\right)\cdot q_k &= \sum_{k\in[K]} \frac{\left(\frac{p_k}{q_k}\right)^\alpha + (\alpha-1) - \alpha\frac{p_k}{q_k}}{\alpha(\alpha-1)}\cdot q_k \\ &= \sum_{k\in[K]} \frac{\left(\frac{p_k}{q_k}\right)^\alpha - 1}{\alpha(\alpha-1)}\cdot q_k = \sum_{k\in[K]} f_\alpha\left(\frac{p_k}{q_k}\right)\cdot q_k = D_\alpha^{\mathrm{TRE}}(p\|q)\ ,\end{aligned}$$

因此

$$D_\alpha(p\|q) \geq D_\alpha^{\mathrm{TRE}}(p\|q)\ , \qquad \textit{if } \alpha\in(-\infty,0)\cup(0,1)\ , \tag{22}$$
$$D_\alpha(p\|q) \leq D_\alpha^{\mathrm{TRE}}(p\|q)\ , \qquad \textit{if } \alpha\in(1,\infty)\ . \tag{23}$$

特别地，对于 $\alpha\in(-\infty,\ 0)\cup(0,\ 1)$，结合公式（22）和适用于 α-Tsallis相对熵的广义 Pinsker 不等式公式（21），我们得到定理 8 对于 $\alpha\in(-\infty,\ 0)\cup(0,\ 1)$ 及 $p,\ q\in\mathsf{relint}(\Delta^K)$ 的以下较宽松的 形式（相差一个$2^{1-\alpha}$因子）

$$D_\alpha(p\|q) \geq D_\alpha^{\mathrm{TRE}}(p\|q) \geq \frac{1}{2}\cdot\|p-q\|_1^2\ , \qquad \text{if } \alpha\in(0,1)\ .$$

由于对于 $\alpha>1$和 $p,\ q\in\mathsf{relint}(\Delta^K)$，反向不等式（公式 23）成立，因此由定理 8 可得适用于 α-Tsallis相对熵的广义 Pinsker 不等式的以下较宽松的 版本

$$D_\alpha^{\mathrm{TRE}}(p\|q) \geq D_\alpha(p\|q) \geq \frac{C_{\alpha,K}}{2}\cdot\|p-q\|_1^2\ , \qquad \text{if } \alpha\in(1,+\infty)\ .$$

$\alpha=0$的情形与 $\alpha\in(-\infty,\ 0)\cup(0,\ 1)$ 的情形类似，其中反向 KL 散度 $(p,\ q)\mapsto D_{\mathrm{KL}}(q\|p)$ 起到 α-Tsallis相对熵的作用。具体而言，对于任意 $p,\ q\in\mathsf{relint}(\Delta^K)$，我们可以如下得到定理 8 的较宽松的形式（相差 2 倍）

$$\begin{aligned}D_0(p\|q) &= \sum_{k\in[K]} \underbrace{\left(\frac{p_k}{q_k} - \ln\frac{p_k}{q_k} - 1\right)}_{\geq 0} \geq \sum_{k\in[K]}\left(\frac{p_k}{q_k} - \ln\frac{p_k}{q_k} - 1\right)\cdot q_k \\ &= \sum_{k\in[K]}\left(-q_k\cdot\ln\frac{p_k}{q_k}\right) = \sum_{k\in[K]} q_k\cdot\ln\frac{q_k}{p_k} = D_{\mathrm{KL}}(q\|p) \geq \frac{1}{2}\cdot\|p-q\|_1^2\ ,\end{aligned}$$

其中，最后一个不等式是经典 Pinsker 不等式。

G 当 $\alpha>2$ 且 $K\geq 3$ 时较弱的不等式

正如我们在定理 8 和附录 E 中所见，对于 $\alpha>2$ 和 $K\geq 3$，不存在满足 $D_\alpha(p\|q)\geq C\cdot\|p-q\|_1^2$ 的正数常量 C，使其对每个 $p,q\in\mathsf{relint}\ (\Delta^K)$都成立。不过，我们在 (9) 中已经看到，

$$2\cdot\frac{D_\alpha(p\|q)}{\|p-q\|_1^2} = \sum_{k\in[K]} v_k^2\gamma_k^{\alpha-2}\ , \tag{24}$$

其中，γ 是 p 和 q 的凸组合，$v := \frac{p-q}{\|p-q\|_1}$ 是 B^K中的一个元素。在某些应用中（尤其是采用对数损失及相关适当评分规则进行序贯预测时），通常会将预测限制在单纯形相对内部的一个真子集内，或采用截断/平滑，使损失和 KL 型散度保持有限且性质良好 [Cesa-Bianchi 和 Lugosi，2006年，2001年，Gneiting 和 Raftery，2007年，van Ommen 等，2016]。例如，我们可以要求存在满足 $0 < \varepsilon < \frac{1}{K}$ 的某个正数 ε，使得对每个 k，$p_k \geq \varepsilon$ 且 $q_k \geq \varepsilon$；这一条件迫使 p和 q 都充分远离 Δ^K的相对边界。由此还可推出，对每个 k，$\gamma_k \geq \varepsilon$，因此 $\sum_{k\in[K]} v_k^2 \gamma_k^{\alpha-2} \geq \varepsilon^{\alpha-2} \|v\|_2^2$。利用 $\|v\|_2^2 \geq C_{2,\ K}$，我们推导得

$$D_\alpha\big(p\|q\big) \geq \frac{\varepsilon^{\alpha-2}}{2} \cdot C_{2,K} \cdot \|p-q\|_1^2 \,. \tag{25}$$

我们还可以在较为宽松的条件下推导出类似的不等式，即只要求 p 和 q 中的一个其坐标下界为 ε。事实上，命题 21 的证明表明，式 (24) 是 S_α 的带拉格朗日余项的一阶泰勒展开的结果。相反，如果我们考虑积分余项——例如 Reid 和 Williamson [2009b, 2011] 用于估计 Bregman 散度的方法正是以此为基础——那么得到的是

$$2 \cdot \frac{D_\alpha\big(p\|q\big)}{\|p-q\|_1^2} = \sum_{k\in[K]} v_k^2 \int_0^1 2(1-x)\big((1-x)q_k + xp_k\big)^{\alpha-2}\,\mathrm{d}x \,,$$

如果约束仅对 p 有效，那么对于每个 k，我们得到

$$\begin{aligned}\int_0^1 2(1-x)\big((1-x)q_k + xp_k\big)^{\alpha-2}\,\mathrm{d}x &\geq \int_0^1 2(1-x)\big(x\varepsilon\big)^{\alpha-2}\,\mathrm{d}x \\ &= \varepsilon^{\alpha-2}\int_0^1 2(1-x)x^{\alpha-2}\,\mathrm{d}x = \varepsilon^{\alpha-2}\cdot\frac{2}{\alpha(\alpha-1)} \,.\end{aligned}$$

这就得到不等式

$$D_\alpha\big(p\|q\big) \geq \frac{\varepsilon^{\alpha-2}}{2} \cdot \frac{2}{\alpha(\alpha-1)} \cdot C_{2,K} \cdot \|p-q\|_1^2 \,. \tag{26}$$

类似地，如果约束仅对 q 有效，那么这次我们有

$$\begin{aligned}\int_0^1 2(1-x)\big((1-x)q_k + xp_k\big)^{\alpha-2}\,\mathrm{d}x \ &\geq \int_0^1 2(1-x)\big((1-x)\varepsilon\big)^{\alpha-2}\,\mathrm{d}x \\ &= \varepsilon^{\alpha-2}\int_0^1 2\big(1-x\big)^{\alpha-1}\,\mathrm{d}x \ = \varepsilon^{\alpha-2}\cdot\frac{2}{\alpha} \,,\end{aligned}$$

因此

$$D_\alpha\big(p\|q\big) \geq \frac{\varepsilon^{\alpha-2}}{2} \cdot \frac{2}{\alpha} \cdot C_{2,K} \cdot \|p-q\|_1^2 \,. \tag{27}$$

我们可以将目前为止的结果改述如下：

命题 29。对于 $\alpha > 2$ 和 $K \geq 3$，考虑不等式

$$D_\alpha\big(p\|q\big) \geq \frac{C}{2} \cdot \|p-q\|_1^2 \,. \tag{28}$$

于是：

1. 对于每个 $p, q \in$ 相对内部$\left(\Delta^K\right)$，不等式 (28) 均满足于

$$C = C(\alpha, K, p, q) = C_{2,K} \cdot \left(\min_{k\in[K]} \{p_k, q_k\}\right)^{\alpha-2};$$

2. 对于每个 $p, q \in$ 相对内部$\left(\Delta^K\right)$，不等式(28) 由

$$C = C(\alpha, K, p) = C_{2,K} \cdot \frac{2}{\alpha(\alpha-1)} \cdot \left(\min_{k\in[K]} \{p_k\}\right)^{\alpha-2};$$

3. 对于每个 $p, q \in$ 相对内部$\left(\Delta^K\right)$，不等式(28) 由

$$C = C(\alpha, K, q) = C_{2,K} \cdot \frac{2}{\alpha} \cdot \left(\min_{k\in[K]} \{q_k\}\right)^{\alpha-2}.$$