arXiv:2302.02497v1 [math.ST] 5 Feb 2023

次伽马向量的范数集中高维位置估计

Shivam Gupta [1] **Jasper C.H. Lee** [2] **Eric Price** [1]

摘要

在位置估计中，给定来自已知分布 f 的 n 个样本，该分布经过未知平移 λ，我们的目标是尽可能精确地估计 λ。渐近情况下，最大似然估计达到误差的 Cramér-Rao 下界 $N(0, \frac{1}{n\mathcal{I}})$，其中 $\mathcal{I}$ 是 f 的费 Fisher 信息。然而，收敛所需的 n 取决于 f，且可能任意大。我们基于使用平滑 估计量的理论，以 $\mathcal{I}_r$（即经过 r 平滑的分布的费 Fisher 信息）为依据，给出有限 n 下的误差界。当 $n \to \infty$ 时，$r \to 0$ 以显式速率收敛，并趋近于 Cramér-Rao 下界。我们（1）改进了先前关于一维 f 的工作，使其不仅以高概率收敛，也能以常数失败概率收敛；（2）将该理论拓展到高维分布。在此过程中，我们证明了一个关于高维随机变量范数的新界，其中该随机变量的一维投影均为次伽马变量；这一结果本身也可能具有独立价值。

1. 引言

位置估计——均值估计的一种变体——是参数统计学中的一个基本问题。假设存在一个平移不变模型 $f^\lambda(x) = f(x-\lambda)$，其对应某个已知分布 f 在 $\mathbb{R}^d$上的情形。统计学家收到n 来自 f^λ 的独立同分布样本，其中真实参数 $\lambda \in \mathbb{R}^d$可任意选取；目标是高精度地估计 λ，并以高概率在这些样本上成功。

与一般的均值估计不同，后者旨在仅依赖对分布的最少假设来估计均值，而这里我们知道分布的确切形状，只是不知道其平移量。这些额外信息使我们能够以更高精度估计 λ。

关于位置估计的经典“教科书式”理论，以及一般而言，关于参数估计的理论，都建议使用最大似然估计（MLE）。MLE具有渐近正态性：如果固定一个分布 f，并令样本数 n 趋于无穷大，那么 MLE 的分布将收敛到多元高斯分布 $\mathcal{N}(\lambda, \frac{1}{n}\mathcal{I}^{-1})$，其中 $\mathcal{I}$ 是费 *Fisher* 信息矩阵，定义为

$$\mathcal{I} = \mathop{\mathbb{E}}_{x\sim f}\left[(\nabla \log f(x))(\nabla \log f(x))^\top\right]$$

作为一项基本性质，若将 f 的协方差矩阵记为 Σ，则总有 $\mathcal{I}^{-1} \preceq \Sigma$，这意味着最大似然估计的渐近表现始终至少与样本均值一样好，而样本均值的表现受协方差 Σ控制。此外，克拉美-罗界指出，任何无偏位置估计量的协方差都不可能小于 $\frac{1}{n}\mathcal{I}^{-1}$，因此最大似然估计在所有无偏估计量中具有最佳的渐近表现。

尽管教科书中的理论令人满意，因为费 Fisher 信息基本上刻画了位置估计的信息论极限，但其预测在实践中可能会产生误导。具体而言，这是由于最大似然估计的性能保证具有渐近性：我们需要将样本数 n 趋于无穷，才能实现次高斯估计误差。就模型 n 的 f 而言，渐近结果对 f (Miao、2010；Spokoiny、2011；Pinelis、2017)的依赖程度可能任意糟糕；对于拉普拉斯分布这类简单例子，相应的界甚至是无穷大。因此，研究目标是建立一种有限样本位置估计理论，明确地将估计误差界定为 n的函数，适用于每一种 f，并且理想情况下还能在估计误差中达到最优常数。

Gupta 等人（2022）近期在一维这一特殊情形下研究了这个问题。他们证明，虽然最大似然估计的有限样本性能可能较差，但可以通过一种简单的调整来改善其表现：在进行最大似然估计之前，给样本和模型都添加半径适当选取的高斯噪声 r，其中 r 随样本数量增加而减小。因此，平滑最大似然估计的理论保证将 f 的费 Fisher 信息替换为平滑分布的费 Fisher 信息 f_r，也称为

[1]德克萨斯大学奥斯汀分校 [2]威斯康星大学麦迪逊分校。

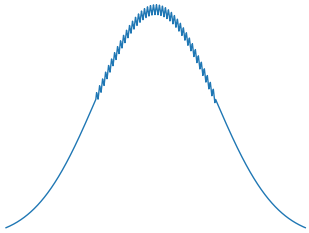

图 1. 高斯分布+锯齿分布

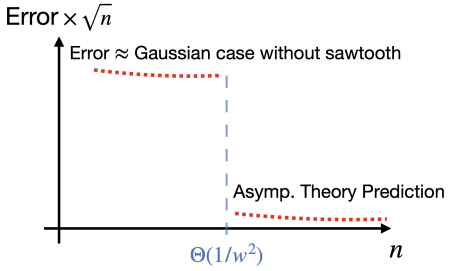

图 2.高斯分布+锯齿分布的常数概率误差下界

平滑费希尔信息 $\mathcal{I}_r$。平滑最大似然估计可获得类似于方差为 $(1+o(1))\mathcal{I}_r^{-1}$ 的高斯分布的有限样本次高斯误差界，其中 $o(1)$ 项可以显式计算，且不依赖于 f。

通过平滑费希尔信息进行刻画。我们的结果将沿用 Gupta 等人（2022）的方法，并以平滑费希尔信息表示有限样本界。这里聚焦于一维情形，简要讨论为什么费 Fisher 信息不足，以及为什么平滑费希尔信息是合适的替代物。

考虑图 +1 所示的“高斯分布+锯齿”分布。该分布是在标准高斯密度的中心部分加入齿宽为 w 、斜率为$\pm\Delta$ 的锯齿。当 $w \to 0$时，该密度收敛于标准高斯分布，但费 Fisher 信息却增长至 $\Theta(\Delta^2)$ ，因为$\Delta \to \infty$。因此，渐近理论以常数概率预测误差为$O(1/(\Delta\sqrt{n}))$ 。

然而，Gupta 等人 (2022) 证明，对于 $n \ll 1/w^2$，任意算法的常数概率误差实际上至少为$\Omega(1/\sqrt{n})$，就如同该分布只是标准高斯分布一样。直观而言，我们需要将模型对齐到与 w 相差不超过一个锯齿波宽度，才能利用锯齿结构实现高精度估计。对于标准高斯分布，要使误差小于$\Omega(1/w^2)$ ，需要w个样本。图 2 展示了高斯锯齿模型的常数概率误差下界图+，误差按 $\sqrt{n}$ 缩放以进行归一化。−

由于样本阈值取决于 w，这个例子表明，不存在能够以分布无关的方式收敛到渐近误差的算法。具体而言，对于分布无关的 $1+o(1)$ 倍 $\mathcal{N}(0,\ 1/(n\mathcal{I}))$ 项误差，不存在任何算法能够达到 $o(1)$误差因子。因此，对于有限样本误差界，我们需要一种替代量来取代 $\mathcal{I}$，使其能够刻画图2中的相变。

平滑费希尔信息呈现出这种相变行为。以半径 $r \gg w$ 进行平滑会模糊锯齿结构——$\mathcal{I}_r$ 较小，且接近标准高斯分布的费 Fisher 信息 1。另一方面，以半径 $r \ll w$ 进行平滑会保留锯齿结构，并使 $\mathcal{I}_r$接近 $\mathcal{I} = \Theta(\Delta^2)$。Gupta 等人 (2022和我们都利用这一行为，证明了与 $(1+o(1))\mathcal{N}(0, 1/(n\mathcal{I}_r))$类似的有限样本界，其中 $o(1)$ 项与分布无关。

我们需要谨慎地选择平滑参数，因为平滑费希尔信息可能对r有微妙的依赖。直观而言，我们预期当 $n \to \infty$时，$r \to 0$；然而，Gupta 等人的结果并非如此。他们选择的平滑参数仅在高概率情形下才趋于零，也就是说，失败概率 δ满足$n \to \infty$ 且 $\delta \to 0$ 时才会趋于零。因此，对于较小的常数 δ，他们的结果可能远非最优。我们的一个新结果消除了 r 对 δ的虚假依赖。

我们的结果。 本文从两个方面改进并扩展了 Gupta 等人（2022）的结果。首先，我们证明该算法的一个变体在一维情形下有更简单、更好的分析。更好的分析支持更小的平滑半径 r，从而带来更高的费希尔信息$\mathcal{I}_r$:

定理 1.1 （一维平滑最大似然估计）。 给定一个模型 f，设分布 f 的 r-平滑费希尔信息为 $\mathcal{I}_r$，并令四分位距 为 f的四分位距。固定失败概率为 $\delta \le 0.5$，并假设对于某个足够大的常数 c，有 $n \ge c \cdot \log \frac{2}{\delta}$。

选择 $r^* = \Omega((\frac{\log \frac{2}{\delta}}{n})^{1/8})$四分位距。那么，以至少$1-\delta$的概率，算法 $\lambda 2$ 的输出满足

$$|\hat{\lambda} - \lambda| \le \left(1 + O\left(\frac{\log \frac{2}{\delta}}{n}\right)^{\frac{1}{10}}\right)\sqrt{\frac{2\log \frac{2}{\delta}}{n\mathcal{I}_{r^*}}}$$

这一结果与（Gupta 等人，2022）的主要区别在于对 δ 的依赖：先前的结果要求$\delta \to 0$ 对于 r 趋近于 0，并且首项常数趋近于 1。而在我们的结果中，对于常数 n，二者都随 δ 多项式衰减。

考虑一下该结果在上文的高斯分布+锯齿示例（图 1）中的表现，其中 δ 为常数。当 n 较小时，我们将选择 $r^* = \frac{1}{\text{poly}(n)} > w$，使误差处于常规高斯尾部的 $1+\frac{1}{\text{poly}(n)}$ 以内；当 n 较大时，$r^* \ll w$ 并且

误差处于 $1+\frac{1}{\mathrm{poly}(n)}$ 渐近最优值 $\mathcal{N}(0,1/(n\mathcal{I}_r))$ 以内。因此，我们得到了与图 2 相同的定性转变行为，尽管转变点不同（$\frac{1}{w^8}$ 而不是 $\frac{1}{w^2}$）。先前的工作（Gupta 等人，2022）还要求 δ 趋于零，大致为 $\delta < 2^{-\mathrm{poly}(n)}$，才能观察到这种行为。

其次，我们更简洁的方法使我们能够将结果推广到高维情形。我们证明了一个与一维结果类似的结论。在理想情况下，由于（未平滑的）最大似然估计渐近满足 $(\widehat{\lambda}-\lambda) \to \mathcal{N}(0, \frac{1}{n}\mathcal{I}^{-1})$ 我们希望达到高斯尾部误差（Boucheron 等人 (2013年，例 5.7）

$$\|\widehat{\lambda}-\lambda\|_2 \le \sqrt{\frac{\mathrm{Tr}(\mathcal{I}^{-1})}{n}} + \sqrt{2\|\mathcal{I}^{-1}\|\frac{\log\frac{1}{\delta}}{n}} \qquad (1)$$

以概率 $1-\delta$ 成立。我们证明这一点几乎成立。令 $d_{\mathrm{eff}}(A) = \mathrm{Tr}(A)/\|A\|$ 表示半正定矩阵 A 的有效维数。若我们对某个 $r^2 \le \|\Sigma\|$ 使用球形高斯分布 $R = r^2 I_d$ 进行平滑，那么对于足够大的 n（其取决于 $\|\Sigma\|/r^2$、$\log\frac{1}{\delta}$、$d_{\mathrm{eff}}(\Sigma)$ 和 $d_{\mathrm{eff}}(\mathcal{I}_R^{-1})$），我们的误差与将（1）中的 $\mathcal{I}$ 替换为平滑费希尔信息 $\mathcal{I}_R$ 后的结果相近。

定理 1.2（高维最大似然估计，非正式表述；参见定理 B.16）。设 f 的协方差矩阵为 Σ。对于任意 $r^2 \le \|\Sigma\|$，令 $R = r^2 I_d$ 和 $\mathcal{I}_R$ 为该分布经 R 平滑后的费希尔信息。对于任意常数 $0 < \eta < 1$，

$$\|\widehat{\lambda}-\lambda\|_2 \le (1+\eta)\sqrt{\frac{\mathrm{Tr}(\mathcal{I}_R^{-1})}{n}} + 5\sqrt{\frac{\|\mathcal{I}_R^{-1}\|\log\frac{4}{\delta}}{n}}$$

以概率 $1-\delta$，对于

$$n > O_\eta\left(\left(\frac{\|\Sigma\|}{r^2}\right)^2\left(\log\frac{2}{\delta} + d_{\mathit{eff}}(\mathcal{I}_R^{-1}) + \frac{d_{\mathit{eff}}(\Sigma)^2}{d_{\mathit{eff}}(\mathcal{I}_R^{-1})}\right)\right)$$

当 $d_{\mathrm{eff}}(\mathcal{I}_R^{-1}) \gg \log\frac{1}{\delta}$ 时，界为 $(1+\eta+o(1))\sqrt{\mathrm{Tr}(\mathcal{I}_R^{-1})}$。这与无偏估计量的 $\sqrt{\mathrm{Tr}(\mathcal{I}^{-1})}$ 的期望误差的克拉美-罗下界非常接近（Bickel 与 Doksum （2015年），定理 3.4.3）。

该定理的正式版本，即定理 B.16，还给出了由对称半正定矩阵 M 所诱导的一般距离 $\|\lambda-\lambda\|_M$ 的界；精确的界以及 n 收敛所需的条件取决于 M。

我们证明中的一个关键部分（或许本身也值得关注），是关于一个高维向量范数的集中界 x，该向量在每个方向上的边缘分布都是次伽马分布。如果一个向量在每个方向上都服从高斯分布，那么它就是一个高维高斯向量，并满足尾界（1）（将 $\mathcal{I}^{-1}$ 替换为协方差矩阵 Σ）。(Hsu 等人，2012年)指出，即使边缘分布仅为次高斯分布，参数为 Σ，同样适用该界。我们将此结果推广，得到次伽马边缘分布的界：

定理 1.3（次伽马随机向量的范数集中；参见定理 5.1）。设 x 是 $\mathbb{R}^d$ 中的均值为零的随机向量，且为 (Σ, C)-次伽马向量，即对任意向量 $v \in \mathbb{R}^d$，

$$\mathbb{E}[e^{\lambda\langle x,v\rangle}] \le e^{\lambda^2 v^T\Sigma v/2}$$

当 $|\lambda| \le \frac{1}{\|Cv\|}$ 时。因此，以 $1-\delta$ 的概率，

$$\|x\| \le \sqrt{\mathrm{Tr}(\Sigma)} + 4\sqrt{\|\Sigma\|\log\frac{2}{\delta}} + 16\|C\|\log\frac{2}{\delta} + \min\left(4\|C\|_F\sqrt{\log\frac{2}{\delta}}, 8\frac{\|C\|_F^2}{\sqrt{\mathrm{Tr}(\Sigma)}}\log\frac{1}{\delta}\right)$$

第一项迹项是期望范数，接下来的两项则是（一维次伽马集中性给出的）紧确界，忽略常数因子。当 x 是 n 个样本的平均值时，Σ 和 C 都会缩小至原来的 n 倍；因此，含有 C 的项以 $1/n$ 的速率衰减，而仅含有 Σ 的项则以 $1/\sqrt{n}$ 的速率衰减。当 $n \to \infty$ 时，与含有 Σ 的高斯项相比，含有 C 的项会消失。

为了更好地理解最后一项，考虑 x 是从球面对称情形（$\Sigma = \sigma^2 I$，$C = cI$）中抽取的 n 个样本 X_i 的平均值。我们还关注高维情形，其中 $d \ge (2/\eta^2)\log(1/\delta)$，对某个较小的 η 成立；此时，式（1）的目标误差界变为 $(1+\eta)\sqrt{\mathrm{tr}(\Sigma)/n}$，也就是说，期望范数误差的误差界在 $(1+\eta)$ 因子以内。在次伽马情形下，只要 $n \gtrsim (c/\sigma)^2 d$，定理 1.3 的界就意味着误差为 $(1+O(\eta))\sqrt{\mathrm{Tr}(\Sigma)/n}$；其中，$n$ 的阈值来自将界中最后一个“min”项与 $\sqrt{\|\Sigma\|\log\frac{2}{\delta}}$ 项进行比较。

在更强的假设下，即随机向量与其期望的距离至多为 c，我们可以将我们的尾界与 Talagrand/Bousquet 的上确界集中不等式进行比较（Boucheron 等人 (2013年)，定理 12.5）。我们再次聚焦于高维球形情形，其中 $d \ge (2/\eta^2)\log(1/\delta)$ 且 $\Sigma = \sigma^2 I$，$C = cI$，Bousquet 不等式给出的误差几乎相同，为 ℓ_2 $(1+O(\eta))\sqrt{\mathrm{Tr}(\Sigma)/n}$，只要 $n \gtrsim (c/\sigma)^2 d$，尽管其隐藏常数更小。鉴于我们的界的 n 阈值源于最后一个“min”项，这类项在定性上很可能是必要的；至少在本文考虑的相关情形中，我们的最后一项不会太大。

1.1. 记号

我们用 f 表示已知分布。在 1 维情况下，f_r 是 r-平滑分布 $f * \mathcal{N}(0, r^2)$，其平滑

费 Fisher 信息为 $\mathcal{I}_r$。在高维情况下，f_R 是 R-平滑分布 $f * \mathcal{N}(0, R)$，其平滑费希尔信息为 $\mathcal{I}_R$——请注意 r 与 R 之间的二次差异，这类似于 1 维与高维高斯分布的（协）方差的常用约定。

真实参数记为 λ。我们的一维和高维算法都会先得到初始估计 $\hat{\lambda}_1$，再将其细化为最终估计 λ。

除非另有说明，对于给定向量 x，$\|x\|$表示 ℓ_2 范数，类似地，$\|A\|$ 是方阵 A的算子范数。给定一个半正定方阵 A，我们将其有效维数定义为 $d_{\text{eff}}(A) = \text{tr}(A)/\|A\|$。矩阵$A$ 的有效维数在其为球形时为 d；但如果其一个或多个特征值偏离最大特征值，有效维数就会降低。

2. 相关工作

关于位置估计及一般参数估计渐近理论的深入教科书式论述，参见（van derVaart，2000）。此外，也有对最大似然估计（MLE）的有限样本分析（Spokoiny，2011）针对高维情形，（Pinelis，2017；Miao，2010）针对一维情形），但除要求强正则条件外，还会在估计误差界中损失（至少）乘法常数。与本文最相关的是Gupta等人此前的工作（2022），该工作在一维位置估计中引入了平滑最大似然估计，并从大 n 和小 δ 的平滑费希尔信息角度，正式分析了其有限样本表现。

近年来，关于密切相关的均值估计问题，研究者开展了大量工作，其所需的最弱假设是方差有限。此时的界取决于方差，而非费 Fisher 信息。在一维情形下，Catoni 的开创性论文（2012年）开启了寻找次高斯均值估计量的研究，该估计量的估计误差在 $1 + o(1)$ 倍因子范围内达到紧界；Devroye 等人（2016）以及 Lee 和 Valiant（2022a）的改进给出了一种一维均值估计量，它适用于所有方差有限（但方差未知）的分布，精度在 $1 + o(1)$ 倍因子范围内达到最优。关键在于，$o(1)$ 项与底层分布无关。

在高维情形下，如何在有界协方差条件下找到具有紧致常数的次高斯均值估计量，仍是一个开放问题。一系列研究（Lugosi 和 Mendelson，2017；Hopkins，2018；Cherapanamjeri 等人，2019）表明，可以在多项式时间内达到次高斯速率（忽略常数）。近期，Lee 和 Valiant（2022b）实现了线性时间和尖锐常数，但要求分布的有效维数远大于 $\log^2 \frac{1}{\delta}$。

我们另一项贡献是针对次伽马随机向量提出了新颖的范数集中界。高斯向量的范数集中性早已得到充分理解，例如可参见教材（Boucheron 等人（2013年），例 5.7）。Hsu 等人（2012年）将这一界推广到了逐方向次高斯向量的情形。范数集中也可以视为经验过程的上确界。Bousquet 对 Talagrand 上确界集中不等式的版本（2002年；2003年）推出了随机向量的范数集中界，条件是这些向量位于其期望的一个 ℓ_2 球内。我们的界将 Bousquet 不等式的这一情形从有界向量推广到了所有次伽马向量。如定理1.3之后所述，球形 Σ 和 C的结果相当相似。

3. 一维位置估计

本节将概述我们的一维位置估计算法及其分析。完整分析请参见附录A 。

算法 1 如下所示，是一个局部 算法，因为它假设我们有一个初始估计 λ_1，且该估计与 ϵ 的距离 λ不超过某个范围，目标是将估计细化到高精度。

算法 1一维局部平滑最大似然估计

输入参数：

- f的描述、平滑参数 r、样本 x_1、……、$x_n \overset{i.i.d.}{\sim} f^\lambda$ 以及 λˆ1 的初始估计 λ_1。令 $s(\lambda)$ 为 f_r 的得分函数，其中 f 的 r-平滑版本。2. 对每个样本 x_i，计算扰动后的样本 $x_i' = x_i + \mathcal{N}(0, r^2)$，其中所有样本的高斯噪声均独立抽取。3. 计算 λ_1 处的经验得分，即 $\hat{s}(\lambda_1) = \frac{1}{n}\sum_{i=1}^n s(x_i' - \lambda_1)$。ˆ4. 返回 $\lambda = \lambda_1 - (\hat{s}(\lambda_1)/\mathcal{I}_r)$。

令 $\mathcal{I}_r$ 为 f_r的费 Fisher 信息，其中 f_r是 f的 r平滑版本。关于得分 $s(x)$ 的基本事实如下：

$$0 = \mathop{\mathbb{E}}_{x\sim f_r}[s(x)]$$
$$\mathcal{I}_r = \mathop{\mathbb{E}}_{x\sim f_r}[-s'(x)] = \mathop{\mathbb{E}}_{x\sim f_r}[s(x)^2].$$

首先，算法 1 向每个 $N(0, r^2)$ 独立添加扰动 x_i 以获得 x_i'，它们的抽取方式为 $(y_1 + \lambda, y_2 +$

λ，……，$y_n + \lambda$) 对于 $y_i \sim f_r$。然后计算

$$\widehat{s}(\lambda_1) := \frac{1}{n}\sum_{i=1}^{n} s(x_i' - \lambda_1) = \frac{1}{n}\sum_{i=1}^{n} s(y_i - \epsilon)$$

其期望为，

$$\mathop{\mathbb{E}}_{x\sim f_r}[s(x-\epsilon)] \approx \mathop{\mathbb{E}}_{x\sim f_r}[s(x) - \epsilon s'(x)] = \epsilon\mathcal{I}_r.$$

因此我们预期 $\lambda = \lambda_1 - \widehat{s}(\lambda_1)/\mathcal{I}_r \approx \lambda$。

这项计算的误差有两个来源：(I) 对 $s(x-\epsilon)$ 的泰勒近似，以及 (II) $s(x-\epsilon)$ 的经验期望与真实期望之间的差异。当 $\epsilon = 0$ 时，泰勒误差为 0，且经验估计量的方差为

$$\frac{\mathrm{Var}(s(x))}{n} = \frac{\mathcal{I}_r}{n}.$$

因此，当 $\lambda_1 = \lambda$ 时，λ 将是 λ 的无偏估计量，方差为 $\frac{1}{n\mathcal{I}_r}$：这恰好就是~~Cramér~~-Rao 界。此外，可以证明 $s(x)$ 是次伽马分布，其方差代理参数为 $\mathcal{I}_r$，尾部参数为 $1/r$，从而在 $\lambda - \lambda$ 上具有与 $\frac{1}{n\mathcal{I}_r}$ 方差的高斯分布相匹配的尾部（在某个由 r 决定的范围内）。因此，我们只需证明，平移 ϵ 在 (I) 和 (II) 中引入的额外误差很小；直观地说，对于 $|\epsilon| \ll r$，这是因为 f_r 已经用半径为 r 的球进行了平滑。

事实上，（Gupta 等人，2022）已经界定了这两个误差：对于（I），其引理 C.2 表明

$$\mathop{\mathbb{E}}_{x\sim f_r}[s(x-\epsilon)] = \mathcal{I}_r\epsilon \pm O(\sqrt{\mathcal{I}_r}\frac{\epsilon^2}{r^2}) \tag{2}$$

对所有 $|\epsilon| \le r/2$，对于（II），其推论 3.3 和引理 C.3 一起表明，存在次伽马集中性

$$|\hat{s}(\lambda_1) - \mathop{\mathbb{E}}_{x\sim f_r}[s(x-\epsilon)]| \lesssim (1+o(1))\sqrt{\frac{\mathcal{I}_r \log\frac{2}{\delta}}{n}} + \frac{\log\frac{2}{\delta}}{nr} \tag{3}$$

当 $r \gg |\epsilon|$时。

因此，对于足够大的 r，$\widehat{s}(\lambda_1)$ 中的总误差由主导项 $\sqrt{\frac{\mathcal{I}_r \log\frac{2}{\delta}}{n}}$ 控制，所得结果与~~最优~~值相差 $1 + o(1)$。

获取初始估计。 我们通过一小部分 λ 的经验 α分位数来估计 κ，其中α；可以证明，其误差至多为 O（四分位距 $\cdot\sqrt{\log_{\kappa n}\frac{1}{\delta}}$），且 $1-\delta$概率成立，其中 四~~分位~~距 表示四分位距。此策略与（Gupta 等人，2022）基本相同，但我们在两个阶段使用新样本，而他们重复使用样本。

算法 2 一维全局平滑最大似然估计

输入参数：

- 失败概率 δ，描述 f, n 从 f^λ 中抽取的独立同分布样本，其中某个 λ未知

1. Let q be $\sqrt{2}(\log\frac{2}{\delta}/n)^{2/5}$.
2. 计算一个 $\alpha \in [q, 1-q]$，以最小化由 $\alpha \pm q$ 分位数定义的 f区间宽度。
3. 取前（$\log\frac{1}{\delta}/n$）$^{1/10}$比例的 n 样本的样本α分位数。
4. Let $r^* = \Omega((\frac{\log\frac{1}{\delta}}{n})^{1/8})\mathrm{IQR}$.
5. 在其余样本上运行算法1，使用初始估计 $\lambda_1 = x_\alpha$ 和 r^*平滑，并返回 λ。

结合上述策略并平衡参数后，得到算法2，作为我们的最终算法。我们在附录A中证明，该算法满足我们的一维结果，即定理1.1。

与先前工作的比较。 我们这一维结果所需的得分函数的全部性质已在（Gupta 等人，2022）中给出，但该论文使用了不同的算法，因此只能证明较弱的结果。（Gupta 等人，2022）中的算法寻找 $\hat{s}$的根，而我们实质上执行一步牛顿法来近似该根。一般的求根需要一致收敛于 $\hat{s}$，但（Gupta 等人，2022）无法在不引入额外损失因子的情况下证明这一点。通过执行一步迭代，并使用（少量）新样本来获得初始估计，我们的算法只需要逐点收敛。

4. 高维位置估计

高维情形在概念上与1维情形类似。完整分析见附录B。主要区别有两点：1）初始估计来自重尾次高斯估计量；2）我们使用针对次伽马向量范数的集中不等式（定理 5.1）来界定我们的估计与真实均值之间的差异。

设 λ 为真实位置，λ 为我们的最终估计。我们首先陈述主定理，它给出了由对称半正定矩阵 M诱导的 $\|\lambda - \lambda\|_M$ 的界。

定理4.1（高维最大似然估计，非正式表述；参见定理 B.16）。设 f 具有协方差矩阵 Σ。对于任意 $r^2 \le \|\Sigma\|$，令 $R = r^2 I_d$ 和 $\mathcal{I}_R$ 分别为该分布的 R-平滑费希尔信息。令 M 为任意对称

半正定矩阵，并令 $T = M^{1/2}\mathcal{I}_R^{-1}M^{1/2}$。对于任意常数$0 < \eta < 1$,

$$\|\widehat{\lambda} - \lambda\|_M \leq (1+\eta)\sqrt{\frac{\mathrm{Tr}(T)}{n}} + 5\sqrt{\frac{\|T\|\log\frac{4}{\delta}}{n}}$$

以概率 $1-\delta$，对于

$$n > O_\eta\left(\left(\frac{\|\Sigma\|}{r^2}\right)^2\left(\log\frac{2}{\delta} + d_{\mathit{eff}}(T) + \frac{d_{\mathit{eff}}(\Sigma)^2}{d_{\mathit{eff}}(T)}\right)\right)$$

作为推论，我们得到定理 1.2，它给出了$\|\lambda - \lambda\|_2$的上界，以及下面这个结果，它给出了马-$\widehat{}$氏距离 $\|\lambda - \lambda\|_{\mathcal{I}_R}$的上界。

推论 4.2。 设 f 的协方差矩阵为 Σ。对于任意$r^2 \leq \|\Sigma\|$，令 $R = r^2 I_d$ 和 $\mathcal{I}_R$ 为该分布的 R- 平滑费希尔信息。对于任意常数 $0 < \eta < 1$,

$$\|\widehat{\lambda} - \lambda\|_{\mathcal{I}_R} \leq (1+\eta)\sqrt{\frac{d}{n}} + 5\sqrt{\frac{\log\frac{4}{\delta}}{n}}$$

以概率 $1-\delta$，对于

$$n > O_\eta\left(\left(\frac{\|\Sigma\|}{r^2}\right)^2\left(\log\frac{2}{\delta} + d + \frac{d_{\mathit{eff}}(\Sigma)^2}{d}\right)\right)$$

现在我们概述分析过程。下面的算法 3以均值的初始估计 λ_1 为起点，并将其细化为精确的$\widehat{}$估计 λ，这与 1 维1维情形的算法类似。

算法 3高维局部最大似然估计

输入参数：

- 分布 f在 $\mathbb{R}^d$上的描述，平滑参数 R，样本 x_1、……、$x_n \overset{i.i.d.}{\sim} f^\lambda$，以及初始估计 λ_1

1. 令 $\mathcal{I}_R$ 为 f的 R平滑版本的费 Fisher 信息矩阵。令 s_R 为 f_R的得分函数。

2. 对于每个样本 x_i，计算一个扰动后的样本 $x_i' = x_i + \mathcal{N}(0, R)$，其中所有高斯噪声在所有样本之间独立抽取。

3. Let $\hat{\epsilon} = \frac{1}{n}\sum_{i=1}^n \mathcal{I}_R^{-1}s_R(x_i' - \lambda_1)$ and return $\hat{\lambda} = \lambda_1 - \hat{\epsilon}$.

令 f 为 $\mathbb{R}^d$ 上的一个分布，并令 $\mathcal{I}_R$ 为 f 的 R 平滑版本 f_R 的费 Fisher 信息矩阵。于是，对于得分 s_R，若$\mathbf{J}_{s_R}$是 s_R 的雅可比矩阵，

$$\mathcal{I}_R = \mathop{\mathbb{E}}_{x\sim f_R}\left[s_R(x)s_R(x)^T\right] = \mathop{\mathbb{E}}_{x\sim f_R}\left[-\mathbf{J}_{s_R}(x)\right]$$

与1维情形类似，算法 3 以初始估计 $\lambda_1 = \lambda + \epsilon$（满足 $\epsilon^T R^{-1}\epsilon \leq 1/4$）开始。算法首先将 $N(0,R)$ 独立地加到每个样本 x_i上，从而得到x_i'，它们按如下方式抽取：$y_i + \lambda$ 对于 $y_i \sim f_R$。然后计算

$$\hat{\epsilon} = \frac{1}{n}\sum_{i=1}^n \mathcal{I}_R^{-1}s_R(x_i' - \lambda_1) = \frac{1}{n}\sum_{i=1}^n \mathcal{I}_R^{-1}s_R(y_i - \epsilon)$$

其期望为

$$\mathop{\mathbb{E}}_{x\sim f_R}\left[\mathcal{I}_R^{-1}s_R(x-\epsilon)\right] \approx \mathop{\mathbb{E}}_{x\sim f_R}\left[-\mathcal{I}_R^{-1}\mathbf{J}_{s_R}(x)\epsilon\right] = \epsilon$$

ˆ因此，我们仍然预期 $\lambda = \lambda_1 - \hat{\epsilon} \approx \lambda$，误差来自 (I) 对 $s_R(x-\epsilon)$ 的泰勒近似，以及 (II) $s_R(x-\epsilon)$ 的经验期望与真实期望之间的差异。

对于 (I)，引理B.3 表明

$$\|\epsilon - \mathop{\mathbb{E}}_{x\sim f_R}\left[\mathcal{I}_R^{-1}s_R(x-\epsilon)\right]\|^2 \lesssim \|\mathcal{I}_R^{-1}\|(\epsilon^T R^{-1}\epsilon)$$

对于 $\epsilon^T R^{-1}\epsilon \leq 1/4$。对于 (II)，推论 B.12 表明，对于任意单位方向 v，$v^T\mathcal{I}_R^{-1}s_R(x-\epsilon)$ 都是次伽马的：

$$v^T\mathcal{I}_R^{-1}s_R(x-\epsilon) \in \Gamma(\mathcal{I}_R^{-1}(1+o(1)), \mathcal{I}_R^{-1}R^{-1/2})$$

when $\frac{\epsilon^T R^{-1}\epsilon}{\sqrt{(\epsilon^T R^{-1}\epsilon)\log\left(\|I_R^{-1}\|\|R^{-1}\|\right)}} \leq 1/4$ and $\ll 1$，因此结合我们的次伽马向量范数集中不等式（定理 5.1），引理 B.13 表明

$$\|\hat{\epsilon} - \mathop{\mathbb{E}}_{x\sim f_R}\left[\mathcal{I}_R^{-1}s_R(x-\epsilon)\right]\| \leq$$
$$(1+o(1))\left(\sqrt{\frac{\mathrm{Tr}(\mathcal{I}_R^{-1})}{n}} + 4\sqrt{\frac{\|\mathcal{I}_R^{-1}\|\log\frac{2}{\delta}}{n}}\right.$$
$$\left. + 16\frac{\|\mathcal{I}_R^{-1}R^{-1/2}\|\log\frac{2}{\delta}}{n} + 8\frac{\|\mathcal{I}_R^{-1}R^{-1/2}\|_F^2\log\frac{2}{\delta}}{n^{3/2}\sqrt{\mathrm{Tr}(\mathcal{I}_R^{-1})}}\right)$$

对于 $R = r^2 I_d$，当 r 较大时，总误差由上述界限中的前两项主导，这两项对应于协方差为 $\mathcal{I}_R^{-1}$ 时的次高斯集中性。

获取初始估计。对于我们的初始估计 λ_1，我们使用重尾估计量（Hopkins，2018；Cherapanamjeri 等人，2019），它能够保证误差具有次高斯特性，且误差取决于协方差 Σ，其中 f，但相差常数因子。

与 1 维情形一样，将初始估计与算法 3 结合起来，即可得到我们的最终定理——定理B.16。下面的算法 4 展示了如何计算初始估计，并将其与局部最大似然估计算法 3 结合起来，从而得到最终估计。

算法 4高维全局最大似然估计

输入参数：

- 失败概率 δ，分布描述 f, n来自 f^λ 的样本，平滑参数 R，近似参数 η

1. 令 Σ 为 f的协方差矩阵。使用定理 λ_1 中的估计量，对 η/C 比例的 n 样本计算初始估计 C，其中 B.15 为较大的常数。

2. 使用剩余 3比例的样本运行算法 $1-\eta/C$，采用 R-平滑和初始估计 $\hat{}\lambda_1$，返回最终估计 λ。

5. 次伽马向量的范数集中

定理 5.1（次伽马向量的范数集中）。设 x 是 $\mathbb{R}^d$ 中的一个均值为零的随机向量，且为（Σ, C）-次伽马向量，即对于所有 $v \in \mathbb{R}^d$, $v^T x \in \Gamma(v^T \Sigma v, \|Cv\|)$ 均成立。换言之，它满足：对于任意向量 $v \in \mathbb{R}^d$,

$$\mathbb{E}[e^{\lambda\langle x,v\rangle}] \le e^{\lambda^2 v^T \Sigma v/2}$$

对于 $|\lambda| \le \frac{1}{\|Cv\|}$。设 $\gamma > 0$。那么，

$$\mathbb{P}\left[\|x\| \ge \sqrt{\mathrm{Tr}(\Sigma)} + t\right] \le 2e^{-\frac{1}{16}\min(\frac{t^2}{\|\Sigma\|}, \frac{t}{\|C\|}, \frac{2t\sqrt{\mathrm{Tr}(\Sigma)}+t^2}{\|C\|_F^2})}.$$

因此，以 $1-\delta$ 的概率，

$$\|x\| \le \sqrt{\mathrm{Tr}(\Sigma)} + 4\sqrt{\|\Sigma\| \log\frac{2}{\delta}} + 16\|C\| \log\frac{2}{\delta} + \min\left(4\|C\|_F\sqrt{\log\frac{2}{\delta}}, 8\frac{\|C\|_F^2}{\sqrt{\mathrm{Tr}(\Sigma)}}\log\frac{2}{\delta}\right)$$

证明思路与次高斯情形下的（Hsu 等人，2012年）类似，具体如下。定义 $v \sim N(0, I)$。我们将 $\mathbb{P}[\|x\| > t]$ 与矩母函数 $\mathbb{E}_x[e^{\lambda^2\|x\|^2}]$ 联系起来，它等于 $\mathbb{E}_{x,\ v}[e^{\lambda\langle x,v\rangle}]$。如果交换期望的顺序，只要 $\|Cv\| \le 1/|\lambda|$，结果至多为 $\mathbb{E}_v[e^{\lambda^2 v^T \Sigma v}]$。由于 v 服从高斯分布，我们可以精确计算最后一个矩母函数。

为处理次伽马情形，我们需要一种方法来控制$\mathbb{E}_{x,\ v}[e^{\lambda\langle x,v\rangle}]$对于那些满足 $\|Cv\| > 1/|\lambda|$的 v。我们通过证明以下两点来做到这一点：(I) 不失一般性，$\|x\|$ 从不会严格大于我们要证明的界；(II) 这类情形对期望的贡献很小。

证明。定义 $\gamma = \frac{t}{\sqrt{\mathrm{Tr}(\Sigma)}}$因此我们希望对 $\mathbb{P}[\|x\| \ge (1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}]$. 进行界定。我们首先证明，不失一般性，$\|x\|$ 永远不会超过此阈值。

引入有界范数假设。我们首先证明，在不失一般性的情况下，可以假设$\|x\| \le (1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}$ 始终成立。令 $s \in \{\pm 1\}$ 服从与 x 独立的均匀分布，并定义

$$y = s \cdot x \cdot \min\left(1, \frac{(1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}}{\|x\|}\right).$$

以截断 x的范数并进行对称化。对于任意 v 和 x,

$$\begin{aligned}\mathop{\mathbb{E}}_s[e^{\lambda\langle y,v\rangle}] &= \cosh\left(\lambda\langle x,v\rangle \cdot \min\left(1, \frac{(1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}}{\|x\|}\right)\right)\\ &\le \cosh(\lambda\langle x,v\rangle)\end{aligned}$$

现在，由于 x 是$(\Sigma,\ C)$-次伽马，

$$\begin{aligned}\mathop{\mathbb{E}}_x[\cosh(\lambda\langle x,v\rangle)] &= \frac{1}{2}\left(\mathop{\mathbb{E}}_x[e^{\lambda\langle x,v\rangle}] + \mathop{\mathbb{E}}_x[e^{\lambda\langle x,-v\rangle}]\right)\\ &\le \frac{1}{2}\left(e^{\lambda^2 v^T\Sigma v/2} + e^{\lambda^2(-v)^T\Sigma(-v)/2}\right)\\ &= e^{\lambda^2 v^T\Sigma v/2}\end{aligned}$$

因此

$$\mathop{\mathbb{E}}_y[e^{\lambda\langle y,v\rangle}] \le e^{\lambda^2 v^T\Sigma v/2}.$$

因此，y 也是$(\Sigma,\ C)$-次伽马。定理中的目标量对于 y 和 x相同：$\mathbb{P}[\|x\| \ge (1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}] = \mathbb{P}[\|y\| \ge (1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}]$。由于 $\|y\| \le (1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}$恒成立，在定理证明中考虑 y 而非 x，我们便可不失一般性地假设 $\|x\| \le (1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}$。

将概率与$\mathbb{E}_{x,\ v}[e^{\lambda\langle x,v\rangle}]$**联系起来。**定义

$$\alpha := \mathbb{P}\left[\|x\| \ge (1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}\right]$$

因此，将马尔可夫不等式应用于 $e^{\lambda^2\|x\|^2/2}$可得，

$$\alpha \le \frac{\mathbb{E}[e^{\lambda^2\|x\|^2/2}]}{e^{\lambda^2(1+\gamma)^2\,\mathrm{Tr}(\Sigma)/2}}$$

对任意 λ。现在，令 $v \sim N(0,\ I_d)$。对任意 x,

$$\mathop{\mathbb{E}}_v[\epsilon^{\lambda\langle x,v\rangle}] = e^{\lambda^2\|x\|^2/2}$$

so

$$\alpha \le \mathop{\mathbb{E}}_{x,v}[e^{\lambda\langle x,v\rangle}]e^{-\lambda^2(1+\gamma)^2\,\mathrm{Tr}(\Sigma)/2}. \tag{4}$$

上界估计$\mathbb{E}_{x,\ v}[e^{\lambda\langle x,v\rangle}]$。我们将通过对 x 取内层期望来给出右端的上界。由于 x 是（$\Sigma,\ C$）-次伽马分布，对于每个 v,

$$\mathop{\mathbb{E}}_x[e^{\lambda\langle x,v\rangle}] \le e^{\lambda^2 v^T\Sigma v/2} \qquad \forall|\lambda| \le \frac{1}{\|Cv\|},$$

因此

$$\begin{aligned}\mathop{\mathbb{E}}_{x,v}[e^{\lambda\langle x,v\rangle}] &= \mathop{\mathbb{E}}_{x,v}[e^{\lambda\langle x,v\rangle}1_{\|Cv\|\le 1/|\lambda|} + e^{\lambda\langle x,v\rangle}1_{\|Cv\|>1/|\lambda|}] \\ &\le \mathop{\mathbb{E}}_{v}[e^{\lambda^2 v^T\Sigma v/2}1_{\|Cv\|\le 1/|\lambda|}] + \mathop{\mathbb{E}}_{x,v}[e^{\lambda\langle x,v\rangle}1_{\|Cv\|>1/|\lambda|}] \\ &\le \mathop{\mathbb{E}}_{v}[e^{\lambda^2 v^T\Sigma v/2}] + \mathop{\mathbb{E}}_{x}[\mathop{\mathbb{E}}_{v}[e^{\lambda\langle x,v\rangle}1_{\|Cv\|>1/|\lambda|}]] \qquad (5)\end{aligned}$$

我们先处理第一项。设 Σ 的特征值为 $\sigma_1^2 \ge \sigma_2^2 \ge \cdots \ge \sigma_d^2$。那么，$v^T\Sigma v/2$ 服从广义卡方分布，其分布与相互独立的高斯变量 $u_i \sim N(0,\ \sigma_i^2/2)$ 对应的 $\sum_i u_i^2$ 相同。容易验证，对于 $u\sim N(0,\ 1)$，u^2 服从 $(4,\ 4)$-次伽马分布，即，

$$\mathbb{E}[e^{\lambda(u^2-\mathbb{E}[u^2])}] = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \le e^{2\lambda^2} \qquad \forall|\lambda|\le\frac{1}{4}.$$

因此 $\sum u_i^2$ 是 $(\sum_i \sigma_i^4,\ 2\max\ \sigma_i^2)$ $=(\|\Sigma\|_F^2,\ 2\|\Sigma\|)$-次伽马。由于 $\|\Sigma\|_F^2 \le \|\Sigma\|\,\mathrm{Tr}(\Sigma)$，$v^T\Sigma v$ 也是 $(\|\Sigma\|\ \mathrm{Tr}(\Sigma)\ 2\|\Sigma\|)$-次伽马。

将均值项也纳入其中 $(\mathbb{E}[v^T\Sigma v/2] = \mathrm{Tr}(\Sigma)/2)$，我们得到

$$\mathop{\mathbb{E}}_{v}[e^{\lambda^2 v^T\Sigma v/2}] \le e^{\lambda^2\,\mathrm{Tr}(\Sigma)/2}\cdot e^{\lambda^4\,\mathrm{Tr}(\Sigma)\|\Sigma\|/2} \quad \forall\lambda^2 \le \frac{1}{2\|\Sigma\|}. \qquad (6)$$

现在我们对每个 的（5）中的第二项进行界定。由于v 是独立同分布高斯变量，$\|Cv\| \le \|C\|_F + \|C\|\sqrt{2\log\ \frac{1}{\delta}}$ 的概率为 $1-\delta$（见公式 1）。因此，对所有 $|\lambda| <1$ $\frac{}{2\|C\|_F}$，

$$\mathbb{P}[\|Cv\| > 1/|\lambda|] \le e^{-\frac{(1/|\lambda|-\|C\|_F)^2}{2\|C\|^2}} \le e^{-\frac{1}{8\lambda^2\|C\|^2}}$$

因此，根据柯西-施瓦茨不等式以及我们对 $\|x\|$ 的界，

$$\begin{aligned}\mathop{\mathbb{E}}_{v}[e^{\lambda\langle x,v\rangle}1_{\|Cv\|>1/|\lambda|}] &\le \sqrt{\mathop{\mathbb{E}}_{v}[e^{2\lambda\langle x,v\rangle}]\,\mathbb{P}[\|Cv\|>1/|\lambda|]} \\ &\le \sqrt{e^{2\lambda^2\|x\|^2}e^{-\frac{1}{8\lambda^2\|C\|^2}}} \\ &= e^{\lambda^2(1+\gamma)^2\,\mathrm{Tr}(\Sigma)-\frac{1}{16\lambda^2\|C\|^2}}.\end{aligned}$$

因此，只要 $\lambda^2 \le \min(\frac{1}{4(1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)\|C\|}}, \frac{1}{4\|C\|_F^2})$，

$$\mathop{\mathbb{E}}_{v}[e^{\lambda\|x\|v_1}1_{\|Cv\|>1/|\lambda|}] \le 1.$$

结合（6）（这是一个始终大于 1 的界）以及（5），

$$\begin{aligned}&\mathop{\mathbb{E}}_{x,v}[e^{\lambda\langle x,v\rangle}] \le 2e^{\lambda^2\,\mathrm{Tr}(\Sigma)/2}\cdot e^{\lambda^4\,\mathrm{Tr}(\Sigma)\|\Sigma\|/2} \\ &\forall\lambda^2 \le \min(\frac{1}{2\|\Sigma\|}, \frac{1}{4(1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}\|C\|}, \frac{1}{4\|C\|_F^2})\end{aligned}$$

并结合（4），

$$\begin{aligned}&\alpha \le 2e^{\frac{1}{2}\lambda^2\,\mathrm{Tr}(\Sigma)(\lambda^2\|\Sigma\|-2\gamma-\gamma^2)} \\ &\forall\lambda^2 \le \min(\frac{1}{2\|\Sigma\|}, \frac{1}{4(1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}\|C\|}, \frac{1}{4\|C\|_F^2})\end{aligned}$$

最终界。 再限制 λ^2 不超过 $\frac{2\gamma+\gamma^2}{2\|\Sigma\|}$，我们得到:

$$\begin{aligned}&\alpha \le 2e^{-\frac{1}{4}\lambda^2\,\mathrm{Tr}(\Sigma)(2\gamma+\gamma^2)} \\ &\forall\lambda^2 \le \min(\frac{1}{2\|\Sigma\|}, \frac{2\gamma+\gamma^2}{2\|\Sigma\|}, \frac{1}{4(1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}\|C\|}, \frac{1}{4\|C\|_F^2})\end{aligned}$$

令 λ^2 取该范围的最大值，得到

$$\alpha \le 2e^{-\frac{1}{4}\min(\frac{1}{2\|\Sigma\|}, \frac{2\gamma+\gamma^2}{2\|\Sigma\|}, \frac{1}{4(1+\gamma)\sqrt{\mathrm{Tr}(\Sigma)}\|C\|}, \frac{1}{4\|C\|_F^2})(2\gamma+\gamma^2)\,\mathrm{Tr}(\Sigma)}$$

前两种情况可以合并：$\min(\frac{2\gamma+\gamma^2}{2}, \frac{(2\gamma+\gamma^2)_2}{\ \ ^2}) \ge \frac{\gamma^2}{2}$。因此:

$$\alpha \le 2e^{-\frac{1}{16}\min(\frac{\gamma^2\,\mathrm{Tr}(\Sigma)}{\|\Sigma\|}, \frac{\gamma\sqrt{\mathrm{Tr}(\Sigma)}}{\|C\|}, \frac{(2\gamma+\gamma^2)\,\mathrm{Tr}(\Sigma)}{\|C\|_F^2})}.$$

代入 $\gamma = \frac{t}{\sqrt{\mathrm{Tr}(\Sigma)}}$ 得到第一个结果；令 t 使指数等于 $\log\ \frac{2}{\delta}$，则得到第二个结果。 □

6. 结论与未来工作

本文提出了一种用于高维位置估计的算法，得到了非渐近误差界，其结果接近于 $\mathcal{N}(0, \frac{\mathcal{I}_R^{-1}}{n})$ 的误差界。其中，$\mathcal{I}_R$ 是我们的分布在使用 $\mathcal{N}(0, R)$ 平滑后得到的费 Fisher 信息矩阵，适用于随 n 衰减的小 R。在证明这一结果的过程中，我们为高维随机变量的范数推导出一个新的集中不等式；这类随机变量的一维投影均为次伽马分布，该不等式本身也可能具有独立的研究价值。即使在一维情形下，我们的结果也改善了常数失败概率下的表现。对于拉普拉斯分布混合等函数类，此前没有工作给出固定 δ 时渐近收敛至 Cramér-Rao 界的速率，其中 $n\to\infty$。

这篇论文是有限样本参数估计理论的一步。我们的定量界还可以改进：我们得到的收敛至 Cramér-Rao 界的收敛速率为 $1+\ ^{1}\mathrm{poly}(n)$，但我们希望能得到更快的收敛速度（一般情况下为 $1+\frac{1}{\sqrt{n}}$，对于某些特定的函数类则为 $1+\frac{1}{n}$）。更一般地，我们可以考虑位置以外的参数估计；Cramér-Rao 界仍将渐近行为与费 Fisher 信息联系起来，但收敛速率仍难以确定。我们认为，理解高维位置估计是理解多个参数估计的重要一步。

7. 致谢

Shivam Gupta 和 Eric Price 获得了美国国家科学基金会项目 CCF-2008868、CCF-1751040（职业奖）以及美国国家科学基金会机器学习基础人工智能研究所（IFML）的资助。Shivam Gupta 曾在加州大学伯克利分校访问期间完成了本研究的部分工作。Jasper C.H. Lee 获得了裘槎博士后研究奖学金的慷慨资助，并获得美国国家科学基金会项目 DMS-2023239 的部分资助。

参考文献

Bickel, P. J. and Doksum, K. A. 《数理统计：基本思想与精选主题》，第一卷. Chapman and Hall/CRC，2015年。Boucheron, S., Lugosi, G., and Massart, P. 《集中不等式——独立性的非渐近理论》. 牛津大学出版社，2013年。ISBN 978-0- 19-953525-5. doi: 10.1093/acprof:oso/9780199535255. 001.0001.
URL
https://doi.org/10.1093/acprof:oso/9780199535255.001.0001.
Bousquet, O. 贝内特集中不等式及其在经验过程上确界中的应用。《数学报告》，334(6):495–500，2002年。Bousquet, O. 使用熵方法研究次可加函数的集中不等式。载于 Gin´e, E., Houdr´e, C., and Nualart, D. (编)，《随机不等式及其应用》，第213–247页，巴塞尔，2003年。Catoni, O. 挑战经验均值与经验方差：一项偏差研究。《亨利·庞加莱研究所年刊´，概率´与统计》，48(4):1148–1185，2012年。doi: 10.1214/11-AIHP454.
URLhttps://doi.org/10.1214/11-AIHP454.
Cherapanamjeri, Y., Flammarion, N., and Bartlett, P. L. 采用次高斯速率的快速均值估计。载于 Beygelzimer, A. and Hsu, D. (编)，第三十二届学习理论会议论文集，第99卷，机器学习研究论文集，第786–806页。PMLR，2019年6月25–28日。URL
https://proceedings.mlr.press/v99/cherapanamjeri19b.html.
Devroye, L., Lerasle, M., Lugosi, G., and Oliveira, R. I. 次高斯均值估计量。《统计年刊》，44(6):2695–2725，2016年。Gupta, S., Lee, J. C. H., Price, E., and Valiant, P. 位置的有限样本最大似然估计。载于神经信息处理系统大会'22论文集，2022年。
Hendeby, G. 线性非高斯系统中的基本估计与检测极限。博士论文，2005年11月。
Hopkins, S. B. 多项式时间的次高斯均值估计。arXiv，abs/1809.07425，2018年。Hsu, D.、Kakade, S. 和 Zhang, T. 次高斯随机向量二次型的尾不等式。《概率电子通信》，17:1–6，2012年。
Lee, J. C. H. 和 Valiant, P.。 $\mathbb{R}$中的最优次高斯均值估计。载于 FOCS'21 会议论文集，第 672–683 页，2022a。Lee, J. C. H. 和 Valiant, P.。超高维中的最优次高斯均值估计。载于 ITCS'22 会议论文集，第 98:1–98:21 页，2022b。Lugosi, G. 和 Mendelson, S. 随机向量均值的次高斯估计量。《统计年刊》，47，2017年2月。doi: 10.1214/17-AOS1639。
Miao, Y. 最大似然估计量的集中不等式。《应用数学快报》，23(10):1305–1309，2010年。Pinelis, I. 最大似然估计量趋于正态性的收敛速率的最优阶一致界与非一致界。《统计电子期刊》，11 (1):1160 – 1179，2017年。doi: 10.1214/17-EJS1264。URL
https://doi.org/10.1214/17-EJS1264。Spokoiny, V. 参数估计：有限样本理论。《统计年刊》，40(6):2877–2909，2011年。doi: 10.1214/12-AOS1054。user940。显示 $\mathbb{E}[f(x)g(x)] \geq \mathbb{E}[f(x)]\mathbb{E}[g(x)]$，对于 f、g 有界且非递减。数学 Stack Exchange，2015年。URL
https://math.stackexchange.com/q/1446526。下载于 2023-01。van der Vaart, A.《渐近统计学》。剑桥大学出版社，2000年。ISBN 9780521784504。

A. 一维位置估计的完整分析

A.1. 一维局部估计

以下算法（算法 1）是 局部一维估计的部分：它假设存在一个接近真实参数的初始估计量 λ。

算法 1一维局部平滑最大似然估计

输入参数：

- f 的描述，平滑参数 r，样本 x_1，……，$x_n \overset{i.i.d.}{\sim} f^\lambda$ 以及 λ 的初始估计 λ_1

^1. 令 $s(\lambda)$ 为 f_r 的得分函数，即 f 的 r 平滑版本。

2. 对每个样本 x_i，计算一个扰动后的样本 $x'_i = x_i + \mathcal{N}(0,\ r^2)$，其中所有样本所添加的高斯噪声均相互独立。

3. 在 λ_1 处计算经验得分，即 $\hat{s}(\lambda_1) = \frac{1}{n}\sum_{i=1}^n s(x'_i - \lambda_1)$。

^4. 返回 $\lambda = \lambda_1 - (\hat{s}(\lambda_1)/\mathcal{I}_r)$。

局部算法采用了平滑最大似然估计的简化视角，这使我们的方法区别于 Gupta 等人此前提出的方法（2022）。

我们将证明算法 1 满足以下保证：如果初始估计 λ_1 与真实参数 λ 的距离至多为 $\epsilon_{\max}$，并且我们选择了足够大的平滑参数 r，那么算法 1 的输出将接近真实参数 λ。

引理 A.1. 在算法 *1* 中，假设 $|\lambda_1 - \lambda| \le \epsilon_{\max}$ 对某个 $\epsilon_{\max} \ge \sqrt{2\log\frac{2}{\delta}}\frac{1}{\mathcal{I}_r}$ 成立。另假设平滑参数为 $r \ge 2\epsilon_{\max}$，且存在参数 $\gamma \ge 1$，使得 *1)* $r^2\sqrt{\mathcal{I}_r} \ge \gamma\epsilon_{\max}$，*2)* $r^2\sqrt{\log\frac{2}{\delta}/n} \ge \gamma\epsilon_{\max}^2$，以及 *3)* $\log\frac{2}{\delta}/n \le 1/\gamma^2$。（解释一下，使用该引理时，$\gamma$ 应当足够大，并且“$\omega(1)$”。）

*T*于是，以至少 $1-\delta$ 的概率，n 对于来自 f^λ 的样本，算法 *1* 的输出满足

$$|\hat{\lambda} - \lambda| \le \left(1 + O\left(\frac{1}{\gamma}\right)\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n\mathcal{I}_r}}$$

引理 A.1 的证明依赖于以下关于平滑分布的经验得分集中性的事实（Gupta 等人，2022）：在接近真实参数的初始参数估计处对其进行评估时。

第一个事实是得分的次伽马集中性。

事实 A.2. 假设我们取 n 独立同分布样本 $y_1, \ldots, y_n \leftarrow f_r^\lambda$，并考虑经验得分函数 $\hat{s}$，它将候选参数 $\hat{\lambda}$ 映射到 $\frac{1}{n}\sum_i s_r(y_i - \hat{\lambda})$，其中 s_r 是 f_r的得分函数。

那么，对任意 $|\epsilon| \le r/2$,

$$\mathop{\mathbb{P}}_{y_i \overset{i.i.d.}{\sim} f_r^\lambda}\left(|\hat{s}(\lambda+\epsilon) - \mathop{\mathbb{E}}_{x\leftarrow f_r}[s(x-\epsilon)]| \ge \sqrt{\frac{2\max(\mathbb{E}_x[s_r^2(x-\epsilon)], \mathcal{I}_r)\log\frac{2}{\delta}}{n}} + \frac{15\log\frac{2}{\delta}}{nr}\right) \le \delta$$

接下来的两个事实给出了得分的期望和二阶矩的界。

事实 A.3. 对于任意 $|\epsilon| \le r/2$，期望得分 $\mathbb{E}_{x\sim f_r}[s_r(x+\epsilon)]$ 满足

$$\mathop{\mathbb{E}}_{x\sim f_r}[s_r(x+\epsilon)] \in \left[-\mathcal{I}_r\epsilon \pm O\left(\sqrt{\mathcal{I}_r}\frac{\epsilon^2}{r^2}\right)\right]$$

Fact A.4. 对于任意 $|\epsilon| \le r/2$，若 $r/\epsilon = \overline{\Omega(\sqrt{\log e/(r^2\mathcal{I}_r)})}$，得分的二阶矩满足

$$\mathop{\mathbb{E}}_{x\sim f_r}\left[s_r^2(x+\epsilon)\right] \le \mathcal{I}_r\left(1+O\left(\frac{\epsilon}{r}\sqrt{\log\frac{e}{r^2\mathcal{I}_r}}\right)\right)$$

此外，我们总有 $\mathcal{I}_r \le 1/r^2$，因此上面的 $\sqrt{\log 1/(r^2\mathcal{I}_r)}$ 是定义良好的。

现在我们可以利用这些事实证明引理 A.1。证明策略很直接：我们利用事实 A.2 和 A.4 证明 $\hat{s}(y)$ 以高概率集中在其期望附近，并利用事实 A.3 证明 $\hat{s}(y)$ 的期望（对于 $y=\lambda+\epsilon$，即 $\mathbb{E}[s(x-\epsilon)]$）非常接近 $\mathcal{I}_r\epsilon$。随后，三角不等式表明，$y-(\hat{s}(y)/\mathcal{I}_r)$ 以高概率必定接近 λ。

引理 *A.1* 的证明。令 $\lambda_1=\lambda+\epsilon$。根据引理假设，$|\epsilon|\le\epsilon_{\max}$。

首先，我们证明，在满足以下引理前提条件的假设下 $r^2\sqrt{\mathcal{I}_r}\ge\gamma\epsilon_{\max}$，事实 A.4 意味着在 $\lambda-\epsilon$ 处得分的二阶矩，即 $\mathbb{E}_{x\sim f_r}[s_r^2(x+\epsilon)],$，上界为 $(1+O(1/\gamma))\mathcal{I}_r$。

为验证事实 A.4 的前提条件成立，注意到 $r^2\sqrt{\mathcal{I}_r}\ge\gamma\epsilon_{\max}\overline{\ge}\gamma\epsilon$ 等价于 $r/\epsilon\ge\gamma/\sqrt{r^2\mathcal{I}_r}$，这意味着

$$\begin{aligned}\frac{r}{\epsilon} &\ge \frac{\gamma}{\sqrt{r^2\mathcal{I}_r}}\\ &= \frac{\gamma}{\sqrt{e}}\sqrt{\frac{e}{r^2\mathcal{I}_r}}\\ &\ge \frac{\gamma}{\sqrt{e}}\sqrt{\log\frac{e}{r^2\mathcal{I}_r}}\end{aligned}$$

满足事实A.4的前提条件。

于是，该事实意味着

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim f_r}[s_r^2(x+\epsilon)] &\le \mathcal{I}_r\left(1+O\left(\frac{\epsilon}{r}\sqrt{\log\frac{e}{r^2\mathcal{I}_r}}\right)\right)\\ &\le \mathcal{I}_r\left(1+O\left(\frac{\epsilon_{\max}}{r}\sqrt{\log\frac{e}{r^2\mathcal{I}_r}}\right)\right)\\ &\le \mathcal{I}_r\left(1+O\left(\frac{\epsilon_{\max}}{r}\sqrt{\frac{e}{r^2\mathcal{I}_r}}\right)\right)\\ &\le \mathcal{I}_r\left(1+O\left(\frac{\epsilon_{\max}}{r^2\sqrt{\mathcal{I}_r}}\right)\right)\\ &\le \mathcal{I}_r\left(1+O\left(\frac{1}{\gamma}\right)\right)\end{aligned}$$

接下来，我们将事实 A.2 的集中界与刚刚推导出的 $\mathbb{E}_x[s_r^2(x+\epsilon)]$ 的二阶矩界结合起来，从而证明 $\hat{s}(\lambda-\epsilon)$ 以高概率接近其期望。

$$\begin{aligned}\left|\hat{s}(y)-\mathop{\mathbb{E}}_{x\leftarrow f_r}[s(\lambda-\epsilon)]\right| &\le \sqrt{\frac{2\log\frac{2}{\delta}}{n}\mathcal{I}_r}\left(1+O\left(\frac{1}{\gamma}\right)\right)+\frac{15\log\frac{2}{\delta}}{nr}\\ &\le \left(1+O\left(\frac{1}{\gamma}\right)\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n}\mathcal{I}_r}+\frac{15}{2\sqrt{\gamma}}\left(\frac{2\log\frac{2}{\delta}}{n}\right)^{\frac{1}{4}}\sqrt{\frac{2\log\frac{2}{\delta}}{n}\mathcal{I}_r}\quad\text{(see below)}\\ &\le \left(1+O\left(\frac{1}{\gamma}\right)\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n}\mathcal{I}_r}+O\left(\frac{1}{\gamma}\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n}\mathcal{I}_r}\quad\text{since }\log\frac{2}{\delta}/n\le 1/\gamma^2\end{aligned}$$

$$= \left(1 + O\left(\frac{1}{\gamma}\right)\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n}\mathcal{I}_r}$$

其中，第二个不等式源于假设 $r^2\sqrt{\mathcal{I}_r} \geq \gamma\epsilon_{\max} \geq \gamma\sqrt{\frac{2\log\frac{1}{\delta}}{n}\frac{1}{\mathcal{I}_r}}$。

进一步利用事实 A.3，这意味着 $\epsilon = y - \lambda$ 可由 $\hat{s}(y)/\mathcal{I}_r$很好地近似，如下所示。

$$\begin{aligned}
|\epsilon - (\hat{s}(y)/\mathcal{I}_r)| &= \frac{1}{\mathcal{I}_r}|\mathcal{I}_r\epsilon - \hat{s}(y)| \\
&= \frac{1}{\mathcal{I}_r}\left|\hat{s}(y) - \mathop{\mathbb{E}}_{x\leftarrow f_r}[s(\lambda-\epsilon)] + \mathop{\mathbb{E}}_{x\leftarrow f_r}[s(\lambda-\epsilon)] - \mathcal{I}_r\epsilon\right| \\
&\leq \frac{1}{\mathcal{I}_r}\left|\hat{s}(y) - \mathop{\mathbb{E}}_{x\leftarrow f_r}[s(\lambda-\epsilon)]\right| + \frac{1}{\mathcal{I}_r}\left|\mathop{\mathbb{E}}_{x\leftarrow f_r}[s(\lambda-\epsilon)] - \mathcal{I}_r\epsilon\right| \\
&= \left(1 + O\left(\frac{1}{\gamma}\right)\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n\mathcal{I}_r}} + O\left(\frac{\epsilon^2}{r^2\sqrt{\mathcal{I}_r}}\right) \\
&\qquad \text{by the previous bound and Fact A.3}
\end{aligned}$$

根据引理的假设，我们有 $\epsilon^2/r^2 \leq \epsilon_{\max}^2/r^2 \leq (1/\gamma)\sqrt{\log\frac{2}{\delta}/n}$，因此我们得到 $|\epsilon - (\hat{s}(y)/\mathcal{I}_r)|$ 的上界为

$$|\epsilon - (\hat{s}(y)/\mathcal{I}_r)| \leq \left(1 + O\left(\frac{1}{\gamma}\right)\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n\mathcal{I}_r}}$$

综上，我们有

$$|\hat{\lambda} - \lambda| = |y - (\hat{s}(y)/\mathcal{I}_r) - \lambda| = |\lambda + \epsilon - (\hat{s}(y)/\mathcal{I}_r) - \lambda| \leq \left(1 + O\left(\frac{1}{\gamma}\right)\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n\mathcal{I}_r}}$$

如所需。 □

A.2 一维全局估计

现在可以给出一维全局估计算法（算法 2）：该算法先从数据的 $o(1)$ 部分中获得真实参数的初步估计，然后在其余数据上调用局部算法 1 。

算法 2一维全局平滑最大似然估计

输入参数：

- 失败概率 δ，f的描述，n从 f^λ 抽取的独立同分布样本，其中某个未知 λ1. 令 q为 $\sqrt{2(\log\frac{2}{\delta}/n)^{2/5}}$。2. 计算一个 $\alpha \in [q, 1-q]$，使由 f的 $\alpha \pm q$ 分位数所界定区间的宽度最小。

3. 取前 $(\log\frac{2}{\delta}/n)^{1/10}$ 部分 n 样本的 α-分位数。

4. 令 $r^* = \Omega((\frac{\log\frac{2}{\delta}}{n})^{1/8})$四分位距。

5. 在其余样本上运行算法 1 ，使用初始估计 $\lambda_1 = x_\alpha$ 和 r^*平滑，并返回最终的 $\hat{}$估计 λ。

算法的全局部分及其分析与 Gupta 等人（2022）的工作基本完全相同，只对某些参数做了细微调整。我们再次指出，算法的改进在于算法的局部部分，具体见算法 1。

定理 1.1 (一维平滑最大似然估计)。 给定一个模型 f，设分布 f 的 r-平滑费希尔信息为 $\mathcal{I}_r$，并设四分位距 为 f 的四分位距。固定失败概率为 $\delta \le 0.5$，并假设 $n \ge c \cdot$ 对数 $\frac{2}{\delta}$，其中 c 为某个充分大的常数。

选择 $r^* = \Omega((\frac{\log \frac{2}{\delta}}{n})^{1/8})$四分位距。那么，以至少 $1-\delta$ 的概率，算法 $\hat{}\lambda 2$ 的输出 满足

$$|\hat{\lambda} - \lambda| \le \left(1 + O\left(\frac{\log \frac{2}{\delta}}{n}\right)^{\frac{1}{10}}\right)\sqrt{\frac{2\log \frac{2}{\delta}}{n\mathcal{I}_{r^*}}}$$

算法 2 的分析还需要用到 (Gupta 等人，2022) 中的另一个技术性事实，即平滑费希尔信息的下界。

事实 A.5。 令 $\mathcal{I}_r$ 表示关于 f_r 的费 *Fisher* 信息，其中 f_r 是分布 f 的 r-平滑版本。令 四分位距 表示 f 的四分位距。那么， $\mathcal{I}_r \gtrsim 1/($四分位距 $+r)^2$。这里，隐藏常数是一个普适常数，与分布 f 无关，也与 r 无关。

定理 *1.1* 的证明。 步骤 2 使用 $(\log \frac{2}{\delta}/n)^{1/10}\, n$ 个样本来计算样本 α-分位数。根据标准的 Chernoff 界，以至少 $(\log \frac{2}{\delta}/n)^2$, 的概率，样本分位数的误差（按其在真实分布中的分位数衡量）至多为

$$\begin{aligned}
&\sqrt{\frac{2\log \frac{2}{\delta(\log \frac{2}{\delta}/n)^2}}{(\log \frac{2}{\delta}/n)^{1/10}\, n}} \\
&\le \sqrt{\frac{2(\log \frac{2}{\delta})(\frac{n}{\log \frac{2}{\delta}})^{1/10}}{(\log \frac{2}{\delta}/n)^{1/10}\, n}} \\
&= \sqrt{2}\left(\frac{\log \frac{2}{\delta}}{n}\right)^{2/5}
\end{aligned}$$

因此，若上述事件发生，步骤 2 将得到一个样本 α-分位数 x_α，使得 $x_\alpha - \lambda$ 位于 f的 $\alpha - \sqrt{2}(\log \frac{2}{\delta}/n)^{2/5}$ 和 $\alpha + \sqrt{2}(\log \frac{2}{\delta}/n)^{2/5}$ 分位数之间。此外，根据 α定义中的最小性条件，这两个分位数之间的距离至多为 $O((\log \frac{2}{\delta}/n)^{2/5})$IQR。

我们将应用引理 A.1 ，使用失败概率 $\delta(1-(\log \frac{2}{\delta}/n)^2)$。我们将验证，**(A)** 在上述意义下以步骤 2 成功为条件时，引理 A.1 的前提条件对 $\lambda_1 = x_\alpha$、所选的 r^* 以及对 γ的适当选择均成立；并且 **(B)** 引理 A.1保证的估计误差蕴含所需的误差界。如果上述确定性检验均成立，那么根据并集界，算法 2 将满足所需的中间界保证，除非发生概率为 δ。

对于 以下计算，注意，$\log \delta\, 1-\ ^2((\log \frac{2}{\delta}/n)^2) \le 1.1\log \frac{2}{\delta}$因为 $n \gg \log \frac{2}{\delta}$且 $\delta \le -\quad 0.5.$

(A)：我们以步骤 2 成功为条件，并检查引理A.1的前提条件。
现在检查前提条件 $r^* \ge 2\epsilon_{\max}$，对于 $\epsilon_{\max} = \max(\sqrt{2\log \delta\, 1-\ ^2((\log \frac{2}{\delta}/n)^2)/(n\mathcal{I}_{r^*})}\, O(\log \frac{2}{\delta}/n)^{2/5}$四分位距)。
首先， $r^* = \Omega((\frac{\log \frac{2}{\delta}}{n})^{1/8})$四分位距 $\gg O((\log \frac{2}{\delta}/n)^{2/5})$四分位距，其中 $\gg$利用了关于 n 大小的假设。我们还可以证明 $O(\log \frac{1}{\delta}/n)^{2/5}$四分位距 $\ge \sqrt{2\log \frac{2}{\delta}/(n\mathcal{I}_{r^*})}$。根据事实 A.5，回顾可知，对于任意 $r>0$， $\mathcal{I}_r \ge \Omega(1/($四分位距 $+r)^2)$。因此，$\sqrt{2\log \delta\, 1-\ ^2((\log \frac{2}{\delta}/n)^2)/(n\mathcal{I}_{r^*})} \le O(\sqrt{\log \frac{2}{\delta}/(n\mathcal{I}_{r^*})}) \le O((\log \frac{2}{\delta}/n)^{1/2}$四分位距) $\ll O((\log \frac{2}{\delta}/n)^{2/5})$ 四分位距，其中 $\ll$ 源于关于 n 大小的定理假设。

现在需要检查引理 A.1 的最后 3 个前提条件。令 $n' = (1-(\log \frac{2}{\delta}/n)^{1/10})n$ 表示算法 1 在步骤 4 中调用时使用的样本数。根据定理假设，我们有 $n' = \Theta(n)$。此外，回顾事实 A.5 可知，$\mathcal{I}_r \ge \Omega(1/($四分位距 $+r)^2)$。取 $\gamma = O(\ ^n\log \frac{2}{\delta})^{1/10}$，我们检查将引理 A.1 应用于步骤 4 中使用的 $n' = \Theta(n)$ 个点时，以下剩余条件是否满足：

1. $(r^*)^2\sqrt{\mathcal{I}_{r^*}} \geq (r^*)^2/$(四分位距 $+ r^*$) $\geq \Omega(\log_n \frac{2}{\delta})^{1/4}$四分位距 $\geq \Omega(\frac{\log \frac{2}{\delta}}{n})^{3/10}$四分位距 $\geq \gamma\epsilon_{\max}$。
2. $(r^*)^2\sqrt{\log \frac{\delta}{1-\frac{2}{((\log \frac{2}{\delta}/n)^2)}}/n'} \geq \Omega(\log_n \frac{2}{\delta})^{1/4}$四分位距$^2\sqrt{\log \frac{1}{\delta}/n} = \Omega(\frac{\log \frac{2}{\delta}}{n})^{7/10}$四分位距 $^2 = \gamma\epsilon_{\max}^2$
3. $\log \frac{\delta}{1-\frac{2}{((\log \frac{2}{\delta}/n)^2)}}/n' \leq O(\log \frac{2}{\delta}/n') \leq O((\log \frac{2}{\delta}/n)^{1/5}) \leq 1/\gamma^2$.

(B)：我们检查引理 A.1 的保证足以推出所需的界。为此，我们需要对 $\log \frac{\delta}{1-\frac{2}{((\log \frac{2}{\delta}/n)^2)}}$给出一个稍微精细一些的界：

$$\begin{aligned}
\log \frac{2}{\delta(1-(\log \frac{2}{\delta}/n)^2)} &= \left(1 + \frac{\log \frac{1}{1-(\log \frac{2}{\delta}/n)^2}}{\log \frac{2}{\delta}}\right) \log \frac{2}{\delta} \\
&\leq \left(1 + O\left(\frac{(\log \frac{2}{\delta}/n)^2}{\log \frac{2}{\delta}}\right)\right) \log \frac{2}{\delta} \quad \text{since } n \gg \log \frac{2}{\delta} \\
&\leq \left(1 + O\left(\frac{\log \frac{2}{\delta}}{n}\right)\right) \log \frac{2}{\delta}
\end{aligned}$$

ˆ当引理 A.1 的前提条件满足时，步骤 4 的成功意味着最终估计 λ 满足 g

$$\begin{aligned}
|\hat{\lambda} - \lambda| &\leq \left(1 + O\left(\frac{1}{\gamma}\right)\right) \sqrt{\frac{2\log \frac{2}{\delta(1-(\log \frac{2}{\delta}/n)^2)}}{n'\mathcal{I}_{r^*}}} \\
&\leq \left(1 + O\left(\frac{1}{\gamma}\right) + O\left(\frac{\log \frac{2}{\delta}}{n}\right)\right) \sqrt{\frac{2\log \frac{2}{\delta}}{n'\mathcal{I}_{r^*}}} \\
&= \left(1 + O\left(\frac{\log \frac{2}{\delta}}{n}\right)^{\frac{1}{10}} + O\left(\frac{\log \frac{2}{\delta}}{n}\right)\right) \sqrt{\frac{2\log \frac{2}{\delta}}{n'\mathcal{I}_{r^*}}} \\
&= \left(1 + O\left(\frac{\log \frac{2}{\delta}}{n}\right)^{\frac{1}{10}}\right) \sqrt{\frac{2\log \frac{2}{\delta}}{n'\mathcal{I}_{r^*}}} \\
&= \left(1 + O\left(\frac{\log \frac{2}{\delta}}{n}\right)^{\frac{1}{10}}\right) \sqrt{\frac{2\log \frac{2}{\delta}}{n\mathcal{I}_{r^*}}} \\
&\quad \text{since } n' = \left(1 - \left(\log \frac{2}{\delta}/n\right)^{1/10}\right) n
\end{aligned}$$

□

B. 高维位置估计

本节对用于估计高维分布的位置的主定理 B.16 进行了完整分析。我们首先在附录B.1中给出一些重要定义。然后，在附录B.2中证明平滑分布的得分的一些关键性质。在附录 B.3 中，我们展示得分函数具有适当的方差和尺度参数的次伽马性质。接着，附录 B.4给出了经验得分估计与真实得分之间偏差的误差界。最后，附录 B.5 和 B.6 分别对我们的 LocalMLE 和全局最大似然估计算法进行了分析。

B.1. 定义

令 f 为 $\mathbb{R}^d$ 上的任意分布，并令 $Y \sim f$。令平滑参数 $R \in \mathbb{R}^{d\times d}$ 为噪声 $Z_R \sim w_R = \mathcal{N}$ 的协方差矩阵，其中噪声服从 $Z_R \sim w_R = \mathcal{N}(0,\ R)$ 分布，且独立于 Y 采样。我们将 R-平滑 分布 f_R 定义为满足

$X = Y + Z_R \sim f_R$。因此， f_R 的概率密度函数为

$$f_R(x) = \mathop{\mathbb{E}}_{Z_R \sim w_R}[f(x + Z_r)]$$

设 s_R 为 f_R的得分函数。我们有

$$s_R(x) = \nabla \log f_R(x) = \frac{\nabla f_R(x)}{f_R(x)}$$

设 $\mathcal{I}_R$ 为 f_R的费 Fisher 信息矩阵。那么，

$$\mathcal{I}_R = \mathop{\mathbb{E}}_{x \sim f_R}[s_R(x) s_R(x)^T]$$

我们将向量 x 的 M-范数定义为

$$\|x\|_M = \sqrt{x^T M x}$$

B.2. 平滑得分的性质

本节将证明得分函数 s_R 的一些性质，其中 R是平滑分布 f_R。这些性质将在全文中用到。首先，我们在引理 B.1 中给出 s_R的一个实用刻画。然后，利用引理 B.2，我们证明引理 B.3：对于位置的良好初始估计，假设其误差为 $\epsilon \in \mathbb{R}^d$，且 ϵ“较小”，将 $s_R(x+\epsilon)$左乘 $-\mathcal{I}_R^{-1}$ 进行“得分反演”，即可在期望意义下得到误差 ϵ 的良好估计。之后，我们利用引理 B.4 证明引理 B.5：当 ϵ较小时，经过适当变换的平移得分 $s_R(x+\epsilon)$，其协方差与费 Fisher 信息矩阵 $\mathcal{I}_R$的相应变换相近。

我们首先给出得分 s_R 的刻画，后文将始终用到这一刻画。

引理 *B.1.* 设 f 是 $\mathbb{R}^d$ 上的任意分布，并令 f_R 为 f 的 R-平滑版本。也就是说， $f_R(x) = \mathbb{E}_{y \sim f}\left[(2\pi)^{-d/2} \det(R)^{-1/2} \exp\left(-\frac{1}{2}(x-Y)^T R^{-1}(x-Y)\right)\right]$。令 s_R为 f_R 的得分函数。设 (X, Y, Z_R) 服从联合分布，使得 $Y \sim f$, $Z_R \sim \mathcal{N}(0, R)$ 相互独立，且 $X = Y + Z_R \sim f_R$。对于 $\epsilon \in \mathbb{R}^d$，有

$$\frac{f_R(x+\epsilon)}{f_R(x)} = \mathop{\mathbb{E}}_{Z_R|x}\left[e^{\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon}\right]$$

因此

$$s_R(x) = \mathop{\mathbb{E}}_{Z_R|x}\left[R^{-1} Z_R\right]$$

证明。 首先，我们证明对于 $\epsilon \in \mathbb{R}$

$$\frac{f_R(x+\epsilon)}{f_R(x)} = \mathop{\mathbb{E}}_{Z_R|x}\left[\frac{w_R(Z_R+\epsilon)}{w_R(Z_R)}\right]$$

注意到

$$p(z|x) = \frac{p(z,x)}{p(x)} = \frac{f(x-z) w_R(z)}{f_R(x)}$$

So,

$$\begin{aligned} f_R(x+\epsilon) &= \int_{[-\infty,\infty]^d} w_R(z) f(x+\epsilon-z) dz \\ &= \int p(z|x) f_R(x) \frac{w_R(z+\epsilon)}{w_R(z)} dz \\ &= f_R(x) \mathop{\mathbb{E}}_{Z_R|x}\left[\frac{w_R(Z_R+\epsilon)}{w_r(Z_R)}\right] \end{aligned}$$

但现在， $w_R(x) = (2\pi)^{-d/2} \det(R)^{-1/2} e^{-\frac{1}{2}x^T R^{-1} x}$因此，

$$\frac{f_R(x+\epsilon)}{f_R(x)} = \mathop{\mathbb{E}}_{Z_R|x}\left[e^{\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon}\right]$$

which 是第一个论断。现在，令 $\epsilon = \gamma e_i$。我们对 γ 求导，并在 $\gamma = 0$ 处取值，以得到 t

$$\frac{\nabla_{e_i} f_R(x)}{f_R(x)} = \mathop{\mathbb{E}}_{Z_R|x}\left[(R^{-1}Z_R)_i\right]$$

So,

$$s_R(x) = \frac{\nabla f_R(x)}{f_R(x)} = \mathop{\mathbb{E}}_{Z_R|x}\left[R^{-1}Z_R\right]$$

□

下一个引理 B.2 是我们在引理中使用的一个效用结果B.3。

引理 B.2。令 f_R 为分布 f 在 $\mathbb{R}^d$上的 R-平滑版本。对于 $\epsilon \in \mathbb{R}^d$，令

$$\Delta_\epsilon(x) := \frac{f_R(x+\epsilon) - f_R(x) - (\nabla f_R(x))^T\epsilon}{f_R(x)}$$

那么，对于任意满足 $|\epsilon^T R^{-1}\epsilon| \le \frac{1}{4}$ 的 ϵ，我们有

$$\mathop{\mathbb{E}}_{x\sim f_R}[\Delta_\epsilon(x)^2] \lesssim (\epsilon^T R^{-1}\epsilon)^2$$

证明。由引理 B.1，我们有

$$\Delta_\epsilon(x) = \frac{f_R(x+\epsilon) - f_R(x) - (\nabla f_R(x))^T\epsilon}{f_R(x)} = \mathop{\mathbb{E}}_{Z_R|x}\left[e^{\epsilon^T R^{-1}Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon} - 1 - Z_R^T R^{-1}\epsilon\right]$$

令 α_ϵ： $\mathbb{R}^d \to \mathbb{R}$ 满足

$$\alpha_\epsilon(z) = e^{\epsilon^T R^{-1}z - \frac{1}{2}\epsilon^T R^{-1}\epsilon} - 1 - z^T R^{-1}\epsilon$$

我们希望对其进行界定

$$\begin{aligned}
\mathop{\mathbb{E}}_x[\Delta_\epsilon(x)^2] &= \mathop{\mathbb{E}}_x[\mathop{\mathbb{E}}_{Z_R|X}[\alpha_\epsilon(Z_r)]^2] \\
&\le \mathop{\mathbb{E}}_{x,Z_R}[(\alpha_\epsilon(Z_R))^2] \\
&= \mathop{\mathbb{E}}_{Z_R\sim\mathcal{N}(0,R)}[(\alpha_\epsilon(Z_R))^2]
\end{aligned} \tag{7}$$

在余下的证明中，令 $W = \epsilon^T R^{-1}Z_R$。由于 $Z_R \sim \mathcal{N}(0,\ R)$，因此有 $W \sim \mathcal{N}(0,\ \epsilon^T R^{-1}\epsilon)$。当 $|W| \le 1$ 时，根据泰勒展开，我们有

$$e^{W - \frac{1}{2}\epsilon^T R^{-1}\epsilon} = 1 + W - \frac{1}{2}\epsilon^T R^{-1}\epsilon + O\left(\left(W - \frac{1}{2}\epsilon^T R^{-1}\epsilon\right)^2\right)$$

因此

$$|\alpha_\epsilon(Z_R)| \lesssim \epsilon^T R^{-1}\epsilon + W^2$$

这意味着 $\alpha_\epsilon(Z_R)^2 \lesssim (\epsilon^T R^{-1}\epsilon)^2 + W^4$，也就是说

$$\begin{aligned}
\mathop{\mathbb{E}}_{Z_R\sim\mathcal{N}(0,R)}\left[(\alpha_\epsilon(Z_R))^2 \cdot \mathbb{1}_{|\epsilon^T R^{-1}Z_R|\le 1}\right] &\lesssim \mathop{\mathbb{E}}_{W\sim\mathcal{N}(0,\epsilon^T R^{-1}\epsilon)}\left[\left((\epsilon^T R^{-1}\epsilon)^2 + W^4\right)\cdot \mathbb{1}_{|\epsilon^T R^{-1}\epsilon|\le 1}\right] \\
&\lesssim (\epsilon^T R^{-1}\epsilon)^2 + \mathop{\mathbb{E}}_{W\sim\mathcal{N}(0,\epsilon^T R^{-1}\epsilon)}[W^4] \\
&\lesssim (\epsilon^T R^{-1}\epsilon)^2
\end{aligned} \tag{8}$$

另一方面，当 $|W| \ge 1$,

$$|\alpha_\epsilon(Z_R)| \le e^{|W|}$$

$$\begin{aligned}
\mathop{\mathbb{E}}_{Z_R\sim\mathcal{N}(0,R)}\left[\alpha_\epsilon(Z_R)^2\cdot\mathbb{1}_{|\epsilon^TR^{-1}Z_r|\geq 1}\right] &\leq \mathop{\mathbb{E}}_{W\sim\mathcal{N}(0,\epsilon^TR^{-1}\epsilon)}[e^{2|W|}\mathbb{1}_{|W|\geq 1}] \\
&= 2\int_1^\infty \frac{1}{\sqrt{2\pi\epsilon^TR^{-1}\epsilon}}e^{2|w|}e^{-\frac{w^2}{2\epsilon^TR^{-1}\epsilon}}dw \\
&= 2e^{2|\epsilon^TR^{-1}\epsilon|}\int_1^\infty \frac{1}{\sqrt{2\pi\epsilon^TR^{-1}\epsilon}}e^{-\frac{(w-2|\epsilon^TR^{-1}\epsilon|)^2}{2\epsilon^TR^{-1}\epsilon}}dw \\
&\leq 2\sqrt{e}\mathop{\mathbb{P}}_{W\sim\mathcal{N}(0,\epsilon^TR^{-1}\epsilon)}\left[W\geq 1-2|\epsilon^TR^{-1}\epsilon|\right] \\
&\lesssim e^{-\frac{(1-2|\epsilon^TR^{-1}\epsilon|)^2}{2\epsilon^TR^{-1}\epsilon}} \\
&\leq e^{-\frac{1}{8\epsilon^TR^{-1}\epsilon}} \lesssim (\epsilon^TR^{-1}\epsilon)^2
\end{aligned} \tag{9}$$

将其与（7）和（8）结合即可得到该结论。 □

下面的引理 B.3 告诉我们，对于良好的初始估计 $\epsilon\in\mathbb{R}^d$，通过左乘 s_R 来 $-\mathcal{I}_R^{-1}$ 对得分进行反演，可以得到 ϵ 的良好期望估计。

引理 B.3（得分反演）。设 f_R 为一个经过 R 平滑的分布，其费 *Fisher* 信息矩阵为 $\mathcal{I}_R$。设 $s_R\colon\mathbb{R}^d\to\mathbb{R}^d$为 f_R的得分函数。设 $M\in\mathbb{R}^{d\times d}$ 为一个对称矩阵，满足 $M\succcurlyeq 0$。那么，对于任意满足$|\epsilon^TR^{-1}\epsilon|\leq 1/4$的 $\epsilon\in\mathbb{R}^d$，我们有

$$\|\mathop{\mathbb{E}}_{x\sim f_R}[-\mathcal{I}_R^{-1}s_R(x+\epsilon)]-\epsilon\|_M^2\lesssim\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|(\epsilon^TR^{-1}\epsilon)^2$$

证明。 根据 s_R的定义，

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim f_R}[s_R(x+\epsilon)] &= \int_{[-\infty,\infty]^d} f_R(x)\frac{\nabla f_R(x+\epsilon)}{f_R(x+\epsilon)}dx \\
&= \int\nabla f_R(x)\left(\frac{f_R(x-\epsilon)-f_R(x)}{f_R(x)}\right)dx
\end{aligned}$$

因为

$$\int\nabla f_R(x)dx=0$$

现在，根据 $\mathcal{I}_R$ 的定义

$$\mathcal{I}_R=\mathop{\mathbb{E}}_{x\sim f_R}[s_R(x)s_R(x)^T]=\int_{[-\infty,\infty]^d}\frac{\nabla f_R(x)(\nabla f_R(x))^T}{f_R(x)}dx$$

So,

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim f_R}[s_R(x+\epsilon)]+\mathcal{I}_R\epsilon &= \int_{-[\infty,\infty]^d}\frac{\nabla f_R(x)}{f_R(x)}\left(f_R(x-\epsilon)-f_R(x)+(\nabla f_R(x))^T\epsilon\right)dx \\
&= \mathop{\mathbb{E}}_{x\sim f_R}[\Delta_{-\epsilon}(x)s_R(x)]
\end{aligned}$$

其中 $\Delta_\epsilon(x):=\frac{f_R(x+\epsilon)-f_R(\quad x\quad -\nabla f_R\ x\quad)^T\epsilon)(())()}{f_R\ x}$。现在，在两边左乘 $-M^{1/2}\mathcal{I}_R^{-1}$，

$$M^{1/2}\left(\mathop{\mathbb{E}}_{x\sim f_R}[-\mathcal{I}_R^{-1}s_R(x+\epsilon)]-\epsilon\right)=\mathop{\mathbb{E}}_{x\sim f_R}\left[\Delta_{-\epsilon}(x)(-M^{1/2}\mathcal{I}_R^{-1}s_R(x))\right]$$

因此，我们有

$$\|\mathop{\mathbb{E}}_{x\sim f_R}[-\mathcal{I}_R^{-1}s_R(x+\epsilon)]-\epsilon\|_M^2=\|\mathop{\mathbb{E}}_{x\sim f_R}[\Delta_{-\epsilon}(x)\left(-M^{1/2}\mathcal{I}_R^{-1}s_R(x)\right)]\|^2$$

现在，根据柯西-施瓦茨不等式

$$\begin{aligned}\| \mathop{\mathbb{E}}_{x\sim f_R}\left[\Delta_{-\epsilon}(x)(-M^{1/2}\mathcal{I}_R^{-1}s_R(x))\right]\|^2 &= \sup_{w\in S^{d-1}} \mathop{\mathbb{E}}_{x\sim f_R}[\Delta_{-\epsilon}(x)(-M^{1/2}\mathcal{I}_R^{-1}s_R(x))^T w]^2 \\ &\le \sup_{w\in S^{d-1}} \mathop{\mathbb{E}}_{x\sim f_R}[\Delta_{-\epsilon}(x)^2] \mathop{\mathbb{E}}_{x\sim f_R}\left[(-M^{1/2}\mathcal{I}_R^{-1}s_R(x))^T w)^2\right] \\ &= \mathop{\mathbb{E}}_{x\sim f_R}\left[\Delta_{-\epsilon}(x)^2\right] \| \mathop{\mathbb{E}}_{x\sim f_R}\left[(-M^{1/2}\mathcal{I}_R^{-1}s_R(x))(s_R(x)^T\mathcal{I}_R^{-1}M^{1/2})\right]\| \\ &= \mathop{\mathbb{E}}_{x\sim f_R}[\Delta_{-\epsilon}(x)^2]\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|\end{aligned}$$

根据引理B.2，我们最终得到

$$\| \mathop{\mathbb{E}}_{x\sim f_R}[-\mathcal{I}_R^{-1}s_R(x+\epsilon)] - \epsilon\|_M^2 \lesssim \|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|(\epsilon^T R^{-1}\epsilon)^2$$

□

接下来，在引理 B.4 中，我们证明一个效用结果，并在引理 B.5中加以使用。

引理 B.4。 令 f_R 为 $\mathbb{R}^d$上的 f 的 R-平滑版本。对于 $\epsilon\in\mathbb{R}^d$，令

$$\zeta_\epsilon(x) = \frac{f_R(x-\epsilon)-f_R(x)}{f_R(x)}$$

那么，对于任意满足 $|\epsilon^T R^{-1}\epsilon|\le 1/4$的 ϵ，以及任意满足 $\alpha^2(\epsilon^T R^{-1}\epsilon)\lesssim 1$ 的 α，我们有

$$\mathop{\mathbb{E}}_{x\sim f_R}[\zeta_\epsilon(x)^2] \lesssim (\epsilon^T R^{-1}\epsilon)(\alpha^2 e^{-\Omega(\alpha^2)} + e^{-\Omega(\alpha^2)})$$

证明。 根据引理 B.1，我们有

$$\zeta_\epsilon(x) = \frac{f_R(x-\epsilon)-f_R(x)}{f_R(x)} = \mathop{\mathbb{E}}_{Z_R|x}\left[e^{-\epsilon^T R^{-1}Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon} - 1\right]$$

对于剩余的证明，令 et $W = \epsilon^T R^{-1}Z_R$。由于 $Z_R\sim\mathcal{N}(0,R)$，我们有 $W\sim\mathcal{N}(0,\epsilon^T R^{-1}\epsilon)$。因此，我们有 t

$$\zeta_\epsilon(x) = \mathop{\mathbb{E}}_{W|x}\left[e^{-W-\frac{1}{2}\epsilon^T R^{-1}\epsilon} - 1\right]$$

令 α 为满足 $\alpha^2(\epsilon^T R^{-1}\epsilon)\lesssim 1$ 的参数。现在，我们有

$$\zeta_\epsilon(x) \le O\left(\alpha\sqrt{\epsilon^T R^{-1}\epsilon}\right) + \mathop{\mathbb{E}}_{W|x}\left[\mathbb{1}_{|W|>\alpha\sqrt{\epsilon^T R^{-1}\epsilon}}\left(e^{-W}-1\right)\right]$$

So,

$$\zeta_\epsilon(x)^2 \lesssim \alpha^2(\epsilon^T R^{-1}\epsilon) + \mathop{\mathbb{E}}_{W|x}\left[\mathbb{1}_{|W|>\alpha\sqrt{\epsilon^T R^{-1}\epsilon}}(e^{-W}-1)\right]^2$$

现在，为了界定第二项，根据詹森不等式，我们有

$$\mathop{\mathbb{E}}_{W|x}\left[\mathbb{1}_{|W|>\alpha\sqrt{\epsilon^T R^{-1}\epsilon}}(e^{-W}-1)\right]^2 \le \mathop{\mathbb{E}}_{W|x}\left[\mathbb{1}_{|W|>\alpha\sqrt{\epsilon^T R^{-1}\epsilon}}(e^{-W}-1)^2\right]$$

因此，我们有

$$\mathop{\mathbb{E}}_{x\sim f_R}\left[\zeta_\epsilon(x)^2\right] \lesssim \alpha^2(\epsilon^T R^{-1}\epsilon) + \mathop{\mathbb{E}}_{W}\left[\mathbb{1}_{|W|>\alpha\sqrt{\epsilon^T R^{-1}\epsilon}}(e^{-W}-1)^2\right]$$

现在，我们将分两种不同情况对上面的第二项 $\mathbb{E}_W\left[\mathbb{1}_{\underline{|W|>\alpha\sqrt{\epsilon^T R^{-1}\epsilon}}}(e^{-W}-1)^2\right]$ 进行界定：当

1. $|W| \le 1$

2. $|W| > 1$

当 $|W| \le 1$ 时，通过对指数函数作线性近似，我们有

$$(e^{-W} - 1)^2 \lesssim W^2$$

So,

$$\begin{aligned}\mathop{\mathbb{E}}_{W}\left[\mathbb{1}_{|W|>\alpha\sqrt{\epsilon^T R^{-1}\epsilon}}\mathbb{1}_{|W|\le 1}(e^{-W}-1)^2\right] &\lesssim \mathop{\mathbb{E}}_{W\sim\mathcal{N}(0,\epsilon^T R^{-1}\epsilon)}\left[\mathbb{1}_{|W|>\alpha\sqrt{\epsilon^T R^{-1}\epsilon}}\cdot W^2\right]\\ &\lesssim \alpha^2(\epsilon^T R^{-1}\epsilon)e^{-\Omega(\alpha^2)}\end{aligned}$$

另一方面，当 $|W| > 1$ 时

$$\begin{aligned}&\mathop{\mathbb{E}}_{W}\left[\mathbb{1}_{|W|>\max(1,\alpha\sqrt{\epsilon^T R^{-1}\epsilon})}(e^{-W}-1)^2\right]\\ &\le \int_{-\infty}^{-(1+\alpha\sqrt{\epsilon^T R^{-1}\epsilon})}\frac{1}{\sqrt{2\pi\epsilon^T R^{-1}\epsilon}}e^{-\frac{w^2}{2\epsilon^T R^{-1}\epsilon}}(e^{-w}-1)^2dw + \int_{1+\alpha\sqrt{\epsilon^T R^{-1}\epsilon}}^{\infty}\frac{1}{\sqrt{2\pi\epsilon^T R^{-1}\epsilon}}e^{-\frac{w^2}{2\epsilon^T R^{-1}\epsilon}}(e^{-w}-1)^2dw\\ &\lesssim \int_{1+\alpha\sqrt{\epsilon^T R^{-1}\epsilon}}^{\infty}\frac{1}{\sqrt{2\pi\epsilon^T R^{-1}\epsilon}}e^{-\frac{w^2}{2\epsilon^T R^{-1}\epsilon}}(e^{w}-1)^2dw\\ &\lesssim \int_{1+\alpha\sqrt{\epsilon^T R^{-1}\epsilon}}^{\infty}\frac{1}{\sqrt{2\pi\epsilon^T R^{-1}\epsilon}}e^{-\frac{w^2}{2\epsilon^T R^{-1}\epsilon}}e^{2w}dw\\ &= e^{2(\epsilon^T R^{-1}\epsilon)}\int_{1+\alpha\sqrt{\epsilon^T R^{-1}\epsilon}}^{\infty}\frac{1}{\sqrt{2\pi\epsilon^T R^{-1}\epsilon}}e^{-\frac{(w-2\epsilon^T R^{-1}\epsilon)^2}{2\epsilon^T R^{-1}\epsilon}}dw\\ &\lesssim e^{-\Omega\left(\frac{1}{\epsilon^T R^{-1}\epsilon}+\alpha^2\right)} \lesssim (\epsilon^T R^{-1}\epsilon)e^{-\Omega(\alpha^2)} \quad \text{since } |\epsilon^T R^{-1}\epsilon| \le 1/4\end{aligned}$$

因此，我们已证明

$$\mathop{\mathbb{E}}_{x\sim f_R}\left[\zeta_\epsilon(x)^2\right] \lesssim \alpha^2(\epsilon^T R^{-1}\epsilon)e^{-\Omega(\alpha^2)} + (\epsilon^T R^{-1}\epsilon)e^{-\Omega(\alpha^2)}$$

该结论得证。 □

下一个引理B.5 表明，当 ϵ较小时，平移得分经适当变换后的协方差$s_R(x+\epsilon)$ 与费 Fisher 信息矩阵 $\mathcal{I}_R$的相应变换相似。

引理 B.5。 设 f_R 是 $\mathbb{R}^d$ 上具有费 *Fisher* 信息矩阵 $\mathcal{I}_R$ 的 R-平滑分布。设 $M \in \mathbb{R}^{d\times d}$ 为满足 $M \succcurlyeq 0$ 的对称矩阵。那么，对于任意满足 $|\epsilon^T R^{-1}\epsilon| \le 1/4$ 的 $\epsilon \in \mathbb{R}^d$，以及任意满足 $\|v\| = 1$ 的 $v \in \mathbb{R}^d$，有:

$$\begin{aligned}&\left|v^T\left(\mathop{\mathbb{E}}_{x\sim f_R}\left[M^{1/2}\mathcal{I}_R^{-1}s_R(x+\epsilon)s_R(x+\epsilon)^T\mathcal{I}_R^{-1}M^{1/2}\right] - M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\right)v\right|\\ &\lesssim \sqrt{\epsilon^T R^{-1}\epsilon}\cdot(v^T M^{1/2}\mathcal{I}_R^{-1}M^{1/2}v)\sqrt{\log\left(\sup_{w\in S^{d-1}}\frac{w^T R^{-1}w}{w^T\mathcal{I}_R w}\right)}\end{aligned}$$

证明。 根据得分的定义，我们有：

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim f_R}[s_R(x+\epsilon)s_R(x+\epsilon)^T] &= \int_{[-\infty,\infty]^d} f_R(x)\frac{\nabla f_R(x+\epsilon)(\nabla f_R(x+\epsilon))^T}{f_R(x+\epsilon)^2}dx \\ &= \int f_R(x-\epsilon)\frac{\nabla f_R(x)(\nabla f_R(x))^T}{f_R(x)^2}dx \\ &= \mathcal{I}_R + \int (f_R(x-\epsilon)-f_R(x))\left(\frac{\nabla f_R(x)(\nabla f_R(x))^T}{f_R(x)^2}\right)dx \\ &= \mathcal{I}_R + \int \zeta_\epsilon(x)\left(\frac{\nabla f_R(x)(\nabla f_R(x))^T}{f_R(x)}\right)dx \\ &= \mathcal{I}_R + \mathop{\mathbb{E}}_{x\sim f_R}\left[\zeta_\epsilon(x)\frac{\nabla f_R(x)(\nabla f_R(x))^T}{f_R(x)^2}\right]\end{aligned}$$

其中 $\zeta_\epsilon(x) = \frac{f_R(x-\epsilon)-f_R(x)}{f_R(x)}$。现在，由于 $s_R(x) = \frac{\nabla f_R(x)}{f_R(x)}$，上式等价于

$$\mathop{\mathbb{E}}_{x\sim f_R}\left[s_R(x+\epsilon)s_R(x+\epsilon)^T\right] - \mathcal{I}_R = \mathop{\mathbb{E}}_{x\sim f_R}\left[\zeta_\epsilon(x)s_R(x)s_R(x)^T\right]$$

将两边分别左乘和右乘 $M^{1/2}\mathcal{I}_R^{-1}$，得

$$\mathop{\mathbb{E}}_{x\sim f_R}\left[M^{1/2}\mathcal{I}_R^{-1}s_R(x+\epsilon)s_R(x+\epsilon)^T\mathcal{I}_R^{-1}M^{1/2}\right] - M^{1/2}\mathcal{I}_R^{-1}M^{1/2} = M^{1/2}\mathcal{I}_R^{-1}\mathop{\mathbb{E}}_{x\sim f_R}\left[\zeta_\epsilon(x)s_R(x)s_R(x)^T\right]\mathcal{I}_R^{-1}M^{1/2}$$

那么，对于 $v\in\mathbb{R}^d$ 且 $\|v\|=1$

$$v^T\left(\mathop{\mathbb{E}}_{x\sim f_R}\left[M^{1/2}\mathcal{I}_R^{-1}s_R(x+\epsilon)s_R(x+\epsilon)^T\mathcal{I}_R^{-1}M^{1/2}\right] - M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\right)v = \mathop{\mathbb{E}}_{x\sim f_R}\left[\zeta_\epsilon(x)(v^TM^{1/2}\mathcal{I}_R^{-1}s_R(x))^2\right]$$

那么，利用柯西-施瓦茨不等式，

$$\begin{aligned}&\left|v^T\left(\mathop{\mathbb{E}}_{x\sim f_R}\left[M^{1/2}\mathcal{I}_R^{-1}s_R(x+\epsilon)s_R(x+\epsilon)^T\mathcal{I}_R^{-1}M^{1/2}\right] - M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\right)v\right| \\ &\le \sqrt{\mathop{\mathbb{E}}_{x\sim f_R}[\zeta_\epsilon(x)^2]\mathop{\mathbb{E}}_{x\sim f_R}\left[(v^TM^{1/2}\mathcal{I}_R^{-1}s_R(x))^4\right]}\end{aligned} \tag{10}$$

为了界定平方根内的第二项，回忆引理 B.1，我们有

$$s_R(x) = \mathop{\mathbb{E}}_{Z_R|x}[R^{-1}Z_R]$$

因此，根据詹森不等式，我们有

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim f_R}[(v^TM^{1/2}\mathcal{I}_R^{-1}s_R(x))^4] &= \mathop{\mathbb{E}}_{x\sim f_R}\left[(v^TM^{1/2}\mathcal{I}_R^{-1}\mathop{\mathbb{E}}_{Z_R|x}[R^{-1}Z_R])^4\right] \\ &= \mathop{\mathbb{E}}_{x\sim f_R}\left[\mathop{\mathbb{E}}_{Z_R|x}[v^TM^{1/2}\mathcal{I}_R^{-1}R^{-1}Z_R]^4\right] \\ &\le \mathop{\mathbb{E}}_{x\sim f_R}\left[\mathop{\mathbb{E}}_{Z_R|x}[(v^TM^{1/2}\mathcal{I}_R^{-1}R^{-1}Z_R)^4]\right] \\ &= \mathop{\mathbb{E}}_{Z_R}\left[(v^TM^{1/2}\mathcal{I}_R^{-1}R^{-1}Z_R)^4\right]\end{aligned}$$

现在，由于 $Z_R\sim\mathcal{N}(0,\ R)$，我们有 $v^\top M^{1/2}\mathcal{I}_R^{-1}R^{-1}Z_R\sim\mathcal{N}(0,\ v^\top M^{1/2}\mathcal{I}_R^{-1}R^{-1}\mathcal{I}_R^{-1}M^{1/2}v)$ 是一个一维高斯分布。因此，利用关于一维高斯分布 4th 阶矩的标准结论，我们有

$$\mathop{\mathbb{E}}_{x\sim f_R}\left[(v^TM^{1/2}\mathcal{I}_R^{-1}s_R(x))^4\right] \le \mathop{\mathbb{E}}_{Z_R}\left[(v^TM^{1/2}\mathcal{I}_R^{-1}R^{-1}Z_R)^4\right] = 3(v^TM^{1/2}\mathcal{I}_R^{-1}R^{-1}\mathcal{I}_R^{-1}M^{1/2}v)^2$$

对于第一项 un 式（10）根号下的部分，由引理 B.4，对于任意 $\alpha \in \mathbb{R}$ 满足 $\alpha^2(\epsilon^T R^{-1}\epsilon) \lesssim 1$，我们有

$$\mathop{\mathbb{E}}_{x\sim f_R}[\zeta_\epsilon(x)^2] \lesssim (\epsilon^T R^{-1}\epsilon)(\alpha^2 e^{-\Omega(\alpha^2)} + e^{-\Omega(\alpha^2)})$$

因此，结合上述结果与（10），可得

$$\begin{aligned}
&\left| v^T \left(\mathop{\mathbb{E}}_{x\sim f_R}\left[M^{1/2}\mathcal{I}_R^{-1} s_R(x+\epsilon) s_R(x+\epsilon)^T \mathcal{I}_R^{-1} M^{1/2} \right] - M^{1/2}\mathcal{I}_R^{-1}M^{1/2} \right) v \right| \\
&\lesssim (v^T M^{1/2}\mathcal{I}_R^{-1} R^{-1} \mathcal{I}_R^{-1} M^{1/2} v) \sqrt{(\epsilon^T R^{-1}\epsilon)(\alpha^2 e^{-\Omega(\alpha^2)} + e^{-\Omega(\alpha^2)})} \\
&\lesssim (v^T M^{1/2}\mathcal{I}_R^{-1} R^{-1} \mathcal{I}_R^{-1} M^{1/2} v) \sqrt{\epsilon^T R^{-1}\epsilon} (\alpha e^{-\Omega(\alpha^2)})
\end{aligned}$$

令 $\alpha = O\left(\sqrt{\log \frac{v^T M^{1/2}\mathcal{I}_R^{-1} R^{-1} \mathcal{I}_R^{-1} M^{1/2} v}{v^T M^{1/2}\mathcal{I}_R^{-1} M^{1/2} v}} \right)$，可得

$$\begin{aligned}
&\left| v^T \left(\mathop{\mathbb{E}}_{x\sim f_R}\left[M^{1/2}\mathcal{I}_R^{-1} s_R(x+\epsilon) s_R(x+\epsilon)^T \mathcal{I}_R^{-1} M^{1/2} \right] - M^{1/2}\mathcal{I}_R^{-1}M^{1/2} \right) v \right| \\
&\lesssim \sqrt{\epsilon^T R^{-1}\epsilon} \cdot (v^T M^{1/2}\mathcal{I}_R^{-1} M^{1/2} v) \sqrt{\log \frac{v^T M^{1/2}\mathcal{I}_R^{-1} R^{-1} \mathcal{I}_R^{-1} M^{1/2} v}{v^T M^{1/2}\mathcal{I}_R^{-1} M^{1/2} v}}
\end{aligned}$$

由于

$$\frac{v^T M^{1/2}\mathcal{I}_R^{-1} R^{-1} \mathcal{I}_R^{-1} M^{1/2} v}{v^T M^{1/2}\mathcal{I}_R^{-1} M^{1/2} v} \le \sup_{w\in S^{d-1}} \frac{w^T R^{-1} w}{w^T \mathcal{I}_R w}$$

结论得证。 □

B.3. 得分函数的次伽马集中

本节将证明，对得分函数 s_R 施加对称半正定线性变换后，其每个一维投影在适当的方差和尺度参数下均服从次伽马分布。我们首先证明得分函数的雅可比矩阵有界，这一结果将在后续引理中用到。

引理 B.6. 设 $s_R\colon \mathbb{R}^d \to \mathbb{R}^d$ 为分布 f 的 R 平滑版本的得分函数。设 $\boldsymbol{J}_{s_R}$ 为 s_R的雅可比矩阵。则

$$\boldsymbol{J}_{s_R} \succcurlyeq -R^{-1}$$

证明。 对引理 B.1 关于 ϵ取梯度，得到

$$\frac{\nabla f_R(x+\epsilon)}{f_R(x)} = \mathop{\mathbb{E}}_{Z_R|x}\left[e^{\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon} \left(R^{-1}Z_R - R^{-1}\epsilon \right) \right]$$

So,

$$s_R(x+\epsilon) = \frac{\nabla f_R(x+\epsilon)}{f_R(x+\epsilon)} \cdot \frac{f_R(x+\epsilon)}{f_R(x)} = \frac{\mathbb{E}_{Z_R|x}\left[e^{\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon} \left(R^{-1}Z_R - R^{-1}\epsilon \right) \right]}{\mathbb{E}_{Z_R|x}\left[e^{\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon} \right]}$$

现在，令 $\epsilon = \gamma v$对于 $\gamma \in \mathbb{R}$， $\gamma > 0$使得 $\|v\| = 1$。现在， $e^{\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon}$和 $v^T R^{-1} Z_R - v^T R^{-1}\epsilon$关于 $v^T R^{-1} Z_R$单调非递减。因此，根据引理 C.1，它们正相关。也就是说，

$$v^T \mathop{\mathbb{E}}_{Z_R|x}\left[e^{\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon} \left(R^{-1}Z_R - R^{-1}\epsilon \right) \right] \ge \mathop{\mathbb{E}}_{Z_R|x}\left[e^{\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon} \right] \cdot \left(v^T \mathop{\mathbb{E}}_{Z_R|x}\left[R^{-1}Z_R - R^{-1}\epsilon \right] \right)$$

So,

$$v^T s_R(x+\epsilon) \geq v^T \mathop{\mathbb{E}}_{Z_R|x}\left[R^{-1}Z_R - R^{-1}\epsilon\right] \tag{11}$$

现在，根据雅可比矩阵的定义

$$\mathbf{J}_{s_R} v = \left[\frac{\partial}{\partial\gamma} s_R(x+\gamma v)\right]_{\gamma=0}$$

因此，在（11）中，对 γ 求导并令 $\gamma = 0$，得到

$$v^T \mathbf{J}_{s_R} v \geq -v^T R^{-1} v$$

如所需。 □

下面的引理表明，得分 $s_R(x)$ 的每个一维投影都是次伽马随机变量，具有适当的方差和尺度参数。作为推论（推论 B.8），我们得到，得分经对称半正定矩阵变换后的每个一维投影也都是次伽马随机变量，其方差和尺度经过了相应变换。

引理 B.7. 设 $s_R\colon \mathbb{R}^d \to \mathbb{R}^d$ 为平滑分布 f_R 的得分函数，其中平滑尺度为 R，费 *Fisher* 信息矩阵为 $\mathcal{I}_R$。对于任意固定的 $v \in \mathbb{R}^d$，满足 $\|v\| = 1$，有

$$\mathop{\mathbb{E}}_{x\sim f_R}[|v^T R^{1/2} s_R(x)|^k] \leq (1.6)^{k-2} k^{k/2} (v^T R^{1/2} \mathcal{I}_R R^{1/2} v)$$

等价地，对于任意 $v \in \mathbb{R}^d$, $v^T R^{1/2} s_R(x)$ 都是次伽马随机变量。

$$v^T R^{1/2} s_R(x) \in \Gamma(v^T R^{1/2} \mathcal{I}_R R^{1/2} v, 1.6\|v\|)$$

证明。 对于 x，$\gamma \in \mathbb{R}^d$，由引理 B.1以及詹森不等式，

$$f_R(x+\gamma) \geq f_R(x) e^{\gamma^T s_R(x) - \frac{1}{2}\gamma^T R^{-1}\gamma}$$

令 $\gamma = R^{1/2} v$。于是，

$$f_R\left(x + \gamma \cdot \mathrm{sign}(\gamma^T s_R(x))\right) \geq f_R(x) e^{|\gamma^T s_R(x)|}/\sqrt{e}$$

现在，由引理 B.6，可得，

$$\begin{aligned}
\gamma^T s_R(x+\gamma) &= \gamma^T s_R(x) + \gamma^T \mathbf{J}_{s_R}\gamma \\
&= \gamma^T s_R(x) + v^T R^{1/2} \mathbf{J}_{s_R} R^{1/2} v \\
&\geq \gamma^T s_R(x) - 1
\end{aligned}$$

类似地，

$$\gamma^T s_R(x-\gamma) \leq \gamma^T s_R(x) + 1$$

结合这两者，可得

$$|\gamma^T s_R(x+\gamma\cdot\mathrm{sign}(\gamma^T s_R(x))| \geq |\gamma^T s_R(x)| - 1$$

因此，对于任意 $k \geq 2$ 和 $|\gamma^T s_R(x)| > \alpha$，对于 $\alpha := 2 + 1.2\sqrt{k}$

$$\begin{aligned}
&f_R(x+\gamma\cdot\mathrm{sign}(\gamma^T s_R(x)))|\gamma^T s_R(x+\gamma\cdot\mathrm{sign}(\gamma^T s_R(x)))|^k \\
&\geq \frac{1}{\sqrt{e}} f_R(x) e^{|\gamma^T s_R(x)|}\left(|\gamma^T s_R(x)| - 1\right)^k \\
&= f_R(x)|\gamma^T s_R(x)|^k \cdot \left(\frac{1}{\sqrt{e}} e^{|\gamma^T s_R(x)|}\left(1 - \frac{1}{|\gamma^T s_R(x)|}\right)^k\right) \\
&\geq f_R(x)|\gamma^T s_R(x)|^k \cdot \left(\frac{1}{\sqrt{e}} e^{\alpha - 1.4\frac{k}{\alpha}}\right) \\
&\geq f_R(x)|\gamma^T s_R(x)|^k \cdot 4
\end{aligned}$$

因此，

$$f_R(x)|\gamma^T s_R(x)|^k \le \frac{1}{4}\left(f_R(x-\gamma)|\gamma^T s_R(x-\gamma)|^k + f_R(x+\gamma)|\gamma^T s_R(x+\gamma)|^k\right)$$

当 $k \ge 2$ 和 $|\gamma^T s_R(x)| \ge \alpha$时。对其积分，

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim f_R}\left[|\gamma^T s_R(x)|^k\right] &= \int_{[-\infty,\infty]^d} f_R(x)|\gamma^T s_R(x)|^k dx \\
&\le 2\int f_R(x)|\gamma^T s_R(x)|^k - \frac{1}{4}f_R(x-\gamma)|\gamma^T s_R(x-\gamma)|^k - \frac{1}{4}f_R(x+\gamma)|\gamma^T s_R(x+\gamma)|dx \\
&\le 2\int f_R(x)|\gamma^T s_R(x)|^k \mathbb{1}_{|\gamma^T s_R(x)|<\alpha} dx \\
&\le 2\int f_R(x)|\gamma^T s_R(x)|^2 \alpha^{k-2} \mathbb{1}_{|\gamma^T s_R(x)|<\alpha} dx \\
&\le 2\alpha^{k-2}\,\mathbb{E}[|\gamma^T s_R(x)|^2] = 2\alpha^{k-2}\gamma^T \mathcal{I}_R \gamma
\end{aligned}$$

最后，对于任意 $k \ge 2$,

$$2\alpha^{k-2} = 2(1.2\sqrt{k}+2)^{k-2} \le k^{k/2}\cdot 1.6^{k-2}$$

结论得证。 □

推论 B.8. 设 $s_R\colon \mathbb{R}^d \to \mathbb{R}^d$ 为一个 R 平滑分布 f_R 的得分函数，其费 *Fisher* 信息矩阵为$\mathcal{I}_R$。设 $M \in \mathbb{R}^{d\times d}$ 为一个对称矩阵，满足 $M \succcurlyeq 0$。对于任意固定的 $v \in \mathbb{R}^d$，若 $\|v\| = 1$，则有

$$\mathop{\mathbb{E}}_{x\sim f_R}\left[|v^T M^{1/2}\mathcal{I}_R^{-1} s_R(x)|^k\right] \le (1.6\|M^{1/2}\mathcal{I}_R^{-1}R^{-1/2}v\|)^{k-2}k^{k/2}(v^T M^{1/2}\mathcal{I}_R^{-1}M^{1/2}v)$$

等价地， $v^T M^{1/2}\mathcal{I}_R^{-1}s_R(x)$ 是次伽马的。

$$|v^T M^{1/2}\mathcal{I}_R^{-1}s_R(x)| \in \Gamma(v^T M^{1/2}\mathcal{I}_R^{-1}M^{1/2}v, 1.6\|M^{1/2}\mathcal{I}_R^{-1}R^{-1/2}v\|)$$

引理 B.9 和 B.10 接下来证明的辅助引理将用于证明本节的主要结果——引理 B.11，该引理表明，对于 $s_R(x+\epsilon)$，任意一维投影都是次伽马分布。 $x \sim f_R$

引理 B.9。设 $s_R\colon \mathbb{R}^d \to \mathbb{R}^d$ 为一个 R-平滑分布 f_R 的得分函数，其费 *Fisher* 信息矩阵为$\mathcal{I}_R$。对于任意固定的 $v \in \mathbb{R}^d$，若 $\|v\| = 1$、 $x \in \mathbb{R}^d$、 $k \ge 3$，且 $\epsilon \in \mathbb{R}^d$满足 $0 \le \epsilon^T R^{-1}\epsilon \le 1/4$，如果 $v^T R^{1/2}s_R(x+\epsilon) \ge \max(2\sqrt{k}+2, 9.5)$，那么对于 $\gamma = R^{1/2}v$,

$$f_R(x)|\gamma^T s_R(x+\epsilon)|^k \le \frac{1}{5}\max\left(f_R(x-\epsilon)|\gamma^T s_R(x-\epsilon)|^k, f_R(x+\epsilon+\gamma)|\gamma^T s_R(x+\epsilon+\gamma)|^k\right)$$

证明。令 $\alpha := \frac{f_R(x)}{f_R(x+\epsilon)}$。由引理 B.1，可得

$$\alpha = \mathop{\mathbb{E}}_{Z_R|x+\epsilon}\left[e^{-\epsilon^T R^{-1}Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon}\right] \tag{12}$$

令 $\gamma = R^{1/2}v$。我们将考虑两种情况

当 $\log\alpha < \frac{3}{4}\gamma^T s_R(x+\epsilon) - 2$。首先，根据引理 B.1 和詹森不等式，我们有

$$\frac{f_R(x+\epsilon+\gamma)}{f_R(x+\epsilon)} \ge e^{\gamma^T s_R(x+\epsilon)-1/2}$$

此外，根据引理 B.6，我们有

$$\gamma^T s_R(x+\epsilon+\gamma) \ge \gamma^T s_R(x+\epsilon) - 1$$

So,

$$\begin{aligned}
f_R(x+\epsilon+\gamma)|\gamma^T s_R(x+\epsilon+\gamma)|^k &\ge f_R(x+\epsilon)|\gamma^T s_R(x+\epsilon)|^k e^{\gamma^T s_R(x+\epsilon)-\frac{1}{2}}\left(1-\frac{1}{\gamma^T(s_R(x+\epsilon)}\right)^k \\
&\ge f_R(x+\epsilon)|\gamma^T s_R(x+\epsilon)|^k e^{\gamma^T s_R(x+\epsilon)-\frac{k}{\gamma^T s_R(x+\epsilon)-1}-\frac{1}{2}}
\end{aligned}$$

由于 $\gamma^T s_R(x+\epsilon) \geq 2\sqrt{k}+2$,

$$f_R(x+\epsilon+\gamma)|s_R(x+\epsilon+\gamma)|^k \geq f_R(x+\epsilon)|s_R(x+\epsilon)|^k e^{\frac{3}{4}\gamma^T s_R(x+\epsilon)}$$

因此，由于

$$\alpha = \frac{f_R(x)}{f_R(x+\epsilon)} \leq e^{\frac{3}{4}\gamma^T s_R(x+\epsilon)-2}$$

我们有

$$f(x)|s_R(x+\epsilon)|^k = \alpha f_R(x+\epsilon)|s_R(x+\epsilon)|^k \leq \frac{1}{5} f_R(x+\epsilon+\gamma)|s_R(x+\epsilon+\gamma)|^k$$

当 $\log\alpha > \frac{3}{4}\gamma^T s_R(x+\epsilon) - 2$。将（12）代入 $x-\epsilon$ 可得

$$\frac{f_R(x-\epsilon)}{f_R(x)} = \mathop{\mathbb{E}}_{Z_R|x}\left[e^{-\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon}\right]$$

对 ϵ 求梯度，我们有

$$\frac{\nabla f_R(x-\epsilon)}{f_R(x)} = \mathop{\mathbb{E}}_{Z_R|x}\left[R^{-1}(Z_R+\epsilon)e^{-\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon}\right]$$

因此在 $x+\epsilon$ 处求值，

$$\frac{\nabla f_R(x)}{f_R(x+\epsilon)} = \mathop{\mathbb{E}}_{Z_R|x+\epsilon}\left[R^{-1}(Z_R+\epsilon)e^{-\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon}\right]$$

特别是，

$$\epsilon^T\frac{\nabla f_R(x)}{f_R(x+\epsilon)} = \mathop{\mathbb{E}}_{Z_R|x+\epsilon}\left[\epsilon^T R^{-1}(Z_R+\epsilon)e^{-\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon}\right]$$

定义 $y = e^{-\epsilon^T R^{-1} Z_R - \epsilon^T R^{-1}\epsilon}$ 使得 $\mathbb{E}_{Z_R|x+\epsilon}[y] = \alpha e^{-\frac{1}{2}\epsilon^T R^{-1}\epsilon}$，并且

$$\epsilon^T R^{-1}(Z_R+\epsilon)e^{-\epsilon^T R^{-1} Z_R - \frac{1}{2}\epsilon^T R^{-1}\epsilon} = -e^{\frac{1}{2}\epsilon^T R^{-1}\epsilon} y\log y$$

是凹函数，因此根据詹森不等式，

$$\epsilon^T\frac{\nabla f_R(x)}{f_R(x+\epsilon)} \leq -e^{\frac{1}{2}\epsilon^T R^{-1}\epsilon}\left(e^{-\frac{1}{2}\epsilon^T R^{-1}\epsilon}\alpha\right)\log\left(e^{-\frac{1}{2}\epsilon^T R^{-1}\epsilon}\alpha\right) = -\alpha\log\alpha + \frac{1}{2}\alpha\epsilon^T R^{-1}\epsilon$$

So,

$$\epsilon^T s_R(x) = \epsilon^T\frac{\nabla f_R(x)}{f_R(x)} \leq -\log\alpha + \frac{1}{2}\epsilon^T R^{-1}\epsilon$$

最后，我们考虑转移到 $x-\epsilon$。根据引理 B.6，可得

$$\epsilon^T s_R(x-\epsilon) \leq s_R(x) + \epsilon^T R^{-1}\epsilon \leq -\log\alpha + \frac{3}{2}\epsilon^T R^{-1}\epsilon$$

根据引理 B.1,

$$\frac{f_R(x-\epsilon)}{f_R(x+\epsilon)} = \mathop{\mathbb{E}}_{Z_R|x+\epsilon}\left[e^{-2\epsilon^T R^{-1} Z_R - 2\epsilon^T R^{-1}\epsilon}\right] = \mathop{\mathbb{E}}_{Z_R|x+\epsilon}[y^2] \geq \mathop{\mathbb{E}}_{Z_R|x+\epsilon}[y]^2 = \alpha^2 e^{-\epsilon^T R^{-1}\epsilon}$$

由于 $\log\alpha \geq \frac{3}{4}\gamma^T s_R(x+\epsilon) - 2$,

$$-\epsilon^T s_R(x-\epsilon) \geq \frac{3}{4}\gamma^T s_R(x+\epsilon) - 2 - \frac{3}{2}\epsilon^T R^{-1}\epsilon \geq \frac{3}{4}\gamma^T s_R(x+\epsilon) - \frac{19}{8} \geq \gamma^T s_R(x)$$

其中，第二个不等式源于 $\frac{3}{4}\gamma^T s_R(x+\epsilon) - 2 > 0$，因此该函数关于 $\epsilon^T R^{-1}\epsilon$ 单调递减，且 $\epsilon^T R^{-1}\epsilon \leq 1/4$。于是，

$$f_R(x-\epsilon)|\gamma^T s_R(x-\epsilon)|^k \geq \alpha e^{-\epsilon^T R^{-1}\epsilon} f_R(x)|s_R(x+\epsilon)|^k$$

由于我们的假设给出 $\alpha e^{-\epsilon^T R^{-1}\epsilon} \geq 5$，因此得到该结果。 □

引理 B.10。 设 s_R： $\mathbb{R}^d \to \mathbb{R}^d$ 为一个 R-平滑分布 f_R 的得分函数，其费 *Fisher* 信息矩阵为$\mathcal{I}_R$。

对于任意固定的 $v \in \mathbb{R}^d$，若 $\|v\| = 1$、 $x \in \mathbb{R}^d$、 $k \geq 3$ 且 $\epsilon \in \mathbb{R}^d$ 满足 $1/4 \leq \epsilon^T R^{-1} \epsilon \leq 0$，如果 $v^T R^{1/2} s_R(x+\epsilon) \geq \alpha$ 对于 $\alpha = 2 + 1.2\sqrt{k}$ 成立，则对于 $\gamma = R^{1/2} v$，我们有：

$$f_R(x)|\gamma^T s_R(x+\epsilon)|^k \leq \frac{1}{4}\left(f_R(x-\gamma)|\gamma^T s_R(x+\epsilon-\gamma)|^k + f_R(x+\gamma)|s_R(x+\epsilon+\gamma)|^k\right)$$

作为直接 推论，该结论在 $0 \leq \epsilon^T R^{-1} \epsilon \leq 1/4$ 和 $v^T R^{1/2} s_R(x) \leq -\alpha$ 时也成立 .

证明。 由引理 B.1和詹森不等式,

$$f_R(x+\gamma) \geq f_R(x) e^{\gamma^T s_R(x)} / \sqrt{e}$$

由引理 B.6，可得

$$\gamma^T s_R(x+\epsilon+\gamma) \geq \gamma^T s_R(x+\epsilon) - 1$$

由于根据假设，右侧为正，因此我们有

$$|\gamma^T s_R(x+\epsilon+\gamma)| \geq |\gamma^T s_R(x+\epsilon)| - 1$$

现在，当 $\epsilon^T R^{-1} \epsilon < 0$ 时，由引理 B.6 可知，且由于 $|\epsilon^T R^{-1} \epsilon| \leq 1$，因此

$$\gamma^T s_R(x) \geq \gamma^T s_R(x+\epsilon) - 1$$

So,

$$\begin{aligned}
f_R(x+\gamma)|\gamma^T s_R(x+\epsilon+\gamma)|^k &\geq \frac{1}{\sqrt{e}} f_R(x) e^{\gamma^T s_R(x)} \left(|\gamma^T s_R(x+\epsilon)| - 1\right)^k \\
&\geq \frac{1}{\sqrt{e}} f_R(x) e^{\gamma^T s_R(x+\epsilon) - 1} \left(|\gamma^T s_R(x+\epsilon)| - 1\right)^k \\
&\geq f_R(x)|\gamma^T s_R(x+\epsilon)|^k \left(\frac{1}{\sqrt{e}} e^{\gamma^T s_R(x+\epsilon)-1} \left(1 - \frac{1}{|\gamma^T s_R(x+\epsilon)|}\right)^k\right) \\
&\geq f_R(x)|\gamma^T s_R(x+\epsilon)|^k \cdot \left(e^{-3/2} e^{\alpha - 1.4k/\alpha}\right) \\
&\geq f_R(x)|\gamma^T s_R(x+\epsilon)|^k \cdot 4
\end{aligned}$$

□

我们现在可以证明，对于 1 维投影 $s_R(x+\epsilon)$，当 $x \sim f_R$ 时，它具有适当的方差和尺度，是次伽马分布。作为推论（推论 B.12），我们得到，对于 1 维投影 $s_R(x+\epsilon)$，经对称半正定矩阵变换后，也具有适当变换后的方差和尺度，是次伽马分布。

引理 B.11。 设 s_R 为一个 R-平滑分布 f_R 的得分函数，其费 *Fisher* 信息矩阵为 $\mathcal{I}_R$。对于$k \geq 3$ 和满足 $|\epsilon^T R^{-1} \epsilon| \leq 1/4$的 $\epsilon \in \mathbb{R}^d$，对于任意满足 $\|v\| = 1$的 $v \in \mathbb{R}^d$，我们有：

$$\mathop{\mathbb{E}}_{x \sim f_R}\left[|v^T R^{1/2} s_R(x+\epsilon)|^k\right] \leq (15)^{k-2} k^{k/2} \max\left(\mathop{\mathbb{E}}_{x \sim f_R}[v^T R^{1/2} s_R(x+\epsilon) s_R(x+\epsilon)^T R^{1/2} v], v^T R^{1/2} \mathcal{I}_R R^{1/2} v\right)$$

等价地， $v^T R^{1/2} s_R(x+\epsilon)$ 是一个次伽马随机变量。

$$v^T R^{1/2} s_R(x+\epsilon) \in \Gamma\left(\max\left(\mathop{\mathbb{E}}_{x \sim f_R}[v^T R^{1/2} s_R(x+\epsilon) s_R(x+\epsilon)^T R^{1/2} v], v^T R^{1/2} \mathcal{I}_R R^{1/2} v\right), 15\right)$$

证明。 不失一般性，我们只证明 $\epsilon^T R^{-1}\epsilon \geq 0$ 情形。与之前一样，设 $\gamma = R^{1/2}v$。利用引理 B.9 和引理 B.7，我们有

$$\begin{aligned}
&\int_{[-\infty,\infty]^d} f_R(x-\epsilon)|\gamma^T s_R(x)|^k \mathbb{1}_{\gamma^T s_R(x) > \max(2\sqrt{k}+2, 9.5)} dx \\
&\leq \int_{[-\infty,\infty]^d} \frac{1}{5}\max\left(f_R(x-2\epsilon)|\gamma^T s_R(x-2\epsilon)|^k, f_R(x+\gamma)|\gamma^T s_R(x+\gamma)|^k\right) dx \\
&= \frac{2}{5}\mathop{\mathbb{E}}_{x\sim f_R}[|\gamma^T s_R(x)|^k] \\
&\leq \frac{2}{5}(1.6)^{k-2}k^{k/2}(\gamma^T \mathcal{I}_R \gamma)
\end{aligned}$$

T然后，我们可以开始界定引理中的 k^{th} 矩量。利用引理B.10，我们有 ve

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim f_R}\left[|\gamma^T s_R(x+\epsilon)|^k\right] = \int_{[-\infty,\infty]^d} f_R(x-\epsilon)|\gamma^T s_R(x)|^k dx \\
&= 2\int f_R(x-\epsilon)|\gamma^T s_R(x)|^k - \frac{1}{4}f_R(x-\epsilon-\gamma)|\gamma^T s_R(x-\gamma)|^k - \frac{1}{4}f_R(x-\epsilon+\gamma)|\gamma^T s_R(x+\gamma)|^k dx \\
&\leq \int f_R(x-\epsilon)|\gamma^T s_R(x)|^k \mathbb{1}_{\gamma^T s_R(x) \geq -\max(2\sqrt{k}+2, 9.5)} dx
\end{aligned}$$

现在，利用前面的论断，我们得到

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim f_R}\left[|\gamma^T s_R(x+\epsilon)|^k\right] \\
&\leq 2\int f_R(x-\epsilon)|\gamma^T s_R(x)|^k \mathbb{1}_{|\gamma^T s_R(x)| \leq \max(2\sqrt{k}+2, 9.5)} dx + \frac{4}{5}(1.6)^{k-2}k^{k/2}(\gamma^T \mathcal{I}_R \gamma) \\
&\leq 2\int f_R(x-\epsilon)|\gamma^T s_R(x)|^2 (\max(2\sqrt{k}+2, 9.5))^{k-2} \mathbb{1}_{|\gamma^T s_R(x)| \leq \max(2\sqrt{k}+2, 9.5)} dx + \frac{4}{5}(1.6)^{k-2}k^{k/2}(\gamma^T \mathcal{I}_R \gamma) \\
&\leq 2\max(2\sqrt{k}+2, 9.5)^{k-2} \mathop{\mathbb{E}}_{x\sim f_R}[|\gamma^T s_R(x+\epsilon)|^2] + \frac{4}{5}(1.6)^{k-2}k^{k/2}(\gamma^T \mathcal{I}_R \gamma) \\
&\leq 2k^{k/2}(2.5)^{k-2} \mathop{\mathbb{E}}_{x\sim f_R}[|\gamma^T s_R(x+\epsilon)|^2] + \frac{4}{5}(1.6)^{k-2}k^{k/2}(\gamma^T \mathcal{I}_R \gamma) \\
&\leq 3k^{k/2}(2.5)^{k-2}\max(\mathop{\mathbb{E}}_{x\sim f_R}[|\gamma^T s_R(x+\epsilon)|^2], \gamma^T \mathcal{I}_R \gamma) \\
&\leq k^{k/2}(15)^{k-2}\max\left(\mathop{\mathbb{E}}_{x\sim f_R}[\gamma^T s_R(x+\epsilon)s_R(x+\epsilon)^T\gamma], \gamma^T \mathcal{I}_R \gamma\right)
\end{aligned}$$

如所要求的。 □

推论 B.12。设 s_R 为一个 R-平滑分布 f_R 的得分函数，其费 *Fisher* 信息矩阵为 $\mathcal{I}_R$。设$M \in \mathbb{R}^{d\times d}$ 为一个满足 $M \succcurlyeq 0$的对称矩阵。对于满足 $|\epsilon^T R^{-1}\epsilon| \leq 1/4$的 $k \geq 3$ 和 $\epsilon \in \mathbb{R}^d$，任取满足 $\|v\| = 1$的 $v \in \mathbb{R}^d$，我们有：

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim f_R}\left[|v^T M^{1/2}\mathcal{I}_R^{-1}s_R(x+\epsilon)|^k\right] \\
&\leq (15\|M^{1/2}\mathcal{I}_R^{-1}R^{-1/2}v\|)^{k-2}k^{k/2}v^T\left(M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\left(1+O\left(\sqrt{\epsilon^T R^{-1}\epsilon}\sqrt{\log\sup_{w\in S^{d-1}}\frac{w^T R^{-1}w}{w^T\mathcal{I}_R w}}\right)\right)\right)v
\end{aligned}$$

换句话说，

$$M^{1/2}\mathcal{I}_R^{-1}s_R(x+\epsilon) \in \Gamma\left(M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\left(1+O\left(\sqrt{\epsilon^T R^{-1}\epsilon}\sqrt{\log\sup_{w\in S^{d-1}}\frac{w^T R^{-1}w}{w^T\mathcal{I}_R w}}\right)\right), M^{1/2}\mathcal{I}_R^{-1}R^{-1/2}\right)$$

证明。 根据引理,

$$\begin{aligned}&\mathop{\mathbb{E}}_{x\sim f_R}\left[|v^TM^{1/2}\mathcal{I}_R^{-1}s_R(x+\epsilon)|^k\right]\\&\le(15\|M^{1/2}\mathcal{I}_R^{-1}R^{-1/2}v\|)^{k-2}k^{k/2}\\&\max\left(\mathop{\mathbb{E}}_{x\sim f_R}\left[v^TM^{1/2}\mathcal{I}_R^{-1}s_R(x+\epsilon)s_R(x+\epsilon)^T\mathcal{I}_R^{-1}M^{1/2}v\right],v^TM^{1/2}\mathcal{I}_R^{-1}M^{1/2}v\right)\end{aligned}$$

于是，利用引理B.5，即可得证。 □

B.4. 逆向得分的估计

在本节中，我们使用推论 B.12中关于 (s_R) 的 $x+\epsilon$ 维投影 $x\sim f_R$ 的次伽马界，以及定理 5.1中关于次伽马向量的范数集中界，来建立 在 $x+\epsilon$ 处的逆向经验得分与其期望之间偏差的界。

引理 B.13。 设 f 是 $\mathbb{R}^d$ 上的任意分布， f_R 是 f 的 R-平滑版本。设 $\mathcal{I}_R$ 是 f_R 的费 *Fisher* 信息矩阵。设 $\epsilon\in\mathbb{R}^d$ 满足 $\epsilon^TR^{-1}\epsilon\le 1/4$。考虑分布参数族$f_R^\lambda(x)=f_R(x-\lambda)$。假设我们有 n 个独立同分布样本 x_1、... ...、 $x_n\sim f_R^\lambda$。设 $M\in\mathbb{R}^{d\times d}$ 是满足 $M\succcurlyeq 0$ 的对称矩阵。令 $\hat{}\epsilon=\frac{1}{n}\sum_{i=1}^n\mathcal{I}_R^{-1}s_R(x_i-\lambda-\epsilon)$。令

$$T:=M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\left(1+O\left(\sqrt{\epsilon^TR^{-1}\epsilon}\sqrt{\log\sup_{w\in S^{d-1}}\frac{w^TR^{-1}w}{w^T\mathcal{I}_Rw}}\right)\right)$$

那么，以概率 $1-\delta$，我们有

$$\begin{aligned}&\|\hat{\epsilon}-\mathop{\mathbb{E}}_{x\sim f_R}[\mathcal{I}_R^{-1}s_R(x-\epsilon)]\|_M\\&\le\sqrt{\frac{\mathrm{Tr}(T)}{n}}+4\sqrt{\frac{\|T\|\log\frac{2}{\delta}}{n}}+16\frac{\|M^{1/2}\mathcal{I}_R^{-1}R^{-1/2}\|\log\frac{2}{\delta}}{n}+8\frac{\|M^{1/2}\mathcal{I}_R^{-1}R^{-1/2}\|_F^2}{n^{3/2}\sqrt{\mathrm{Tr}(T)}}\log\frac{2}{\delta}\end{aligned}$$

证明。 根据推论 B.12， $M^{1/2}\mathcal{I}_R^{-1}s_R(x)$ 是 $(T,\ M^{1/2}\mathcal{I}_R^{-1}R^{-1/2})$-次伽马分布。然后，应用定理5.1中的次伽马范数集中界可得

$$\begin{aligned}&\|\hat{\epsilon}-\mathop{\mathbb{E}}_{x\sim f_R}\left[\mathcal{I}_R^{-1}s_R(x-\epsilon)\right]\|_M\\&=\left\|M^{1/2}\left(\frac{1}{n}\sum_{i=1}^n\mathcal{I}_R^{-1}s_R(x_i-\lambda-\epsilon)\right)-M^{1/2}\mathop{\mathbb{E}}_{x\sim f_R}\left[\mathcal{I}_R^{-1}s_R(x-\epsilon)\right]\right\|\\&=\left\|\left(\frac{1}{n}\sum_{i=1}^nM^{1/2}\mathcal{I}_R^{-1}s_R(x_i-\lambda-\epsilon)\right)-\mathop{\mathbb{E}}_{x\sim f_R}\left[M^{1/2}\mathcal{I}_R^{-1}s_R(x-\epsilon)\right]\right\|\\&\le\sqrt{\frac{\mathrm{Tr}(T)}{n}}+4\sqrt{\frac{\|T\|\log\frac{2}{\delta}}{n}}+16\frac{\|M^{1/2}\mathcal{I}_R^{-1}R^{-1/2}\|\log\frac{2}{\delta}}{n}+8\frac{\|M^{1/2}\mathcal{I}_R^{-1}R^{-1/2}\|_F^2}{n^{3/2}\sqrt{\mathrm{Tr}(T)}}\log\frac{2}{\delta}\end{aligned}$$

□

B.5. 局部最大似然估计

本节将展示，在给定来自 f_λ 的样本以及一个保证包含 λ 的初始不确定区域 S 时，如何以取决于 $\mathcal{I}_R$ 的速率估计我们的位置 λ。

算法 3高维局部最大似然估计

输入参数：

• 分布 f 在 $\mathbb{R}^d$ 上的描述，平滑 R，样本 x_1、……、$x_n \overset{i.i.d.}{\sim} f^\lambda$，以及初始估计 λ_1

1. 令 $\mathcal{I}_R$ 为 f_R 的费 Fisher 信息矩阵，即 f 的 R-平滑版本。令 s_R 为 f_R 的得分函数。

2. 对每个样本 x_i，计算一个扰动后的样本 $x_i' = x_i + \mathcal{N}(0, R)$，其中所有高斯噪声均在各个样本之间独立抽取。

3. 令 $\hat{\epsilon} = \frac{1}{n}\sum_{i=1}^n \mathcal{I}_R^{-1} s_R(x_i' - \lambda_1)$ 并返回 $\hat{\lambda} = \lambda_1 - \hat{\epsilon}$。

引理 *B.14*（局部最大似然估计）。设我们已知模型 f *on* $\mathbb{R}^d$，且 f_R 是 f 的 R-平滑版本，*for* $R = r^2 I_d$，其中标量为 $r > 0$。设 f_R 的费 *Fisher* 信息矩阵为 $\mathcal{I}_R$。此外，设未知的真实参数为 λ，并且我们可以获得初始估计 $\lambda_1 = \lambda + \epsilon$，且保证 $\epsilon^T R^{-1} \epsilon \le \tau$ 对于 $\tau \le 1/4$ 成立。设存在一个足够大的参数 $\gamma \ge 1$，使得 $\tau \le \frac{1}{\gamma^2 \log^2 \frac{\|\mathcal{I}_R^{-1}\|}{r^2}}$。此外，设 $r^2 \ge 4\gamma^2 \|\mathcal{I}_R^{-1}\| \log \frac{2}{\delta}$。那么，对于从 f^λ 抽取的 n 个样本，以概率 $1-\delta$，算法 *3* 的输出满足

$$\|\hat{\lambda} - \lambda\|_M \le \left(1 + O\left(\frac{1}{\gamma}\right)\right)\left(\sqrt{\frac{\operatorname{Tr}(M^{1/2}\mathcal{I}_R^{-1}M^{1/2})}{n}} + 4\sqrt{\frac{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\| \log \frac{2}{\delta}}{n}}\right) + O\left(\tau\sqrt{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|}\right)$$

证明。根据 关于 $\lambda_1 = \lambda + \epsilon$，我们有 $\epsilon^T R^{-1} \epsilon \le \tau$。令 T如引理 B.13 中所定义。没有 w

$$\sup_{w \in S^{d-1}} \frac{w^T R^{-1} w}{w^T \mathcal{I}_R w} = \frac{\|\mathcal{I}_R^{-1}\|}{r^2}$$

因此，由于 $\tau \le \frac{1}{\gamma^2 \text{对数}^2 \frac{\|\mathcal{I}_R^{-1}\|}{r^2}}$，

$$\sqrt{\tau} \log\left(\sup_{w \sim S^{d-1}} \frac{w^T R^{-1} w}{w^T \mathcal{I}_R w}\right) \le \frac{1}{\gamma}$$

因此，我们有

$$\operatorname{Tr}(T) \le \operatorname{Tr}(M^{1/2}\mathcal{I}_R^{-1}M^{1/2})\left(1 + \frac{1}{\gamma}\right)$$

and

$$\|T\| \le \|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|\left(1 + \frac{1}{\gamma}\right)$$

因此，由引理B.13

$$\begin{aligned}
&\|\hat{\epsilon} - \mathop{\mathbb{E}}_{x \sim f_R}[\mathcal{I}_R^{-1} s_R(x - \epsilon)]\|_M \\
&\le \left(1 + O\left(\frac{1}{\gamma}\right)\right)\left(\sqrt{\frac{\operatorname{Tr}(M^{1/2}\mathcal{I}_R^{-1}M^{1/2})}{n}} + 4\sqrt{\frac{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\| \log \frac{2}{\delta}}{n}} + \frac{8\|M^{1/2}\mathcal{I}_R^{-1}\|_F^2}{r^2 n^{3/2}\sqrt{\operatorname{Tr}(M^{1/2}\mathcal{I}_R^{-1}M^{1/2})}} \log \frac{2}{\delta}\right) \\
&+ 16\frac{\|M^{1/2}\mathcal{I}_R^{-1}\| \log \frac{2}{\delta}}{rn}
\end{aligned}$$

由于 $r^2 \geq 4\gamma^2\|\mathcal{I}_R^{-1}\|\log\frac{2}{\delta}/n$，$\frac{1}{r} \leq \left(\frac{n}{\log\frac{2}{\delta}}\right)^{1/2}\frac{1}{2\gamma\sqrt{\|\mathcal{I}_R^{-1}\|}}$。因此，

$$16\frac{\|M^{1/2}\mathcal{I}_R^{-1}\|\log\frac{2}{\delta}}{rn} \leq \frac{8\|M^{1/2}\mathcal{I}_R^{-1}\|}{\gamma\sqrt{\|\mathcal{I}_R^{-1}\|}}\left(\frac{\log\frac{2}{\delta}}{n}\right)^{1/2}$$
$$\leq \frac{8}{\gamma}\sqrt{\frac{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|\log\frac{2}{\delta}}{n}}$$
$$\text{since } \frac{(\|M^{1/2}\mathcal{I}_R^{-1}\|)^2}{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|} = \left(\frac{\|M^{1/2}\mathcal{I}_R^{-1}\|}{\|M^{1/2}\mathcal{I}_R^{-1/2}\|}\right)^2 \leq \|\mathcal{I}_R^{-1}\|$$

同理，$\frac{1}{r^2} \leq \frac{n}{4\gamma^2\|\mathcal{I}_R^{-1}\|\log\frac{2}{\delta}}$。因此，

$$\frac{8\|M^{1/2}\mathcal{I}_R^{-1}\|_F^2}{r^2n^{3/2}\sqrt{\mathrm{Tr}(M^{1/2}\mathcal{I}_R^{-1}M^{1/2})}}\log\frac{2}{\delta} \leq \frac{8\,\mathrm{Tr}(M\mathcal{I}_R^{-2})}{r^2n^{3/2}\sqrt{\mathrm{Tr}(M\mathcal{I}_R^{-1})}}\log\frac{2}{\delta}$$
$$\leq \frac{2\,\mathrm{Tr}(M\mathcal{I}_R^{-2})}{\gamma\|\mathcal{I}_R^{-1}\|n\sqrt{\mathrm{Tr}(M\mathcal{I}_R^{-1})}}\sqrt{\log\frac{2}{\delta}}$$
$$\leq \frac{8}{\gamma}\sqrt{\frac{\mathrm{Tr}(M^{1/2}\mathcal{I}_R^{-1}M^{1/2})}{n}} \quad \text{using Lemma C.3}$$

因此，我们有

$$\left\|\hat{\epsilon} - \mathop{\mathbb{E}}_{x\sim f_R}\left[\mathcal{I}_R^{-1}s_R(x-\epsilon)\right]\right\|_M \leq \left(1+O\left(\frac{1}{\gamma}\right)\right)\left(\sqrt{\frac{\mathrm{Tr}(M^{1/2}\mathcal{I}_R^{-1}M^{1/2})}{n}} + \sqrt{\frac{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|\log\frac{2}{\delta}}{n}}\right)$$

现在，利用引理 B.3

$$\left\|\epsilon - \mathop{\mathbb{E}}_{x\sim f_R}\left[\mathcal{I}_R^{-1}s_R(x-\epsilon)\right]\right\|_M \lesssim \sqrt{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|}(\epsilon^T R^{-1}\epsilon) \leq \tau\sqrt{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|}$$

因此，我们有

$$\|\hat{\epsilon}-\epsilon\|_M \leq \|\hat{\epsilon} - \mathop{\mathbb{E}}_{x\sim f_R}\left[\mathcal{I}_R^{-1}s_R(x-\epsilon)\right]\|_M + \|\epsilon - \mathop{\mathbb{E}}_{x\sim f_R}\left[\mathcal{I}_R^{-1}s_R(x-\epsilon)\right]\|_M$$
$$\leq \left(1+O\left(\frac{1}{\gamma}\right)\right)\left(\sqrt{\frac{\mathrm{Tr}(M^{1/2}\mathcal{I}_R^{-1}M^{1/2})}{n}} + 4\sqrt{\frac{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|\log\frac{2}{\delta}}{n}}\right)$$
$$+ O\left(\tau\sqrt{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|}\right)$$

^^ 现在，由于 $\lambda = \lambda_1 - \hat{\epsilon}$ 和 $\lambda = \lambda_1 - \epsilon, \lambda - \lambda = \hat{\epsilon} - \epsilon$。结论成立。 □

B.6. 全局最大似然估计

本节陈述并证明我们的主定理，该定理展示了如何根据来自 f^λ 的 n 个样本，估计依赖于$\mathcal{I}_R$的位置 λ on 率。

我们首先陈述一项重尾估计领域的结果，并将利用它生成一个初始估计 $\lambda+\epsilon$。随后，我们将应用上一节的结果来改进此估计，从而得到最终估计。

定理 *B.15*（（*Hopkins*，*2018*；*Cherapanamjeri* 等人，*2019*））。存在普适常数 C_0、C_1、C_2，使得对于任意 n、$d \in \mathbb{N}$ 和 $\delta > 2^{-n/C_2}$，都存在一个运行时间为 $O(nd) + (d\log(1/\delta))^{C_0}$ 的算法；对于任意定义在 $\mathbb{R}^d$ 上的随机变量 X，给定其独立同分布副本 X_1、……、X_n，该算法输出一个向量 $\hat{\mu}_\delta(X_1, \ldots\ldots, X_n)$，使得

$$\mathbb{P}\left[\|\mu - \hat{\mu}_\delta\| > C_1\left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} + \sqrt{\frac{\|\Sigma\|\log(1/\delta)}{n}}\right)\right] \le \delta$$

其中 $\mathbb{E}[X] = \mu$ 和 $\mathbb{E}\left[(X-\mu)(X-\mu)^T\right] = \Sigma$

算法 4高维全局最大似然估计

输入参数：

- 失败概率 δ、分布描述 f，n 个来自 f^λ 的样本、平滑 R、近似参数 γ

1. 设 Σ 为 f的协方差矩阵。使用定理 λ_1，利用前 $1/\gamma$ 比例的 n 个样本，采用定理B.15中的估计量计算初始估计。

2. 使用剩余的 算法 3 样本比例 $1 - 1/\gamma$，结合 R-平滑和初始估计 λ_1，$\hat{}$返回最终估计 λ。

定理 *B.16*（全局最大似然估计）。设 f 是定义在 $\mathbb{R}^d$ 上的给定模型，且我们已获得来自 f^λ 的 n 个样本，用于估计未知的 λ。设 $R = r^2 I_d$ 满足 $0 < r^2 < \|\Sigma\|$，使得 $\mathcal{I}_R$ 是 f 的 R-平滑费希尔信息矩阵；并设 Σ 是 f 的协方差。设 $M \in \mathbb{R}^{d\times d}$ 是任意一个满足 $M \succcurlyeq 0$ 的对称矩阵，并令 $d_R := d_{\mathit{eff}}(M^{1/2}\mathcal{I}_R^{-1}M^{1/2})$。固定失败概率 $\delta > 0$，并令 $2 \le \gamma \le \binom{n}{d_R+}$ 为某个 $\alpha \ge 0$ 的 $\log \frac{1}{\delta}$ 。令 $n \ge C\gamma^4(\frac{\|\Sigma\|}{r^2})^2$ 为大于某个足够大的常数 $C > 0$ 时的 $\log \frac{4}{\delta} + d_R + \binom{d_{\mathit{eff}}(\Sigma^2)}{d_R}$。那么，以概率 $1-\delta$，算法 *4* 的输出 λ 满足

$$\|\hat{\lambda} - \lambda\|_M \le \left(1 + O\left(\frac{1}{\gamma}\right)\right)\left(\sqrt{\frac{\mathrm{Tr}(M^{1/2}\mathcal{I}_R^{-1}M^{1/2})}{n}} + 4\sqrt{\frac{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|\log\frac{4}{\delta}}{n}}\right)$$

证明。 根据定理 B.15给出的保证，步骤 1 得到的初始估计 $\lambda_1 = \lambda + \epsilon$ 具有如下性质：以概率 $1 - \delta/2$,

$$\|\epsilon\|^2 \lesssim \frac{\mathrm{Tr}(\Sigma)}{n/\gamma} + \frac{\|\Sigma\|\log\frac{2}{\delta}}{n/\gamma}$$

我们以步骤 1 成功为条件。令 $n' = n(1-1/\gamma) \ge n/2$ 表示步骤 2 中用于调用局部最大似然估计算法 3的样本数。根据我们对 n 的下界,

$$r^2 \ge \sqrt{C}\gamma^2\frac{\|\Sigma\|}{\sqrt{n}}\left(\frac{d_{\mathrm{eff}}(\Sigma)}{\sqrt{d_R}} + d_R + \sqrt{\log\frac{4}{\delta}}\right) \ge \sqrt{C}\gamma^2\frac{\mathrm{Tr}(\Sigma) + \|\Sigma\|\log\frac{4}{\delta}}{\sqrt{2n'}} \cdot \frac{(n')^{1/2}}{\sqrt{d_R + \log\frac{4}{\delta}}}$$

因此，当 C 足够大时，由于 $\gamma > 1$，令

$$\tau = \frac{1}{\gamma}\sqrt{\frac{d_R + \log\frac{4}{\delta}}{n'}}$$

可得

$$\epsilon^T R^{-1}\epsilon = \frac{\|\epsilon\|^2}{r^2} \le \tau$$

此外 $\tau \le 1/4$ 由于 $n' \ge \frac{C}{2}(\log\frac{4}{\delta} + d_R)$。因此，置信集 S 满足引理 B.14 中的条件。

根据对 n的约束，我们有

$$r^2 \geq \sqrt{C}\gamma^2\|\Sigma\|\sqrt{\frac{\log\frac{4}{\delta}+d_R+\left(\frac{d_{\text{eff}}(\Sigma)^2}{d_R}\right)}{n}} \geq \sqrt{\frac{C}{2}}\gamma^3\|\Sigma\|\tau$$

根据引理C.2，我们还可得 $\|\mathcal{I}_R^{-1}\|/\|\Sigma\| \leq \frac{\|\Sigma+R\|}{\|\Sigma\|} \leq 2$，因此

$$\log\frac{\|\mathcal{I}_R^{-1}\|}{r^2} \leq \log\frac{2\sqrt{2}}{\sqrt{C}\gamma^3\tau}$$

因此， τ约束为

$$\gamma^2\tau\log^2\frac{2\sqrt{2}}{\sqrt{C}\gamma^3\tau} \leq 1$$

左侧至多为

$$O(\tau^{0.99}\gamma^{2.01}) < 1$$

因为 $\tau < 1/\gamma^5$，其中的常数可以任意小，只要 C。因此，引理 τ的约束B.14 得以满足。利用 $n' \geq \frac{C}{2}(\log\frac{4}{\delta})$这一事实，

$$\begin{aligned}
\frac{r^2n'}{\|\mathcal{I}_R^{-1}\|\log\frac{4}{\delta}} &\geq \frac{C}{2}\gamma^2\frac{\|\Sigma\|\sqrt{n}}{\|\mathcal{I}_R^{-1}\|\log\frac{4}{\delta}}\left(\frac{d_{\text{eff}}(\Sigma)}{\sqrt{d_R}}+\sqrt{\log\frac{4}{\delta}}\right)\\
&\geq \frac{C}{2}\gamma^2\frac{\|\Sigma\|\sqrt{n}}{\|\Sigma+R\|\log\frac{4}{\delta}}\sqrt{\log\frac{4}{\delta}} \quad \text{by Lemma C.2}\\
&\geq \gamma^2 \quad \text{since } R=r^2I_d \text{ so that } \|R\|=r^2<\|\Sigma\|
\end{aligned}$$

因此，引理B.14 的条件均得到满足，并且以 $1-\delta/2$的概率，

$$\begin{aligned}
\|\hat{\lambda}-\lambda\|_M &\leq \left(1+O\left(\frac{1}{\gamma}\right)\right)\left(\sqrt{\frac{\text{Tr}(M^{1/2}\mathcal{I}_R^{-1}M^{1/2})}{n'}}+4\sqrt{\frac{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|\log\frac{4}{\delta}}{n'}}\right)\\
&\quad+O\left(\tau\sqrt{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|}\right)\\
&\leq \left(1+O\left(\frac{1}{\gamma}\right)\right)\left(\sqrt{\frac{\text{Tr}(M^{1/2}\mathcal{I}_R^{-1}M^{1/2})}{n}}+4\sqrt{\frac{\|M^{1/2}\mathcal{I}_R^{-1}M^{1/2}\|\log\frac{4}{\delta}}{n}}\right)
\end{aligned}$$

由于 $n'=n(1-1/\gamma)$ 且 $\tau=\frac{1}{\gamma}\sqrt{\frac{d_R+\log\frac{4}{\delta}}{n'}}$。因此，我们的总失败概率为 δ。结论得证。 □

定理 B.17 (全局最大似然估计，非正式表述)。设 f 的协方差矩阵为 Σ。对于任意 $r^2 \leq \|\Sigma\|$，令 $R=r^2I_d$ 和 $\mathcal{I}_R$ 为该分布的 R-平滑费希尔信息。对于任意常数 $0<\epsilon<1$，

$$\|\widehat{\lambda}-\lambda\|_2 \leq (1+\epsilon)\sqrt{\frac{\text{Tr}(\mathcal{I}_R^{-1})}{n}}+5\sqrt{\frac{\|\mathcal{I}_R^{-1}\|\log\frac{4}{\delta}}{n}}$$

以概率 $1-\delta$，对于 $n > O_\epsilon\left(\left(\frac{\|\Sigma\|}{r^2}\right)^2\left(\log\frac{2}{\delta}+d_{\text{eff}}(\mathcal{I}_R^{-1})+\frac{d_{\text{eff}}(\Sigma^2)}{d_{\text{eff}}\mathcal{I}_R^{-1}}\right)\right)$。

证明。首先，若 $\epsilon > 1/4$，我们重新设定 $\epsilon = 1/4$。令 $M=I_d$ 使得 $d_R = d_{\text{eff}}(\mathcal{I}_R^{-1})$，并令 $\gamma = \frac{C_0}{\epsilon}$取足够大的常数 C_0，则定理 B.16 可得所述结论。 □

C. 有用结果

下面给出重排不等式的一个连续版本（user940，2015年）：

引理 C.1. 设 f、$g\colon \mathbb{R}\to\mathbb{R}$ 为单调非减函数，且 X 是定义在 $\mathbb{R}$上的随机变量。那么

$$\mathbb{E}[f(X)]\,\mathbb{E}[g(X)] \le \mathbb{E}[f(X)g(X)]$$

证明. 设 Y是 X的一个独立副本。由单调性可得,

$$(f(X)-f(Y))(g(X)-g(Y)) \ge 0$$

恒成立。对两边取期望,

$$2\,\mathbb{E}[f(X)g(X)] - 2\,\mathbb{E}[f(X)g(Y)] \ge 0.$$

由于 Y与 X相互独立，因此得到所需结果。 □

引理 C.2. 设 f 是 $\mathbb{R}^d$ 上的任意分布， Σ 是其协方差矩阵。设 f_R 是 f 的 R-平滑版本，其费 *Fisher* 信息矩阵为 $\mathcal{I}_R$。那么，

$$\mathcal{I}_R \succcurlyeq (\Sigma+R)^{-1}$$

证明。 由此可得 事实可知， f_R 的协方差为 $\Sigma+R$，再利用（Hendeby，2005年）中的定理 1.2 . □

引理 C.3. 设 A、B 是对称半正定矩阵。那么

$$Tr(AB) \le Tr(A)\|B\|$$

证明。 设 B 的特征向量为 v_1、.....、 v_d。那么

$$\operatorname{Tr}(AB) = \sum_{i=1}^{d} v_i^T A(Bv_i) \le \|B\| \sum_{i=1}^{d} v_i^T A v_i = \|B\|\operatorname{Tr}(A).$$

□