arXiv:2603.03673v1 [cs.LG] 4 Mar 2026

具有有界支撑的 q-高斯分布的 Stein 恒等式

Sophia Sklaviadis[*,†,1], Thomas Möllenhoff[*,2], André F. T. Martins[†,3], MárioA. T.Figueiredo[†,4], Mohammad Emtiyaz Khan[*,5]

[*]理化学研究所人工智能项目中心，东京，日本
[†]电信研究所，高等技术学院，里斯本大学，葡萄牙

[1]ssklaviadis@gmail.com，[2]thomas.moellenhoff@riken.jp,[3] andre.t.martins@tecnico.ulisboa.pt，[4]mario.figueiredo@tecnico.ulisboa.pt,[5] emtiyaz.khan@riken.jp

Stein 恒等式是机器学习中的一项基础工具，可应用于生成模型、随机优化，以及其他涉及高斯分布下期望梯度的问题。相比之下，非高斯期望问题受到的关注较少。本文研究有界支撑 q-高斯分布这一类分布，并推导出一种新的 Stein 恒等式，由此得到的梯度估计量形式与高斯分布的梯度估计量几乎相同，而且实现起来同样简单。为此，我们扩展 Landsman、Vanduffel 和 Yao（2013）先前的研究结果，证明了新的 q-高斯分布 Bonnet 型和 Price 型定理。我们还利用伴随分布 简化了这些定理的形式。实验表明，有界支撑分布可以降低梯度估计量的方差，这可能对贝叶斯深度学习和锐度感知最小化有所帮助。总的来说，我们的工作简化了 Stein 恒等式在一类重要非高斯分布中的应用。

1. 引言

Stein 恒等式在机器学习中广受欢迎，可用于估计 $\mathbb{E}_p[f(\mathbf{x})]$,，其中实值可微函数 $f\colon \mathbb{R}^D \to \mathbb{R}$的期望是相对于高斯密度 $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ 计算的，其均值为 $\boldsymbol{\mu}$，协方差为 $\boldsymbol{\Sigma}$。其应用广泛，涵盖随机优化、 [1], 深度生成模型 [2], 和变分推断 [3, 4]等各类机器学习问题。Stein 恒等式指出，以下等式成立：

$$\mathbb{E}_p\left[(\mathbf{x}-\boldsymbol{\mu})f(\mathbf{x})\right] = \mathrm{Cov}_p(\mathbf{x})\,\mathbb{E}_p\left[\nabla_{\mathbf{x}} f(\mathbf{x})\right]. \tag{1}$$

该恒等式最早由 Charles Stein 证明 [5, 6], ，源自分部积分，并在关于 f 的温和条件下成立（参见 [7, 附录 A2]）。

Stein 恒等式可用于将关于 $\boldsymbol{\mu}$和 $\boldsymbol{\Sigma}$的梯度分别表示为 $f(\mathbf{x})$的梯度和 Hessian矩阵（推导见附录 A）：

$$\nabla_{\boldsymbol{\mu}}\,\mathbb{E}_p\left[f(\mathbf{x})\right] = \mathbb{E}_p\left[\nabla_{\mathbf{x}} f(\mathbf{x})\right], \qquad \nabla_{\boldsymbol{\Sigma}}\,\mathbb{E}_p\left[f(\mathbf{x})\right] = \tfrac{1}{2}\mathbb{E}_p\left[\nabla^2_{\mathbf{x}} f(\mathbf{x})\right]. \tag{2}$$

这些表达式分别称为 Bonnet 定理 [8] 和 Price 定理 [9] ，因形式简洁而在深度学习中受到广泛关注。关于 $\boldsymbol{\mu}$ 和$\boldsymbol{\Sigma}$ 的随机梯度，可通过先采样$\mathbf{x} \sim p(\mathbf{x})$，再在该样本处计算 $\nabla_{\mathbf{x}} f(\mathbf{x})$和 $\nabla^2_{\mathbf{x}} f(\mathbf{x})$来得到。这种思路催生了“重参数化技巧”的若干变体以及其他路径梯度估计器，这些方法的方差通常低于得分函数估计器 [10]。总的来说，Stein 恒等式对随机梯度估计技术产生了深远影响 [11]。

将 Stein 恒等式推广到非高斯分布所受到的关注远少于高斯情形。推广确实存在，例如推广到椭圆分布族，其中包含许多

第三届简约与学习会议（CPAL 2026）。

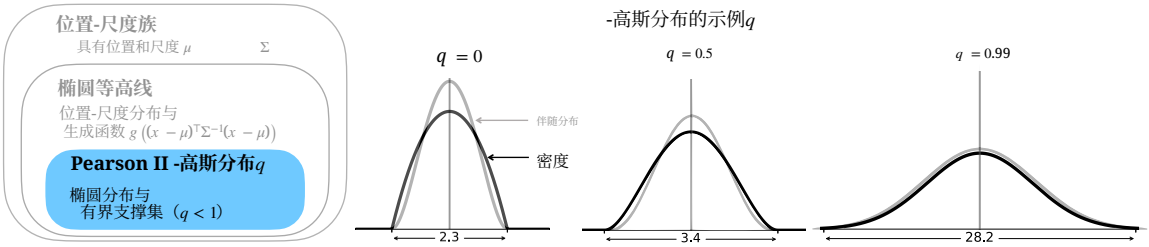

图 1：有界支撑的 q-高斯分布是椭圆等高线位置-尺度族的一个子类，也就是说，它们是通过将二次型 $s(\mathbf{x}) = (\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ 与生成函数 g 复合而得到的，因此 $p(\mathbf{x}) = g\big(s(\mathbf{x})\big)$。右侧展示了 q-高斯分布的三个示例，其参数分别为 $q = 0, 0.5,$ 和 0.99。可以看到，q 越大，基准密度（黑色曲线）的峰值越低，支撑集越大。我们还展示了相应的第一阶伴随分布 $2-q$-高斯密度（灰色曲线），它们的峰值略高于基准密度。当 $q \to 1$时，q-高斯分布收敛于高斯分布。

有趣的位置-尺度族，包括高斯分布。特别是，Landsman 和 Nešlehová [12] 以及 Landsman 等人 [13, 14] 研究了此类推广，重点关注 Pearson VII 类，其中包含各种重尾分布。据我们所知，此类推广对梯度估计的用处尚未得到探讨。针对非高斯分布族导出的梯度估计量能否同样具有简单形式，并且同样易于实现？本文的目标是回答这些问题。

本文为 Pearson II 类椭圆分布族推导了一个新的 Stein 恒等式，称为有界支撑 q-高斯分布（图 1）。这类有界支撑分布对梯度估计很有意义，因为与高斯分布不同，从中抽取的样本始终落在一个有界区间内，这也自然意味着梯度方差有界。我们为这类 q-高斯分布推导了 Bonnet 型和 Price 型定理，并表明其梯度估计量的形式与高斯随机梯度的几乎完全相同，实现起来也同样简单。这意味着，我们只需在样本 $f(\mathbf{x})$ 处计算 $\mathbf{x} \sim p(\mathbf{x})$的梯度和 Hessian 矩阵，即可得到无偏梯度估计。这些梯度表达式是通过使用一种与基准分布相关的分布得出的；这种分布在信息几何和统计物理中称为 [15–17]，即 伴随分布（图 1）。我们通过数值实验验证了有界方差性质，并将其与贝叶斯深度学习方法和锐度感知最小化进行比较 [18]。总体而言，我们的工作简化了将 Stein 恒等式应用于涉及一类重要有界支撑 q-高斯分布的梯度估计。

2. 有界支撑集 q-高斯分布

我们提出了一种新的 Stein 恒等式，并将其应用于涉及有界支撑集 q-高斯分布的梯度估计。这类分布是位置-尺度分布的一种特殊类型，可通过将特定的 生成函数 函数 g 应用于类似高斯的二次型而得到。我们将位置记为 $\boldsymbol{\mu}$，将椭圆尺度矩阵记为 $\boldsymbol{\Sigma}$。D元椭圆密度的形式为

$$p(\mathbf{x}) = |\boldsymbol{\Sigma}|^{-1/2} g\big(s(\mathbf{x})\big), \quad \text{where } s(\mathbf{x}) = (\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}), \tag{3}$$

且 $g: \mathbb{R} \to \mathbb{R}_+$ 是非负的密度生成函数，经过适当缩放，使得 $p(\mathbf{x})$ 归一化，也就是说，其积分为 1。例如，将 $g(s)$ 设为 $\propto \exp(-s/2)$，即可得到高斯分布。归一化条件可写为

$$\int_0^\infty s^{D/2-1} g(s)\, ds = \frac{\Gamma(D/2)}{\pi^{D/2}},$$

其中 Γ 是伽马函数。如果二阶矩存在，则协方差矩阵 $\mathrm{Cov}(\mathbf{x})$ 与 $\boldsymbol{\Sigma}$ 成正比。在高斯情形下，$\mathrm{Cov}(\mathbf{x}) = \boldsymbol{\Sigma}$，但在其他情况下，应将 $\boldsymbol{\Sigma}$ 视为尺度或离散参数，它与协方差成正比。

有界支撑集 q-高斯分布属于一种称为 II 型皮尔逊分布的分布类 [19]，，定义在半径为R 的椭球体 $\{\mathbf{x} \in \mathbb{R}^D : (\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) < R^2\}$上，其中 $R>0$。给定 R，对于区间 s 中的所有 $0 \le s < R^2$，生成函数具有以下形式：

$$g(s) = \frac{Z}{R^{D+2m}}\left(R^2 - s\right)^m, \quad \text{where } Z = \frac{1}{\pi^{D/2}}\frac{\Gamma(\frac{D}{2}+m+1)}{\Gamma(m+1)}, \tag{4}$$

和 m 是生成函数的形状参数。我们用另一个标量 $m = 1/(1-q)$ 来定义$q<1$，由此得到 q-高斯这一称谓，该称谓在信息几何和统计物理学中广为人知 [15–17]。最大支撑半径 R 仅取决于 q 和维数 D，如下引理给出了 $p(\mathbf{x})$ (证明见附录 B)。

引理 1。 具有 $q<1$ 的广义 *II* 型皮尔逊子族的密度可写成 q-高斯分布：

$$p(\mathbf{x}) = \mathcal{N}_q(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \exp_q\left[\tfrac{1}{2}|\boldsymbol{\Sigma}|^{-\frac{1}{2m}}\left(R^2 - (\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) - m\right], \tag{5}$$

其中，$\exp_q(t) := [1+t/m]_+^m$ 表示 q-变形指数函数 [20]，，其中 $m = 1/(1-q)$ 且$[u]_+ = \max(0,\ u)$。支撑半径 R 是 q 和 D 的函数，定义为

$$R^2 = \left[(2m)^m Z\right]^{2/(2m+D)}. \tag{6}$$

Martins 等人 [21] 最近将这些分布作为由 Tsallis 熵 [22] 诱导的稀疏连续分布进行了研究。利用 $\exp_q$ 的定义，同一密度可简洁地表示为

$$p(\mathbf{x}) \propto |\boldsymbol{\Sigma}|^{-1/2}\left(R^2 - s(\mathbf{x})\right)_+^m, \tag{7}$$

这将在后文中用于凸显与所需的伴随分布之间的差异。

Landsman 和 Nešlehová [12], Landsman 等人 [14**?**] 证明了 Stein 恒等式向绝对连续椭圆分布类的扩展，并特别关注属于 VII 型皮尔逊分布族的重尾分布。他们的结果基于关于椭圆分布族性质的经典研究 [23–26]。具体而言，Landsman 等人 [14, 命题 2] 给出了椭圆 Stein 恒等式的一般性证明，但没有讨论本文关注的有界支撑情形；另见 [13, 引理 2] 和 [12]。值得注意的是，基于广义椭圆 Stein 恒等式的随机梯度估计量尚未得到研究，我们也未发现任何工作表明此类估计量能够转化为类似于高斯随机梯度的形式。本文旨在填补这一空白。

3. 有界支撑集 q-高斯分布的 Stein 型恒等式 s

本节将推导一种专为有界支撑集 q-高斯分布（$q<1$）设计的 Stein 型恒等式。我们采用 Landsman 和 Nešlehová [12], Landsman 等人 [13, 14], 的方法，利用伴随密度 p^* 推导 Stein 型恒等式。我们在附录 B 中证明，第一伴随律与（$2-q$）-伴随密度 $p^\star(\mathbf{x}) \propto p(\mathbf{x})^{2-q}$ [27]相同。据我们所知，经典统计学文献中椭圆分布族所定义的伴随分布 与统计物理和信息几何中研究的 [24, 28], 伴随分布之间的这种联系此前尚未被注意到。使用伴随分布有助于以优雅的方式推广 Stein 恒等式，使其呈现出与高斯斯坦因恒等式相仿的形式，并凸显出概率结构；否则，这一结构会被反复的分部积分所掩盖。

伴随密度 $p^\star$ 通过对区间（s, R^2）上的原始生成函数 g 进行积分所得到的生成函数来定义，

$$G(s) = \int_s^{R^2} g(t)dt. \tag{8}$$

正如我们在附录 B中所述，使用该生成函数会得到另一个指数增加 1 的皮尔逊 II 分布，

$$p^\star(\mathbf{x}) \propto |\boldsymbol{\Sigma}|^{-1/2}\left(R^2 - s(\mathbf{x})\right)_+^{m+1}. \tag{9}$$

将这一表达式与公式 7 进行比较可以看出，唯一的区别在于，指数中的 m 被替换为 $m+1$。注意，p 和 $p^\star$ 具有相同的位置-尺度参数($\boldsymbol{\mu}$，$\boldsymbol{\Sigma}$)以及有界最大支撑半径 R，但 $p^\star$ 由于指数更大，在 $\boldsymbol{\mu}$ 处具有更尖锐的峰值。这一差异在图 1中清晰可见。

在附录 C 中，我们证明以下定理，该定理为支撑集有界的 q-高斯分布建立了一种新的 Stein 型恒等式。

定理 1（支撑集有界的 q-高斯 Stein 恒等式）。对于任意几乎处处可微的 $f\colon \mathbb{R}^D \to \mathbb{R}$，满足 $\mathbb{E}_{p^\star}\|\nabla f(\mathbf{x})\| < \infty$，且二阶矩 $\mathrm{Cov}p(\mathbf{x})$ 存在：

$$\mathbb{E}_p\left[(\mathbf{x}-\boldsymbol{\mu})f(\mathbf{x})\right] = \mathrm{Cov}_p(\mathbf{x})\,\mathbb{E}_{p^\star}\left[\nabla_{\mathbf{x}} f(\mathbf{x})\right]. \tag{10}$$

公式 10 的形式 与公式 1几乎完全相同，区别在于右侧的期望使用伴随分布 $p^\star$ 而非原始的 p（如红色标注所示）。附录 C 中的证明通过定义$\mathbf{z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})$，应用迭代的一维分部积分，并利用 Pearson II 密度及其伴随分布在边界处均趋于零这一事实，以及下文引理 2 中所述的性质。

引理 2。对于任何支撑集有界的 q-高斯分布，若满足 $q<1$，以下性质均成立，

1. 定义 $r(\mathbf{x}) = \sqrt{s(\mathbf{x})}$，则在 p 和 $p^\star$ 下分别有以下结论，

$$\frac{r(\mathbf{x})^2}{R^2} \sim \mathrm{Beta}\left(\tfrac{D}{2}, m+1\right), \qquad \textit{and} \qquad \frac{r(\mathbf{x})^2}{R^2} \sim \mathrm{Beta}\left(\tfrac{D}{2}, m+2\right).$$

2. $s(\mathbf{x}) = r(\mathbf{x})^2$ 的期望可用 R、D和 m表示为闭式表达式，

$$\mathbb{E}_p\left[s(\mathbf{x})\right] = \frac{D\,R^2}{D+2(m+1)}, \qquad \textit{and} \qquad \mathbb{E}_{p^\star}\left[s(\mathbf{x})\right] = \frac{D\,R^2}{D+2\,(m+2)}.$$

3. 二阶矩可用 $\boldsymbol{\Sigma}$表示，

$$\mathrm{Cov}_p(\mathbf{x}) = \frac{1}{D}\mathbb{E}_p[s(\mathbf{x})]\boldsymbol{\Sigma}, \qquad \textit{and} \qquad \mathrm{Cov}_{p^\star}(\mathbf{x}) = \frac{1}{D}\mathbb{E}_{p^\star}[s(\mathbf{x})]\boldsymbol{\Sigma}.$$

4. 最后，我们有 $p^\star$的以下重新加权表示，

$$p^\star(\mathbf{x}) = \frac{(R^2 - s(\mathbf{x}))\,p(\mathbf{x})}{\mathbb{E}_p\left[R^2 - s(\mathbf{x})\right]}.$$

所有这些结果均可由 Beta 期望和协方差的定义推得。附录 B包含更多细节。

该恒等式还有另外两种有用的变体，

$$\mathbb{E}_p\left[(\mathbf{x}-\boldsymbol{\mu})f(\mathbf{x})\right] = \frac{1}{D}\mathbb{E}_p\left[s(\mathbf{x})\right]\,\boldsymbol{\Sigma}\,\mathbb{E}_{p^\star}\left[\nabla_{\mathbf{x}} f(\mathbf{x})\right] \tag{11}$$

$$= \frac{1}{D}\mathbb{E}_p\left[s(\mathbf{x})\right]\,\boldsymbol{\Sigma}\,\frac{\mathbb{E}_p\left[(R^2 - s(\mathbf{x}))\nabla_{\mathbf{x}} f(\mathbf{x})\right]}{\mathbb{E}_p\left[R^2 - s(\mathbf{x})\right]}. \tag{12}$$

在第一行中，我们将二阶矩展开为 $\boldsymbol{\Sigma}$的形式；在第二行中，我们利用引理 2的最后一个性质，将 $p^\star$ 期望改写为基准密度 $p(\mathbf{x})$ 的形式。公式 12 的形式特别便于实现，因为它将恒等式表示为 $p(\mathbf{x})$的函数。接下来我们将展示，从 $p(\mathbf{x})$ 采样与从高斯分布采样的效率相当。

利用引理 2，可以高效地从 $p(\mathbf{x})$中采样。我们定义 $\mathbf{z}(\mathbf{x}) = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})$，它可以用独立随机变量 $\mathbf{u}$ 表示，其中 $\mathbf{u}$ 在 $\mathbb{R}^D$ 中的球面上均匀分布：$\mathbf{z}(\mathbf{x}) = r(\mathbf{x})\mathbf{u}$，其中 $\mathbf{u} \sim \mathrm{Unif}(S^{D-1})$，且 $S^{D-1} = \{\mathbf{u} \in \mathbb{R}^D : \|\mathbf{u}\|_2 = 1\}$。此外，$r(\mathbf{x})^2/R^2$ 是一个独立的 Beta 随机变量。因此，采样分为以下四个步骤：

$$\mathbf{u} \sim \mathrm{Unif}(S^{D-1}), \qquad \frac{r^2}{R^2} \sim \mathrm{Beta}\left(\tfrac{D}{2}, m+1\right), \qquad \mathbf{z} \leftarrow r\mathbf{u}, \qquad \mathbf{x} \leftarrow \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{z}. \tag{13}$$

附录 B 中回顾的径向参数化附录 B。

4. Bonnet 型和 Price 型定理

我们给出了有界支撑集 q-高斯分布的 Bonnet 型和 Price 型定理，证明见 附录 D。这些定理被转化为与相应的高斯定理完全相同或在结构上类似的形式。我们推导出的随机梯度表达式易于估计，因为利用 式 13，可以轻松地从定义期望的分布中进行采样，从而进行蒙特卡罗采样。

我们先从 Bonnet 型定理开始，它对应于公式 2中的第一个等式。

定理 2 (q-Bonnet)。 对于有界支撑集的 $p(\mathbf{x}) = \mathcal{N}_q(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$，假设 $f\colon \mathbb{R}^D \to \mathbb{R}$ 在包含 $\{\mathbf{x}\colon s(\mathbf{x}) \leq R^2\}$的开集上是 C^1，且 $\mathbb{E}_p \|\nabla f(\mathbf{x})\| < \infty$。那么，

$$\nabla_{\boldsymbol{\mu}} \mathbb{E}_p [f(\mathbf{x})] = \mathbb{E}_p [\nabla f(\mathbf{x})]. \tag{14}$$

该恒等式与公式 2中的第一个等式形式完全相同。证明见附录 D.1，由 $p(\mathbf{x})$ 在公式 7 中的形式以及以下事实可得：$\nabla_{\boldsymbol{\mu}} s(\mathbf{x}) = -2\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ 这使我们可以将梯度写为

$$\nabla_{\boldsymbol{\mu}} \log p(\mathbf{x}) = \frac{2m}{R^2 - s(\mathbf{x})} \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}).$$

对积分号下求导后，公式 14 左侧可用以下形式表示：$\mathbb{E}_p \left[f(\mathbf{x}) \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) / (R^2 - s(\mathbf{x})) \right]$。最后，结合公式 12 中的 Stein 型恒等式和一些代数运算，即可得到所需结果。

接下来，我们陈述 Price 型定理，它与式 2中的第二个等式相对应：

定理 3 (q-Price)。 对于支撑集有界的 $p(\mathbf{x}) = \mathcal{N}_q(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$，假设 $f\colon \mathbb{R}^D \to \mathbb{R}$ 在包含 $\{s(\mathbf{x}) \leq R^2\}$ 的开集上是 C^2，并且 $\mathbb{E}_p [\|\nabla f(\mathbf{x})\|] + \mathbb{E}_p \left[\|\nabla^2 f(\mathbf{x})\|_F\right] < \infty$。那么，

$$\nabla_{\boldsymbol{\Sigma}} \mathbb{E}[f(\mathbf{x})] = \frac{1}{D} \mathbb{E}_p[s(\mathbf{x})] \frac{1}{2} \mathbb{E}_{p^\star}[\nabla^2_{\mathbf{x}} f(\mathbf{x})]. \tag{15}$$

在高斯极限下 $q \uparrow 1$(即 $m \to \infty$ 且 $R \to \infty$)，$p^\star \to p$ 且 $\mathbb{E}_p[r(\mathbf{x})^2]/D \to 1$，这意味着 式 15化简为经典 *Price* 定理 $\frac{\partial}{\partial \boldsymbol{\Sigma}_{ij}} \mathbb{E}[f(\mathbf{x})] = \frac{1}{2}\mathbb{E}[\partial_{x_i}\partial_{x_j} f(\mathbf{x})]$。

证明见D.2 节。这个 Price 型定理与式 2第二个等式中的高斯 Price 定理非常相似，差异用红色标出。右侧的期望是针对$(2-q)$-伴随分布 $p^\star$计算的，此外还有一个因子 $\mathbb{E}_p[s(\mathbf{x})]/D$。这个表达式很容易求值，因为从 $p^\star$ 中采样的效率很高，如式 13所示。

5. 应用

5.1. 有界方差蒙特卡罗估计量

有界支撑分布的一个优势在于，它们自然会导出具有有界方差的梯度估计量。我们分别基于 公式 12 和 公式 15 给出关于这些估计量方差上界的形式化陈述。 定理 1 中的 Stein 型恒等式以及 q-Price 定理 3 将所关注的梯度表示为伴随分布期望：$\mathbb{E}_{p^\star}[\nabla f(\mathbf{x})]$ 和 $\mathbb{E}_{p^\star}[\nabla^2 f(\mathbf{x})]$。由于我们希望使用蒙特卡罗来近似这些期望，所得估计量的方差会直接影响随机梯度噪声。利用 p-only 表达式（见 公式 25，位于 附录 B），在对 ∇f 和 $\nabla^2 f$施加温和的有界性假设后，我们可以对随机梯度蒙特卡罗估计量的方差进行界定。

命题 1 （有界方差蒙特卡罗估计量）。 设 $\mathbf{x}_1, \ldots \mathbf{x}_S$ 为来自 q-高斯分布 $p(\mathbf{x})$的独立同分布样本。对于任意几乎处处可微的 $t\colon \mathbb{R}^D \to \mathbb{R}$，满足 $\mathbb{E}_{p^\star}[\|\nabla t(\mathbf{x})\|] < \infty$，以及在包含 $\{s(\mathbf{x}) \leq R^2\}$ 的开集上C^2 的 $f\colon \mathbb{R}^D \to \mathbb{R}$，使得 $\mathbb{E}_p [\|\nabla f(\mathbf{x})\|] + \mathbb{E}_p \left[\|\nabla^2 f(\mathbf{x})\|_F\right] < \infty$，定义以下蒙特卡罗估计量：

$$\mathbb{E}_{p^\star}[\nabla t(\mathbf{x})] \approx \widehat{\mathbf{g}} = \frac{1}{S} \sum_{k=1}^{S} \frac{(R^2 - s(\mathbf{x}_k)) \nabla t(\mathbf{x}_k)}{\mathbb{E}_p [R^2 - s(\mathbf{x})]}, \qquad \mathbb{E}_{p^\star}[\nabla^2 f(\mathbf{x})] \approx \widehat{\mathbf{H}} = \frac{1}{S} \sum_{k=1}^{S} \frac{(R^2 - s(\mathbf{x}_k)) \nabla^2 f(\mathbf{x}_k)}{\mathbb{E}_p [R^2 - s(\mathbf{x})]}.$$

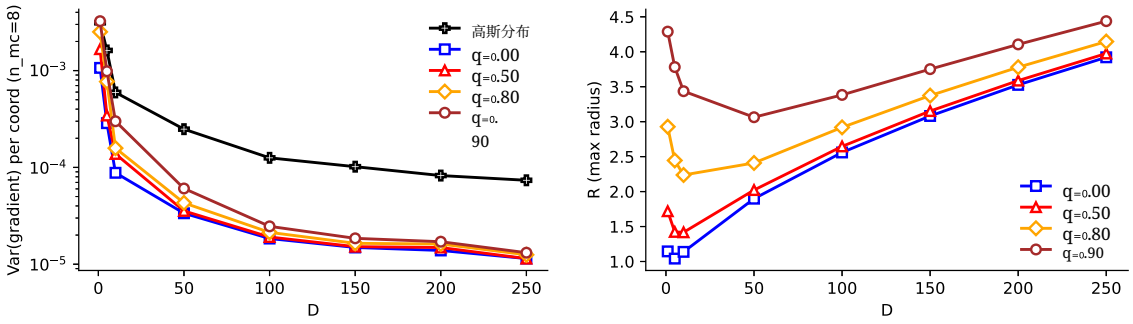

图2：合成逻辑回归。左图：对于 $D \in \{10, 50, 200\}$ 和 $q \in \{0.0, 0.5, 0.8, 1\}$，我们抽取 8 个蒙特卡洛样本，并计算经验逐坐标梯度方差 $\frac{1}{D}\sum_{j=1}^{D} \mathrm{Var}(\widehat{\nabla} F(\mathbf{w}^\star)_j)$，对 50 次独立重复试验取平均；所有标准误差均为 <.003。右图：q-高斯分布支撑集的最大半径随 D变化。

注意 $M = \mathbb{E}_p\left[R^2 - s(\mathbf{x})\right]$ 可由引理 2以闭式形式给出。假设存在有限常数C_1，C_2 使得

$$\sup_{\{s(\mathbf{x})<R^2\}} \|\nabla t(\mathbf{x})\| \le C_1, \qquad \textit{and} \qquad \sup_{\{s(\mathbf{x})<R^2\}} \|\nabla^2 f(\mathbf{x})\|_{\mathrm{op}} \le C_2,$$

其中 $\|\cdot\|$ 是矩阵上的谱范数（算子范数）。那么，对于每个元素 $\widehat{g}_j$ 和 H_{ij}，我们有：

$$\mathrm{Var}\left(\widehat{g}_j\right) \le \frac{1}{S}\left(\frac{R^2 C_1}{M}\right)^2, \qquad \textit{and} \qquad \mathrm{Var}\left((\widehat{H})_{ij}\right) \le \frac{1}{S}\left(\frac{R^2 C_2}{M}\right)^2.$$

因此，

$$\mathbb{E}\left[\left\|\widehat{\mathbf{H}} - \mathbb{E}\left[\widehat{\mathbf{H}}\right]\right\|_F^2\right] \le D^2 \frac{1}{S}\left(\frac{R^2 C_2}{M}\right)^2, \qquad \textit{and} \qquad \mathbb{E}\left[\left\|\widehat{\mathbf{H}} - \mathbb{E}\left[\widehat{\mathbf{H}}\right]\right\|_{\mathrm{op}}\right] \le \frac{C_3 R^2 C_2}{M}\sqrt{\frac{\log D}{S}}$$

对于 $M = \mathbb{E}_p\left[R^2 - s(\mathbf{x})\right]$ 以及有限常数 C_3。

证明。 证明见 第D.3节。它由有界范围方差界（波波维丘不等式）和矩阵霍夫丁不等式得出。 □

5.2. 数值实验

合成逻辑回归实验。 对于维度 $D \in \{10, 50, 200\}$，我们抽取一个数据集$\{(\mathbf{x}_i, y_i)\}_{i=1}^N$，其中 $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$，真实权重向量 $\mathbf{w}^\star \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$，以及标签 $y_i \mid \mathbf{x}_i \sim$伯努利分布$(\sigma(\mathbf{x}_i^\top \mathbf{w}^\star))$，其中 σ 是Sigmoid函数。损失 $f(\mathbf{w})$ 是二元交叉熵损失。我们考虑在 $\mathbf{w} = \mathbf{w}^\star$ 处求值的目标函数梯度的蒙特卡洛路径导数估计量，

$$F(\mathbf{w}) = \mathbb{E}_{\boldsymbol{\epsilon}}\left[f(\mathbf{w} + \boldsymbol{\epsilon})\right].$$

对于高斯基线，我们取 $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$。对于有界支撑集的 q-平滑，我们从各向同性的 D 维有界支撑集 q-高斯分布中抽取 $\boldsymbol{\epsilon}$，使得 $\|\boldsymbol{\epsilon}\|_2 \le R(D, q)$。对于每个$D \in \{10, 50, 200\}$ 和 $q \in \{0.0, 0.5, 0.8, 1\}$，我们抽取 8 个蒙特卡洛样本 $\boldsymbol{\epsilon}$并计算

$$\widehat{\nabla} F(\mathbf{w}^\star) = \frac{1}{S}\sum^{S} \nabla_{\mathbf{w}} f(\mathbf{w}),$$

对于 $\mathbf{w} = \mathbf{w}^\star + \boldsymbol{\epsilon}$。在 图 2 （左）中，对于每个 q，我们根据 50 次独立重复试验，绘制了逐坐标梯度方差的经验值 $\frac{1}{D}\sum_{j=1}^{D} \mathrm{Var}(\widehat{\nabla} F(\mathbf{w}^\star)_j)$ 与 D的关系。 图 2 （左）中的结果证实，较小的 q会带来方差更低的梯度估计量。我们还在 D 的条件下，于 图 2 （右）中绘制了支撑集的最大半径，以展示半径如何随维度增大，以及 q 趋近于 1 而增大；维度对 R 的大小有显著影响，而 q 的影响只有在 D非常小时才显著。

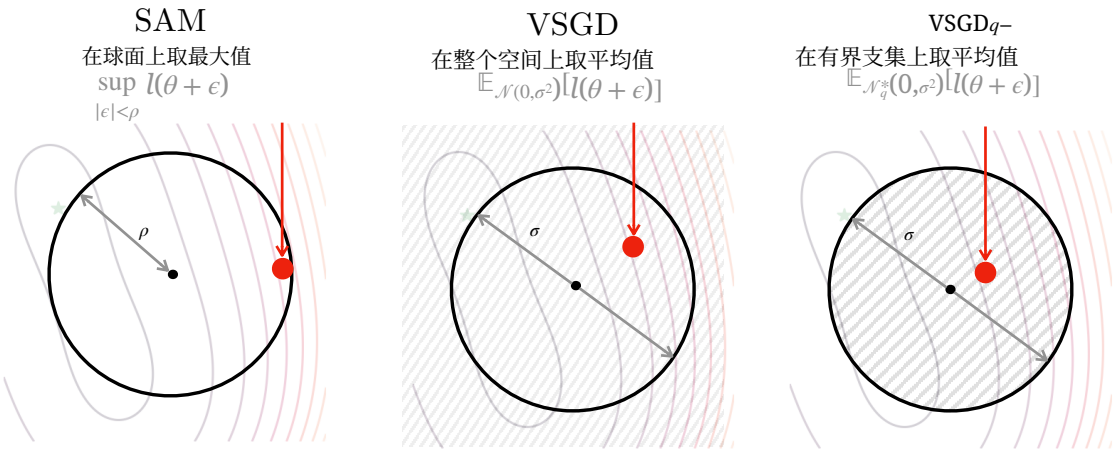

图3：锐度感知最小化（SAM）[18] 考虑紧致球上的对抗性扰动。采用高斯权重扰动的变分随机梯度下降（VSGD）在整个空间（无界支集）上对扰动取平均值。我们提出的 q-VSGD使用 q-高斯权重扰动，其与SAM类似具有有界支撑，但与VSGD类似使用平均值。该方法结合了SAM和VSGD的两个互补特征。

采用 q-高斯分布噪声的变分随机梯度下降。 接下来，我们探讨将我们的梯度估计量用于贝叶斯深度学习的潜在用途，以估计 q-高斯后验，而非通常使用的高斯后验。我们将其与几种变分训练方法以及锐度感知最小化（SAM）算法进行比较。 图 3展示了可视化结果，说明了主要差异；这些差异与计算梯度时所选取的点有关。

在SAM中，我们考虑当前参数周围圆内具有最差损失值的一个点。形式上，令 $f(\mathbf{w};\ \mathcal{B})$为小批量交叉熵，其中 $\mathcal{B}$ 是小批量。然后，我们将扰动设为 $\delta = \rho\mathbf{g}/\|\mathbf{g}\|_2$，其中$\mathbf{g} = \nabla_{\mathbf{w}} f(\mathbf{w};\ \mathcal{B})$。随后通过在该点计算梯度来更新参数，

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla_{\mathbf{w}} f(\mathbf{w}+\delta;\mathcal{B}).$$

总体而言，这需要两次梯度计算，总共使用两次前向和反向传播。在变分随机梯度下降（VSGD）中，目标是估计各向同性高斯后验的均值。该均值通过取期望损失的梯度进行更新，至少需要一个蒙特卡洛样本。这意味着更新方式与上述相同，但我们需要使用 $\delta \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$。

我们提出一种新的 VSGD 变体，旨在估计 q-高斯后验的均值，同时将 $\mathbf{\Sigma}$ 固定为$\mathbf{I}_D$。只需对 VSGD 做一个简单的改动即可实现：不再从标准正态分布中采样，而是像下文所示，从 q-高斯分布中采样：

$$\delta \leftarrow \rho \frac{\boldsymbol{\epsilon}}{R(q,D)},\ \text{where } \boldsymbol{\epsilon} \sim \mathcal{N}_q(\mathbf{0}, \mathbf{I}_D),$$

而 $R(q,\ D)$ 是用于归一化的、依赖于分布的半径。请注意，当$q = 1$ 时，q-VSGD 等价于 VSGD。我们还将 q-VSGD 与一种名为改进变分在线牛顿法（IVON）的优化器进行比较；该优化器采用了更灵活的对角高斯后验 [29]。

我们在 CIFAR-10 数据集上使用标准的 ResNet-20 模型比较这些方法，发现改变 q 相较于 VSGD 能带来小幅改进。实验设置沿用提出 IVON 的论文中的设置 [29]。我们使用批量大小 50，采用带动量的 SGD 训练 200 轮 0。9，并使用权重衰减 10^{-4}；先进行 5 轮线性预热，再采用余弦退火，并按照 $q \leq 1$ 中报告的设置对取值进行遍历表 1。在表 1 中，我们报告了测试准确率、负对数似然、布里尔分数、期望校准误差（20 个区间）、受试者工作特征曲线下面积（详见 [30]），以及每轮墙钟时间（秒）（均值 ± 10 个随机种子上的标准误差）。对于 $q = 0.6$，我们观察到准确率有所提升。增加蒙特卡洛样本数可以进一步提高准确率，但会使算法变慢。总的来说，结果并不明确，表明调整 q 可能无法直接

方法	准确率 (%) ↑	负对数似然 ↓	期望校准误差 (%) ↓	布里尔分数 ↓	受试者工作特征曲线下面积 ↑	时间 ↓
SGD	92.0 ± 0.1	0.28 ± 0.00	3.84 ± 0.11	0.12 ± 0.00	92.1 ± 0.1	37s
IVON（1-MC） [29]	92.5 ± 0.1	0.26 ± 0.00	3.43 ± 0.09	0.12 ± 0.00	92.5 ± 0.1	43s
SAM [18]	92.6 ± 0.1	0.22 ± 0.00	1.56 ± 0.05	0.11 ± 0.00	92.5 ± 0.1	70s
1 蒙特卡洛样本						
q-VSGD（$q = 0.0$）	92.1 ± 0.0	0.25 ± 0.00	3.01 ± 0.05	0.12 ± 0.00	92.3 ± 0.0	44s
q-VSGD（$q = 0.2$）	92.1 ± 0.0	0.26 ± 0.00	2.99 ± 0.04	0.12 ± 0.00	92.3 ± 0.1	43s
q-VSGD（$q = 0.4$）	92.0 ± 0.0	0.26 ± 0.00	3.10 ± 0.05	0.12 ± 0.00	92.3 ± 0.1	43s
q-VSGD（$q = 0.6$）	92.2 ± 0.0	0.25 ± 0.00	2.93 ± 0.04	0.12 ± 0.00	92.3 ± 0.1	43s
q-VSGD（$q = 0.8$）	92.1 ± 0.0	0.26 ± 0.00	3.02 ± 0.04	0.12 ± 0.00	92.2 ± 0.1	43s
VSGD	92.1 ± 0.0	0.25 ± 0.00	2.79 ± 0.05	0.12 ± 0.00	92.4 ± 0.0	42s
5 蒙特卡罗样本						
q-VSGD（$q = 0.0$）	92.6 ± 0.1	0.25 ± 0.00	3.06 ± 0.08	0.11 ± 0.00	92.5 ± 0.1	173s
q-VSGD ($q = 0.2$)	92.5 ± 0.0	0.25 ± 0.00	3.13 ± 0.07	0.11 ± 0.00	92.6 ± 0.1	173s
q-VSGD ($q = 0.4$)	92.4 ± 0.0	0.25 ± 0.00	3.17 ± 0.06	0.11 ± 0.00	92.7 ± 0.1	174s
q-VSGD ($q = 0.6$)	92.6 ± 0.1	0.25 ± 0.00	3.04 ± 0.07	0.11 ± 0.00	92.5 ± 0.1	174s
q-VSGD（$q = 0.8$）	92.4 ± 0.0	0.25 ± 0.00	3.17 ± 0.05	0.11 ± 0.00	92.7 ± 0.1	182s
VSGD	92.4 ± 0.1	0.25 ± 0.00	3.13 ± 0.09	0.12 ± 0.00	92.6 ± 0.1	173s

表 1：ResNet-20 在 CIFAR-10 上的测试性能。均值 ±10 个随机种子上的标准误差。基线（顶部的第一组）中，SGD 的表现最差，但速度最快。SAM 的准确率最高，但由于每次迭代需要计算两次梯度，因此也较慢；而 IVON 的表现尚可，计算成本也合理。在第二组中，我们展示了使用 1 个蒙特卡洛样本的 q-VSGD；通过增加 $q = 0.6$，其准确率略高于对应的 VSGD，但其他指标相当。第三组采用 5次蒙特卡洛后，性能有所提升，但算法速度变慢。结果喜忧参半，表明还需要进一步研究如何通过调整 q 来提升性能。

带来提升　　d性能立刻显现，尽管梯度估计器的方差
有界。
性能未能进一步提升的一个潜在原因是，在非常大的维数下，改变 q 的效果会急剧减弱。这是因为支撑集受到维数的强烈影响，如图2所示，在高维情况下，所有 q 取值都具有非常相似的支撑集。我们的数值结果表明，应使用更灵活的 q-高斯分布形式，例如估计尺度参数。我们预期 q-IVON算法的表现将优于 q-VSGD，而后者确实优于VSGD。或许最有效的改进是采用低维 q-高斯分布分解。我们希望在未来的工作中探索这些替代方案。

6. 结论

本文为支撑集有界的 q-高斯分布证明了一种新的 Stein 型恒等式；这类分布属于椭圆分布 Pearson II 类的子族。我们证明，相应的（累积生成函数）分布律具有简单的伴随分布解释，从而可以推导出 Bonnet 型和 Price 型恒等式，其形式与高斯随机梯度估计器的形式十分相似。我们将这些结果应用于实用的路径梯度估计器；这类估计器只需从 q-高斯密度中采样，便易于实现。支撑集有界的一个重要结果是，所得蒙特卡洛估计器具有简单的有界方差保证。我们的实验在受控的合成场景以及基于 CIFAR-10 训练的深度神经模型中展示了这些效果；在这些场景中，支撑集有界的扰动具有竞争力，能够与 SAM 媲美，并为有界半径扰动方法提供了一种原理清晰的分布式类比。

本文的研究结果为进一步研究开辟了新方向，即在更广泛的随机梯度应用中利用广义 Stein 恒等式，例如变分推断和鲁棒优化；相关研究还可扩展到支撑集有界的 q-高斯分布之外（尤其是重尾 $3 > q > 1$ 情形），并探索学习或自适应 R，以及考虑各向异性 $\mathbf{\Sigma}$。

致谢

SS、MEK 和 TM 获得了贝叶斯对偶项目和 JST CREST 项目（资助编号：JPMJCR2112）的支持。AM 获得了 DECOLLAGE 项目（ERC-2022-CoG 101088763）的支持。AM 和 MF 获得了葡萄牙复苏与韧性计划通过项目 C645008882000000055（负责任人工智能中心）提供的支持，以及葡萄牙科学与技术基金会通过合同 UIDB/50008/2020 提供的支持。

参考文献

[1] Shakir Mohamed、Mihaela Rosca、Michael Figurnov 和 Andriy Mnih。机器学习中的蒙特卡洛梯度估计。机器学习研究期刊，21(132):1–62，2020。[2] Danilo Jimenez Rezende、Shakir Mohamed 和 Daan Wierstra。深度生成模型中的随机反向传播与近似推断。载于国际机器学习会议 *(ICML)*，2014。[3] Charles Blundell、Julien Cornebise、Koray Kavukcuoglu 和 Daan Wierstra。神经网络中的权重不确定性。载于国际机器学习会议 *(ICML)*，2015。[4] Mohammad Emtiyaz Khan、Didrik Nielsen、Voot Tangkaratt、Wu Lin、Yarin Gal 和 Akash Srivastava。通过 Adam 中的权重扰动实现快速、可扩展的贝叶斯深度学习。载于国际机器学习会议 *(ICML)*，2018。[5] Charles M. Stein。多元正态分布均值的估计。载于布拉格渐近统计研讨会论文集，第345–381页，1973。[6] Charles M. Stein。多元正态分布均值的估计。统计学年刊，第1135–1151页，1981。[7] Wu Lin、Mohammad Emtiyaz Khan 和 Mark Schmidt。指数族混合的重参数化技巧的 Stein 引理。*arXiv:1910.13398*，2019。[8] Georges Bonnet。无记忆非线性系统中随机信号的变换。载于电信年鉴，第19卷，第203–220页，1964。[9] Robert Price。关于具有高斯分布输入的非线性器件的一个有用定理。*IRE*信息论汇刊，4(2):69–72，2003。[10] Ronald J Williams。用于联结主义强化学习的简单统计梯度跟随算法。机器学习，8(3):229–256，1992。[11] Qiang Liu 和 Dilin Wang。Stein变分梯度下降：一种通用贝叶斯推断算法。神经信息处理系统进展 *29*，2016。[12] Zinoviy Landsman 和 Johanna Nešlehová。椭圆随机向量的 Stein 引理。多元分析期刊，99(5):912–927，2008。[13] Zinoviy Landsman、Steven Vanduffel 和 Jing Yao。关于多元椭圆分布 Stein 引理的注记。统计规划与推断期刊，143(11):2016–2022，2013。[14] Zinoviy Landsman、Steven Vanduffel 和 Jing Yao。多元椭圆分布的一些 Stein 型不等式及其应用。统计与概率快报，97:54–62，2015。

[15] Jan Naudts。统计物理中的 q 指数族。中欧物理学期刊，7(3):405–413，2009。[16] Shun-ichi Amari 和 Atsumi Ohara。概率分布 q 指数族的几何。熵，13(6):1170–1185，2011。[17] Hiroshi Matsuzoe 和 Atsumi Ohara。q 指数族的几何。微分几何及相关领域的最新进展，第55–71页，2011。[18] Pierre Foret、Ariel Kleiner、Hossein Mobahi 和 Behnam Neyshabur。用于高效提升泛化能力的尖锐度感知最小化。载于国际学习表征会议 *(ICLR)*，2021。[19] Mohammad Khalafi 和 Majid Azimmohseni。多元 II 型皮尔逊分布：统计与数学特征。概率与数理统计，34，2014。[20] Jan Naudts。广义热力学统计中的变形指数与对数。物理学*A*：统计力学及其应用，316:323–334，2002。[21] André FT Martins、Marcos Treviso、António Farinhas、Pedro MQ Aguiar、Mário AT Figueiredo、Mathieu Blondel 和 Vlad Niculae。稀疏连续分布与 Fenchel-Young 损失。机器学习研究期刊，23(257):1–74，2022。[22] Constantino Tsallis。玻尔兹曼-吉布斯统计的一种可能推广。统计物理学期刊，52:479–487，1988。[23] Stamatis Cambanis、Steel Huang 和 Gordon Simons。椭圆等高线分布理论。多元分析期刊，11(3):368–385，1981。[24] Kai Tai Fang、Samuel Kotz 和 Kai Wang Ng。对称多元分布及相关分布。Springer-Science+Business Media，B.V.，1990。[25] Reinaldo B Arellano-Valle。关于球形分布的一些刻画。统计与概率快报，54(3):227–232，2001。[26] Reinaldo B Arellano-Valle、Guido del Pino 和 Pilar Iglesias。球形线性模型中的贝叶斯推断：稳健性与共轭分析。多元分析期刊，97(1):179–197，2006。[27] Jan Naudts。统计物理中的估计量、伴随分布概率与 φ 指数族。纯粹与应用数学不等式期刊，5:1443–5756，2004。[28] Mark E Johnson。多元统计模拟：连续多元分布的选择与生成指南。John Wiley & Sons，1987。[29] Yuesong Shen、Nico Daheim、Bai Cong、Peter Nickl、Gian Maria Marconi、Clement Bazan、Rio Yokota、Iryna Gurevych、Daniel Cremers、Mohammad Emtiyaz Khan 和 Thomas Möllenhoff。变分学习对大型深度网络有效。载于国际机器学习会议 *(ICML)*，2024。[30] Kazuki Osawa、Siddharth Swaroop、Mohammad Emtiyaz Khan、Anirudh Jain、Runa Eschenhagen、Richard E Turner 和 Rio Yokota。基于贝叶斯原理的实用深度学习。载于神经信息处理系统进展 *(NeurIPS)*，2019。[31] Amir Rezaei、Maryam Sharafi、Javad Behboodian 和 Atefeh Zamani。基于 GLD5 分布的应力–强度参数推断。统计学通讯：模拟与计算，47(5):1251–1263，2018。

A. 高斯分布下邦内定理与Price定理的证明

高斯斯坦因恒等式、邦内定理与Price定理。 设 $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 的密度为 $p(\mathbf{x})$。一般的 Stein 恒等式指出，对于任意可微测试函数 $t\colon\ \mathbb{R}^D \to \mathbb{R}$，其中 $\mathbb{E}\,\|t(\mathbf{x})\| + \mathbb{E}\,\|\nabla_{\mathbf{x}} t(\mathbf{x})\| < \infty$,

$$\mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu})t(\mathbf{x})\right] = \boldsymbol{\Sigma}\mathbb{E}\left[\nabla_{\mathbf{x}} t(\mathbf{x})\right]. \tag{16}$$

这可由得分恒等式 $\mathbb{E}\left[t(\mathbf{x})\nabla_{\mathbf{x}} \log\, p(\mathbf{x})\right] = -\mathbb{E}\left[\nabla_{\mathbf{x}} t(\mathbf{x})\right]$ 通过分部积分得到，同时还用到了高斯分布特有的性质：$\nabla_{\mathbf{x}} \log\, p(\mathbf{x}) = -\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$。

得分恒等式意味着:

$$\left.\begin{aligned} \nabla_{\boldsymbol{\mu}}\mathbb{E}\left[f(\mathbf{x})\right] &= \mathbb{E}\left[f(\mathbf{x})\nabla_{\boldsymbol{\mu}} \log p(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\right] \\ \nabla_{\boldsymbol{\mu}} \log p(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \end{aligned}\right\} \implies \nabla_{\boldsymbol{\mu}}\mathbb{E}\left[f(\mathbf{x})\right] = \boldsymbol{\Sigma}^{-1}\mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu})f(\mathbf{x})\right].$$

将 Stein 恒等式应用于 $t(\mathbf{x}) = f(\mathbf{x})$，可得邦内定理

$$\nabla_{\boldsymbol{\mu}}\mathbb{E}\left[f(\mathbf{x})\right] = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\mathbb{E}\left[\nabla_{\mathbf{x}} f(\mathbf{x})\right] = \mathbb{E}\left[\nabla_{\mathbf{x}} f(\mathbf{x})\right].$$

令 $\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\epsilon}$，其中 $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$。由于 $\partial\mathbf{x}/\partial\boldsymbol{\mu} = \mathbf{I}$，我们看到，位置-尺度变换或“重参数化技巧”的一个常见情形，就是邦内定理的路径式实现。类似地，Price定理指出，对于可微的 f，有 $\nabla_{\boldsymbol{\Sigma}}\mathbb{E}\left[f(\mathbf{x})\right] = \frac{1}{2}\mathbb{E}\left[\nabla^2_{\mathbf{x}} f(\mathbf{x})\right]$；将相同的 Stein 恒等式应用于 $\nabla_{\mathbf{x}} f(\mathbf{x})$ 的各个分量，并使用关于 $\boldsymbol{\Sigma}$ 的导数的得分恒等式，即可得出该定理。

高斯斯坦因与普赖斯定理。 令 $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，定义在 $\mathbb{R}^D$ 上，并令 $f\colon\ \mathbb{R}^D \to \mathbb{R}$ 二阶连续可微，且满足 $\mathbb{E}\,\|\nabla_{\mathbf{x}} f(\mathbf{x})\| < \infty$ 和 $\mathbb{E}\left\|\nabla^2_{\mathbf{x}} f(\mathbf{x})\right\|_{\mathrm{F}} < \infty$。标量高斯斯坦因恒等式指出，对于任意足够规则的 $t\colon\ \mathbb{R}^D \to \mathbb{R}$ 以及每个坐标 $i = 1,\ldots,D$,

$$\mathbb{E}\left[(\mathbf{x}_i - \boldsymbol{\mu}_i)t(\mathbf{x})\right] = \sum_{\ell=1}^{D} \Sigma_{i\ell}\mathbb{E}\left[\partial_{x_\ell} t(\mathbf{x})\right]. \tag{Stein}$$

*Price*定理的证明。 为对 $\boldsymbol{\Sigma}$ 求导，我们使用高斯密度的矩阵得分恒等式:

$$\nabla_{\boldsymbol{\Sigma}} \log p(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\tfrac{1}{2}\boldsymbol{\Sigma}^{-1} + \tfrac{1}{2}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}, \tag{17}$$

因此，逐项可得,

$$\frac{\partial}{\partial\boldsymbol{\Sigma}_{ij}}\mathbb{E}\left[f(\mathbf{x})\right] = \mathbb{E}\left[f(\mathbf{x})\frac{\partial}{\partial\boldsymbol{\Sigma}_{ij}} \log p(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\right] \tag{18}$$

$$= -\tfrac{1}{2}(\boldsymbol{\Sigma}^{-1})_{ji}\mathbb{E}\left[f(\mathbf{x})\right] + \tfrac{1}{2}\left(\boldsymbol{\Sigma}^{-1}\mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\top}f(\mathbf{x})\right]\boldsymbol{\Sigma}^{-1}\right)_{ji}. \tag{19}$$

对每个固定的 $j = 1,\ldots,D$，将 Stein 恒等式应用于标量检验函数 $t(\mathbf{x}) := (\mathbf{x}^j - \boldsymbol{\mu}_j)f(\mathbf{x})$。对于每个$i$,

$$\mathbb{E}\left[(\mathbf{x}_i - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_j)f(\mathbf{x})\right] = \sum_{\ell=1}^{D} \boldsymbol{\Sigma}_{i\ell}\mathbb{E}\left[\partial_{\mathbf{x}_\ell}\left((\mathbf{x}_j - \boldsymbol{\mu}_j)f(\mathbf{x})\right)\right].$$

由于

$$\partial_{\mathbf{x}_\ell}\left((\mathbf{x}_j - \boldsymbol{\mu}_j)f(\mathbf{x})\right) = \delta_{\ell j}f(\mathbf{x}) + (\mathbf{x}_j - \boldsymbol{\mu}_j)\partial_{\mathbf{x}_\ell} f(\mathbf{x}),$$

其中 $\delta_{\ell j}$ 是克罗内克 delta，我们得到

$$\mathbb{E}\left[(\mathbf{x}_i - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_j)f(\mathbf{x})\right] = \sum_{\ell} \boldsymbol{\Sigma}_{i\ell}\delta_{\ell j}\mathbb{E}\left[f(\mathbf{x})\right] + \sum_{\ell} \boldsymbol{\Sigma}_{i\ell}\mathbb{E}\left[(\mathbf{x}_j - \boldsymbol{\mu}_j)\partial_{\mathbf{x}_\ell} f(\mathbf{x})\right] \tag{20}$$

$$= \boldsymbol{\Sigma}_{ij}\mathbb{E}\left[f(\mathbf{x})\right] + \sum_{\ell} \boldsymbol{\Sigma}_{i\ell}\mathbb{E}\left[(\mathbf{x}_j - \boldsymbol{\mu}_j)\partial_{\mathbf{x}_\ell} f(\mathbf{x})\right]. \tag{21}$$

接下来，使用标量检验函数 $t(\mathbf{x}) := \partial_{\mathbf{x}_\ell} f(\mathbf{x})$ 再次应用 Stein。对于每个 $\ell = 1, \ldots\ldots D$,

$$\mathbb{E}\left[(\mathbf{x}_j - \boldsymbol{\mu}_j)\partial_{\mathbf{x}_\ell} f(\mathbf{x})\right] = \sum_{k=1}^{D} \boldsymbol{\Sigma}_{jk}\mathbb{E}\left[\partial_{\mathbf{x}_k}\partial_{\mathbf{x}_\ell} f(\mathbf{x})\right].$$

代入前面的表达式可得

$$\mathbb{E}\left[(\mathbf{x}_i-\boldsymbol{\mu}_i)(\mathbf{x}_j-\boldsymbol{\mu}_j)f(\mathbf{x})\right]=\boldsymbol{\Sigma}_{ij}\mathbb{E}\left[f(\mathbf{x})\right]+\sum_{\ell,k}\boldsymbol{\Sigma}_{i\ell}\boldsymbol{\Sigma}_{jk}\mathbb{E}\left[\partial_{\mathbf{x}_k}\partial_{\mathbf{x}_\ell}f(\mathbf{x})\right].$$

用矩阵记法表示为

$$\mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top f(\mathbf{x})\right]=\boldsymbol{\Sigma}\mathbb{E}\left[f(\mathbf{x})\right]+\boldsymbol{\Sigma}\mathbb{E}\left[\nabla^2_{\mathbf{x}}f(\mathbf{x})\right]\boldsymbol{\Sigma}. \tag{22}$$

在左侧和右侧分别乘以 (22)： $\boldsymbol{\Sigma}^{-1}$:

$$\boldsymbol{\Sigma}^{-1}\mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top f(\mathbf{x})\right]\boldsymbol{\Sigma}^{-1}=\mathbb{E}\left[f(\mathbf{x})\right]\boldsymbol{\Sigma}^{-1}+\mathbb{E}\left[\nabla^2_{\mathbf{x}}f(\mathbf{x})\right].$$

将其代回 (18。（$\mathbf{x}$）] 项相互抵消，得到

$$\frac{\partial}{\partial\boldsymbol{\Sigma}_{ij}}\mathbb{E}\left[f(\mathbf{x})\right]=\tfrac{1}{2}\mathbb{E}\left[\partial_{\mathbf{x}_i}\partial_{\mathbf{x}_j}f(\mathbf{x})\right].$$

等价地，用矩阵形式表示为，

$$\nabla_{\boldsymbol{\Sigma}}\mathbb{E}\left[f(\mathbf{x})\right]=\tfrac{1}{2}\mathbb{E}\left[\nabla^2_{\mathbf{x}}f(\mathbf{x})\right],$$

这就是 Price定理。 □

注 1（ $\boldsymbol{\Sigma}$中的对称性）。 在 Price定理的证明中，我们对各项 $\boldsymbol{\Sigma}_{ij}$ 求导时，将 $\boldsymbol{\Sigma}$ 视为无约束矩阵。梯度最终表达式右侧是对称的，因为对于 C^2 函数 f，Hessian矩阵 $\nabla^2_{\mathbf{x}}f$（$\mathbf{x}$）是对称的。因此，梯度自然属于对称矩阵空间，并且该公式符合协方差约束 $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}^\top$。

B. 椭圆分布

广义Pearson II型。 椭圆分布的一个重要子类是Pearson II型分布。这些分布的一个基本特征是，D维球面Pearson II型随机向量$\mathbf{z}$ 的随机表示为$\mathbf{z}\overset{d}{=}r\mathbf{u}$，其中$\mathbf{u}$ 在 $\mathbb{R}^D$的单位球面上服从均匀分布，符号 d 表示同一分布。随机变量 $r\overset{d}{=}\|\mathbf{z}\|$和$\mathbf{u}\overset{d}{=}\mathbf{z}/\|\mathbf{z}\|$ 相互独立，因此$\mathbf{z}$的分布完全由其平方长度 $r^2=\mathbf{z}^\top\mathbf{z}$的分布决定。随机变量 r 和

—

$$s(\mathbf{x})=r^2=(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

称为随机向量$\mathbf{x}$的径向部分和平方式径向部分。它们的分布由式（4）中的特定Pearson II型密度生成函数定义，如下:

$$f_r(r)=\frac{2\pi^{D/2}}{\Gamma(D/2)}r^{D-1}g(r^2),\quad r>0,\qquad f_s(s)=\frac{\pi^{D/2}}{\Gamma(D/2)}s^{D/2-1}g(s).$$

更多细节请参见 [24, 25, 28, 31]。

引理*1*的证明。 将Pearson II型密度的指数 m 用 $q<1$ 的熵指数表示，其中 q-高斯密度为 $m=1/1-q$，并选取支撑半径 $R=R(D,q)$为

$$R^2=\left[\frac{\Gamma\left(\frac{D}{2}+\frac{2-q}{1-q}\right)}{\pi^{D/2}\Gamma\left(\frac{2-q}{1-q}\right)}\left(\frac{2}{1-q}\right)^{\frac{1}{1-q}}\right]^{\frac{2(1-q)}{2+D(1-q)}}, \tag{23}$$

式（4）中给出的 II 型皮尔逊分布的密度生成函数形式为

$$g(s)=\left(\frac{1-q}{2}\left(R^2-s\right)\right)_+^{1/(1-q)}. \tag{24}$$

相应的 II 型皮尔逊密度为

$$p(\mathbf{x}) = |\mathbf{\Sigma}|^{-1/2}\left[\frac{1-q}{2}\left(R^2 - (\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)\right]^{1/(1-q)} \quad q<1, (\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) < R^2.$$

这也可以等价地表示为 q-高斯密度的形式:

$$\begin{aligned} p(\mathbf{x}) &= \left[1 + (1-q)\left(-\frac{1}{1-q} + \frac{|\mathbf{\Sigma}|^{-(1-q)/2}}{2}\left(R^2 - (\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)\right)\right]_+^{1/(1-q)} \\ &= \exp_q\left[-\frac{1}{1-q} + \frac{|\mathbf{\Sigma}|^{-(1-q)/2}}{2}\left(R^2 - (\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)\right]. \end{aligned}$$

为方便起见，引入尺度参数的变换 $\widetilde{\mathbf{\Sigma}} = |\mathbf{\Sigma}|^{\frac{1-q}{2}}\mathbf{\Sigma}$ 并利用以下事实 $|\mathbf{\Sigma}| = |\widetilde{\mathbf{\Sigma}}|^{\frac{2}{1-q+D+2}}$，Martins 等人 [21,定义 15] 将密度参数化为以下形式 $\boldsymbol{\mu}$ 和 $\mathbf{\Sigma}$：

$$\mathbf{x} \sim N_q(\boldsymbol{\mu}, \widetilde{\mathbf{\Sigma}}) \Rightarrow p(\mathbf{x}) = \exp_q\left[-\frac{1}{1-q} + \frac{R^2|\widetilde{\mathbf{\Sigma}}|^{-\frac{1-q}{D(1-q)+2}}}{2} - \frac{(\mathbf{x}-\boldsymbol{\mu})^\top\widetilde{\mathbf{\Sigma}}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}\right].$$

Johnson [28] 详细讨论了形状参数 q的影响。对于 $q \to 1$，半径 $R \to \infty$，$\exp_q$ 收敛到普通指数函数，从而得到多元正态分布$N(\boldsymbol{\mu}, \mathbf{\Sigma})$。

□

作为伴随分布的伴随律。 利用有界支撑 $g(\cdot)$的密度生成函数q-高斯分布，该函数见式（8），累积生成函数定义为原始生成函数的积分:

$$G(s) := \int_s^{R^2} g(t)dt = \left(\tfrac{1-q}{2}\right)^{\frac{1}{1-q}}\frac{1-q}{2-q}(R^2-s)^{\frac{2-q}{1-q}} \propto (R^2-s)^{m+1}, \quad m := \frac{1}{1-q},$$

这会导出密度为

$$p^\star(\mathbf{x}) = \frac{|\mathbf{\Sigma}|^{-1/2}G\left(s(\mathbf{x})\right)}{\frac{\pi^{D/2}}{\Gamma(\frac{D}{2})}\int_0^{R^2} s^{\frac{D}{2}-1}G(s)ds} \propto |\mathbf{\Sigma}|^{-1/2}\left(R^2 - s(\mathbf{x})\right)_+^{m+1},$$

其中 $s(\mathbf{x}) := (\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$。根据构造， p和 $p^\star$ 具有相同的位置—尺度参数（$\boldsymbol{\mu}, \mathbf{\Sigma}$）以及相同的有界支撑集 supp $= \{\mathbf{x}: s(\mathbf{x}) < R^2\}$；其伴随分布是在相同支撑集上的一种新的 II 型皮尔逊分布，幂律型 II 类皮尔逊形式使其指数增加了 1。

我们也可以将其等价地表示为对基准分布 $p(\mathbf{x})$ 的归一化重加权:

$$p^\star(\mathbf{x}) = \frac{\left(R^2 - s(\mathbf{x})\right)p(\mathbf{x})}{\mathbb{E}_p\left[R^2 - s(\mathbf{x})\right]}, \tag{25}$$

伴随分布的伴随解释。第一伴随律与基准密度的（$2-q$）阶伴随分布一致。事实上，

$$p^\star(\mathbf{x}) \propto (R^2 - s(\mathbf{x}))_+^{m+1} = \left((R^2 - s(\mathbf{x}))_+^m\right)^{2-q} \propto p(\mathbf{x})^{2-q},$$

因为 $m+1 = \frac{2-q}{1-q} = (2-q)m$。更一般地，第 k 伴随律满足

$$p^{\langle k\rangle}(\mathbf{x}) \propto (R^2 - s(\mathbf{x}))_+^{m+k} \propto p(\mathbf{x})^{1+k(1-q)}, \qquad k = 0, 1, 2, \ldots,$$

因此，伴随分布的阶数每一步增加 $(1-q)$。所有这些分布具有相同的有界支撑集，并且逐渐更加集中于中心，在 $\boldsymbol{\mu}$ 处峰值更尖锐，且在边界附近比 $p(\mathbf{x})$衰减得更快。

C. 有界支撑 Stein 恒等式的证明

我们证明，适用于一般测试函数 $t(\mathbf{x})$的有界支撑 Stein 结论，而非更具特定性的 $f(\mathbf{x})$ ；后者是我们在定理1的主要陈述中为简便起见所采用的。我们假设 p 是一个参数为 q的有界支撑($\boldsymbol{\mu}$， $\boldsymbol{\Sigma}$， q) -高斯密度，如引理1所述；并且 $p^\star$ 是其第一个伴随（累积生成函数）分布。对于任意几乎处处可微的 $t:\ \mathbb{R}^D \to \mathbb{R}$，满足$\mathbb{E}_{p^\star}\|\nabla t(\mathbf{x})\| < \infty$,

$$\mathbb{E}_p\left[(\mathbf{x}-\boldsymbol{\mu})t(\mathbf{x})\right] = \mathrm{Cov}_p(\mathbf{x})\mathbb{E}_{p^\star}\left[\nabla t(\mathbf{x})\right] = \frac{\mathbb{E}_p\left[r^2\right]}{D}\boldsymbol{\Sigma}\mathbb{E}_{p^\star}\left[\nabla t(\mathbf{x})\right], \tag{26}$$

其中 $r^2 = s(\mathbf{x}) = (\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ 且 $\mathbb{E}_p\left[r^2\right]$ 由引理 2 给出。

定理 *1* 的证明。 设 $\mathbf{z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})$，从而

$$p(\mathbf{z}) \propto (R^2-\|\mathbf{z}\|^2)^m \mathbb{1}_{\|\mathbf{z}\|<R}, \qquad p^\star(\mathbf{z}) = \frac{(R^2-\|\mathbf{z}\|^2)p(\mathbf{z})}{\mathbb{E}_p\left[R^2-\|\mathbf{z}\|^2\right]} \propto (R^2-\|\mathbf{z}\|^2)^{m+1}\mathbb{1}_{\|\mathbf{z}\|<R}, \tag{27}$$

并定义 $T(\mathbf{z}) := t(\boldsymbol{\mu}+\boldsymbol{\Sigma}^{1/2}\mathbf{z})$。为计算 $\mathbb{E}_{p^\star}\left[\partial_{\mathbf{z}_i}T(\mathbf{z})\right]$，我们首先对 $\mathbf{z}_i$ 积分。固定指标 i。对每个满足 $\|\mathbf{z}_{-i}\| < R$ 的$\mathbf{z}_{-i} \in \mathbb{R}^{D-1}$，令

$$\rho(\mathbf{z}_{-i}) := \sqrt{R^2-\|\mathbf{z}_{-i}\|^2} \in (0, R],$$

从而 $\mathbf{z}_i$ 的容许范围恰好为

$$\mathbf{z}_i \in I(\mathbf{z}_{-i}) := [-\rho(\mathbf{z}_{-i}), \rho(\mathbf{z}_{-i})],$$

因为 $\|\mathbf{z}\|^2 < R^2 \implies \|\mathbf{z}_{-i}\|^2 + \mathbf{z}_i^2 < R^2 \implies |\mathbf{z}_i| < \rho(\mathbf{z}_{-i})$。对每个固定的 $\mathbf{z}_{-i}$，我们在区间 $I(\mathbf{z}_{-i})$ 上关于变量 $\mathbf{z}_i$ 应用一维分部积分:

$$\begin{aligned}
&\int_{-\rho(\mathbf{z}_{-i})}^{\rho(\mathbf{z}_{-i})} \partial_{\mathbf{z}_i}T(\mathbf{z}_{-i},\mathbf{z}_i)(R^2-\|\mathbf{z}_{-i}\|^2-\mathbf{z}_i^2)^{m+1}d\mathbf{z}_i \\
&= \left[T(\mathbf{z}_{-i},\mathbf{z}_i)(R^2-\|\mathbf{z}_{-i}\|^2-\mathbf{z}_i^2)^{m+1}\right]_{\mathbf{z}_i=-\rho(\mathbf{z}_{-i})}^{\rho(\mathbf{z}_{-i})} + \int_{-\rho(\mathbf{z}_{-i})}^{\rho(\mathbf{z}_{-i})} T(\mathbf{z}_{-i},\mathbf{z}_i)\partial_{\mathbf{z}_i}(R^2-\|\mathbf{z}_{-i}\|^2-\mathbf{z}_i^2)^{m+1}d\mathbf{z}_i.
\end{aligned}$$

由于 $m+1 > 0$，多项式 $(R^2-\|\mathbf{z}_{-i}\|^2-\mathbf{z}_i^2)^{m+1}$ 在 $\mathbf{z}_i = \pm\rho(\mathbf{z}_{-i})$ 处为零，因此每个切片的边界项都恰好为零。对多项式项求导可得

$$\partial_{\mathbf{z}_i}(R^2-\|\mathbf{z}_{-i}\|^2-\mathbf{z}_i^2)^{m+1} = -2(m+1)\mathbf{z}_i(R^2-\|\mathbf{z}_{-i}\|^2-\mathbf{z}_i^2)^m.$$

因此,

$$\int_{-\rho(\mathbf{z}_{-i})}^{\rho(\mathbf{z}_{-i})} \partial_{\mathbf{z}_i}T(\mathbf{z}_{-i},\mathbf{z}_i)(R^2-\|\mathbf{z}_{-i}\|^2-\mathbf{z}_i^2)^{m+1}d\mathbf{z}_i = \int_{-\rho(\mathbf{z}_{-i})}^{\rho(\mathbf{z}_{-i})} T(\mathbf{z}_{-i},\mathbf{z}_i)2(m+1)\mathbf{z}_i(R^2-\|\mathbf{z}_{-i}\|^2-\mathbf{z}_i^2)^m d\mathbf{z}_i. \tag{28}$$

接下来，我们对所有满足$\mathbf{z}_{-i}$ 且 $\|\mathbf{z}_{-i}\| < R$的情况积分。(28) 的左侧变为

$$\begin{aligned}
\int_{\mathbb{R}^{D-1}}\int_{\mathbb{R}} \left(\partial_{\mathbf{z}_i}T(\mathbf{z}_{-i},\mathbf{z}_i)\right)(R^2-\|\mathbf{z}_{-i}\|^2-\mathbf{z}_i^2)_+^{m+1}d\mathbf{z}_i d\mathbf{z}_{-i} &= \int_{\mathbb{R}^D}(\partial_{\mathbf{z}_i}T(\mathbf{z}))(R^2-\|\mathbf{z}\|^2)_+^{m+1}d\mathbf{z} \\
&= C^\star\mathbb{E}_{p^\star}\left[\partial_{\mathbf{z}_i}T(\mathbf{z})\right],
\end{aligned}$$

w此处 $C^\star$ 是 $p^\star$的归一化常数。最后，(28) 的右侧变为

$$2(m+1)\int_{\|\mathbf{z}\|<R} T(\mathbf{z})\mathbf{z}_i(R^2-\|\mathbf{z}\|^2)^m d\mathbf{z} = 2(m+1)C\mathbb{E}_p\left[\mathbf{z}_i T(\mathbf{z})\right],$$

其中， C 是 p的归一化常数。因此

$$\mathbb{E}_{p^\star}\left[\partial_{\mathbf{z}_i}T(\mathbf{z})\right] = \frac{2(m+1)C}{C^\star}\mathbb{E}_p\left[\mathbf{z}_i T(\mathbf{z})\right]. \tag{29}$$

利用关系式

$$\mathbb{E}_p\left[R^2-\|\mathbf{z}\|^2\right]=\frac{1}{C}\int(R^2-\|\mathbf{z}\|^2)(R^2-\|\mathbf{z}\|^2)^m d\mathbf{z}=\frac{C^\star}{C},$$

我们得到 $\frac{C}{C^\star}=\frac{1}{\mathbb{E}_p\left[R^2-\|\mathbf{z}\|^2\right]}$，因此（29）化简为

$$\mathbb{E}_{p^\star}\left[\partial_{\mathbf{z}_i}T(\mathbf{z})\right]=\frac{2(m+1)}{\mathbb{E}_p\left[R^2-\|\mathbf{z}\|^2\right]}\mathbb{E}_p\left[\mathbf{z}_i T(\mathbf{z})\right]. \tag{30}$$

对 i 进行向量化，我们得到

$$\mathbb{E}_{p^\star}\left[\nabla_{\mathbf{z}}T(\mathbf{z})\right]=\frac{2(m+1)}{\mathbb{E}_p\left[R^2-\|\mathbf{z}\|^2\right]}\mathbb{E}_p\left[\mathbf{z}T(\mathbf{z})\right].$$

回到 x-空间。 回想一下，$\mathbf{z}=\boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})$ 且 $\nabla_{\mathbf{z}}T(\mathbf{z})=\boldsymbol{\Sigma}^{1/2}\nabla_{\mathbf{x}}t(\boldsymbol{\mu}+\boldsymbol{\Sigma}^{1/2}\mathbf{z})$。取期望并乘以 $\boldsymbol{\Sigma}^{1/2}$，

$$\mathbb{E}_p\left[(\mathbf{x}-\boldsymbol{\mu})t(\mathbf{x})\right]=\boldsymbol{\Sigma}^{1/2}\mathbb{E}_p\left[\mathbf{z}T(\mathbf{z})\right]=\frac{\mathbb{E}_p\left[R^2-\|\mathbf{z}\|^2\right]}{2(m+1)}\boldsymbol{\Sigma}^{1/2}\mathbb{E}_{p^\star}\left[\nabla_{\mathbf{z}}T(\mathbf{z})\right]=\frac{\mathbb{E}_p\left[R^2-\|\mathbf{z}\|^2\right]}{2(m+1)}\boldsymbol{\Sigma}\mathbb{E}_{p^\star}\left[\nabla_{\mathbf{x}}t(\mathbf{x})\right].$$

由引理 2，

$$\frac{\mathbb{E}_p\left[r^2\right]}{D}=\frac{\mathbb{E}_p\left[R^2-r^2\right]}{2(m+1)},$$

因此系数等于 $\frac{\mathbb{E}_p\left[r^2\right]}{D}$，我们得到（10）。利用（25）将 $\mathbb{E}_{p^\star}$ 表示为 $\mathbb{E}_p$ 的形式，得到（12）。

□

D. q-Bonnet、q-Price 公式及方差界的证明

与附录 C 类似，在以下证明中，我们明确给出所使用测试函数的具体形式 $t(\mathbf{x})$，并指出在定理的主要陈述中，为简便起见，我们改用 $f(\mathbf{x})$。

D.1. q-邦内特

设 p 是引理 1 中的一个具有有界支撑集的 q-高斯分布密度 $N_q(\boldsymbol{\mu},\boldsymbol{\Sigma})$，其中 $q<1$ 且 $m:=1/(1-q)>0$。设 $f\colon\mathbb{R}^D\to\mathbb{R}$ 在包含 $\{s(\mathbf{x})\le R^2\}$ 的开集上满足 C^1，其中 $s(\mathbf{x}):=(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$，并假设 $\mathbb{E}_p\|\nabla f(\mathbf{x})\|<\infty$。我们证明

$$\nabla_{\boldsymbol{\mu}}\mathbb{E}_p\left[f(\mathbf{x})\right]=\mathbb{E}_p\left[\nabla f(\mathbf{x})\right].$$

定理 2 的证明。由于 $p(\mathbf{x})\propto|\boldsymbol{\Sigma}|^{-1/2}\left(R^2-s(\mathbf{x})\right)_+^m$，我们有

$$\log p(\mathbf{x})=-\tfrac{1}{2}\log|\boldsymbol{\Sigma}|+m\log\left(R^2-s(\mathbf{x})\right)+\text{const},\quad \nabla_{\boldsymbol{\mu}}s(\mathbf{x})=-2\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}).$$

因此

$$\nabla_{\boldsymbol{\mu}}\log p(\mathbf{x})=m\frac{\nabla_{\boldsymbol{\mu}}\left(R^2-s(\mathbf{x})\right)}{R^2-s(\mathbf{x})}=\frac{2m}{R^2-s(\mathbf{x})}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}).$$

由于有界支撑集上的控制收敛，且 $m>0$，在积分号下求导是合理的，因此我们得到得分恒等式

$$\nabla_{\boldsymbol{\mu}}\mathbb{E}_p\left[f(\mathbf{x})\right]=\mathbb{E}_p\left[f(\mathbf{x})\nabla_{\boldsymbol{\mu}}\log p(\mathbf{x})\right]=2m\mathbb{E}_p\left[\frac{f(\mathbf{x})}{R^2-s(\mathbf{x})}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]. \tag{31}$$

应用 p-only Stein 恒等式（12），其中函数为 $t(\mathbf{x})=\frac{f}{}(\mathbf{x})R^2-s(\mathbf{x})$。通过商法则代入（12）的右端，并利用 $\nabla s(\mathbf{x})=2\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$，得到

$$(R^2-s(\mathbf{x}))\nabla\left(\frac{f(\mathbf{x})}{R^2-s(\mathbf{x})}\right)=\nabla f(\mathbf{x})+2\frac{f(\mathbf{x})}{R^2-s(\mathbf{x})}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}).$$

因此（12）得到

$$\boldsymbol{\Sigma}\mathbb{E}_p\left[\frac{f(\mathbf{x})}{R^2-s(\mathbf{x})}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]=\frac{\mathbb{E}_p\left[\nabla f(\mathbf{x})\right]+2\mathbb{E}_p\left[\frac{f(\mathbf{x})}{R^2-s(\mathbf{x})}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]}{\frac{\mathbb{E}_p\left[r^2\right]}{D}}\boldsymbol{\Sigma}\mathbb{E}_p\left[R^2-s(\mathbf{x})\right]. \tag{32}$$

根据引理中的矩2公式，我们有

$$\frac{\mathbb{E}_p\left[r^2\right]}{D\mathbb{E}_p\left[R^2-s(\mathbf{x})\right]}=\frac{1}{2(m+1)}.$$

因此，32) 得出

$$\begin{aligned}\mathbb{E}_p\left[\frac{f(\mathbf{x})}{R^2-s(\mathbf{x})}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]&=\frac{1}{2(m+1)}\left(\mathbb{E}_p\left[\nabla f(\mathbf{x})\right]+2\mathbb{E}_p\left[\frac{f(\mathbf{x})}{R^2-s(\mathbf{x})}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]\right)\\&=\frac{1}{2m}\mathbb{E}_p\left[\nabla f(\mathbf{x})\right].\end{aligned}$$

将其代入（31）：

$$\nabla_{\boldsymbol{\mu}}\mathbb{E}_p\left[f(\mathbf{x})\right]=2m\mathbb{E}_p\left[\frac{f(\mathbf{x})}{R^2-s(\mathbf{x})}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]=\mathbb{E}_p\left[\nabla f(\mathbf{x})\right].$$

□

D.2. q-Price

引理 3（矩阵微积分）。设 $\boldsymbol{\Sigma}\succ 0$。我们使用 *Frobenius* 内积 $A:\ B:=\mathrm{tr}(A^\top B)$。对于 $s(\mathbf{x})=(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ 以及在 (i,j) 位置为 *1*、其余位置为 *0* 的基本矩阵 $\mathbf{E}^{ij}$：

$$\frac{\partial}{\partial\boldsymbol{\Sigma}_{ij}}\log|\boldsymbol{\Sigma}|=(\boldsymbol{\Sigma}^{-1})_{ji}, \tag{33}$$

$$\frac{\partial s(\mathbf{x})}{\partial\boldsymbol{\Sigma}_{ij}}=-\left(\boldsymbol{\Sigma}^{-1}\mathbf{E}_{ji}\boldsymbol{\Sigma}^{-1}\right):\left((\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\right)=-\left(\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}\right)_{ji}. \tag{34}$$

此外，对于任意矩阵 $\mathbf{H}$，

$$\left(\boldsymbol{\Sigma}^{-1}\mathbf{E}_{ji}\boldsymbol{\Sigma}^{-1}\right):\boldsymbol{\Sigma}=(\boldsymbol{\Sigma}^{-1})_{ji}, \tag{35}$$

$$\left(\boldsymbol{\Sigma}^{-1}\mathbf{E}_{ji}\boldsymbol{\Sigma}^{-1}\right):(\boldsymbol{\Sigma}\mathbf{H}\boldsymbol{\Sigma})=\mathbf{H}_{ji}. \tag{36}$$

与之前一样，设 p 为有界支撑集的 q-高斯密度 $N_q(\boldsymbol{\mu},\boldsymbol{\Sigma})$，其中 $q<1$ 且形状参数为 $m=\frac{1}{1-q}>0$，支撑集为 $\{s(\mathbf{x})<R^2\}$。在定理 3 的证明中，我们假设 $f:\ \mathbb{R}^D\to\mathbb{R}$在包含 $\{s(\mathbf{x})\le R^2\}$的开集上是 C^2，并且 $\mathbb{E}_p\|\nabla f(\mathbf{x})\|+\mathbb{E}_p\|\nabla^2 f(\mathbf{x})\|_F<\infty$。我们定义 M，并回顾协方差的形式

$$M:=\mathbb{E}_p\left[R^2-s(\mathbf{x})\right],\qquad \mathrm{Cov}_p(\mathbf{x})=\frac{\mathbb{E}_p[r^2]}{D}\boldsymbol{\Sigma}.$$

通常，$p^\star$ 表示 p的第一伴随律，因此对于任意可积的h，$\mathbb{E}_{p^\star}[h]=\mathbb{E}_p[(R^2-s)h]/M$。将条目 $\boldsymbol{\Sigma}_{ij}$ 视为独立参数，对于每个 $i,\ j\in\{1,\dots,D\}$，以下恒等式成立：

$$\frac{\partial}{\partial\boldsymbol{\Sigma}_{ij}}\mathbb{E}_p\left[f(\mathbf{x})\right]=\frac{\mathbb{E}_p[r^2]}{2D}\mathbb{E}_{p^\star}\left[\frac{\partial^2 f(\mathbf{x})}{\partial x_i\partial x_j}\right], \tag{37}$$

等价地，用矩阵形式表示为，

$$\nabla_{\boldsymbol{\Sigma}}\mathbb{E}[f(\mathbf{x})]=\frac{\mathbb{E}_p[r^2]}{2D}\mathbb{E}_{p^\star}[\nabla^2_{\mathbf{x}}f(\mathbf{x})] \tag{38}$$

在高斯极限下 $q\uparrow 1$(即 $m\to\infty,\ R\to\infty$)，$p^\star=p$且 $\mathbb{E}_p[r^2]/D\to 1$，因此（37）化为经典 Price 定理 $\frac{\partial}{\partial\boldsymbol{\Sigma}_{ij}}\mathbb{E}[f(\mathbf{x})]=\frac{1}{2}\mathbb{E}[\partial_{x_i}\partial_{x_j}f(\mathbf{x})]$。

定理3的证明。 **关于 $\boldsymbol{\Sigma}$ 的得分技巧**。 根据积分号下求导（由假设和有界支撑集保证其合理性），

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_{ij}} \mathbb{E}_p[f(\mathbf{x})] = \mathbb{E}_p\left[f(\mathbf{x}) \frac{\partial}{\partial \boldsymbol{\Sigma}_{ij}} \log p(\mathbf{x}) \right].$$

对于 $p(\mathbf{x}) \propto |\boldsymbol{\Sigma}|^{-1/2}(R^2 - s(\mathbf{x}))_+^m$,

$$\log p(\mathbf{x}) = -\tfrac{1}{2} \log |\boldsymbol{\Sigma}| + m \log(R^2 - s(\mathbf{x})) + \text{const.}$$

利用引理 3,

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_{ij}} \log p(\mathbf{x}) = -\tfrac{1}{2} (\boldsymbol{\Sigma}^{-1})_{ji} + \frac{m}{R^2 - s(\mathbf{x})} \left(\boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} \right)_{ji}.$$

定义矩阵

$$B := \mathbb{E}_p\left[\frac{(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top}{R^2 - s(\mathbf{x})} f(\mathbf{x}) \right].$$

于是，以矩阵形式表示,

$$\nabla_{\boldsymbol{\Sigma}} \mathbb{E}_p[f(\mathbf{x})] = -\tfrac{1}{2} \boldsymbol{\Sigma}^{-1} \mathbb{E}_p[f(\mathbf{x})] + m \boldsymbol{\Sigma}^{-1} B \boldsymbol{\Sigma}^{-1}. \tag{39}$$

通过 Stein 恒等式识别 B。 以仅含 p 的形式应用有界支撑 Stein 恒等式:

$$\mathbb{E}_p\left[(\mathbf{x} - \boldsymbol{\mu}) t(\mathbf{x})^\top \right] = \frac{\mathbb{E}_p[r^2]}{D} \boldsymbol{\Sigma} \frac{\mathbb{E}_p\left[(R^2 - s(\mathbf{x})) \nabla t(\mathbf{x})^\top \right]}{M}, \qquad M := \mathbb{E}_p[R^2 - s(\mathbf{x})].$$

选择向量测试函数

$$t(\mathbf{x}) := \frac{f(\mathbf{x})}{R^2 - s(\mathbf{x})} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}).$$

左侧。根据 B的定义,

$$\mathbb{E}_p\left[(\mathbf{x} - \boldsymbol{\mu}) t(\mathbf{x})^\top \right] = \mathbb{E}_p\left[\frac{(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top}{R^2 - s(\mathbf{x})} f(\mathbf{x}) \right] \boldsymbol{\Sigma}^{-1} = B \boldsymbol{\Sigma}^{-1}.$$

右侧。计算

$$(R^2 - s) \nabla t = \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \nabla f(\mathbf{x})^\top + 2 \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} \frac{f(\mathbf{x})}{R^2 - s(\mathbf{x})} + f(\mathbf{x}) \boldsymbol{\Sigma}^{-1},$$

利用 $\nabla s(\mathbf{x}) = 2\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ 以及乘积法则。取 $\mathbb{E}_p$ 并代入 Stein 恒等式，得到

$$B \boldsymbol{\Sigma}^{-1} = c \left(\boldsymbol{\Sigma} \mathbb{E}_p[\nabla f(\mathbf{x})(\mathbf{x} - \boldsymbol{\mu})^\top] \boldsymbol{\Sigma}^{-1} + 2B \boldsymbol{\Sigma}^{-1} + \mathbb{E}_p[f(\mathbf{x})] \mathbf{I} \right), \qquad c := \frac{\mathbb{E}_p[r^2]}{DM}.$$

在右侧乘以 $\boldsymbol{\Sigma}$ 并重新整理，得到

$$(1 - 2c) B = c \left(\boldsymbol{\Sigma} \mathbb{E}_p[\nabla f(\mathbf{x})(\mathbf{x} - \boldsymbol{\mu})^\top] + \mathbb{E}_p[f(\mathbf{x})] \boldsymbol{\Sigma} \right).$$

对于 有界支撑 q-高斯分布，引理 2 给出 $c = \mathbb{E}_p[r^2]/(DM) = 1/(2(m+1))$，因此

$$B = \frac{1}{2m} \left(\boldsymbol{\Sigma} \mathbb{E}_p[\nabla f(\mathbf{x})(\mathbf{x} - \boldsymbol{\mu})^\top] + \mathbb{E}_p[f(\mathbf{x})] \boldsymbol{\Sigma} \right). \tag{40}$$

代入 B **得分形式中**。将（40）代入（39）：

$$\nabla_{\boldsymbol{\Sigma}} \mathbb{E}_p[f(\mathbf{x})] = -\tfrac{1}{2} \boldsymbol{\Sigma}^{-1} \mathbb{E}_p[f(\mathbf{x})] + \frac{m}{2m} \left(\boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma} \mathbb{E}_p[\nabla f(\mathbf{x})(\mathbf{x} - \boldsymbol{\mu})^\top] \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1} \mathbb{E}_p[f(\mathbf{x})] \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-1} \right).$$

$\pm \frac{1}{2} \boldsymbol{\Sigma}^{-1} \mathbb{E}_p[f(\mathbf{x})]$ 项相消，剩下

$$\nabla_{\boldsymbol{\Sigma}} \mathbb{E}_p[f(\mathbf{x})] = \frac{1}{2} \mathbb{E}_p\left[(\mathbf{x} - \boldsymbol{\mu}) \nabla f(\mathbf{x})^\top \right] \boldsymbol{\Sigma}^{-1}. \tag{41}$$

进行第二次 Stein 步以得到 Hessian矩阵。 对标量检验函数$t_j(\mathbf{x})$ 应用（向量）Stein 恒等式：$= \partial_{x_j} f(\mathbf{x})$、 $j = 1, \ldots, D$，并将它们堆叠起来：

$$\mathbb{E}_p\left[(\mathbf{x}-\boldsymbol{\mu})\nabla f(\mathbf{x})^\top\right] = \frac{\mathbb{E}_p[r^2]}{D}\boldsymbol{\Sigma}\mathbb{E}_{p^\star}\left[\nabla^2_{\mathbf{x}} f(\mathbf{x})\right].$$

将其代入（41）可得

$$\nabla_{\boldsymbol{\Sigma}}\mathbb{E}_p[f(\mathbf{x})] = \frac{\mathbb{E}_p[r^2]}{2D}\mathbb{E}_{p^\star}\left[\nabla^2_{\mathbf{x}} f(\mathbf{x})\right],$$

这就是（15）。

□

备注 2（ $\boldsymbol{\Sigma}$中的对称性）。 我们对元素 $\boldsymbol{\Sigma}_{ij}$ 进行求导时，将 $\boldsymbol{\Sigma}$视为无约束矩阵。（15）右端是对称的，因为 $\nabla^2_{\mathbf{x}} f$（$\mathbf{x}$）对于 C^2函数 f是对称的。因此，梯度 $\nabla_{\boldsymbol{\Sigma}}\mathbb{E}_p[f$（$\mathbf{x}$）] 自然属于对称矩阵空间，并且与约束 $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^\top$一致。

D.3. 方差界

命题 *1* 的证明。 无偏性很容易验证： $\mathbb{E}\left[\widehat{\ }\mathbf{g}\right] = \mathbb{E}_{p^\star}\left[\nabla t\right.$（$\mathbf{x}$）] 和 $\mathbb{E}\left[\widehat{\ }\mathbf{H}\right] = \mathbb{E}_{p^\star}\left[\nabla^2 f\right.$（$\mathbf{x}$）]。

注意， $\widehat{g}_j$ 是标量，令

$$Y_k := \frac{(R^2 - s(\mathbf{x}_k))\partial_j t(\mathbf{x}_k)}{\mathbb{E}_p\left[R^2 - s(\mathbf{x})\right]}.$$

由于 $0 < R^2 - s(\mathbf{x}) \le R^2$ 且 $|\partial_j t(\mathbf{x})| \le \|\nabla t(\mathbf{x})\| \le C_1$，因此

$$|Y_k| \le \frac{R^2 C_1}{\mathbb{E}_p\left[R^2 - s(\mathbf{x})\right]} =: B_g.$$

因此 $Y_k \in [-B_g, B_g]$ 几乎必然成立，并由有界值域方差的波波维丘不等式可得，

$$\mathrm{Var}(Y_k) \le B_g^2.$$

由于 $\widehat{g}_j = \frac{1}{S}\sum_{k=1}^n Y_k$ 且 Y_k独立同分布，

$$\mathrm{Var}(\widehat{g}_j) = \frac{\mathrm{Var}(Y_1)}{S} \le \frac{B_g^2}{S} = \frac{1}{S}\left(\frac{R^2 C_1}{M}\right)^2.$$

接下来，固定 (i, j) 并定义

$$Z_k := \frac{(R^2 - s(\mathbf{x}_k))(\nabla^2 f(\mathbf{x}_k))_{ij}}{M}.$$

由于 $0 < (R^2 - s(\mathbf{x})) \le R^2$ 且 $|(\nabla^2 f(\mathbf{x}))_{ij}| \le \|\nabla^2 f(\mathbf{x})\|_{\mathrm{op}} \le C_2$，因此

$$|Z_k| \le \frac{R^2 C_2}{M} =: B_H,$$

因此 $Z_k \in [-B_H, B_H]$ 几乎必然成立；再次由 Popoviciu 定理可得，$\mathrm{Var}(Z_k) \le B_H^2$，并且

$$\mathrm{Var}\left((\widehat{H})_{ij}\right) = \mathrm{Var}\left(\frac{1}{S}\sum_{k=1}^n Z_k\right) = \frac{\mathrm{Var}(Z_1)}{S} \le \frac{B_H^2}{S} = \frac{(R^2 C_2)^2}{SM^2}.$$

对各元素的方差求和可得

$$\mathbb{E}\left[\left\|\widehat{\mathbf{H}} - \mathbb{E}\left[\widehat{\mathbf{H}}\right]\right\|_F^2\right] = \sum_{i,j}\mathrm{Var}\left((\widehat{\mathbf{H}})_{ij}\right) \le \frac{D^2 B_H^2}{S}.$$

由于 $\|\cdot\|_{\mathrm{op}} \le \|\cdot\|_F$，同样的界意味着

$$\mathbb{E}\left[\left\|\widehat{\mathbf{H}} - \mathbb{E}\left[\widehat{\mathbf{H}}\right]\right\|_{\mathrm{op}}\right] \le \frac{DB_H}{\sqrt{S}}.$$

通过矩阵 Hoeffding 不等式，可以得到更精细的维度依赖：定义中心化的

$$\mathbf{A}_k := \frac{(R^2 - s(\mathbf{x}_k))\nabla^2 f(\mathbf{x}_k)}{M} - \mathbb{E}\left[\widehat{\mathbf{H}}\right]$$

意味着 $\|\mathbf{A}_k\|_{\mathrm{op}} \le 2B_H$。因此，由于 $\frac{1}{S}\sum_{k=1}^{S}\mathbf{A}_k = \widehat{\mathbf{H}} - \mathbb{E}\left[\widehat{\mathbf{H}}\right]$，我们有

$$\Pr\left(\left\|\widehat{\mathbf{H}} - \mathbb{E}\left[\widehat{\mathbf{H}}\right]\right\|_{\mathrm{op}} \ge t\right) \le 2D\exp\left(-\frac{St^2}{8(2B_H)^2}\right),$$

积分后得到 $\mathbb{E}\left[\left\|\widehat{\mathbf{H}} - \mathbb{E}\left[\widehat{\mathbf{H}}\right]\right\|_{\mathrm{op}}\right] \le C_3 B_H\sqrt{\frac{\log(2D)}{S}}$，其中 C_3 为普适常数：沿用上文相同的尾部积分论证，并在 $t_0 > 0$ 处分段，

$$\mathbb{E}\left[\left\|\widehat{\mathbf{H}} - \mathbb{E}\left[\widehat{\mathbf{H}}\right]\right\|_{\mathrm{op}}\right] = \int_0^\infty \Pr(\left\|\widehat{\mathbf{H}} - \mathbb{E}\left[\widehat{\mathbf{H}}\right]\right\|_{\mathrm{op}} \ge t)dt \le t_0 + \int_{t_0}^\infty 2D\exp\left(-\frac{St^2}{8(2B_H)^2}\right)dt.$$

令 $a := \frac{S}{8(2B_H)^2}$，并利用高斯尾界 $\int_x^\infty e^{-at^2}dt \le \frac{1}{2ax}e^{-ax^2}$，

$$\mathbb{E}\left[\left\|\widehat{\mathbf{H}} - \mathbb{E}\left[\widehat{\mathbf{H}}\right]\right\|_{\mathrm{op}}\right] \le t_0 + 2D\cdot\frac{1}{2at_0}e^{-at_0^2}.$$

选择 $t_0 := \sqrt{\frac{\log(2D)}{a}}$，使得 $e^{-at_0^2} = 1/(2D)$。于是

$$\mathbb{E}\left[\left\|\widehat{\mathbf{H}} - \mathbb{E}\left[\widehat{\mathbf{H}}\right]\right\|_{\mathrm{op}}\right] \le \sqrt{\frac{\log(2D)}{a}} + \frac{1}{2\sqrt{a}}\frac{1}{\sqrt{\log(2D)}} \le C_4\frac{1}{\sqrt{a}}\sqrt{\log(2D)} = C_3 B_H\sqrt{\frac{\log(2D)}{S}},$$

对于通用常数 C_4，$C_3 > 0$，如所述。 □