A Stein Identity for q-Gaussians with Bounded Support

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

[*]RIKEN Center for AI Project, Tokyo, Japan
[†]Instituto de Telecomunicações, Instituto Superior Técnico, Universidade de Lisboa, Portugal

[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's identity is a fundamental tool in machine learning with applications in generative models, stochastic optimization, and other problems involving gradients of expectations under Gaussian distributions. Less attention has been paid to problems with non-Gaussian expectations. Here, we consider the class of bounded-support q-Gaussians and derive a new Stein identity leading to gradient estimators which have nearly identical forms to the Gaussian ones, and which are similarly easy to implement. We do this by extending the previous results of Landsman, Vanduffel, and Yao (2013) to prove new Bonnet- and Price-type theorems for q-Gaussians. We also simplify their forms by using *escort* distributions. Our experiments show that bounded-support distributions can reduce the variance of gradient estimators, which can potentially be useful for Bayesian deep learning and sharpness-aware minimization. Overall, our work simplifies the application of Stein's identity for an important class of non-Gaussian distributions.

1. Introduction

Stein's identity has been popular in machine learning for estimating gradients of $\mathbb{E}_p[f(\mathbf{x})]$, where the expectation of a real-valued differentiable function $f : \mathbb{R}^D \to \mathbb{R}$ is taken with respect to a Gaussian density $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Applications arise in a wide-variety of machine-learning problems including stochastic optimization [1], deep generative models [2], and variational inference [3, 4]. Stein's identity states that the following equality holds:

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

First proved by Charles Stein [5, 6], the identity results from integration by parts and holds under mild conditions on f (see [7, App. A.2]).

The Stein identity is used to express gradients with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in terms of the gradient and Hessian of $f(\mathbf{x})$, respectively (derivations are in App. A):

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

These expressions, known as Bonnet's [8] and Price's [9] theorems have received considerable attention in deep learning due to their convenient forms. Stochastic gradients with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are obtained by first sampling $\mathbf{x} \sim p(\mathbf{x})$ and then computing $\nabla_{\mathbf{x}} f(\mathbf{x})$ and $\nabla^2_{\mathbf{x}} f(\mathbf{x})$ at the sample. This line of thought has led to several variants of the "reparameterization trick" and other path-wise gradient estimators, which often have lower variance than score-function estimators [10]. Overall, Stein's identity has had a profound impact on stochastic-gradient estimation techniques [11].

Extensions of Stein's identity to non-Gaussian distributions have received much less attention than the Gaussian case. Generalizations do exist, for instance to elliptical families, which contain many

Third Conference on Parsimony and Learning (CPAL 2026).

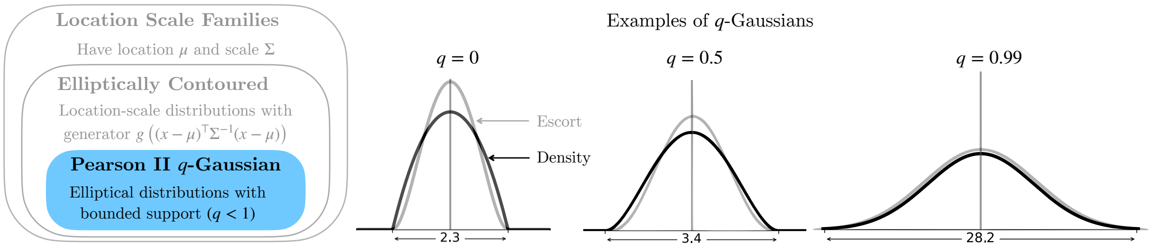

Figure 1: Bounded support q-Gaussians are a subclass of location-scale families that are elliptically contoured, that is, they are obtained by composing a quadratic $s(\mathbf{x}) = (\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ with a generator function g, thus $p(\mathbf{x}) = g\big(s(\mathbf{x})\big)$. On the right, we show three examples of q-Gaussians for $q = 0, 0.5$, and 0.99. We see for larger q the base densities (black curves) are less peaked and have larger support. We also show the first associated *escort* $2-q$-Gaussian densities (gray curves), which are slightly more peaked than their base densities. As $q \to 1$, q-Gaussians converge to Gaussians.

interesting location-scale families, including the Gaussian distribution. In particular, Landsman and Nešlehová [12] and Landsman et al. [13, 14] have studied such generalizations focusing on the Pearson VII class, which contains various heavy-tailed distributions. The usefulness of such generalizations for gradient estimation has, to the best of our knowledge, not been explored. Can gradient estimators derived for non-Gaussian families also take simple forms and are they as easy to implement? The goal of this paper is answer these questions.

In this paper, we derive a new Stein identity for the Pearson II class of elliptical families called bounded-support q-Gaussian distributions (Fig. 1). Such bounded-support distributions are interesting for gradient estimation because, unlike Gaussians, samples drawn from them always land inside a bounded interval, which also naturally implies a bound on the gradient variance. We derive Bonnet- and Price-type theorems for this class of q-Gaussian distributions and show that the gradient estimators have nearly-identical form to those of Gaussian stochastic gradients, and are similarly easy to implement. This means we can compute unbiased gradient estimates by evaluating gradients and Hessians of $f(\mathbf{x})$ at samples $\mathbf{x} \sim p(\mathbf{x})$. These gradient expressions are obtained by using a distribution associated with the base distribution known in information geometry and statistical physics [15–17] as the *escort* distribution (Fig. 1). We present numerical experiments confirming the bounded variance property, and compare them to Bayesian deep-learning methods and to sharpness-aware minimization [18]. Overall, our work simplifies the application of Stein's identity to gradient estimation involving an important class of bounded-support q-Gaussian distributions.

2. Bounded-Support q-Gaussian Distributions

We present a new Stein identity applied to gradient estimation involving bounded-support q-Gaussian distributions. Such distributions are a special type of location-scale distributions, obtained by applying a specific *generator* function g to a Gaussian-like quadratic form. We denote the location by $\boldsymbol{\mu}$ and the elliptical scale matrix by $\boldsymbol{\Sigma}$. A D-variate elliptical density has the form

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

and $g : \mathbb{R} \to \mathbb{R}_+$ is a non-negative, density-generator function, appropriately scaled so that $p(\mathbf{x})$ is normalized, that is, it integrates to 1. For example, Gaussians are obtained by setting $g(s) \propto \exp(-s/2)$. The normalization condition can be written as

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

where Γ is the Gamma function. If the second moments exists, the covariance matrix $\mathrm{Cov}(\mathbf{x})$ is proportional to $\boldsymbol{\Sigma}$. In the Gaussian case, $\mathrm{Cov}(\mathbf{x}) = \boldsymbol{\Sigma}$, but otherwise $\boldsymbol{\Sigma}$ should be thought of as a scale or dispersion parameter, which is proportional to the covariance.

Bounded-support q-Gaussians belong to a class of distribution known as Pearson Type II [19], defined on the radius-R ellipsoid $\{\mathbf{x} \in \mathbb{R}^D : (\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) < R^2\}$, where $R > 0$. Given R, for all s in the interval $0 \leq s < R^2$, the generator function takes the following form,

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

and m is the shape parameter of the generator. We define $m = 1/(1-q)$ in terms of another scalar $q < 1$, which gives rise to q-Gaussian designation that is well-known in information geometry and statistical physics [15–17]. The max support radius R depends only on q and the dimensionality D, as shown by the following lemma that gives an explicit form of $p(\mathbf{x})$ (the proof is in App. B).

Lemma 1. *The density of the generalized Pearson Type II subfamily with $q < 1$ can be written as a q-Gaussian:*

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

where $\exp_q(t) := [1 + t/m]_+^m$ denotes the q-deformed exponential function [20], with $m = 1/(1-q)$ and $[u]_+ = \max(0, u)$. The support radius R is a function of q and D and is given by

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

These distributions have recently been studied by Martins et al. [21] as sparse continuous distributions induced by Tsallis entropies [22]. By using the definition of $\exp_q$, the same density can be expressed compactly as

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

which will be useful later to highlight differences relative to the required *escort* distribution.

Landsman and Nešlehová [12], Landsman et al. [14**?**] prove an extension of Stein's identity to the class of absolutely continuous elliptical distributions, with a special focus on heavy-tailed distributions belonging to Pearson Type VII family. Their results are based on canonical work on the properties of elliptical families [23–26]. Specifically, Landsman et al. [14, Prop. 2] give a general proof of elliptical Stein, but do not treat the bounded-support case on which we focus in this paper; see also [13, Lemma 2] and [12]. Notably, stochastic gradient estimators based on generalized elliptical Stein identities have not been studied and we are not aware of any works showing that such estimators can be brought into a form similar to those of Gaussian stochastic gradients. Our goal in this paper is to fill this gap.

3. A Stein-type identity for bounded-support q-Gaussians

In this section we derive a Stein-type identity tailored to bounded-support q-Gaussians ($q < 1$). We follow the approach of Landsman and Nešlehová [12], Landsman et al. [13, 14], using the *associated* density p^* to derive a Stein-type identity. We show in App. B that the first associated law coincides with the $(2-q)$-*escort* density $p^\star(\mathbf{x}) \propto p(\mathbf{x})^{2-q}$ [27]. As far as we know this connection between the *associated laws* defined in the classical statistical literature on elliptical families [24, 28], and the *escort* distributions studied in statistical physics and information geometry [17, 27] has not been noticed before. The use of escort distributions is instrumental in extending Stein's identity in an elegant way, mirroring the Gaussian Stein expression, and highlights the probabilistic structure that would otherwise be buried in repeated integration by parts.

The associated density $p^\star$ is defined through the generator function that is obtained by integrating the original generator g over the interval (s, R^2),

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

As we review in App. B, using this generator yields another Pearson II distribution with exponent increased by 1,

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

Comparing this expression to Eq. 7 shows that the only difference is that m in the exponent is replaced by $m+1$. Note that p and $p^\star$ share the same location-scale parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and bounded max support radius R, but $p^\star$ has a sharper peak at $\boldsymbol{\mu}$ because of the larger exponent. This difference is visible in Fig. 1.

In App. C we prove the following theorem which establishes a new Stein-type identity for bounded-support q-Gaussian distributions.

Theorem 1 (Bounded-support q-Gaussian Stein identity)**.** *For any almost everywhere differentiable $f : \mathbb{R}^D \to \mathbb{R}$, with $\mathbb{E}_{p^\star} \|\nabla f(\mathbf{x})\| < \infty$, when the second moment $\mathrm{Cov}_p(\mathbf{x})$ exists:*

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

The form of Eq. 10 is nearly identical to that of Eq. 1, with the difference that the expectation on the right-hand side uses the escort $p^\star$ instead of the original p (as highlighted in red). The proof in App. C proceeds by defining $\mathbf{z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})$, applying iterative one-dimensional integration by parts, and exploiting the fact that the Pearson II density and its associated law vanish at the boundary, as well as the facts described in Lemma 2 below.

Lemma 2. *The following facts hold for any bounded-support q-Gaussian with $q < 1$,*

1. *Defining $r(\mathbf{x}) = \sqrt{s(\mathbf{x})}$, the following holds under p and $p^\star$ respectively,*
$$\frac{r(\mathbf{x})^2}{R^2} \sim \mathrm{Beta}\left(\tfrac{D}{2}, m+1\right), \qquad \textit{and} \qquad \frac{r(\mathbf{x})^2}{R^2} \sim \mathrm{Beta}\left(\tfrac{D}{2}, m+2\right).$$
2. *The expectations of $s(\mathbf{x}) = r(\mathbf{x})^2$ have a closed-form expression in terms of R, D, and m,*
$$\mathbb{E}_p\left[s(\mathbf{x})\right] = \frac{D\,R^2}{D+2(m+1)}, \qquad \textit{and} \qquad \mathbb{E}_{p^\star}\left[s(\mathbf{x})\right] = \frac{D\,R^2}{D+2\,(m+2)}.$$
3. *The second moment can be written in terms of $\boldsymbol{\Sigma}$,*
$$\mathrm{Cov}_p(\mathbf{x}) = \frac{1}{D}\mathbb{E}_p[s(\mathbf{x})]\boldsymbol{\Sigma}, \qquad \textit{and} \qquad \mathrm{Cov}_{p^\star}(\mathbf{x}) = \frac{1}{D}\mathbb{E}_{p^\star}[s(\mathbf{x})]\boldsymbol{\Sigma}.$$
4. *Finally, we have the following reweighted representation of $p^\star$,*
$$p^\star(\mathbf{x}) = \frac{(R^2 - s(\mathbf{x}))\,p(\mathbf{x})}{\mathbb{E}_p\left[R^2 - s(\mathbf{x})\right]}.$$

All of these results follow from the application of the definitions of the Beta expectation and the covariance. App. B contains additional details.

There are two other useful variants of the identity,

$$\begin{aligned} \mathbb{E}_p\left[(\mathbf{x}-\boldsymbol{\mu}) f(\mathbf{x})\right] &= \frac{1}{D}\mathbb{E}_p\left[s(\mathbf{x})\right]\,\boldsymbol{\Sigma}\,\mathbb{E}_{p^\star}\left[\nabla_{\mathbf{x}} f(\mathbf{x})\right] && (11)\\ &= \frac{1}{D}\mathbb{E}_p\left[s(\mathbf{x})\right]\,\boldsymbol{\Sigma}\,\frac{\mathbb{E}_p\left[(R^2 - s(\mathbf{x}))\nabla_{\mathbf{x}} f(\mathbf{x})\right]}{\mathbb{E}_p\left[R^2 - s(\mathbf{x})\right]}. && (12) \end{aligned}$$

In the first line, we expand the second moment in terms of $\boldsymbol{\Sigma}$, and in the second line, we rewrite the $p^\star$-expectation in terms of the base density $p(\mathbf{x})$ by using the last fact of Lemma 2. The form of Eq. 12 is useful specifically for implementation because it expresses the identity in terms of $p(\mathbf{x})$ only. Sampling from $p(\mathbf{x})$ is comparably efficient to sampling from a Gaussian, as we show next.

Efficient sampling from $p(\mathbf{x})$ is possible by using Lemma 2. We define $\mathbf{z}(\mathbf{x}) = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})$, which can be rewritten in terms of an independent random variable $\mathbf{u}$ that is uniformly distributed on a sphere in $\mathbb{R}^D$ as $\mathbf{z}(\mathbf{x}) = r(\mathbf{x})\mathbf{u}$, where $\mathbf{u} \sim \mathrm{Unif}(S^{D-1})$ and $S^{D-1} = \{\mathbf{u} \in \mathbb{R}^D : \|\mathbf{u}\|_2 = 1\}$. Further, $r(\mathbf{x})^2/R^2$ is an independent Beta random variable. Thus sampling proceeds in the following four steps:

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

The radial parametrization reviewed in App. B.

4. Bonnet- and Price-type Theorems

We state Bonnet- and Price-type theorems for bounded-support q-Gaussians, for which proofs can be found in App. D. These theorems are brought into forms that are either identical or structurally analogous to the corresponding Gaussian theorems. The stochastic gradient expressions we derive are easy to estimate using Monte-Carlo sampling since the distributions over which the expectations are defined are easy to sample from by exploiting Eq. 13.

We begin with the Bonnet-type theorem, paralleling the first equality in Eq. 2.

Theorem 2 (q-Bonnet). *For bounded-support $p(\mathbf{x}) = \mathcal{N}_q(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, assume $f : \mathbb{R}^D \to \mathbb{R}$ to be C^1 on an open set containing $\{\mathbf{x} : s(\mathbf{x}) \leq R^2\}$, and $\mathbb{E}_p \|\nabla f(\mathbf{x})\| < \infty$. Then,*

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

This identity has exactly the same form as the first equality in Eq. 2. The proof is in App. D.1 and follows from the form of $p(\mathbf{x})$ given in Eq. 7 and the fact that $\nabla_{\boldsymbol{\mu}} s(\mathbf{x}) = -2\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ which allows us to write the gradient as

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

Differentiating under the integral, the left-hand side in Eq. 14 can be written in terms of $\mathbb{E}_p \left[f(\mathbf{x}) \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})/(R^2 - s(\mathbf{x})) \right]$. Finally, the Stein-type identity in Eq. 12 and some algebra lead to the desired result.

Next, we state the Price-type theorem which parallels the second equality in Eq. 2:

Theorem 3 (q-Price). *For bounded-support $p(\mathbf{x}) = \mathcal{N}_q(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, assume $f : \mathbb{R}^D \to \mathbb{R}$ is C^2 on an open set containing $\{s(\mathbf{x}) \leq R^2\}$ and that $\mathbb{E}_p [\|\nabla f(\mathbf{x})\|] + \mathbb{E}_p \left[\|\nabla^2 f(\mathbf{x})\|_F \right] < \infty$. Then,*

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

In the Gaussian limit $q \uparrow 1$ (thus $m \to \infty$ and $R \to \infty$), $p^\star \to p$ and $\mathbb{E}_p[r(\mathbf{x})^2]/D \to 1$, which implies that Eq. 15 reduces to the classical Price theorem $\frac{\partial}{\partial \boldsymbol{\Sigma}_{ij}} \mathbb{E}[f(\mathbf{x})] = \frac{1}{2} \mathbb{E}[\partial_{x_i} \partial_{x_j} f(\mathbf{x})]$.

A proof is provided in App. D.2. This Price-type theorem is very similar to the Gaussian Price theorem in the second equation of Eq. 2, and differences are highlighted in red. The right-hand expectation is taken with respect to the $(2-q)$-escort $p^\star$, and there is an additional factor $\mathbb{E}_p[s(\mathbf{x})]/D$. The expression is easy to evaluate because sampling from $p^\star$ is efficient as shown in Eq. 13.

5. Applications

5.1. Bounded-variance Monte Carlo estimators

An advantage of bounded-support distributions is that they lead naturally to gradient estimators with bounded variance. We give a formal statement of bounds on the variance of the estimators based on Eq. 12 and Eq. 15 respectively. The Stein-type identity in Theorem 1 and the q-Price Theorem 3 express the gradients of interest in terms of escort expectations as $\mathbb{E}_{p^\star}[\nabla f(\mathbf{x})]$ and $\mathbb{E}_{p^\star}[\nabla^2 f(\mathbf{x})]$. Since we are interested in approximating these expectations by Monte Carlo, the variance of the resulting estimators directly affects the stochastic gradient noise. Using the reweighted p-only expressions from Eq. 25 in App. B, under mild boundedness assumptions on ∇f and $\nabla^2 f$, we can bound the variance of the stochastic gradient MC estimators.

Proposition 1 (Bounded variance MC estimators). *Let $\mathbf{x}_1, ...\mathbf{x}_S$ be iid samples from a q-Gaussian $p(\mathbf{x})$. For any almost everywhere differentiable $t : \mathbb{R}^D \to \mathbb{R}$ with $\mathbb{E}_{p^\star} [\|\nabla t(\mathbf{x})\|] < \infty$, and $f : \mathbb{R}^D \to \mathbb{R}$ that is C^2 on an open set containing $\{s(\mathbf{x}) \leq R^2\}$ such that $\mathbb{E}_p [\|\nabla f(\mathbf{x})\|] + \mathbb{E}_p \left[\|\nabla^2 f(\mathbf{x})\|_F \right] < \infty$, define the following Monte-Carlo estimators:*

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

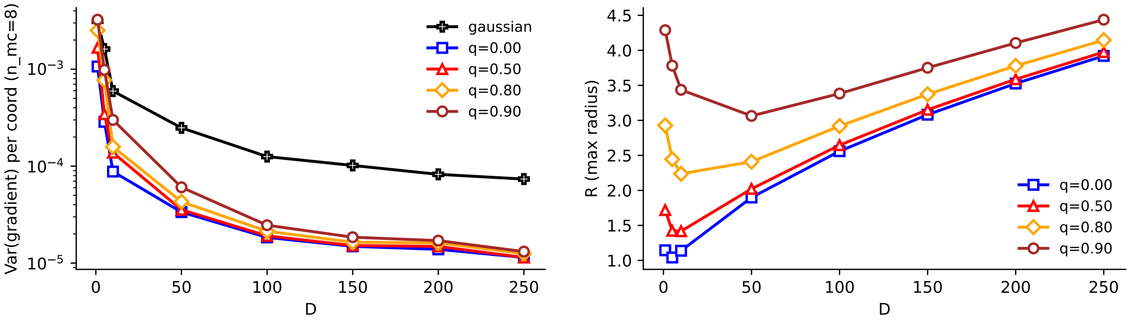

Figure 2: Synthetic logistic regression. Left: For $D \in \{10, 50, 200\}$ and $q \in \{0.0, 0.5, 0.8, 1\}$, we draw 8 Monte Carlo samples and compute the empirical per-coordinate gradient variance $\frac{1}{D}\sum_{j=1}^{D} \mathrm{Var}(\widehat{\nabla} F(\mathbf{w}^\star)_j)$, averaged over 50 independent repetitions; standard errors are all <.003. Right: Maximum radius of the q-Gaussian support against D.

Note that $M = \mathbb{E}_p\left[R^2 - s(\mathbf{x})\right]$ is available in closed form from Lemma 2. Assume there exist finite constants C_1, C_2 such that

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

where $\|\cdot\|$ is the spectral (operator) norm on matrices. Then, for each entry $\widehat{g}_j$ and $\widehat{H_{ij}}$ we have,

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

Consequently,

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

for $M = \mathbb{E}_p\left[R^2 - s(\mathbf{x})\right]$ and finite constant C_3.

Proof. The proof is in App. D.3. It follows from bounded-range variance bounds (Popoviciu's inequality) and matrix Hoeffding. □

5.2. Numerical experiments

Synthetic logistic regression experiment. For dimensions $D \in \{10, 50, 200\}$, we draw a dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ with $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{I}_D)$, a ground-truth weight vector $\mathbf{w}^\star \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$, and labels $y_i \mid \mathbf{x}_i \sim \mathrm{Bernoulli}(\sigma(\mathbf{x}_i^\top \mathbf{w}^\star))$, where σ is the sigmoid function. The loss $f(\mathbf{w})$ is the binary cross-entropy loss. We consider Monte Carlo pathwise estimators of the gradient of the objective evaluated at $\mathbf{w} = \mathbf{w}^\star$,

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

For the Gaussian baseline we take $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_D)$. For bounded-support q-smoothing, we draw $\boldsymbol{\epsilon}$ from an isotropic D-dimensional bounded-support q-Gaussian, so that $\|\boldsymbol{\epsilon}\|_2 \le R(D, q)$. For each $D \in \{10, 50, 200\}$ and $q \in \{0.0, 0.5, 0.8, 1\}$, we draw 8 Monte Carlo samples $\boldsymbol{\epsilon}$ and compute

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

for $\mathbf{w} = \mathbf{w}^\star + \boldsymbol{\epsilon}$. In Fig. 2 (left), for each q, over 50 independent repetitions we plot the empirical per-coordinate gradient variance $\frac{1}{D}\sum_{j=1}^{D} \mathrm{Var}(\widehat{\nabla} F(\mathbf{w}^\star)_j)$ against D. The results in Fig. 2 (left) confirm that smaller values of q lead to lower-variance gradient estimators. We also plot the maximum radius of the support against D in Fig. 2 (right) to visualize how the radius increases with the dimension and as q gets closer to one; the dimension noticeably dominates the magnitude of R and the effect of q is significant only with very small D.

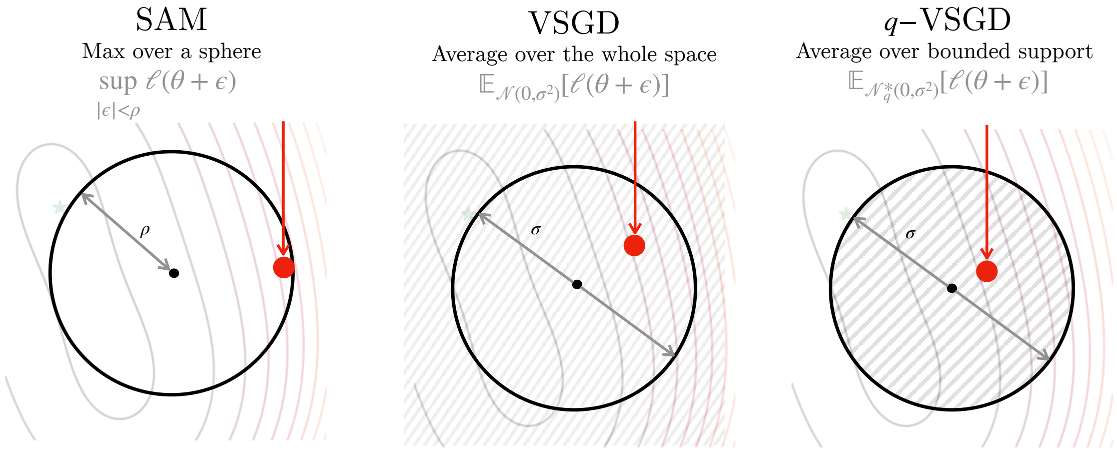
SAM
Max over a sphere
$\sup_{|\epsilon|<\rho} \ell(\theta+\epsilon)$
ρ
VSGD
Average over the whole space
$\mathbb{E}_{\mathcal{N}(0,\sigma^2)}[\ell(\theta+\epsilon)]$
σ
$q-$VSGD
Average over bounded support
$\mathbb{E}_{\mathcal{N}_q^*(0,\sigma^2)}[\ell(\theta+\epsilon)]$
σ

Figure 3: Sharpness-Aware Minimization (SAM) [18] considers an adversarial perturbation over a compact ball. Variational stochastic gradient descent (VSGD) with Gaussian weight perturbation averages perturbations over whole space (unbounded support). Our proposed q-VSGD uses q-Gaussian weight perturbations, which have bounded support similarly to SAM but uses averages similarly to VSGD. The method combines the two complementary features of SAM and VSGD.

Variational SGD with q-Gaussian noise. Next, we explore the potential use of our gradient estimators in Bayesian deep learning to estimate a q-Gaussian posteriors instead of the usual Gaussian ones. We compare with several variational training methods and also with the Sharpness-Aware Minimization (SAM) algorithm. A visualization is in Fig. 3, illustrating the main differences which are related to the choice of the point at which gradients are evaluated.

In SAM, we consider a point in a circle around the current parameter that has the worst loss value. Formally, let $f(\mathbf{w}; \mathcal{B})$ be the minibatch cross-entropy, where $\mathcal{B}$ is the minibatch. Then, we set the perturbation to be $\delta = \rho\mathbf{g}/\|\mathbf{g}\|_2$ with $\mathbf{g} = \nabla_{\mathbf{w}} f(\mathbf{w}; \mathcal{B})$. The parameters are then updated by evaluating the gradient at this point,

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

Overall, this requires two gradient evaluations, using in total two forward and backward pass. In Variational SGD (VSGD), the goal is to estimate the mean of an isotropic Gaussian posterior. The mean is updated by taking the gradient of the expected loss, requiring at least one MC sample. This means that the update is the same as above but we need to use $\delta \sim \mathcal{N}(0, \mathbf{I}_D)$.

We propose a new variant of VSGD where we aim to estimate the mean of a q-Gaussian posterior, while fixing $\mathbf{\Sigma}$ to $\mathbf{I}_D$. This can be done by making a simple change to VSGD: instead of sampling from a standard normal, we sample from a q-Gaussian as shown below:

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

and $R(q, D)$ is the distribution-dependent radius used for normalization. Note that q-VSGD with $q = 1$ is equivalent to VSGD. We also compare q-VSGD to an optimizer called Improved Variational Online Newton (IVON) which uses a more flexible diagonal-Gaussian posterior [29].

We compare the methods on a standard ResNet-20 model on the CIFAR-10 dataset, where we find that changing q leads to small improvements over VSGD. The experimental setup follows the one used in the paper that proposed IVON [29]. We train for 200 epochs with batch size 50 using SGD with momentum 0.9 and weight decay 10^{-4}, a 5-epoch linear warmup followed by cosine annealing, and sweep over values of $q \leq 1$ as reported in Table 1. In Table 1 we report test accuracy, NLL, Brier score, ECE (20 bins), AUROC (as described in [30]), and wall-clock seconds per epoch (mean ± standard error over 10 seeds). For $q = 0.6$, we observe some improvement in the accuracy. We can further increase the accuracy by increasing the number of MC samples, but the algorithm becomes slower. Overall, the results are not conclusive and indicate that tuning q may not directly

Method	Acc. (%) ↑	NLL ↓	ECE (%) ↓	Brier ↓	AUROC ↑	Time ↓
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 Monte-Carlo Sample						
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 Monte-Carlo Samples						
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

Table 1: ResNet-20 on CIFAR-10 test performance. Mean ± standard error over 10 seeds. Among the baselines (the first block at the top), SGD is the worst but fastest. SAM achieves the highest accuracy but is also slow due to two gradients needed in each iteration, while IVON performs reasonably well with a reasonable cost. In the second block, we show q-VSGD with 1 Monte-Carlo sample which achieves slightly better accuracy than its counterpart VSGD by increasing $q = 0.6$, although it is comparable in other metrics. Going to 5-MC in the third block improves the performance but slows down the algorithm. The results are mixed and indicate more work is needed in improving performance via tweaking q.

yield increased performance right away, despite the fact that the variance of gradient estimator is bounded.

One potential reason behind the lack of better performance is that for very large dimensions the effect of varying q reduces drastically. This happens because the support is heavily influenced by dimensionality, as shown in Fig. 2 where for high dimensions all q values have very similar support. Our numerical results suggest using a more flexible q-Gaussian form, for example, estimating the scale parameter. We expect a q-IVON algorithm will perform better than q-VSGD, which does outperform VSGD. Perhaps the most effective modification is to use a low-dimensional q-Gaussian factorization. We hope to explore these alternatives in future work.

6. Conclusion

This paper proves a new Stein-type identity for bounded-support q-Gaussians, which belong to the subfamily of the Pearson II class of elliptical distributions. We show that the associated (cumulative-generator) law has a simple escort interpretation, allowing for the derivation of Bonnet- and Price-type identities with forms closely resembling those of the Gaussian stochastic gradient estimators. These results are applied to practical pathwise gradient estimators that are easy to implement by sampling from the q-Gaussian density. A key consequence of the support's boundedness is that the resulting Monte Carlo estimators have simple bounded variance guarantees. Our experiments illustrate these effects in a controlled synthetic setting, and in deep neural models trained on CIFAR-10, where bounded-support perturbations are competitive with SAM and offer a principled distributional analogue of bounded-radius perturbation methods.

The results in this paper open new directions for research to exploit generalized Stein identities in broader stochastic gradient applications, such as variational inference and robust optimization, including extensions beyond bounded-support q-Gaussians (notably to the heavy-tailed $3 > q > 1$ regime), learning or adapting R, and considering anisotropic $\mathbf{\Sigma}$.

Acknowledgements

SS, MEK, and TM were supported by the Bayes duality project, JST CREST Grant Number JPMJCR2112. AM was supported by the project DECOLLAGE (ERC-2022-CoG 101088763). AM and MF were supported by the Portuguese Recovery and Resilience Plan through project C645008882000000055 (Center for Responsible AI) and by Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020.

References

[1] Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte Carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(132):1–62, 2020.

[2] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning (ICML)*, 2014.

[3] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning (ICML)*, 2015.

[4] Mohammad Emtiyaz Khan, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, and Akash Srivastava. Fast and scalable Bayesian deep learning by weight-perturbation in Adam. In *International Conference on Machine Learning (ICML)*, 2018.

[5] Charles M. Stein. Estimation of the mean of a multivariate normal distribution. In *Proceedings of Prague Symposium on Asymptotic Statistics*, pages 345–381, 1973.

[6] Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, pages 1135–1151, 1981.

[7] Wu Lin, Mohammad Emtiyaz Khan, and Mark Schmidt. Stein's lemma for the reparameterization trick with exponential family mixtures. *arXiv:1910.13398*, 2019.

[8] Georges Bonnet. Transformations des signaux aléatoires a travers les systemes non linéaires sans mémoire. In *Annales des Télécommunications*, volume 19, pages 203–220, 1964.

[9] Robert Price. A useful theorem for nonlinear devices having gaussian inputs. *IRE Transactions on Information Theory*, 4(2):69–72, 2003.

[10] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

[11] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. *Advances in Neural Information Processing Systems 29*, 2016.

[12] Zinoviy Landsman and Johanna Nešlehová. Stein's lemma for elliptical random vectors. *Journal of Multivariate Analysis*, 99(5):912–927, 2008.

[13] Zinoviy Landsman, Steven Vanduffel, and Jing Yao. A note on stein's lemma for multivariate elliptical distributions. *Journal of Statistical Planning and Inference*, 143(11):2016–2022, 2013.

[14] Zinoviy Landsman, Steven Vanduffel, and Jing Yao. Some stein-type inequalities for multivariate elliptical distributions and applications. *Statistics & Probability Letters*, 97:54–62, 2015.

[15] Jan Naudts. The q-exponential family in statistical physics. *Central European Journal of Physics*, 7(3):405–413, 2009.

[16] Shun-ichi Amari and Atsumi Ohara. Geometry of q-exponential family of probability distributions. *Entropy*, 13(6):1170–1185, 2011.

[17] Hiroshi Matsuzoe and Atsumi Ohara. Geometry for q-exponential families. *Recent progress in differential geometry and its related fields*, pages 55–71, 2011.

[18] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations (ICLR)*, 2021.

[19] Mohammad Khalafi and Majid Azimmohseni. Multivariate Pearson type II distribution: Statistical and mathematical features. *Probability and Mathematical Statistics*, 34, 2014.

[20] Jan Naudts. Deformed exponentials and logarithms in generalized thermostatistics. *Physica A: Statistical Mechanics and its Applications*, 316:323–334, 2002.

[21] André FT Martins, Marcos Treviso, António Farinhas, Pedro MQ Aguiar, Mário AT Figueiredo, Mathieu Blondel, and Vlad Niculae. Sparse continuous distributions and Fenchel-Young losses. *Journal of Machine Learning Research*, 23(257):1–74, 2022.

[22] Constantino Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52:479–487, 1988.

[23] Stamatis Cambanis, Steel Huang, and Gordon Simons. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11(3):368–385, 1981.

[24] Kai Tai Fang, Samuel Kotz, and Kai Wang Ng. *Symmetric multivariate and related distributions*. Springer-Science+Business Media, B.V., 1990.

[25] Reinaldo B Arellano-Valle. On some characterizations of spherical distributions. *Statistics & probability letters*, 54(3):227–232, 2001.

[26] Reinaldo B Arellano-Valle, Guido del Pino, and Pilar Iglesias. Bayesian inference in spherical linear models: robustness and conjugate analysis. *Journal of Multivariate Analysis*, 97(1):179–197, 2006.

[27] Jan Naudts. Estimators, escort probabilities, and phi-exponential families in statistical physics. *Journal of Inequalities in Pure and Applied Mathematics*, 5:1443–5756, 2004.

[28] Mark E Johnson. *Multivariate statistical simulation: A guide to selecting and generating continuous multivariate distributions*. 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, and Thomas Möllenhoff. Variational learning is effective for large deep networks. In *International Conference on Machine Learning (ICML)*, 2024.

[30] Kazuki Osawa, Siddharth Swaroop, Mohammad Emtiyaz Khan, Anirudh Jain, Runa Eschenhagen, Richard E Turner, and Rio Yokota. Practical deep learning with Bayesian principles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[31] Amir Rezaei, Maryam Sharafi, Javad Behboodian, and Atefeh Zamani. Inferences on stress–strength parameter based on GLD5 distribution. *Communications in Statistics-Simulation and Computation*, 47(5):1251–1263, 2018.

A. Proof of Bonnet's and Price's theorems with Gaussians

Gaussian Stein, Bonnet, and Price. Let $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with density $p(\mathbf{x})$. The general Stein identity states that for any differentiable test function $t : \mathbb{R}^D \to \mathbb{R}$ with $\mathbb{E}\,\|t(\mathbf{x})\| + \mathbb{E}\,\|\nabla_{\mathbf{x}} t(\mathbf{x})\| < \infty$,

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

This follows from the score identity $\mathbb{E}\left[t(\mathbf{x})\nabla_{\mathbf{x}} \log p(\mathbf{x})\right] = -\mathbb{E}\left[\nabla_{\mathbf{x}} t(\mathbf{x})\right]$ through integration by parts, and the Gaussian-specific fact that $\nabla_{\mathbf{x}} \log p(\mathbf{x}) = -\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$.

The score identity implies:

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

Applying the Stein identity with $t(\mathbf{x}) = f(\mathbf{x})$, we obtain Bonnet's theorem

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

Writing $\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, since $\partial\mathbf{x}/\partial\boldsymbol{\mu} = \mathbf{I}$, we see that a popular case of the location-scale transform or "reparameterization trick" is the pathwise implementation of Bonnet. Similarly, Price's theorem, which states that for differentiable f we have $\nabla_{\boldsymbol{\Sigma}}\mathbb{E}\left[f(\mathbf{x})\right] = \frac{1}{2}\mathbb{E}\left[\nabla^2_{\mathbf{x}} f(\mathbf{x})\right]$, follows by applying the same Stein identity to the components of $\nabla_{\mathbf{x}} f(\mathbf{x})$ and using the score identity for derivatives with respect to $\boldsymbol{\Sigma}$.

Gaussian Stein and Price's theorem. Let $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ on $\mathbb{R}^D$ and let $f : \mathbb{R}^D \to \mathbb{R}$ be twice continuously differentiable with $\mathbb{E}\,\|\nabla_{\mathbf{x}} f(\mathbf{x})\| < \infty$ and $\mathbb{E}\,\left\|\nabla^2_{\mathbf{x}} f(\mathbf{x})\right\|_{\mathrm{F}} < \infty$. The scalar Gaussian Stein identity states that for any sufficiently regular $t : \mathbb{R}^D \to \mathbb{R}$ and each coordinate $i = 1, ..., D$,

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

Proof of Price. To differentiate with respect to $\boldsymbol{\Sigma}$, we use the matrix score identity for the Gaussian density:

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

so, entrywise,

$$\begin{aligned} \frac{\partial}{\partial\boldsymbol{\Sigma}_{ij}}\mathbb{E}\left[f(\mathbf{x})\right] &= \mathbb{E}\left[f(\mathbf{x})\frac{\partial}{\partial\boldsymbol{\Sigma}_{ij}} \log p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\right] && (18) \\ &= -\tfrac{1}{2}(\boldsymbol{\Sigma}^{-1})_{ji}\mathbb{E}\left[f(\mathbf{x})\right] + \tfrac{1}{2}\left(\boldsymbol{\Sigma}^{-1}\mathbb{E}\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top f(\mathbf{x})\right]\boldsymbol{\Sigma}^{-1}\right)_{ji}. && (19) \end{aligned}$$

For each fixed $j = 1, ..., D$, apply Stein with the scalar test function $t(\mathbf{x}) := (\mathbf{x}_j - \boldsymbol{\mu}_j)f(\mathbf{x})$. For each i,

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

Since

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

where $\delta_{\ell j}$ is the Kronecker delta, we get

$$\begin{aligned} \mathbb{E}\left[(\mathbf{x}_i - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_j)f(\mathbf{x})\right] &= \sum_{\ell} \boldsymbol{\Sigma}_{i\ell}\delta_{\ell j}\mathbb{E}\left[f(\mathbf{x})\right] + \sum_{\ell} \boldsymbol{\Sigma}_{i\ell}\mathbb{E}\left[(\mathbf{x}_j - \boldsymbol{\mu}_j)\partial_{\mathbf{x}_\ell} f(\mathbf{x})\right] && (20) \\ &= \boldsymbol{\Sigma}_{ij}\mathbb{E}\left[f(\mathbf{x})\right] + \sum_{\ell} \boldsymbol{\Sigma}_{i\ell}\mathbb{E}\left[(\mathbf{x}_j - \boldsymbol{\mu}_j)\partial_{\mathbf{x}_\ell} f(\mathbf{x})\right]. && (21) \end{aligned}$$

Next, apply Stein again with the scalar test function $t(\mathbf{x}) := \partial_{\mathbf{x}_\ell} f(\mathbf{x})$. For each $\ell = 1, ..., D$,

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

Substituting into the previous expression yields

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

In matrix notation this is

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

Multiply (22) on the left and right by $\boldsymbol{\Sigma}^{-1}$:

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

Substitute this back into (18). The $\boldsymbol{\Sigma}^{-1}\mathbb{E}\left[f(\mathbf{x})\right]$ terms cancel, and we obtain

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

Equivalently, in matrix form,

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

which is Price's theorem. □

Remark 1 (Symmetry in $\boldsymbol{\Sigma}$). In the proof of Price's theorem we differentiate with respect to the entries $\boldsymbol{\Sigma}_{ij}$ treating $\boldsymbol{\Sigma}$ as an unconstrained matrix. The right-hand side of the final expression for the gradient is symmetric because the Hessian $\nabla^2_{\mathbf{x}} f(\mathbf{x})$ is symmetric for C^2 functions f. Thus the gradient naturally lies in the space of symmetric matrices, and the formula is consistent with the covariance constraint $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^\top$.

B. Elliptical Laws

Generalized Pearson Type II. An important subclass of elliptical distributions are the Pearson Type II distributions. A fundamental property characterizing these distributions, is the stochastic representation of a D-dimensional spherical Pearson Type II random vector $\mathbf{z}$ as $\mathbf{z} \overset{d}{=} r\mathbf{u}$, where $\mathbf{u}$ has uniform distribution on the unit sphere surface in $\mathbb{R}^D$, and the sign d indicates the same distribution. The random variables $r \overset{d}{=} \|\mathbf{z}\|$ and $\mathbf{u} \overset{d}{=} \frac{\mathbf{z}}{\|\mathbf{z}\|}$ are independent and therefore the distribution of $\mathbf{z}$ is fully determined by that of its squared length $r^2 = \mathbf{z}^\top\mathbf{z}$. The random variables r and

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

are called the radial and squared radial parts of the random vector $\mathbf{x}$. Their distributions are defined by the specific Pearson Type II density generator function in (4), as

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

For further details see [24, 25, 28, 31].

Proof of Lemma 1. Expressing the exponent m of the Pearson type II density in terms of the entropic index $q < 1$ of the q-Gaussian density as $m = 1/1 - q$, and choosing the support radius $R = R(D, q)$ to be

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

the density generator of the Pearson type II distribution, given in (4), takes the form

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

The corresponding Pearson type II density is given by

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

This can be equivalently expressed in the form of the q-Gaussian density as

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

Introducing for convenience a transformation of the scale parameter $\widetilde{\mathbf{\Sigma}} = |\mathbf{\Sigma}|^{\frac{1-q}{2}}\mathbf{\Sigma}$ and using the fact that $|\mathbf{\Sigma}| = |\widetilde{\mathbf{\Sigma}}|^{\frac{2}{(1-q)D+2}}$, Martins et al. [21, Def. 15] define the density parametrized in terms of $\boldsymbol{\mu}$ and $\widetilde{\mathbf{\Sigma}}$ as

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

Johnson [28] provides a detailed discussion of the effect of the shape parameter q. For $q \to 1$, the radius $R \to \infty$, $\exp_q$ converges to the ordinary exponential, and we recover the multivariate normal $N(\boldsymbol{\mu}, \mathbf{\Sigma})$. □

Associated laws as escort distributions. Using the density generator $g(\cdot)$ of the bounded-support q-Gaussian, given in (8), the cumulative generator function is defined as the integral of the original generator:

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

which induces the first associated law with density

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

where $s(\mathbf{x}) := (\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$. By construction, p and $p^\star$ share the same location-scale parameters $(\boldsymbol{\mu}, \mathbf{\Sigma})$ and the same bounded support $\mathrm{supp} = \{\mathbf{x} : s(\mathbf{x}) < R^2\}$; the associated law is a new Pearson Type II distribution on the same support set, with exponent increased by 1 as a consequence of the power-law Pearson II form.

We can equivalently express it as normalized reweighting of the base law $p(\mathbf{x})$:

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

Escort interpretation of the associated laws. The first associated law coincides with the $(2-q)$-escort of the base density. Indeed,

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

since $m+1 = \frac{2-q}{1-q} = (2-q)m$. More generally, the k-th associated law satisfies

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

so the escort order increases by $(1-q)$ at each step. All these laws share the same bounded support and are progressively more centrally concentrated, with a sharper peak at $\boldsymbol{\mu}$ and vanishing faster near the boundary than $p(\mathbf{x})$.

C. Proof of the bounded-support Stein identity

We prove the bounded-support Stein for a general test function $t(\mathbf{x})$ rather than the more specific $f(\mathbf{x})$ which we use for simplicity in the main statement of Theorem 1. We assume p is a bounded-support q-Gaussian density with parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, q)$ as in Lemma 1, and $p^\star$ is its first associated (cumulative-generator) law. For any almost everywhere differentiable $t : \mathbb{R}^D \to \mathbb{R}$ with $\mathbb{E}_{p^\star} \|\nabla t(\mathbf{x})\| < \infty$,

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

where $r^2 = s(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ and $\mathbb{E}_p\left[r^2\right]$ are given in Lemma 2.

Proof of Theorem 1. Let $\mathbf{z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$, so that

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

and define $T(\mathbf{z}) := t(\boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{z})$. To compute $\mathbb{E}_{p^\star}\left[\partial_{\mathbf{z}_i} T(\mathbf{z})\right]$ first we integrate out $\mathbf{z}_i$. Fix an index i. For each $\mathbf{z}_{-i} \in \mathbb{R}^{D-1}$ with $\|\mathbf{z}_{-i}\| < R$, set

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

so that the admissible range for $\mathbf{z}_i$ is exactly

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

since $\|\mathbf{z}\|^2 < R^2 \implies \|\mathbf{z}_{-i}\|^2 + \mathbf{z}_i^2 < R^2 \implies |\mathbf{z}_i| < \rho(\mathbf{z}_{-i})$. For each fixed $\mathbf{z}_{-i}$, we apply one-dimensional integration by parts in the variable $\mathbf{z}_i$ on the interval $I(\mathbf{z}_{-i})$:

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

Because $m + 1 > 0$, the polynomial $(R^2 - \|\mathbf{z}_{-i}\|^2 - \mathbf{z}_i^2)^{m+1}$ vanishes at $\mathbf{z}_i = \pm\rho(\mathbf{z}_{-i})$, and the boundary term is exactly zero for each slice. Differentiating the polynomial term gives

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

Hence,

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

Next, we integrate over all $\mathbf{z}_{-i}$ with $\|\mathbf{z}_{-i}\| < R$. The left-hand side of (28) becomes

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

where $C^\star$ is the normalizing constant of $p^\star$. Finally, the right-hand side of (28) becomes

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

where C is the normalizing constant of p. Hence

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

Using the relation

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

we get $\frac{C}{C^\star} = \frac{1}{\mathbb{E}_p\left[R^2 - \|\mathbf{z}\|^2\right]}$, so (29) simplifies to

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

Vectorizing over i, we obtain

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

Return to x-space. Recall $\mathbf{z} = \mathbf{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ and $\nabla_{\mathbf{z}} T(\mathbf{z}) = \mathbf{\Sigma}^{1/2}\nabla_{\mathbf{x}} t(\boldsymbol{\mu} + \mathbf{\Sigma}^{1/2}\mathbf{z})$. Taking expectations and multiplying by $\mathbf{\Sigma}^{1/2}$,

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

From Lemma 2,

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

so the coefficient equals $\frac{\mathbb{E}_p\left[r^2\right]}{D}$ and we obtain (10). Using (25) to express $\mathbb{E}_{p^\star}$ in terms of $\mathbb{E}_p$ yields (12). □

D. Proofs of q-Bonnet, q-Price, and variance bounds

Similarly to Appendix C we specify in the following proofs the exact form of the test function $t(\mathbf{x})$ that we use, and note that in the main statements of the Theorems we use $f(\mathbf{x})$ instead for simplicity.

D.1. q-Bonnet

Let p be a bounded-support q-Gaussian density $N_q(\boldsymbol{\mu}, \mathbf{\Sigma})$ from Lemma 1, with $q < 1$ and $m := 1/(1-q) > 0$. Let $f : \mathbb{R}^D \to \mathbb{R}$ be C^1 on an open set containing $\{s(\mathbf{x}) \le R^2\}$, where $s(\mathbf{x}) := (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$, and assume $\mathbb{E}_p \|\nabla f(\mathbf{x})\| < \infty$. We show that

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

Proof of Theorem 2. Since $p(\mathbf{x}) \propto |\mathbf{\Sigma}|^{-1/2}\left(R^2 - s(\mathbf{x})\right)_+^m$, we have

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

Hence

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

Differentiating under the integral is justified by dominated convergence on the bounded support together with $m > 0$, so we obtain the score identity

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

Apply the p-only Stein identity (12) with $t(\mathbf{x}) = \frac{f(\mathbf{x})}{R^2 - s(\mathbf{x})}$. Substituting into the right-hand side of (12) by the quotient rule and using $\nabla s(\mathbf{x}) = 2\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ gives

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

Hence (12) yields

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

From the moment formulas in Lemma 2, we have

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

Thus (32) yields

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

Substitute this into (31):

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

□

D.2. q-Price

Lemma 3 (Matrix calculus)**.** *Let $\mathbf{\Sigma}\succ 0$. We use the Frobenius inner product $A:B:=\operatorname{tr}(A^\top B)$. For $s(\mathbf{x})=(\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ and the elementary matrix $\mathbf{E}_{ij}$ that has 1 in the (i,j) position and 0 elsewhere:*

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

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

Moreover, for any matrix $\mathbf{H}$,

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

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

As before, let p be the bounded-support q-Gaussian density $N_q(\boldsymbol{\mu},\mathbf{\Sigma})$ with $q<1$ and shape $m=\frac{1}{1-q}>0$, supported on $\{s(\mathbf{x})<R^2\}$. For the proof of Theorem 3 we assume $f:\mathbb{R}^D\to\mathbb{R}$ is C^2 on an open set containing $\{s(\mathbf{x})\le R^2\}$ and that $\mathbb{E}_p\|\nabla f(\mathbf{x})\|+\mathbb{E}_p\|\nabla^2 f(\mathbf{x})\|_F<\infty$. We define M and recall the form of the covariance

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

As usual, $p^\star$ denotes the first associated law of p, and so $\mathbb{E}_{p^\star}[h]=\mathbb{E}_p[(R^2-s)h]/M$ for any integrable h. Treating the entries $\mathbf{\Sigma}_{ij}$ as independent parameters, the following identities hold for every $i,j\in\{1,\dots,D\}$:

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

Equivalently, in matrix form,

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

In the Gaussian limit $q\uparrow 1$ (so $m\to\infty$, $R\to\infty$), $p^\star=p$ and $\mathbb{E}_p[r^2]/D\to 1$, so (37) reduces to the classical Price theorem $\frac{\partial}{\partial\mathbf{\Sigma}_{ij}}\mathbb{E}[f(\mathbf{x})]=\frac{1}{2}\mathbb{E}[\partial_{x_i}\partial_{x_j}f(\mathbf{x})]$.

Proof of Theorem 3. **Score trick w.r.t. $\mathbf{\Sigma}$.** By differentiation under the integral (justified by the assumptions and bounded support),

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

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

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

Using Lemma 3,

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

Define the matrix

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

Then, in matrix form,

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

Identify B via the Stein identity. Invoke the bounded-support Stein identity in its p-only form:

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

Choose the vector test function

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

Left-hand side. Using the definition of B,

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

Right-hand side. Compute

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

using $\nabla s(\mathbf{x}) = 2\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ and the product rule. Taking $\mathbb{E}_p$ and substituting into Stein's identity yields

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

Right-multiplying by $\mathbf{\Sigma}$ and rearranging gives

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

For the bounded-support q-Gaussian, Lemma 2 gives $c = \mathbb{E}_p[r^2]/(DM) = 1/(2(m+1))$, hence

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

Insert B back into the score form. Substitute (40) into (39):

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

The $\pm\frac{1}{2}\mathbf{\Sigma}^{-1}\mathbb{E}_p[f(\mathbf{x})]$ terms cancel, leaving

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

Second Stein step to reach the Hessian. Apply the (vector) Stein identity with the scalar tests $t_j(\mathbf{x}) := \partial_{x_j} f(\mathbf{x})$, $j = 1, \ldots, D$, and stack:

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

Substitute this into (41) to obtain

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

which is (15).

□

Remark 2 (Symmetry in $\boldsymbol{\Sigma}$). We differentiated with respect to the entries $\boldsymbol{\Sigma}_{ij}$ treating $\boldsymbol{\Sigma}$ as an unconstrained matrix. The right-hand side of (15) is symmetric because $\nabla^2_{\mathbf{x}} f(\mathbf{x})$ is symmetric for C^2 functions f. Hence the gradient $\nabla_{\boldsymbol{\Sigma}} \mathbb{E}_p[f(\mathbf{x})]$ naturally lies in the space of symmetric matrices and is consistent with the constraint $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^\top$.

D.3. Variance bounds

Proof of Proposition 1. It is straightforward to verify unbiasedness: $\mathbb{E}[\widehat{\mathbf{g}}] = \mathbb{E}_{p^\star}[\nabla t(\mathbf{x})]$ and $\mathbb{E}\left[\widehat{\mathbf{H}}\right] = \mathbb{E}_{p^\star}\left[\nabla^2 f(\mathbf{x})\right]$.

Note that $\widehat{g}_j$ is scalar, and let

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

Since $0 < R^2 - s(\mathbf{x}) \leq R^2$ and $|\partial_j t(\mathbf{x})| \leq \|\nabla t(\mathbf{x})\| \leq C_1$, hence

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

Therefore $Y_k \in [-B_g, B_g]$ almost surely, and by Popoviciu's inequality on bounded-range variances,

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

Since $\widehat{g}_j = \frac{1}{S}\sum_{k=1}^n Y_k$ with iid Y_k,

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

Next, fix (i, j) and define

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

Since $0 < (R^2 - s(\mathbf{x})) \leq R^2$ and $|(\nabla^2 f(\mathbf{x}))_{ij}| \leq \|\nabla^2 f(\mathbf{x})\|_{\mathrm{op}} \leq C_2$, hence

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

so $Z_k \in [-B_H, B_H]$ almost surely, and again by Popoviciu, $\mathrm{Var}(Z_k) \leq B_H^2$, and

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

Summing entrywise variances gives

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

Since $\|\cdot\|_{\mathrm{op}} \leq \|\cdot\|_F$, the same bound implies

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

A sharper dimension dependence follows from matrix Hoeffding: defining centered

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

implies $\|\mathbf{A}_k\|_{\mathrm{op}} \leq 2B_H$. Hence, since $\frac{1}{S}\sum_{k=1}^{S} \mathbf{A}_k = \widehat{\mathbf{H}} - \mathbb{E}\left[\widehat{\mathbf{H}}\right]$, we have

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

which integrates to $\mathbb{E}\left[\left\|\widehat{\mathbf{H}} - \mathbb{E}\left[\widehat{\mathbf{H}}\right]\right\|_{\mathrm{op}}\right] \leq C_3 B_H \sqrt{\frac{\log(2D)}{S}}$ for a universal C_3: by the same tail-integration argument as above, and splitting at $t_0 > 0$,

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

With $a := \dfrac{S}{8(2B_H)^2}$ and the Gaussian tail bound $\int_x^\infty e^{-at^2} dt \leq \dfrac{1}{2ax} e^{-ax^2}$,

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

Choose $t_0 := \sqrt{\dfrac{\log(2D)}{a}}$ so that $e^{-at_0^2} = 1/(2D)$. Then

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

for universal constants $C_4, C_3 > 0$, as claimed. □