Stein's Lemma for the Reparameterization Trick with Exponential Family Mixtures

Wu Lin [1] **Mohammad Emtiyaz Khan** [2] **Mark Schmidt** [1]

Abstract

Stein's method (Stein, 1973; 1981) is a powerful tool for statistical applications, and has had a significant impact in machine learning. Stein's lemma plays an essential role in Stein's method. Previous applications of Stein's lemma either required strong technical assumptions or were limited to Gaussian distributions with restricted covariance structures.

In this work, we extend Stein's lemma to exponential-family mixture distributions including Gaussian distributions with full covariance structures. Our generalization enables us to establish a connection between Stein's lemma and the reparamterization trick to derive gradients of expectations of a large class of functions under weak assumptions. Using this connection, we can derive many new reparameterizable gradient-identities that goes beyond the reach of existing works. For example, we give gradient identities when expectation is taken with respect to Student's t-distribution, skew Gaussian, exponentially modified Gaussian, and normal inverse Gaussian.

1. Introduction

Stein's lemma (Stein, 1973; 1981; Liu, 1994) plays an essential role in Stein's method. The lemma gives a first-order identity to estimate the mean of a multivariate Gaussian distribution. In machine learning, Fan et al. (2015); Erdogdu (2015); Rezende et al. (2014) use integration by parts to extend the lemma without giving technical conditions. Another applications of Stein's lemma are De Bruijn's identity (Park et al., 2012) and the heat equation identity (Brown et al., 2006). These two works give the same second-order identity to estimate the covariance of a multivariate Gaussian distribution. In practices, the second-order identity gives a better unbiased estimation than the one obtained from the first-order identity. (Khan & Lin, 2017; Khan et al., 2018; Salimans & Knowles, 2013). However, Park et al. (2012); Brown et al. (2006) use stronger assumptions to simplify proofs where the authors either assume diagonal covariance structure or twice continuous differentiability. In gradient estimation, the first-order identity is known as Bonnet's theorem (Bonnet, 1964). Bonnet's theorem gives the reparameterization gradient for the mean (Kingma & Welling, 2013). The second-order identity is known as Price's theorem (Price, 1958). However, Price (1958); Bonnet (1964) use characteristic functions as the proof technique. This technique is not easy to be extended to Gaussian mixture and to be used to identify weak assumptions. Beyond Gaussian distribution, Stein's lemma is proposed by Hudson et al. (1978); Brown (1986). In machine learning, Salimans & Knowles (2013); Figurnov et al. (2018) propose an implicit reparameterization trick under continuous differentiability for a class of distributions.

In this work, we generalize Stein's lemma to Gaussian variance-mean mixtures with arbitrary covariance structure and exponentially family mixtures while keeping assumptions week and proofs simple. Moreover, we present a second-order identity for the covariance estimation of the Gaussian mixtures. Our theory also shows a direct connection between Stein's lemma and the reparameterizable gradient estimation. Furthermore, we show that we can obtain the implicit reparameterization trick via Stein's lemma under weaker assumptions than Salimans & Knowles (2013); Figurnov et al. (2018). Last but not least, we give examples of gradient identities derived from our theory such as multivariate Student's t distribution, multivariate skew Gaussian, multivariate exponentially modified Gaussian and multivariate normal inverse Gaussian. We find out that these identities are useful in variational inference with Gaussian variance-mean mixture approximations as shown in Lin et al. (2019).

2. Related Works

There are many existing works about Stein's lemma. Stein (1973; 1981) give a first-order identity for a multivariate

[1]University of British Columbia, Vancouver, Canada. [2]RIKEN Center for Advanced Intelligence Project, Tokyo, Japan. Correspondence to: Wu Lin <wlin2018@cs.ubc.ca>.

Technical Report, Working in Progress.

Gaussian with diagonal covariance structure. Liu (1994) extends the first-order identity to a multivariate Gaussian with arbitrary covariance structure. For gradient estimation, Stein's lemma indeed recovers Bonnet's theorem (Bonnet, 1964) . Price's theorem (Price, 1958) gives the second-order identity for a multivariate Gaussian with arbitrary covariance structure. However, Price (1958); Opper & Archambeau (2009) use the characteristic function of Gaussian to prove Price's theorem, which is not easy to extend to the Gaussian mixture case. Hudson et al. (1978); Brown (1986); Arnold et al. (2001); Landsman (2006); Landsman & Nešlehová (2008); Kattumannil (2009); Kattumannil & Dewan (2016); Adcock (2007); Adcock & Shutes (2012) further extend Stein's lemma to exponential family and beyond. Unfortunately, these works neither show the connection between the gradient identity and the implicit reparameterization trick nor give any second-order identity.

3. Smoothness Assumptions

We first give smoothness conditions. These conditions will be used in the gradient identities. The key definition is the absolute continuity (AC) of a function: $\mathbf{h}(\cdot) : [a,b] \mapsto \mathcal{R}^m$, where $[a,b]$ is a compact interval in $\mathcal{R}$.

Definition 1 *A vector function: $\mathbf{h}(\cdot) : [a,b] \mapsto \mathcal{R}^m$ is AC if the following assumptions are satisfied.*

- *Its derivative $\nabla_z \mathbf{h}(z)$ exists almost everywhere for $z \in [a,b]$.*
- *The derivative is Lebesgue integrable. In other words, $\int_a^b ||\nabla_z \mathbf{h}(z)||\, dz < \infty$, where $|| \cdot ||$ denotes the Euclidean norm.*
- *The fundamental theorem of calculus holds, that is, $\mathbf{h}(z) = \mathbf{h}(a) + \int_a^z \nabla_z \mathbf{h}(t) dt$ for any $z \in [a,b]$.*

Since we want to deal with a class of functions whose domain is $\mathcal{R}$, we define the locally AC of this class of functions.

Definition 2 *Let $\mathbf{h}(\cdot) : \mathcal{R} \mapsto \mathcal{R}^m$ be a vector function. If $\mathbf{h}(\cdot)$ is AC at every compact interval of its domain $\mathcal{R}$, we say that the function is locally AC.*

The property below is essential in the following sections.

Property 1 *A product of two locally AC functions is also locally AC.*

Now, we extend the definition of locally AC to a set of functions whose domain is $\mathcal{R}^n$. It is known as the absolute continuity on almost every straight line (ACL) (Leoni, 2017).

Definition 3 *Let $\mathbf{h}(\mathbf{z}) : \mathcal{R}^d \mapsto \mathcal{R}^m$ be a vector function. Given $\mathbf{z}_{-i} \in \mathcal{R}^{d-1}$ is fixed, let's define a function $\mathbf{h}_i(\cdot) := \mathbf{h}(\cdot, \mathbf{z}_{-i}) : \mathcal{R} \mapsto \mathcal{R}^m$. We say the function $\mathbf{h}(\cdot)$ is locally ACL if for all i and almost every point $\mathbf{z}_{-i} \in \mathcal{R}^{d-1}$, $\mathbf{h}_i(\cdot)$ is locally AC.*

A locally ACL function is a member of the Sobolev family (Leoni, 2017). Obviously, the derivative $\nabla_z \mathbf{h}(\mathbf{z})$ exists almost everywhere if $\mathbf{h}(\mathbf{z})$ is locally ACL. Due to Royen & Fitzpatrick (2010), a locally Lipschitz-continuous function is locally ACL. The immediate consequence is that a function is locally ACL and continuous if it is either locally Lipschitz-continuous or continuously differentiable.

In the following sections, we assume that the regular conditions for swapping of differentiation and integration are satisfied so that the following identify holds.

$$\nabla_\lambda \mathbb{E}_{q(z|\lambda)} \left[h(\mathbf{z}) \right] = \mathbb{E}_{q(z|\lambda)} \left[h(\mathbf{z}) \frac{\nabla_\lambda q(\mathbf{z}|\boldsymbol{\lambda})}{q(\mathbf{z}|\boldsymbol{\lambda})} \right]$$

The regular conditions are required to use the dominated convergence theorem, which allows us to interchange differentiation and integration. One particular condition is

$$\mathbb{E}_{q(z|\lambda)} \left[\left|\left| h(\mathbf{z}) \frac{\nabla_\lambda q(\mathbf{z}|\boldsymbol{\lambda})}{q(\mathbf{z}|\boldsymbol{\lambda})} \right|\right| \right] < \infty.$$

For simplicity, we assume the regular conditions hold without explicitly mentioning them.

4. Expectation and Conditional Expectation

By definition, an expectation can be either non-existent or non-finite. To avoid such cases, we say an expectation $\mathbb{E}_{q(z)} \left[\mathbf{h}(\mathbf{z}) \right]$ is well-defined if

$$\mathbb{E}_{q(z)} \left[||\mathbf{h}(\mathbf{z})|| \right] < \infty,$$

where $||\cdot||$ is an appropriate norm. Due to Fubini's theorem, the following identity holds for a random vector $\mathbf{z}$ on a product measure when the expectation is well-defined.

$$\mathbb{E}_{q(z)} \left[\mathbf{h}(\mathbf{z}) \right] = \mathbb{E}_{q(z_{-i})} \left[\mathbb{E}_{q(z_i|z_{-i})} \left[\mathbf{h}(\mathbf{z}) \right] \right]$$

The above expression shows that conditional expectation $\mathbb{E}_{q(z_i|z_{-i})} \left[\mathbf{h}(\mathbf{z}) \right]$ is also well-defined for almost every $\mathbf{z}_{-i}$.

For simplicity, we implicitly assume expectations are well-defined in the following sections.

5. Identities for Gaussian Distribution

Now, we describe the univariate case of Stein's lemma.

Lemma 1 **(Stein's Lemma):** *Let $h(\cdot) : \mathcal{R} \mapsto \mathcal{R}$ be locally AC. $q(z)$ is an univariate Gaussian distribution denoted* *by $\mathcal{N}(z|\mu,\sigma)$ with mean μ and variance σ. The following first-order identity holds.*

$$\mathbb{E}_q\left[\frac{-\nabla_z q(z)}{q(z)} h(z)\right] = \mathbb{E}_q\left[\nabla_z h(z)\right].$$

where $\frac{-\nabla_z q(z)}{q(z)} = \sigma^{-1}\left(z-\mu\right)$.

The proof of this lemma is given at Appendix A.1.

Let's consider Bonnet's theorem given below. This theorem establishes the connection between this lemma and the reparameterizable gradient for the mean μ.

Theorem 1 (Bonnet's Theorem) : *Let $h(\cdot) : \mathcal{R} \mapsto \mathcal{R}$ be locally AC. $q(z)$ is a univariate Gaussian distribution with mean μ and variance σ. We have the following gradient identity.*

$$\nabla_\mu \mathbb{E}_q\left[h(z)\right] = \mathbb{E}_q\left[\nabla_z h(z)\right] = \mathbb{E}_q\left[\sigma^{-1}\left(z-\mu\right)h(z)\right]$$

The proof of Bonnet's Theorem is fairly simple. First, we swap differentiation and integration. We obtain the following expression $\nabla_\mu \mathbb{E}_q\left[h(z)\right] = \mathbb{E}_q\left[\sigma^{-1}\left(z-\mu\right)h(z)\right]$. By applying Lemma 1 to $h(z)$, we obtain the identity $\mathbb{E}_q\left[\nabla_z h(z)\right] = \mathbb{E}_q\left[\sigma^{-1}\left(z-\mu\right)h(z)\right]$. Clearly, the reparameterizable gradient for the mean μ is directly derived from Stein's lemma. Now, we discuss the reparameterizable gradient for the variance. First, we give the following lemma.

Lemma 2 *Let $h(\cdot) : \mathcal{R} \mapsto \mathcal{R}$ be locally AC. We assume $\mathbb{E}_q\left[h(z)\right]$ is well-defined. The following identity holds.*

$$\mathbb{E}_q\left[\sigma^{-2}\left(\left(z-\mu\right)^2-\sigma\right)h(z)\right] = \mathbb{E}_q\left[\sigma^{-1}\left(z-\mu\right)\nabla_z h(z)\right]$$

The key idea of the proof is that we define an auxiliary function $f(z) := \sigma^{-1}\left(z-\mu\right)h(z)$ and apply Lemma 1 to $f(z)$. By the lemma, we have the following result.

$$\begin{aligned}\mathbb{E}_q\left[\sigma^{-2}\left(z-\mu\right)^2 h(z)\right] &= \mathbb{E}_q\left[\sigma^{-1}\left(z-\mu\right)f(z)\right]\\ &= \mathbb{E}_q\left[\nabla_z f(z)\right]\\ &= \mathbb{E}_q\left[\nabla_z\left(\sigma^{-1}\left(z-\mu\right)h(z)\right)\right]\\ &= \mathbb{E}_q\left[\sigma^{-1}h(z) + \sigma^{-1}\left(z-\mu\right)\nabla_z h(z)\right]\end{aligned}$$

From this expression, we can easily obtain the gradient identity. Now, we discuss the identity to compute the reparameterizable gradient for the variance.

Lemma 3 *Let $h(z) : \mathcal{R} \mapsto \mathcal{R}$ be locally AC. We assume the conditions of Lemma 2 are satisfied. The following gradient identity holds.*

$$\nabla_\sigma \mathbb{E}_q\left[h(z)\right] = \tfrac{1}{2}\mathbb{E}_q\left[\sigma^{-1}\left(z-\mu\right)\nabla_z h(z)\right],$$

Likewise, the proof of this lemma is trivial. First, we swap differentiation and integration. We obtain the following expression $\nabla_\sigma \mathbb{E}_q\left[h(z)\right] = \frac{1}{2}\mathbb{E}_q\left[\sigma^{-2}\left(\left(z-\mu\right)^2-\sigma\right)h(z)\right]$. After that, we obtain the above identity by lemma 2. At this point, we can see that the reparameterizable gradient for Gaussian can be derived from Stein's lemma. Furthermore, Stein's lemma empirically gives a low-variance and unbiased gradient estimator if we allow to use the second-order information. This idea is known as Price's theorem. Before we discuss Price's theorem, we first describe the following lemma.

Lemma 4 *Let $h(z) : \mathcal{R} \mapsto \mathcal{R}$ be continuously differentiable. Additionally, its derivative $\nabla_z h(z) : \mathcal{R} \mapsto \mathcal{R}$ is locally AC. $q(z)$ is an univariate Gaussian distribution denoted by $\mathcal{N}(z|\mu,\sigma)$ with mean μ and variance σ. We have the following identity.*

$$\mathbb{E}_q\left[\nabla_z^2 h(z)\right] = \mathbb{E}_q\left[\sigma^{-1}\left(z-\mu\right)\nabla_z h(z)\right].$$

The key idea of the proof that we define auxiliary functions $f(z) := \nabla h(z)$ and apply Lemma 1 to $f(z)$.

Using the above lemmas, we obtain Price's theorem as shown below.

Theorem 2 (Price's Theorem): *Let $h(z) : \mathcal{R} \mapsto \mathcal{R}$ be continuously differentiable and its derivative $\nabla h(\mathbf{z})$ be locally AC. We further assume $\mathbb{E}_q\left[h(z)\right]$ is well-defined. The following second-order identity holds.*

$$\nabla_\sigma \mathbb{E}_q\left[h(z)\right] = \tfrac{1}{2}\mathbb{E}_q\left[\sigma^{-1}\left(z-\mu\right)\nabla_z^T h(z)\right] = \tfrac{1}{2}\mathbb{E}_q\left[\nabla_z^2 h(z)\right]$$

The above theorem can be readily shown by Lemma 4 and Lemma 3.

Now, we describe Stein's lemma for a multivariate Gaussian with arbitrary covariance structure.

Lemma 5 (Stein's Lemma): *Let $h(\mathbf{z}) : \mathcal{R}^d \mapsto \mathcal{R}$ be locally ACL and continuous. $q(\mathbf{z})$ be a multivariate Gaussian distribution denoted by $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. The following identity holds.*

$$\mathbb{E}_q\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)h(\mathbf{z})\right] = \mathbb{E}_q\left[\nabla_z h(\mathbf{z})\right].$$

The proof can be found at Appendix A.2. Bonnet's theorem and Price's theorem are given below. The proof of Bonnet's theorem and Price's theorem can be found at Appendix A.3 and A.4 respectively.

Theorem 3 (Bonnet's Theorem): *Let $h(\mathbf{z}) : \mathcal{R}^d \mapsto \mathcal{R}$ be locally ACL and continuous. $q(\mathbf{z})$ be a multivariate Gaussian distribution denoted by $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})$. The following first-order identity holds.*

$$\nabla_\mu \mathbb{E}_q\left[h(\mathbf{z})\right] = \mathbb{E}_q\left[\nabla_z h(\mathbf{z})\right] = \mathbb{E}_q\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)h(\mathbf{z})\right]$$

Theorem 4 (Price's Theorem): *Let $h(\mathbf{z}) : \mathcal{R}^d \mapsto \mathcal{R}$ be continuously differentiable and its derivative $\nabla h(\mathbf{z})$ be locally ACL. Furthermore, we assume $\mathbb{E}_q [h(\mathbf{z})]$ is well-defined. The following second-order identity holds.*

$$\begin{aligned} \nabla_\Sigma \mathbb{E}_q [h(\mathbf{z})] &= \tfrac{1}{2}\mathbb{E}_q \left[\mathbf{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})\nabla_z^T h(\mathbf{z})\right] \\ &= \tfrac{1}{2}\mathbb{E}_q \left[\nabla_z^2 h(\mathbf{z})\right] \end{aligned}$$

6. Identities for Univariate Continuous Exponential-family

We can generalize Stein's identity to a class of exponential family. First of all, we say a function is locally AC with its domain (l, u), where $-\infty \leq l < u \leq \infty$ if the function is AC at every compact interval inside its domain. We consider the following exponential-family (EF) distribution with $z \in (l, u)$, where $-\infty \leq l < u \leq \infty$. Furthermore, we assume $q(z|\boldsymbol{\lambda})$ is locally AC w.r.t z and differentiable w.r.t. $\boldsymbol{\lambda}$. In the case when $\boldsymbol{\phi}_z(\boldsymbol{\lambda}) = \boldsymbol{\lambda}$, it can be shown that $q(z|\boldsymbol{\lambda})$ is differentiable w.r.t. $\boldsymbol{\lambda}$.

$$q(z|\boldsymbol{\lambda}) = h_z(z)\exp\left\{\langle \mathbf{T}_z(z), \boldsymbol{\phi}_z(\boldsymbol{\lambda})\rangle - A_z(\boldsymbol{\lambda})\right\}$$

where l and u do not depend on $\boldsymbol{\lambda}$.

We denote the CDF of $q(z|\boldsymbol{\lambda})$ by $\psi(z, \boldsymbol{\lambda}) := \int_l^z q(t|\boldsymbol{\lambda})dt$. The following assumption is known as the boundary condition in the literature.

$$\lim_{z\downarrow l} h(z)q(z|\boldsymbol{\lambda}) = 0, \;\; \lim_{z\uparrow u} h(z)q(z|\boldsymbol{\lambda}) = 0$$

Lemma 6 (Stein's Lemma): *Let $h(\cdot) : (l, u) \mapsto \mathcal{R}$ be locally AC. If the boundary condition is satisfied, the following identity holds.*

$$-\mathbb{E}_q \left[h(z)\frac{\nabla_z q(z|\boldsymbol{\lambda})}{q(z|\boldsymbol{\lambda})}\right] = \mathbb{E}_q \left[\nabla_z h(z)\right].$$

In the Gaussian case, we can further exploit the structure of Gaussian as shown in Lemma 8 so that the boundary condition is implicitly satisfied. For general cases, we have to explicitly assume that the boundary condition is satisfied. The proof is exactly the same as the proof for Lemma 1 as shown in A.1.

Applying Lemma 6 to $\tilde{f}_i(z)$ defined below, we obtain the implicit reparameterization trick .

Theorem 5 (Implicit Reparametrization Trick) : *Let $h(\cdot) : (l, u) \mapsto \mathcal{R}$ be a locally AC function. We define $f_i(z) := \frac{\nabla_{\lambda_i}\psi(z,\boldsymbol{\lambda})}{q(z|\boldsymbol{\lambda})}$ where λ_i is a scalar. If the conditions of Lemma 6 for $\tilde{f}_i(z) := h(z)f_i(z)$ are satisfied, we have the following identity.*

$$\nabla_{\lambda_i}\mathbb{E}_q [h(z)] = -\mathbb{E}_q \left[f_i(z)\nabla_z h(z)\right],$$

7. Identities for Exponential-family Mixtures

7.1. Identities for Gaussian Variance-mean Mixtures

We consider the following Gaussian mixture.

$$\begin{aligned} q(w, \mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\alpha}, \mathbf{\Sigma}) :=& \mathcal{N}(\mathbf{z}|\boldsymbol{\mu} + u(w)\boldsymbol{\alpha}, v(w)\mathbf{\Sigma})q(w) \\ q(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\alpha}, \mathbf{\Sigma}) :=& \int q(w, \mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\alpha}, \mathbf{\Sigma})dw \end{aligned}$$

where $v(w) > 0$.

Theorem 6 (Bonnet's Theorem): *Let $h(\mathbf{z}) : \mathcal{R}^d \mapsto \mathcal{R}$ be locally ACL and continuous. $q(\mathbf{z})$ is a Gaussian variance-mean mixture and $q(w, \mathbf{z})$ is the joint distribution. The following gradient identity holds.*

$$\begin{aligned} \nabla_\mu \mathbb{E}_{q(z)} [h(\mathbf{z})] =& \mathbb{E}_{q(z)} [\nabla_z h(\mathbf{z})] \\ \nabla_\alpha \mathbb{E}_{q(z)} [h(\mathbf{z})] =& \mathbb{E}_{q(w,z)} [u(w)\nabla_z h(\mathbf{z})] \end{aligned}$$

Corollary 6.1 *If $u(w)$ has the following property,*

$$\int u(w)q(w, \mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\alpha}, \mathbf{\Sigma})dw = \sum_j^k u_j(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \mathbf{\Sigma})\hat{q}_j(\mathbf{z}),$$

where each $\hat{q}_j(\mathbf{z})$ is a normalized distribution and k is finite, the following identity also holds.

$$\nabla_\alpha \mathbb{E}_{q(z)} [h(\mathbf{z})] = \sum_j^k \mathbb{E}_{\hat{q}_j(z)} [u_j(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \mathbf{\Sigma})\nabla_z h(\mathbf{z})]$$

Example 6.1 *A concrete example is the multivariate skew Gaussian distribution, which can be found at Appendix C.2.*

Example 6.2 *Another example is the multivariate exponentially modified Gaussian distribution, which can be found at Appendix C.3.*

Theorem 7 (Price's Theorem): *Let $h(\mathbf{z}) : \mathcal{R}^d \mapsto \mathcal{R}$ be continuously differentiable that its derivative $\nabla h(\mathbf{z})$ be locally ACL. If $\mathbb{E}_{q(w,z)} [v(w)h(\mathbf{z})]$ is well-defined, the following identity holds.*

$$\begin{aligned} \nabla_\Sigma \mathbb{E}_{q(z)} [h(\mathbf{z})] &= \tfrac{1}{2}\mathbb{E}_{q(w,z)} \left[v(w)\nabla_z^2 h(\mathbf{z})\right] \\ &= \tfrac{1}{2}\mathbb{E}_{q(w,z)} \left[\mathbf{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu}-u(w)\boldsymbol{\alpha})\nabla_z^T h(\mathbf{z})\right] \end{aligned}$$

Corollary 7.1 *If $v(w)$ has the following property,*

$$\int v(w)q(w, \mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\alpha}, \mathbf{\Sigma})dw = \sum_j^k v_j(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \mathbf{\Sigma})\hat{q}_j(\mathbf{z}),$$

where each $\hat{q}_j(\mathbf{z})$ is a normalized distribution and k is finite, the following identity also holds.

$$\begin{aligned} \nabla_\Sigma \mathbb{E}_{q(z)} [h(\mathbf{z})] &= \tfrac{1}{2}\mathbb{E}_{q(w,z)} \left[v(w)\nabla_z^2 h(\mathbf{z})\right] \\ &= \tfrac{1}{2}\sum_j^k \mathbb{E}_{\hat{q}_j(z)} \left[v_j(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \mathbf{\Sigma})\nabla_z^2 h(\mathbf{z})\right] \end{aligned}$$

Example 7.1 *A concrete example is the multivariate Student's t-distribution, which can be found at Appendix C.5.*

Example 7.2 *Another example is the multivariate normal inverse-Gaussian distribution, which can be found at Appendix C.6.*

7.2. Identities for Continuous Exponential-family Mixtures

We consider the following EF mixtures in a product space $\mathbf{z} = (z_1, z_2) \in (l_1, u_1) \times (l_2, u_2)$, where $-\infty \leq l_1 < u_1 \leq \infty$ and $-\infty \leq l_2 < u_2 \leq \infty$. Moreover, we assume $q(z_1)$ $q(z_2|z_1)$, and $q(z_1, z_2)$ are locally AC, locally ACL, and continuous, respectively.

$$q(z_1|\boldsymbol{\lambda}) = h_1(z_1) \exp\left\{\langle \mathbf{T}_1(z_1), \boldsymbol{\phi}_1(\boldsymbol{\lambda})\rangle - A_1(\boldsymbol{\lambda})\right\}$$
$$q(z_2|z_1, \boldsymbol{\lambda}) = h_2(z_2, z_1) \exp\left\{\langle \mathbf{T}_2(z_2, z_1), \boldsymbol{\phi}_2(\boldsymbol{\lambda})\rangle - A_2(\boldsymbol{\lambda}, z_1)\right\}$$

Let's denote the CDF of $q(z_1|\boldsymbol{\lambda})$ and the conditional CDF of $q(z_2|z_1, \boldsymbol{\lambda})$ by

$$\psi_1(z_1, \boldsymbol{\lambda}) = \int_{l_1}^{z_1} q(t_1|\boldsymbol{\lambda}) dt_1$$
$$\psi_2(z_1, z_2, \boldsymbol{\lambda}) = \int_{l_2}^{z_2} q(t_2|z_1, \boldsymbol{\lambda}) dt_2.$$

We define the following functions:

$$\boldsymbol{\Psi}(\mathbf{z}, \boldsymbol{\lambda}) := \left[\psi_1(z_1, \boldsymbol{\lambda}), \psi_2(z_1, z_2, \boldsymbol{\lambda})\right]^T$$
$$\nabla_{\lambda_i} \boldsymbol{\Psi}(\mathbf{z}, \boldsymbol{\lambda}) := \left[\nabla_{\lambda_i} \psi_1(z_1, \boldsymbol{\lambda}), \nabla_{\lambda_i} \psi_2(z_1, z_2, \boldsymbol{\lambda})\right]^T$$
$$\nabla_z \boldsymbol{\Psi}(\mathbf{z}, \boldsymbol{\lambda}) := \begin{bmatrix} q(z_1|\boldsymbol{\lambda}) & 0 \\ \nabla_{z_1} \psi_2(z_1, z_2, \boldsymbol{\lambda}) & q(z_2|z_1, \boldsymbol{\lambda}) \end{bmatrix}.$$

Applying Lemma 6 to $\tilde{f}_{i,j}(z_j)$ defined below, we obtain the following identity.

Theorem 8 (Bivariate Implicit Reparametrization Trick): *Let $h(\cdot) : (l_1, u_1) \times (l_2, u_2) \mapsto \mathcal{R}$ be locally ACL and continuous. First, we define function $f_{i,j}(\mathbf{z}) := \mathbf{e}_j^T \left[\nabla_z \boldsymbol{\Psi}(\mathbf{z}, \boldsymbol{\lambda})\right]^{-1} \nabla_{\lambda_i} \boldsymbol{\Psi}(\mathbf{z}, \boldsymbol{\lambda})$ and function $\tilde{f}_{i,j}(z_j) := f_{i,j}(z_j, \mathbf{z}_{-j}) h(z_j, \mathbf{z}_{-j}) \prod_{k \geq j+1} q(z_k|\mathbf{z}_{1:(k-1)}, \boldsymbol{\lambda})$. If the conditions of Lemma 6 for each $\tilde{f}_{i,j}(z_j)$ are satisfied, we have the following identity.*

$$\nabla_{\lambda_i} \mathbb{E}_q\left[h(\mathbf{z})\right] = -\mathbb{E}_q\left[\sum_j f_{i,j}(\mathbf{z}) \nabla_{z_j} h(\mathbf{z})\right]$$

The proof of this theorem can be found at Appendix D.1. The identity can be easily extended to multivariate version for the implicit reparametrization trick. Figurnov et al. (2018) assume that $h(\mathbf{z})$ is continuously differentiable. As shown in Theorem 8, the identity holds even when $h(\mathbf{z})$ is not continuously differentiable.

References

Adcock, C. Extensions of Stein's lemma for the skew-normal distribution. *Communications in Statistics-Theory and Methods*, 36(9):1661–1671, 2007.

Adcock, C. and Shutes, K. On the multivariate extended skew-normal, normal-exponential, and normal-gamma distributions. *Journal of Statistical Theory and Practice*, 6(4):636–664, 2012.

Arnold, B. C., Castillo, E., and Sarabia, J. M. A multivariate version of stein's identity with applications to moment calculations and estimation of conditionally specified distributions. 2001.

Bonnet, G. Transformations des signaux aléatoires a travers les systemes non linéaires sans mémoire. In *Annales des Télécommunications*, volume 19, pp. 203–220. Springer, 1964.

Border, K. C. Lecture Notes: integration by parts. `www.its.caltech.edu/~kcborder/Notes/IntegrationByParts.pdf`, 1996. Accessed: 2019/06.

Brown, L., DasGupta, A., Haff, L. R., and Strawderman, W. E. The heat equation and Stein's identity: Connections, applications. *Journal of Statistical Planning and Inference*, 136(7):2254–2278, 2006.

Brown, L. D. Fundamentals of statistical exponential families: with applications in statistical decision theory. Ims, 1986.

Erdogdu, M. A. Newton-Stein method: a second order method for GLMs via Stein's Lemma. In *Advances in Neural Information Processing Systems*, pp. 1216–1224, 2015.

Fan, K., Wang, Z., Beck, J., Kwok, J., and Heller, K. A. Fast second order stochastic backpropagation for variational inference. In *Advances in Neural Information Processing Systems*, pp. 1387–1395, 2015.

Figurnov, M., Mohamed, S., and Mnih, A. Implicit Reparameterization Gradients. 2018.

Hudson, H. M. et al. A natural identity for exponential families with applications in multiparameter estimation. *The Annals of Statistics*, 6(3):473–484, 1978.

Jia, R.-Q. Lecture Notes: Honors Real Variable II. `sites.ualberta.ca/~rjia/Math418/Notes/chap3.pdf`, 2010. Accessed: 2019/03/25.

Kattumannil, S. K. On Steins identity and its applications. *Statistics & Probability Letters*, 79(12):1444–1449, 2009.

Kattumannil, S. K. and Dewan, I. On generalized moment identity and its applications: a unified approach. *Statistics*, 50(5):1149–1160, 2016.

Khan, M. and Lin, W. Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. In *Artificial Intelligence and Statistics*, pp. 878–887, 2017.

Khan, M. E., Nielsen, D., Tangkaratt, V., Lin, W., Gal, Y., and Srivastava, A. Fast and Scalable Bayesian Deep Learning by Weight-Perturbation in Adam. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2611–2620, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Landsman, Z. On the generalization of Stein's Lemma for elliptical class of distributions. *Statistics & probability letters*, 76(10):1012–1016, 2006.

Landsman, Z. and Nešlehová, J. Stein's Lemma for elliptical random vectors. *Journal of Multivariate Analysis*, 99(5): 912–927, 2008.

Leoni, G. *A first course in Sobolev spaces*, volume 181. American Mathematical Soc., 2017.

Lin, W., Khan, M. E., and Schmidt, M. Fast and simple natural-gradient variational inference with mixture of exponential-family approximations. In *International Conference on Machine Learning*, pp. 3992–4002, 2019.

Liu, J. S. Siegel's formula via Stein's identities. *Statistics & Probability Letters*, 21(3):247–251, 1994.

Opper, M. and Archambeau, C. The variational Gaussian approximation revisited. *Neural computation*, 21(3):786–792, 2009.

Park, S., Serpedin, E., and Qaraqe, K. On the equivalence between Stein and de Bruijn identities. *IEEE Transactions on Information Theory*, 58(12):7045–7067, 2012.

Price, R. A useful theorem for nonlinear devices having Gaussian inputs. *IRE Transactions on Information Theory*, 4(2):69–72, 1958.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Royen, H. and Fitzpatrick, P. Real analysis. 2010.

Salimans, T. and Knowles, D. Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8(4):837–882, 2013.

Stein, C. Estimation of the Mean of a Multivariate Normal Distribution. *Proc. Prague Sympos. Asymptotic Statistics*, 1973.

Stein, C. M. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pp. 1135–1151, 1981.

A. Gradient Identities for Gaussian Distribution

For completeness, we give a proof of integration by parts for AC functions, which is essential for many proofs in this paper.

Lemma 7 (Integration by parts): *Let $h(\cdot), q(\cdot) : [a, b] \mapsto \mathcal{R}$ be AC functions, where $-\infty < a < b < \infty$. The following identity holds.*

$$h(b)q(b) - h(a)q(a) = \int_a^b q(z)\nabla_z h(z)dz + \int_a^b h(z)\nabla_z q(z)dz$$

Proof: Since $h(z)$ and $q(z)$ are AC in $[a, b]$, we know that the product $h(z)q(z)$ is AC in $[a, b]$. By the product rule for AC, the following identity holds almost everywhere for $z \in [a, b]$.

$$\nabla_z \left(h(z)q(z)\right) = q(z)\nabla_z h(z) + h(z)\nabla_z q(z) \tag{1}$$

Since $q(z)$ is continuous and $h(z)$ is AC in $[a, b]$, we know that $q(z)\nabla_z h(z)$ is integrable over $[a, b]$. Similarly, we can show $h(z)\nabla_z q(z)$ is integrable over $[a, b]$. Integrating both sides of (1) over $[a, b]$, we obtain the identity.

$$h(b)q(b) - h(a)q(a) = \int_a^b q(z)\nabla_z h(z)dz + \int_a^b h(z)\nabla_z q(z)dz$$

□

An alternative proof of Lemma 7 using Fubini's theorem can be found at Theorem 6 of Border (1996). Note that the condition of Fubini's theorem shown below is satisfied.

$$\int_a^b \int_a^b \left|\nabla_x h(x)\nabla_y q(y)\right| dxdy < \infty,$$

since by the definition of AC, we have

$$\int_a^b \left|\nabla_x h(x)\right| dx < \infty, \quad \int_a^b \left|\nabla_y q(y)\right| dy < \infty.$$

To use integration by parts in the proof of Lemma 1, we first prove the following lemma.

Lemma 8 *Let $h(\cdot) : \mathcal{R} \mapsto \mathcal{R}$ be a locally AC function and $q(z) := \mathcal{N}(z|\mu, \sigma)$ be an univariate Gaussian distribution with mean μ and variance σ. If $\mathbb{E}_q\left[\nabla_z h(z)\right]$ is well-defined ($\mathbb{E}_q\left[\left|\nabla_z h(z)\right|\right] < \infty$), then the following boundary conditions are satisfied:*

$$\lim_{z \uparrow \infty} h(z)q(z) = 0$$
$$\lim_{z \downarrow -\infty} h(z)q(z) = 0$$

Proof: We show that $\lim_{z\uparrow\infty} h(z)q(z) = 0$ by showing that $\lim_{z\uparrow\infty} \left|h(z)\right| q(z) = 0$, where $q(z) = \mathcal{N}(z|\mu, \sigma)$. Given a compact interval, since $h(z)$ is AC, by Theorem 3.1 of Jia (2010), we know that $\left|h(z)\right|$ is also AC and the following identity holds almost everywhere for t in the interval.

$$-\left|\nabla_t h(t)\right| \le \nabla_t \left|h(t)\right| \le \left|\nabla_t h(t)\right|. \tag{2}$$

Since $\left|h(z)\right|$ is AC in the interval, given c in the interval, by the fundamental theorem of calculus, we have

$$\left|h(z)\right| = \left|h(c)\right| + \int_c^z \nabla_t \left|h(t)\right| dt. \tag{3}$$

Given any $z \geq c \geq \mu$, by (3) and then (2), we have

$$\begin{aligned}|h(z)|\, q(z) &= q(z)\left(|h(c)| + \int_c^z \nabla_t \,|h(t)|\, dt\right) \\ &\leq q(z)\left(|h(c)| + \int_c^z |\nabla_t h(t)|\, dt\right) \\ &\leq q(z)\,|h(c)| + \int_c^z q(t)\,|\nabla_t h(t)|\, dt\end{aligned}$$

where we use the monotonicity of Gaussian: $q(z) \leq q(t)$ when $\mu \leq c \leq t \leq z$.

Therefore, by the assumption $\mathbb{E}_{q(t)}\left[|\nabla_t h(t)|\right] < \infty$, we have

$$\lim_{z\uparrow\infty} \sup |h(z)|\, q(z) \leq \underbrace{\lim_{z\uparrow\infty} q(z)\, \overbrace{|h(c)|}^{\text{constant}}}_{0} + \underbrace{\int_c^\infty q(t)\,|\nabla_t h(t)|\, dt}_{<\infty} = \int_c^\infty q(t)\,|\nabla_t h(t)|\, dt,$$

where we use the Gaussian identity that $\lim_{z\uparrow\infty} q(z) = 0$.

Taking c to the positive infinity, we obtain the following result, which implies that $\lim_{z\uparrow\infty} |h(z)|\, q(z) = 0$.

$$0 \leq \lim_{z\uparrow\infty} \inf |h(z)|\, q(z) \leq \lim_{z\uparrow\infty} \sup |h(z)|\, q(z) \leq \underbrace{\lim_{c\uparrow\infty} \int_c^\infty q(t)\,|\nabla_t h(t)|\, dt}_{0}$$

Likewise, we can show that $\lim_{z\downarrow-\infty} h(z)q(z) = 0$. □

A.1. Proof of Lemma 1 and Lemma 6

Proof: First, we denote the support by (l, u). In the Gaussian case, $l = -\infty$ and $u = \infty$. Since $\mathbb{E}_q\left[\nabla_z h(z)\right]$ is well-defined, we use the following expression to prove the claim.

$$\mathbb{E}_q\left[\nabla_z h(z)\right] = \lim_{r_1\downarrow l} \int_{r_1}^c q(z)\nabla_z h(z) dz + \lim_{r_2\uparrow u} \int_c^{r_2} q(z)\nabla_z h(z) dz$$

where $c \in (l, u)$ is a constant number.

Given any compact interval $[r_1, c]$, we know that $h(z)$ and $q(z)$ are AC in this interval. By integration by parts (Lemma 7), we have

$$h(c)q(c) - h(r_1)q(r_1) = \int_{r_1}^c q(z)\nabla_z h(z) dz + \int_{r_1}^c h(z)\nabla_z q(z) dz$$

In the Gaussian case, we have $\lim_{r_1\downarrow l} h(r_1)q(r_1) = 0$ due to Lemma 8. Taking r_1 to l, we have

$$\begin{aligned}h(c)q(c) &= \lim_{r_1\downarrow l}\left[\int_{r_1}^c q(z)\nabla_z h(z) dz + \int_{r_1}^c h(z)\nabla_z q(z) dz\right] \\ &= \lim_{r_1\downarrow l} \int_{r_1}^c q(z)\nabla_z h(z) dz + \lim_{r_1\downarrow l} \int_{r_1}^c h(z)\nabla_z q(z) dz. \qquad (4)\end{aligned}$$

Note that $\lim_{r_1\downarrow l} \int_{r_1}^c q(z)\nabla_z h(z) dz$ exists since $\mathbb{E}_q\left[|\nabla_z h(z)|\right] < \infty$ ($\mathbb{E}_q\left[\nabla_z h(z)\right]$ is well-defined). Since $h(c)q(c)$ is finite, we know that $\lim_{r_1\downarrow l} \int_{r_1}^c h(z)\nabla_z q(z) dz$ is also finite. Therefore, (4) is valid.

Likewise, given any compact interval $[c, r_2]$, by integration by parts and taking r_2 to u, we have

$$-h(c)q(c) = \lim_{r_2\uparrow u} \int_c^{r_2} q(z)\nabla_z h(z) dz + \lim_{r_2\uparrow u} \int_c^{r_2} h(z)\nabla_z q(z) dz. \qquad (5)$$

where $\lim_{r_2 \uparrow u} h(r_2)q(r_2) = 0$.

By (4) and (5), we have

$$\mathbb{E}_q \left[\nabla_z h(z)\right] = -\mathbb{E}_q \left[h(z) \frac{\nabla_z q(z)}{q(z)}\right],$$

In the Gaussian case, we have $\nabla_z q(z) = \sigma^{-1}(\mu - z)q(z)$, which shows that the above expression is the identity. □

A.2. Proof of Lemma 5

Proof: We denote the i-th element of $\mathbf{z}$ by z_i. Given $\mathbf{z}_{-i}$ is known, we define a function $h_i(z_i) := h(z_i, \mathbf{z}_{-i})$. Since $\mathbb{E}_q \left[\nabla_z h(\mathbf{z})\right]$ is well-defined, we have

$$\mathbb{E}_q \left[\nabla_{z_i} h(\mathbf{z})\right] = \mathbb{E}_{q(z_{-i})q(z_i|z_{-i})} \left[\nabla_{z_i} h(z_i, \mathbf{z}_{-i})\right] = \mathbb{E}_{q(z_{-i})} \left[\mathbb{E}_{q(z_i|z_{-i})} \left[\nabla_{z_i} h_i(z_i)\right]\right]$$

Without loss of generality, we assume z_i is the last element of $\mathbf{z}$. It is possible since we can permute the elements of $\mathbf{z}$ to achieve that. Therefore, we can re-express the mean and the covariance matrix as below.

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_{-i} \\ \mu_i \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{-i,-i} & \boldsymbol{\Sigma}_{-i,i} \\ \boldsymbol{\Sigma}_{-i,i}^T & \Sigma_{i,i} \end{bmatrix}$$

Note that $q(z_i|\mathbf{z}_{-i})$ is an univariate Gaussian distribution denote by $\mathcal{N}(z_i|m, \sigma)$, where

$$m = \mu_i + \boldsymbol{\Sigma}_{-i,i}^T \boldsymbol{\Sigma}_{-i,-i}^{-1} \left(\mathbf{z}_{-i} - \boldsymbol{\mu}_{-i}\right) \tag{6}$$

$$\sigma = \Sigma_{i,i} - \boldsymbol{\Sigma}_{-i,i}^T \boldsymbol{\Sigma}_{-i,-i}^{-1} \boldsymbol{\Sigma}_{-i,i}. \tag{7}$$

Since $h_i(z_i)$ is locally AC, we have the following result by applying Lemma 1 to $h_i(z_i)$.

$$\begin{aligned} \mathbb{E}_q \left[\nabla_{z_i} h(\mathbf{z})\right] &= \mathbb{E}_{q(z_{-i})} \left[\mathbb{E}_{q(z_i|z_{-i})} \left[\nabla_{z_i} h_i(z_i)\right]\right] \\ &= \mathbb{E}_{q(z_{-i})} \left[\mathbb{E}_{q(z_i|z_{-i})} \left[\sigma^{-1} \left(z_i - m\right) h_i(z_i)\right]\right] \\ &= \mathbb{E}_q \left[\sigma^{-1} \left(z_i - m\right) h_i(z_i)\right] \\ &= \mathbb{E}_q \left[\sigma^{-1} \left(z_i - m\right) h(\mathbf{z})\right] \end{aligned}$$

Recall that by assumptions the above expectations are well-defined. It can be verified that

$$\sigma^{-1} \left(z_i - m\right) = \mathbf{e}_i^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{z} - \boldsymbol{\mu}\right) \tag{8}$$

where $\mathbf{e}_i$ is an one-hot vector where it has all zero elements except the i-th element with value 1.

Using the result at (8), we have

$$\mathbb{E}_q \left[\nabla_{z_i} h(\mathbf{z})\right] = \mathbb{E}_q \left[\sigma^{-1} \left(z_i - m\right) h(\mathbf{z})\right] = \mathbb{E}_q \left[\mathbf{e}_i^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{z} - \boldsymbol{\mu}\right) h(\mathbf{z})\right] \tag{9}$$

Therefore, we have

$$\mathbb{E}_q \left[\nabla_z h(\mathbf{z})\right] = \mathbb{E}_q \left[\boldsymbol{\Sigma}^{-1} \left(\mathbf{z} - \boldsymbol{\mu}\right) h(\mathbf{z})\right]$$

□

A.3. Proof of Theorem 3

Proof: We swap integration and differentiation and obtain the following result.

$$\begin{aligned} \nabla_\mu \mathbb{E}_q \left[h(\mathbf{z})\right] &= \int h(\mathbf{z}) \nabla_\mu \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{z} \\ &= \int h(\mathbf{z}) \boldsymbol{\Sigma}^{-1} \left(\mathbf{z} - \boldsymbol{\mu}\right) \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{z} \\ &= \mathbb{E}_q \left[\boldsymbol{\Sigma}^{-1} \left(\mathbf{z} - \boldsymbol{\mu}\right) h(\mathbf{z})\right] \end{aligned}$$

which is known as the score function estimator.

Using Lemma 5 to move from line 1 to line 2, we obtain the gradient identity given below.

$$\begin{aligned}\nabla_\mu \mathbb{E}_q\left[h(\mathbf{z})\right] &= \mathbb{E}_q\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)h(\mathbf{z})\right]\\ &= \mathbb{E}_q\left[\nabla_z h(\mathbf{z})\right]\end{aligned}$$

which is known as the re-parametrization trick for the mean $\boldsymbol{\mu}$.

□

A.4. Proof of Theorem 4

To prove Theorem 4, we first prove the following lemma, which is a multivariate extension of Lemma 2. By convention, $\mathbf{e}_i$ is an one-hot vector where it has all zero elements except the i-th element with value 1.

Lemma 9 *Let $h(\mathbf{z}) : \mathcal{R}^d \mapsto \mathcal{R}$ be locally ACL and continuous. We define an auxiliary vector function $\mathbf{f}(\mathbf{z}) = \boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)h(\mathbf{z})$. If $\mathbb{E}_q\left[h(\mathbf{z})\right]$ is well-defined, the following identity holds.*

$$\mathbb{E}_q\left[\boldsymbol{\Sigma}^{-1}\left[\left(\mathbf{z}-\boldsymbol{\mu}\right)\left(\mathbf{z}-\boldsymbol{\mu}\right)^T-\boldsymbol{\Sigma}\right]\boldsymbol{\Sigma}^{-1}h(\mathbf{z})\right] = \mathbb{E}_q\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)\nabla_z^T h(\mathbf{z})\right]$$

Proof: We define an auxiliary function $f_i(\mathbf{z}) := \mathbf{e}_i^T\mathbf{f}(\mathbf{z})$. By applying Lemma 5 to $f_i(\mathbf{z})$, we have

$$\mathbb{E}_q\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)\mathbf{e}_i^T\mathbf{f}(\mathbf{z})\right] = \mathbb{E}_q\left[\nabla_z\left(\mathbf{e}_i^T\mathbf{f}(\mathbf{z})\right)\right]$$

Recall that $\mathbb{E}_q\left[\nabla_z\left(\mathbf{e}_i^T\mathbf{f}(\mathbf{z})\right)\right] = \mathbb{E}_q\left[\boldsymbol{\Sigma}^{-1}\mathbf{e}_i^T h(\mathbf{z}) + \mathbf{e}_i^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)\nabla_z h(\mathbf{z})\right]$.

Therefore, we know that

$$\begin{aligned}\mathbb{E}_q\left[\mathbf{e}_j^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\mathbf{e}_i h(\mathbf{z})\right] =& \mathbb{E}_q\left[\mathbf{e}_j^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)\mathbf{e}_i^T\mathbf{f}(\mathbf{z})\right]\\ =& \mathbb{E}_q\left[\mathbf{e}_j^T\nabla_z\left(\mathbf{e}_i^T\mathbf{f}(\mathbf{z})\right)\right]\\ =& \mathbb{E}_q\left[\mathbf{e}_j^T\boldsymbol{\Sigma}^{-1}\mathbf{e}_i h(\mathbf{z}) + \mathbf{e}_i^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)\mathbf{e}_j^T\nabla_z h(\mathbf{z})\right]\end{aligned}$$

Since $\mathbb{E}_q\left[h(\mathbf{z})\right]$ is also well-defined, we have

$$\begin{aligned}\mathbb{E}_q\left[\mathbf{e}_i^T\boldsymbol{\Sigma}^{-1}\left(\left(\mathbf{z}-\boldsymbol{\mu}\right)\left(\mathbf{z}-\boldsymbol{\mu}\right)^T-\boldsymbol{\Sigma}\right)\boldsymbol{\Sigma}^{-1}\mathbf{e}_j h(\mathbf{z})\right] =& \mathbb{E}_q\left[\mathbf{e}_j^T\boldsymbol{\Sigma}^{-1}\left(\left(\mathbf{z}-\boldsymbol{\mu}\right)\left(\mathbf{z}-\boldsymbol{\mu}\right)^T-\boldsymbol{\Sigma}\right)\boldsymbol{\Sigma}^{-1}\mathbf{e}_i h(\mathbf{z})\right]\\ =& \mathbb{E}_q\left[\mathbf{e}_i^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)\mathbf{e}_j^T\nabla_z h(\mathbf{z})\right]\\ =& \mathbb{E}_q\left[\mathbf{e}_i^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)\nabla_z^T h(\mathbf{z})\mathbf{e}_j\right],\end{aligned}$$

which implies that

$$\mathbb{E}_q\left[\boldsymbol{\Sigma}^{-1}\left(\left(\mathbf{z}-\boldsymbol{\mu}\right)\left(\mathbf{z}-\boldsymbol{\mu}\right)^T-\boldsymbol{\Sigma}\right)\boldsymbol{\Sigma}^{-1}h(\mathbf{z})\right] = \mathbb{E}_q\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)\nabla_z^T h(\mathbf{z})\right].$$

□

Next, we prove the following lemma, which is a multivariate extension of Lemma 3.

Lemma 10 *Let $h(\mathbf{z}) : \mathcal{R}^d \mapsto \mathcal{R}$ be locally ACL and continuous. We assume the conditions of Lemma 9 are satisfied. The following gradient identity holds.*

$$\nabla_\Sigma \mathbb{E}_q\left[h(\mathbf{z})\right] = \tfrac{1}{2}\mathbb{E}_q\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)\nabla_z^T h(\mathbf{z})\right]$$

Proof: By the assumptions, we can interchange the integration and differentiation to obtain the following result.

$$\begin{aligned}\nabla_\Sigma \mathbb{E}_q\left[h(\mathbf{z})\right] &= \int h(\mathbf{z}) \nabla_\Sigma \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{z} \\ &= \tfrac{1}{2} \int h(\mathbf{z}) \boldsymbol{\Sigma}^{-1} \left[\left(\mathbf{z}-\boldsymbol{\mu}\right)\left(\mathbf{z}-\boldsymbol{\mu}\right)^T - \boldsymbol{\Sigma} \right] \boldsymbol{\Sigma}^{-1} \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{z} \\ &= \tfrac{1}{2} \mathbb{E}_q \left[\boldsymbol{\Sigma}^{-1} \left[\left(\mathbf{z}-\boldsymbol{\mu}\right)\left(\mathbf{z}-\boldsymbol{\mu}\right)^T - \boldsymbol{\Sigma} \right] \boldsymbol{\Sigma}^{-1} h(\mathbf{z}) \right]\end{aligned}$$

which is known as the score function estimator.

By Lemma 9, we have

$$\begin{aligned}\nabla_\Sigma \mathbb{E}_q\left[h(\mathbf{z})\right] &= \tfrac{1}{2} \mathbb{E}_q \left[\boldsymbol{\Sigma}^{-1} \left[\left(\mathbf{z}-\boldsymbol{\mu}\right)\left(\mathbf{z}-\boldsymbol{\mu}\right)^T - \boldsymbol{\Sigma} \right] \boldsymbol{\Sigma}^{-1} h(\mathbf{z}) \right] \\ &= \tfrac{1}{2} \mathbb{E}_q \left[\boldsymbol{\Sigma}^{-1} \left(\mathbf{z}-\boldsymbol{\mu}\right) \nabla_z^T h(\mathbf{z}) \right].\end{aligned}$$

□

The following lemma is also useful when we prove Theorem 4. This lemma is a multivariate extension of Lemma 4.

Lemma 11 *Let a function $h(\mathbf{z}) : \mathcal{R}^d \mapsto \mathcal{R}$ be continuously differentiable and its derivative $\nabla_z h(\mathbf{z}) : \mathcal{R}^d \mapsto \mathcal{R}^d$ be locally ACL. $q(\mathbf{z})$ is a multivariate Gaussian distribution denoted by $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. The following identity holds.*

$$\mathbb{E}_q\left[\nabla_z^2 h(\mathbf{z})\right] = \mathbb{E}_q\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)\nabla_z^T h(\mathbf{z})\right], \tag{10}$$

Proof: We define an auxiliary function $g_i(\mathbf{z}) = \nabla_z^T h(\mathbf{z}) \mathbf{e}_i$. By applying Lemma 5 to $g_i(\mathbf{z})$, we have

$$\mathbb{E}_q\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right) g_i(\mathbf{z})\right] = \mathbb{E}_q\left[\nabla_z g_i(\mathbf{z})\right].$$

Therefore, we have

$$\begin{aligned}\mathbb{E}_q\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)\nabla_z^T h(\mathbf{z})\mathbf{e}_i\right] =& \mathbb{E}_q\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right) g_i(\mathbf{z})\right] \\ =& \mathbb{E}_q\left[\nabla_z g_i(\mathbf{z})\right] \\ =& \mathbb{E}_q\left[\nabla_z\left(\nabla_z^T h(\mathbf{z})\mathbf{e}_i\right)\right] \\ =& \mathbb{E}_q\left[\nabla_z^2 h(\mathbf{z})\mathbf{e}_i\right],\end{aligned}$$

which implies that

$$\mathbb{E}_q\left[\nabla_z^2 h(\mathbf{z})\right] = \mathbb{E}_q\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)\nabla_z^T h(\mathbf{z})\right] \tag{11}$$

□

Now, it is time for us to prove Theorem 4.

Proof: Note that all conditions of Lemma 11 are satisfied. By Lemma 11, we have

$$\mathbb{E}_q\left[\nabla_z^2 h(\mathbf{z})\right] = \mathbb{E}_q\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)\nabla_z^T h(\mathbf{z})\right], \tag{12}$$

Since all conditions of Lemma 10 are satisfied, by Lemma 10, we have

$$\nabla_\Sigma \mathbb{E}_q\left[h(\mathbf{z})\right] = \tfrac{1}{2}\mathbb{E}_q\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)\nabla_z^T h(\mathbf{z})\right] \tag{13}$$

Therefore, by (12) and (13) we have

$$\nabla_\Sigma \mathbb{E}_q\left[h(\mathbf{z})\right] = \tfrac{1}{2}\mathbb{E}_q\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)\nabla_z^T h(\mathbf{z})\right] = \tfrac{1}{2}\mathbb{E}_q\left[\nabla_z^2 h(\mathbf{z})\right]$$

□

B. Gradient Identities for Univariate Continuous Exponentially-family Distributions

B.1. Proof of Theorem 5

Proof: It is easy to verify that $\tilde{f}_i(z) = f_i(z)h(z)$ is locally AC since $f_i(z)$ and $h(z)$ are both locally AC. By Lemma 6, we have

$$-\mathbb{E}_q\left[\tilde{f}_i(z)\frac{\nabla_z q(z|\boldsymbol{\lambda}_z)}{q(z|\boldsymbol{\lambda}_z)}\right] = \mathbb{E}_q\left[\nabla_z \tilde{f}_i(z)\right] \tag{14}$$

Notice that

$$\nabla_z \tilde{f}_i(z) = -\Big[\underbrace{\frac{\nabla_{\lambda_i}\psi(z,\boldsymbol{\lambda}_z)}{q(z|\boldsymbol{\lambda}_z)}}_{=f_i(z)}\nabla_z h(z) + \frac{h(z)\nabla_{\lambda_i} q(z|\boldsymbol{\lambda}_z)}{q(z|\boldsymbol{\lambda}_z)} + \frac{\tilde{f}_i(z)\nabla_z q(z|\boldsymbol{\lambda}_z)}{q(z|\boldsymbol{\lambda}_z)}\Big]$$

The expression (14) can be re-expressed as

$$-\mathbb{E}_q\Big[\cancel{\tilde{f}_i(z)\frac{\nabla_z q(z|\boldsymbol{\lambda}_z)}{q(z|\boldsymbol{\lambda}_z)}}\Big] = -\mathbb{E}_q\Big[f_i(z)\nabla_z h(z) + \frac{h(z)\nabla_{\lambda_i} q(z|\boldsymbol{\lambda}_z)}{q(z|\boldsymbol{\lambda}_z)} + \cancel{\frac{\tilde{f}_i(z)\nabla_z q(z|\boldsymbol{\lambda}_z)}{q(z|\boldsymbol{\lambda}_z)}}\Big] \tag{15}$$

By (15), we have the following identity

$$\mathbb{E}_q\left[\frac{h(z)\nabla_{\lambda_i} q(z|\boldsymbol{\lambda}_z)}{q(z|\boldsymbol{\lambda}_z)}\right] = -\mathbb{E}_q\left[f_i(z)\nabla_z h(z)\right]$$

Since we can interchange the integration and differentiation, we know that

$$\nabla_{\lambda_i}\mathbb{E}_q\left[h(z)\right] = \mathbb{E}_q\left[\frac{h(z)\nabla_{\lambda_i} q(z|\boldsymbol{\lambda}_z)}{q(z|\boldsymbol{\lambda}_z)}\right] = -\mathbb{E}_q\left[f_i(z)\nabla_z h(z)\right]$$

□

C. Gradient Identities for Gaussian Variance-mean Mixtures

C.1. Proof of Theorem 6

Proof: Let's consider the gradient identity for $\boldsymbol{\alpha}$. By the assumptions, we can the integration and differentiation to obtain the following expression.

$$\begin{aligned}
\nabla_\alpha \mathbb{E}_{q(z)}\left[h(\mathbf{z})\right] &= \int \nabla_\alpha q(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma})h(\mathbf{z})d\mathbf{z} \\
&= \int \nabla_\alpha \left[\int q(w,\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma})dw\right] h(\mathbf{z})d\mathbf{z} \\
&= \int \left[\int \frac{u(w)}{v(w)}\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}-u(w)\boldsymbol{\alpha}\right) q(w,\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma})dw\right] h(\mathbf{z})d\mathbf{z} \\
&= \mathbb{E}_{q(w,z)}\left[\frac{u(w)}{v(w)}\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}-u(w)\boldsymbol{\alpha}\right)h(\mathbf{z})\right]
\end{aligned}$$

Recall that $q(\mathbf{z}|w)$ is Gaussian denoted by $q(\mathbf{z}|w) := \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}+u(w)\boldsymbol{\alpha}, v(w)\boldsymbol{\Sigma})$. By applying Lemma 5 to $u(w)h(\mathbf{z})$, we have

$$\begin{aligned}
\mathbb{E}_{q(w,z)}\left[\nabla_z\left(u(w)h(\mathbf{z})\right)\right] &= \mathbb{E}_{q(w)}\left[\mathbb{E}_{q(z|w)}\left[\nabla_z\left(u(w)h(\mathbf{z})\right)\right]\right] \\
&= \mathbb{E}_{q(w)}\left[\mathbb{E}_{q(z|w)}\left[\left(v(w)\boldsymbol{\Sigma}\right)^{-1}\left(\mathbf{z}-\boldsymbol{\mu}-u(w)\boldsymbol{\alpha}\right)\left(u(w)h(\mathbf{z})\right)\right]\right] \\
&= \mathbb{E}_{q(w,z)}\left[\frac{u(w)}{v(w)}\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}-u(w)\boldsymbol{\alpha}\right)h(\mathbf{z})\right]
\end{aligned}$$

Therefore, we have

$$\nabla_{\alpha}\mathbb{E}_{q(z)}\left[h(\mathbf{z})\right] = \mathbb{E}_{q(w,z)}\left[u(w)\nabla_z h(\mathbf{z})\right]$$

Similarly, we can show that

$$\nabla_{\mu}\mathbb{E}_{q(z)}\left[h(\mathbf{z})\right] = \mathbb{E}_{q(w,z)}\left[\nabla_z h(\mathbf{z})\right] = \mathbb{E}_{q(z)}\left[\nabla_z h(\mathbf{z})\right]$$

□

C.2. Example 6.1

A concrete example is the multivariate skew Gaussian distribution.

$$\begin{aligned} q(w,\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) :=& \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}+|w|\boldsymbol{\alpha},\boldsymbol{\Sigma})\mathcal{N}(w|0,1) \\ q(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) :=& \int q(w,\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma})dw \\ =& 2\Phi\left(\frac{\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}{\sqrt{1+\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}}\right)\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma}+\boldsymbol{\alpha}\boldsymbol{\alpha}^T) \end{aligned}$$

where $u(w) = |w|$ and $v(w) = 1$.

Furthermore, we have

$$\begin{aligned} &\int |w|q(w,\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma})dw \\ &= u_1(\mathbf{z},\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma})\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma}) + u_2(\mathbf{z},\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma})q(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) \end{aligned}$$

where $u_1(\mathbf{z},\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) = \frac{\sqrt{2/\pi}}{1+\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}$, $u_2(\mathbf{z},\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) = \frac{(\mathbf{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}{1+\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}$.

C.3. Example 6.2

Another example is the multivariate exponentially modified Gaussian distribution.

$$\begin{aligned} q(w,\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) :=& \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}+w\boldsymbol{\alpha},\boldsymbol{\Sigma})\mathrm{Exp}(w|1) \\ q(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) :=& \int_0^{+\infty} q(w,\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma})dw \\ =& \frac{\sqrt{2\pi}\det\left(2\pi\boldsymbol{\Sigma}\right)^{-\frac{1}{2}}}{\sqrt{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}}\Phi\left(\frac{\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}-1}{\sqrt{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}}\right)\exp\left\{\frac{1}{2}\left[\frac{\left(\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}-1\right)^2}{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}} - \left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)\right]\right\} \end{aligned}$$

where $u(w) = w$ and $v(w) = 1$.

Furthermore, we have

$$\int_0^{+\infty} wq(w,\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma})dw = u_1(\mathbf{z},\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma})\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma}) + u_2(\mathbf{z},\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma})q(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) \quad (16)$$

where $u_1(\mathbf{z},\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) = \frac{1}{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}$ and $u_2(\mathbf{z},\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) = \frac{(\mathbf{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}-1}{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}$.

C.4. Proof of Theorem 7

Proof: Firstly, note that

$$\mathbb{E}_{q(w,z)}\left[v(w)\nabla_z^2 h(\mathbf{z})\right] = \mathbb{E}_{q(w)}\left[\mathbb{E}_{q(z|w)}\left[\nabla_z^2\left(v(w)h(\mathbf{z})\right)\right]\right]$$

Conditioned on w, we know that $q(\mathbf{z}|w) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu} + u(w)\boldsymbol{\alpha}, v(w)\boldsymbol{\Sigma})$ is Gaussian with mean $\boldsymbol{\mu} + u(w)\boldsymbol{\alpha}$ and variance $v(w)\boldsymbol{\Sigma}$. By applying Lemma 11 to $v(w)h(\mathbf{z})$, we have

$$\begin{aligned}
\mathbb{E}_{q(w,z)}\left[v(w)\nabla_z^2 h(\mathbf{z})\right] &= \mathbb{E}_{q(w)}\left[\mathbb{E}_{q(z|w)}\left[\nabla_z^2\left(v(w)h(\mathbf{z})\right)\right]\right]\\
&= \mathbb{E}_{q(w)}\left[\mathbb{E}_{q(z|w)}\left[\left(v(w)\boldsymbol{\Sigma}\right)^{-1}\left(\mathbf{z}-\boldsymbol{\mu}-u(w)\boldsymbol{\alpha}\right)\nabla_z^T\left(v(w)h(\mathbf{z})\right)\right]\right]\\
&= \mathbb{E}_{q(w)}\left[\mathbb{E}_{q(z|w)}\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}-u(w)\boldsymbol{\alpha}\right)\nabla_z^T h(\mathbf{z})\right]\right]\\
&= \mathbb{E}_{q(w,z)}\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}-u(w)\boldsymbol{\alpha}\right)\nabla_z^T h(\mathbf{z})\right]
\end{aligned} \tag{17}$$

Recall that $q(\mathbf{z}|w) = \mathcal{N}(\mathbf{z}|\underbrace{\boldsymbol{\mu} + u(w)\boldsymbol{\alpha}}_{\hat{\boldsymbol{\mu}}}, \underbrace{v(w)\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\Sigma}}})$ is Gaussian. By applying Lemma 9 to $v(w)h(\mathbf{z})$, we have the following expression.

$$\begin{aligned}
&\mathbb{E}_{q(w,z)}\left[\boldsymbol{\Sigma}^{-1}\left[v^{-1}(w)\left(\mathbf{z}-\boldsymbol{\mu}-u(w)\boldsymbol{\alpha}\right)\left(\mathbf{z}-\boldsymbol{\mu}-u(w)\boldsymbol{\alpha}\right)^T-\boldsymbol{\Sigma}\right]\boldsymbol{\Sigma}^{-1}h(\mathbf{z})\right]\\
=&\mathbb{E}_{q(w)}\left[\mathbb{E}_{q(z|w)}\left[\left(v(w)\boldsymbol{\Sigma}\right)^{-1}\left[\left(\mathbf{z}-\boldsymbol{\mu}-u(w)\boldsymbol{\alpha}\right)\left(\mathbf{z}-\boldsymbol{\mu}-u(w)\boldsymbol{\alpha}\right)^T-v(w)\boldsymbol{\Sigma}\right]\left(v(w)\boldsymbol{\Sigma}\right)^{-1}\left(v(w)h(\mathbf{z})\right)\right]\right]\\
=&\mathbb{E}_{q(w)}\left[\mathbb{E}_{q(z|w)}\left[\left(\hat{\boldsymbol{\Sigma}}\right)^{-1}\left[\left(\mathbf{z}-\hat{\boldsymbol{\mu}}\right)\left(\mathbf{z}-\hat{\boldsymbol{\mu}}\right)^T-\hat{\boldsymbol{\Sigma}}\right]\left(\hat{\boldsymbol{\Sigma}}\right)^{-1}\left(v(w)h(\mathbf{z})\right)\right]\right]\\
=&\mathbb{E}_{q(w)}\left[\mathbb{E}_{q(z|w)}\left[\left(\hat{\boldsymbol{\Sigma}}\right)^{-1}\left(\mathbf{z}-\hat{\boldsymbol{\mu}}\right)\nabla_z^T\left(v(w)h(\mathbf{z})\right)\right]\right]\\
=&\mathbb{E}_{q(w,z)}\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}-u(w)\boldsymbol{\alpha}\right)\nabla_z^T h(\mathbf{z})\right]
\end{aligned} \tag{18}$$

By the regular assumptions, we can swap the integration and differentiation to get the following expression.

$$\begin{aligned}
\nabla_\Sigma\mathbb{E}_{q(z)}\left[h(\mathbf{z})\right] &= \int\nabla_\Sigma q(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})h(\mathbf{z})d\mathbf{z}\\
&= \int\int\left[\nabla_\Sigma\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},v(w)\boldsymbol{\Sigma})q(w)dw\right]h(\mathbf{z})d\mathbf{z}\\
&= \tfrac{1}{2}\mathbb{E}_{q(w,z)}\left[\boldsymbol{\Sigma}^{-1}\left[v^{-1}(w)\left(\mathbf{z}-\boldsymbol{\mu}-u(w)\boldsymbol{\alpha}\right)\left(\mathbf{z}-\boldsymbol{\mu}-u(w)\boldsymbol{\alpha}\right)^T-\boldsymbol{\Sigma}\right]\boldsymbol{\Sigma}^{-1}h(\mathbf{z})\right]
\end{aligned} \tag{19}$$

Finally, by Eq. (17) , (18), and (19), we have

$$\nabla_\Sigma\mathbb{E}_{q(z)}\left[h(\mathbf{z})\right] = \tfrac{1}{2}\mathbb{E}_{q(w,z)}\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}-u(w)\boldsymbol{\alpha}\right)\nabla_z^T h(\mathbf{z})\right] = \tfrac{1}{2}\mathbb{E}_{q(w,z)}\left[v(w)\nabla_z^2 h(\mathbf{z})\right]$$

□

C.5. Example 7.1

A concrete example is the multivariate Student's t-distribution with fixed degree of freedom 2β. We consider a case when $\beta > 1$, since the variance does not exist when $\beta \leq 1$.

$$\begin{aligned}
q(w,\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) :=& \mathcal{N}(\mathbf{z}|\boldsymbol{\mu},w\boldsymbol{\Sigma})\mathrm{IG}(w|\beta,\beta)\\
q(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) :=& \int q(w,\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma})dw\\
=& \det\left(\pi\boldsymbol{\Sigma}\right)^{-1/2}\frac{\Gamma(\beta+d/2)\left(2\beta+\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)\right)^{-\beta-d/2}}{\Gamma(\beta)\left(2\beta\right)^{-\beta}}.
\end{aligned}$$

where $u(w) = 0$ and $v(w) = w > 0$ since w is generate from inverse Gamma distribution $\mathrm{IG}(w|\beta,\beta)$.

When $\beta > 1$, we have

$$\int wq(w,\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma})dw = v_1(\mathbf{z},\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma})q(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) \tag{20}$$

where $v_1(\mathbf{z},\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) := \frac{\beta}{(\beta+d/2-1)}\left(1+(2\beta)^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)\right)$.

C.6. Example 7.2

Another example is the multivariate normal inverse-Gaussian distribution, where $\beta > 0$ is fixed.

$$\begin{aligned} q(w, \mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) :=& \mathcal{N}(\mathbf{z}|\boldsymbol{\mu} + w\boldsymbol{\alpha}, w\boldsymbol{\Sigma})\text{InvGauss}(w|1, \beta) \\ q(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) :=& \int q(w, \mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) dw \\ =& \frac{\beta^{\frac{1}{2}}}{(2\pi)^{\frac{d+1}{2}}} \det(\boldsymbol{\Sigma})^{-1/2} \exp\left[(\mathbf{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha} + \beta \right] \frac{2\mathcal{K}_{\frac{d+1}{2}} \left(\sqrt{\left(\boldsymbol{\alpha}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha} + \beta\right) \left((\mathbf{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu}) + \beta \right)} \right)}{\left(\sqrt{\frac{\boldsymbol{\alpha}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha} + \beta}{(\mathbf{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu}) + \beta}} \right)^{\frac{-d-1}{2}}} \end{aligned}$$

where $u(w) = v(w) = w > 0$ since w is generate from an inverse Gaussian distribution $\text{InvGauss}(w|1, \beta) = \left(\frac{\beta}{2\pi w^3}\right)^{\frac{1}{2}} \exp\left\{ -\frac{\beta}{2} \left(w + w^{-1} \right) + \beta \right\}$

We have

$$\int w q(w, \mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) dw = v_1(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) q(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \tag{21}$$

where

$$v_1(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) := \sqrt{\frac{(\mathbf{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu}) + \beta}{\boldsymbol{\alpha}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha} + \beta}} \frac{\mathcal{K}_{\frac{d-1}{2}} \left(\sqrt{\left(\boldsymbol{\alpha}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha} + \beta\right) \left((\mathbf{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu}) + \beta \right)} \right)}{\mathcal{K}_{\frac{d+1}{2}} \left(\sqrt{\left(\boldsymbol{\alpha}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha} + \beta\right) \left((\mathbf{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu}) + \beta \right)} \right)}$$

D. Gradient Identities for Continuous Exponential-family Mixtures

D.1. Proof of Theorem 8

Proof: It is easy to verify that $\tilde{f}_{i,j}(z_j) := f_{i,j}(z_j, \mathbf{z}_{-j}) h(z_j, \mathbf{z}_{-j}) \prod_{k \geq j+1} q(z_k|\mathbf{z}_{1:(k-1)}, \boldsymbol{\lambda})$ is locally AC since $h(z_j, z_{-j})$, $f_{i,j}(z_j, z_{-j})$, and $q(z_k|\mathbf{z}_{1:(k-1)}, \boldsymbol{\lambda})$ are all locally AC w.r.t. z_j for almost every z_{-j}.

By the definitions, we have

$$(\nabla_z \boldsymbol{\Psi}(\mathbf{z}, \boldsymbol{\lambda}))^{-1} = \begin{bmatrix} \frac{1}{q(z_1|\boldsymbol{\lambda})} & 0 \\ -\frac{\nabla_{z_1} \psi_2(z_1, z_2, \boldsymbol{\lambda})}{q(z_1|\boldsymbol{\lambda}) q(z_2|z_1, \boldsymbol{\lambda})} & \frac{1}{q(z_2|z_1, \boldsymbol{\lambda})} \end{bmatrix}$$

$$\begin{aligned} f_{i,1}(\mathbf{z}) &= \frac{\nabla_{\lambda_i} \psi_1(z_1, \boldsymbol{\lambda})}{q(z_1|\boldsymbol{\lambda})} \\ f_{i,2}(\mathbf{z}) &= -\frac{f_{i,1}(\mathbf{z}) \nabla_{z_1} \psi_2(z_1, z_2, \boldsymbol{\lambda})}{q(z_2|z_1, \boldsymbol{\lambda})} + \frac{\nabla_{\lambda_i} \psi_2(z_1, z_2, \boldsymbol{\lambda})}{q(z_2|z_1, \boldsymbol{\lambda})} \end{aligned}$$

Note that the following expression holds almost everywhere due to the product rule for locally AC functions.

$$\nabla_{z_1} f_{i,1}(\mathbf{z}) = \frac{\nabla_{\lambda_i} q(z_1|\boldsymbol{\lambda})}{q(z_1|\boldsymbol{\lambda})} - f_{i,1}(\mathbf{z}) \frac{\nabla_{z_1} q(z_1|\boldsymbol{\lambda})}{q(z_1|\boldsymbol{\lambda})} \tag{22}$$

Recall that $\tilde{f}_{i,1}(z_1) = f_{i,1}(\mathbf{z}) h(\mathbf{z}) q(z_2|z_1, \boldsymbol{\lambda})$.

Therefore, by Lemma 6, we have

$$\begin{aligned}
&-\mathbb{E}_{q(z_1)}\big[\ \cancel{q(z_2|z_1,\boldsymbol{\lambda})h(\mathbf{z})f_{i,1}(\mathbf{z})\frac{\nabla_{z_1}q(z_1|\boldsymbol{\lambda})}{q(z_1|\boldsymbol{\lambda})}}\ \big]\\
=&\mathbb{E}_{q(z_1)}\left[\nabla_{z_1}\left[q(z_2|z_1,\boldsymbol{\lambda})h(\mathbf{z})f_{i,1}(\mathbf{z})\right]\right]\\
=&\mathbb{E}_{q(z_1)}\left[h(\mathbf{z})f_{i,1}(\mathbf{z})\nabla_{z_1}q(z_2|z_1,\boldsymbol{\lambda})+q(z_2|z_1,\boldsymbol{\lambda})f_{i,1}(\mathbf{z})\nabla_{z_1}h(\mathbf{z})+q(z_2|z_1,\boldsymbol{\lambda})h(\mathbf{z})\nabla_{z_1}f_{i,1}(\mathbf{z})\right]\\
=&\mathbb{E}_{q(z_1)}\big[h(\mathbf{z})f_{i,1}(\mathbf{z})\nabla_{z_1}q(z_2|z_1,\boldsymbol{\lambda})+q(z_2|z_1,\boldsymbol{\lambda})f_{i,1}(\mathbf{z})\nabla_{z_1}h(\mathbf{z})+q(z_2|z_1,\boldsymbol{\lambda})h(\mathbf{z})\big(\frac{\nabla_{\lambda_i}q(z_1|\boldsymbol{\lambda})}{q(z_1|\boldsymbol{\lambda})}-\cancel{f_{i,1}(\mathbf{z})\frac{\nabla_{z_1}q(z_1|\boldsymbol{\lambda})}{q(z_1|\boldsymbol{\lambda})}}\big)\big]
\end{aligned}$$

where we obtain the last equation by (22).

The above expression gives the following identity.

$$0=\mathbb{E}_{q(z_1)}\left[h(\mathbf{z})f_{i,1}(\mathbf{z})\nabla_{z_1}q(z_2|z_1,\boldsymbol{\lambda})+q(z_2|z_1,\boldsymbol{\lambda})f_{i,1}(\mathbf{z})\nabla_{z_1}h(\mathbf{z})+q(z_2|z_1,\boldsymbol{\lambda})h(\mathbf{z})\frac{\nabla_{\lambda_i}q(z_1|\boldsymbol{\lambda})}{q(z_1|\boldsymbol{\lambda})}\right] \tag{23}$$

Likewise, the following expression holds for almost every z_1 due to the product rule for locally AC functions.

$$\nabla_{z_2}f_{i,2}(\mathbf{z})=-\frac{f_{i,1}(\mathbf{z})\nabla_{z_1}q(z_2|z_1,\boldsymbol{\lambda})}{q(z_2|z_1,\boldsymbol{\lambda})}-f_{i,2}(\mathbf{z})\frac{\nabla_{z_2}(z_2|z_1,\boldsymbol{\lambda})}{q(z_2|z_1,\boldsymbol{\lambda})}+\frac{\nabla_{\lambda_i}q(z_2|z_1,\boldsymbol{\lambda})}{q(z_2|z_1,\boldsymbol{\lambda})} \tag{24}$$

Note that $\tilde{f}_{i,2}(z_1)=f_{i,2}(\mathbf{z})h(\mathbf{z})$. By Lemma 6, we have

$$\begin{aligned}
&-\mathbb{E}_{q(z_1)}\big[\mathbb{E}_{q(z_2|z_1)}\big[\cancel{h(\mathbf{z})f_{i,2}(\mathbf{z})\frac{\nabla_{z_2}(z_2|z_1,\boldsymbol{\lambda})}{q(z_2|z_1,\boldsymbol{\lambda})}}\big]\big]\\
=&\mathbb{E}_{q(z_1)}\left[\mathbb{E}_{q(z_2|z_1)}\left[\nabla_{z_2}\left[h(\mathbf{z})f_{i,2}(\mathbf{z})\right]\right]\right]\\
=&\mathbb{E}_{q(z_1)}\left[\mathbb{E}_{q(z_2|z_1)}\left[f_{i,2}(\mathbf{z})\nabla_{z_2}h(\mathbf{z})+h(\mathbf{z})\nabla_{z_2}f_{i,2}(\mathbf{z})\right]\right]\\
=&\mathbb{E}_{q(z_1)}\big[\mathbb{E}_{q(z_2|z_1)}\big[f_{i,2}(\mathbf{z})\nabla_{z_2}h(\mathbf{z})+h(\mathbf{z})\big(-\frac{f_{i,1}(\mathbf{z})\nabla_{z_1}q(z_2|z_1,\boldsymbol{\lambda})}{q(z_2|z_1,\boldsymbol{\lambda})}-\cancel{f_{i,2}(\mathbf{z})\frac{\nabla_{z_2}(z_2|z_1,\boldsymbol{\lambda})}{q(z_2|z_1,\boldsymbol{\lambda})}}+\frac{\nabla_{\lambda_i}q(z_2|z_1,\boldsymbol{\lambda})}{q(z_2|z_1,\boldsymbol{\lambda})}\big)\big]\big]
\end{aligned}$$

where we obtain the last equation by (24).

The above expression gives the following identity.

$$0=\mathbb{E}_{q(z_1)q(z_2|z_1)}\left[f_{i,2}(\mathbf{z})\nabla_{z_2}h(\mathbf{z})+h(\mathbf{z})\left(\frac{-f_{i,1}(\mathbf{z})\nabla_{z_1}q(z_2|z_1,\boldsymbol{\lambda})}{q(z_2|z_1,\boldsymbol{\lambda})}+\frac{\nabla_{\lambda_i}q(z_2|z_1,\boldsymbol{\lambda})}{q(z_2|z_1,\boldsymbol{\lambda})}\right)\right] \tag{25}$$

By (23) and (25), we have

$$0=\mathbb{E}_{q(z_1)q(z_2|z_1)}\left[f_{i,1}(\mathbf{z})\nabla_{z_1}h(\mathbf{z})+f_{i,2}(\mathbf{z})\nabla_{z_2}h(\mathbf{z})+h(\mathbf{z})\left(\frac{\nabla_{\lambda_i}q(z_2|z_1,\boldsymbol{\lambda})}{q(z_2|z_1,\boldsymbol{\lambda})}+\frac{\nabla_{\lambda_i}q(z_1|\boldsymbol{\lambda})}{q(z_1|\boldsymbol{\lambda})}\right)\right] \tag{26}$$

Therefore, by (26), we have the following identity

$$\mathbb{E}_{q(z_1,z_2)}\left[h(\mathbf{z})\left(\frac{\nabla_{\lambda_i}q(z_2|z_1,\boldsymbol{\lambda})}{q(z_2|z_1,\boldsymbol{\lambda})}+\frac{\nabla_{\lambda_i}q(z_1|\boldsymbol{\lambda})}{q(z_1|\boldsymbol{\lambda})}\right)\right]=-\mathbb{E}_{q(z_1,z_2)}\left[f_{i,1}(\mathbf{z})\nabla_{z_1}h(\mathbf{z})+f_{i,2}(\mathbf{z})\nabla_{z_2}h(\mathbf{z})\right]$$

Since we can interchange the integration and differentiation, we know that

$$\nabla_{\lambda_i}\mathbb{E}_q\left[h(\mathbf{z})\right]=\mathbb{E}_q\left[h(z)\left(\frac{\nabla_{\lambda_i}q(z_1|\boldsymbol{\lambda}_z)}{q(z_1|\boldsymbol{\lambda}_z)}+\frac{\nabla_{\lambda_i}q(z_2|z_1,\boldsymbol{\lambda}_z)}{q(z_2|z_1,\boldsymbol{\lambda}_z)}\right)\right],$$

which gives the desired identity. □