Stein's Lemma for the Reparameterization Trick with Exponential-family Mixtures

Wu Lin (UBC), Mohammad Emtiyaz Khan (AIP, RIKEN), Mark Schmidt (UBC)

1 Introduction

Stein's lemma is a powerful tool for statistical applications but many applications either required strong technical assumptions or were limited to Gaussian distributions with restricted covariance structures.

- We establish a connection between Stein's lemma and the reparamterization trick (Suri et al (1988), Salimans et al (2012), Kingma et al (2013), Figurnov et al (2018)).
- We show that the implicit reparameterization trick can work with locally absolutely continuous functions.
- We derive new reparameterizable gradient-identities for Gaussian mixtures.
- We demonstrate applications of these identities in variational inference with Gaussian mixtures.

2 Absolute Continuity

Since we will use integration by parts to prove Stein's lemma, we first introduce a smoothness condition when integration by parts is valid.

Definition 1: Absolutely Continuous (AC)

A function $\mathbf{h}(\cdot) : [a,b] \subset \mathcal{R} \mapsto \mathcal{R}^d$ is absolutely continuous if

- its derivative $\nabla \mathbf{h}$ exists almost everywhere in $[a,b]$,
- the derivative is Lebesgue integrable: $\int_a^b \left| \mathbf{e}_i^T \nabla \mathbf{h}(z) \right| dz < \infty$ for all i,
- the fundamental theorem of calculus holds: $\mathbf{h}(z) = \mathbf{h}(a) + \int_a^z \nabla \mathbf{h}(t) dt$ for $z \in [a,b]$.

The following condition is essential for multivariate extensions of Stein's lemma.

Definition 2: Absolutely Continuous on Lines (ACL)

A function $\mathbf{h}(\cdot) : [a_1,b_1] \times \cdots \times [a_k,b_k] \subset \mathcal{R}^k \mapsto \mathcal{R}^d$ is absolutely continuous on lines if given each coordinate index $j \in [1,\dots,k]$, $\mathbf{h}_i(z_j) := \mathbf{h}(z_j, \mathbf{z}_{-j}) : [a_j,b_j] \mapsto \mathcal{R}^d$ is AC for almost every $\mathbf{z}_{-j}$ in its domain.

Definition 3: Locally AC & Locally ACL

A function $\mathbf{h}(\cdot) : \mathcal{R} \mapsto \mathcal{R}^d$ is locally AC if it is AC in $[a,b]$ for every compact interval $[a,b] \subset \mathcal{R}$.
A function $\mathbf{h}(\cdot) : \mathcal{R}^k \mapsto \mathcal{R}^d$ is locally ACL if it is ACL in $[a_1,b_1] \times \cdots \times [a_k,b_k]$ for every k-cell $[a_1,b_1] \times \cdots \times [a_k,b_k] \subset \mathcal{R}^k$.

Examples of locally ACL functions:

- A continuously differentiable function is locally ACL.
- A (locally) Lipschitz-continuous function is locally ACL.

The following theorem is valid due to Fubini's theorem.

Theorem 1: Integration by Parts

Let $h(\cdot), q(\cdot) : [a,b] \subset \mathcal{R} \mapsto \mathcal{R}$ be AC functions. The following identity holds.

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

3 Gradient Identities For Gaussian

We show applications of Stein's lemma for Gaussian distributions with arbitrary covariance structures and how the lemma connects to existing reparameterizable gradient identities.

Let $h(\mathbf{z}) : \mathcal{R}^k \mapsto \mathcal{R}$ be a locally ACL function and $q(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a multivariate Gaussian with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. We define the following functions

$$f_j(\mathbf{z}) := \mathbf{e}_j^T \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu}) \, h(\mathbf{z}) : \mathcal{R}^k \mapsto \mathcal{R} \text{ for } j \in [1,\dots,k],$$

We assume we can swap the differentiation and integration in the following results.

Stein (1973) gives the following identity but under diagonal covariance structures. The identity with arbitrary covariance structures can be shown due to Theorem 1. Bonnet (1964) gives the same identity for gradient estimation, which is now known as the reparameterization trick for the mean.

Lemma 1: Stein's Lemma for the Mean

If $\mathbb{E}_q \left[|\nabla_{z_i} h(\mathbf{z})| \right] < \infty$ for all i, the following identity holds.

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

Bonnet (1964): $\nabla_\mu \mathbb{E}_q [h(\mathbf{z})] = \mathbb{E}_q \left[\boldsymbol{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu}) \, h(\mathbf{z}) \right] = \mathbb{E}_q [\nabla_z h(\mathbf{z})]$.

Applying Stein's lemma to $f_j(\mathbf{z})$, we obtain the following identity, which is known as the reparameterization trick for the variance.

Lemma 2: Stein's Lemma for the Variance

If the conditions of Stein's lemma are satisfied for all $f_j(\mathbf{z})$ and $\mathbb{E}_q [h(\mathbf{z})]$ exists, the following gradient identity holds.

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

Applying Stein's lemma to $g_j(\mathbf{z})$ defined below and using Lemma 2, we obtain a second-order identity, which is known as Price's theorem (1958). Price proves the theorem using characteristic functions, which is non-trivial to extend to Gaussian mixtures. Due to Stein's lemma, we can readily extend Price's theorem to Gaussian mixtures as shown in next section.

Theorem 2: Price's Theorem

We further assume $h(\mathbf{z})$ is continuously differentiable and its derivative $\nabla h(\mathbf{z})$ is locally ACL. We define the following functions

$$g_j(\mathbf{z}) := \nabla_{z_j} h(\mathbf{z}) : \mathcal{R}^k \mapsto \mathcal{R} \text{ for } j \in [1,\dots,k].$$

If the conditions of Stein's lemma are satisfied for all $f_j(\mathbf{z})$ and $g_j(\mathbf{z})$, we have the following identity.

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

References:

- Price, A useful theorem for nonlinear devices having Gaussian inputs, 1958.
- Bonnet, Transformation des signaux aléatoires a travers les systemes non linéaires sans mémoire, 1964.
- Stein, Estimation of the mean of a multivariate normal distribution, 1973.
- Hudson et al, A natural identity for exponential families with applications in multiparameter estimation, 1978.
- Brown, Fundamentals of statistical exponential families: with applications in statistical decision theory, 1986.
- Suri et al, Perturbation analysis gives strongly consistent sensitivity estimates for the M/G/1 queue, 1988
- Arnold et al. A multivariate version of Stein's identity with applications to moment calculations and estimation of conditionally specified distributions, 2001.
- Salimans et al, Fixed-form variational posterior approximation through stochastic linear regression, 2012.
- Kingma et al, Auto-encoding variational Bayes, 2013.
- Figurnov et al, Implicit reparameterization gradients, 2018

4 Gradient Identities For Non-Gaussian

We will show that the implicit reparameterizable gradient can be used in more general settings and then propose new reparameterizable gradient identities to Gaussian mixtures.

Univariate Continuous Exponential Family

We consider the following exponential-family (EF) distribution. with $z \in (l,u)$, where $-\infty \leq l < u \leq \infty$. Furthermore, we assume $q(z)$ is locally AC.

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

Let's denote the CDF of $q(z|\boldsymbol{\lambda})$ by $\psi(z,\boldsymbol{\lambda}) := \int_l^z q(t|\boldsymbol{\lambda}) dt$.
Hudson et al (1978), Brown (1986), and Arnold et al (2001) extend Stein's lemma to EF.

Lemma 3: Stein's Lemma for Univariate Exponential Family

Let $h(\cdot) : (l,u) \mapsto \mathcal{R}$ be a locally AC function. If $\mathbb{E}_q [|h(z)|] < \infty$, $\lim_{z \downarrow l} h(z)q(z|\boldsymbol{\lambda}) = 0$ and $\lim_{z \uparrow u} h(z)q(z|\boldsymbol{\lambda}) = 0$, the following identity holds.

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

In Gaussian case, $\mathbb{E}_q [|h(z)|] < \infty$ implies $\lim_{z \downarrow -\infty} h(z)q(z|\boldsymbol{\lambda}) = 0$ and $\lim_{z \uparrow \infty} h(z)q(z|\boldsymbol{\lambda}) = 0$.

Applying Lemma 3 to $\tilde{f}_i(z)$ defined below, we obtain the implicit reparameterization trick .

Theorem 3: Univariate Implicit Reparametrization Trick

Let $h(\cdot) : (l,u) \mapsto \mathcal{R}$ be a locally AC function. We define $f_i(z) := \frac{\nabla_{\lambda_i} \psi(z,\boldsymbol{\lambda})}{q(z|\boldsymbol{\lambda})}$. We can show that $\tilde{f}_i(z) := h(z) f_i(z)$ is locally AC. If $\mathbb{E}_q [|f_i(z) \nabla_z h(z)|] < \infty$ and the conditions of Lemma 3 for $\tilde{f}_i(z)$ are satisfied, we have the following identity.

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

Continuous Exponential-family Mixture

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

$$q(z_1|\boldsymbol{\lambda}) = h_1(z_1) \exp \left\{ \langle \mathbf{T}_1(z_1), \phi_1(\boldsymbol{\lambda}) \rangle - A_1(\boldsymbol{\lambda}) \right\}$$
$$q(z_2|z_1,\boldsymbol{\lambda}) = h_2(z_2, z_1) \exp \left\{ \langle \mathbf{T}_2(z_2, z_1), \phi_2(\boldsymbol{\lambda}) \rangle - A_2(\boldsymbol{\lambda}, z_1) \right\}$$

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

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

We define the following functions: $\boldsymbol{\Psi}(\mathbf{z},\boldsymbol{\lambda}) := [\psi_1(z_1,\boldsymbol{\lambda}), \psi_2(z_1,z_2,\boldsymbol{\lambda})]^T$,

$$\nabla_{\lambda_i} \boldsymbol{\Psi}(\mathbf{z},\boldsymbol{\lambda}) := [\nabla_{\lambda_i}\psi_1(z_1,\boldsymbol{\lambda}), \nabla_{\lambda_i}\psi_2(z_1,z_2,\boldsymbol{\lambda})]^T , \quad \nabla_z \boldsymbol{\Psi}(\mathbf{z},\boldsymbol{\lambda}) := \begin{bmatrix} q(z_1|\boldsymbol{\lambda}) & 0 \\ \int_{l_2}^{z_2} \nabla_{z_1} q(t_2|z_1,\boldsymbol{\lambda}) dt_2 & q(z_2|z_1,\boldsymbol{\lambda}) \end{bmatrix}.$$

Applying Lemma 3 to $\tilde{f}_{i,j}(z_j)$ defined below, we obtain the following identity.

Theorem 4: Bivariate Implicit Reparametrization Trick

Let $h(\cdot) : (l_1, u_1) \times (l_2, u_2) \mapsto \mathcal{R}$ be a locally ACL function. First, we define function $f_{i,j}(\mathbf{z}) := \mathbf{e}_j^T [\nabla_z \boldsymbol{\Psi}(\mathbf{z},\boldsymbol{\lambda})]^{-1} \nabla_{\lambda_i} \boldsymbol{\Psi}(\mathbf{z},\boldsymbol{\lambda})$. We can show that $\tilde{f}_{i,j}(z_j) := h(z_j, z_{-j}) f_{i,j}(z_j, z_{-j})$ is locally AC. If $\mathbb{E}_q \left[\left| \sum_j f_{i,j}(\mathbf{z}) \nabla_{z_j} h(\mathbf{z}) \right| \right] < \infty$ and the conditions of Lemma 3 for $\tilde{f}_{i,j}(z_j)$ are satisfied, we have the following identity.

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

Figurnov et al (2018) assume that $h(\mathbf{z})$ is continuously differentiable. As shown in Theorem 4, the identity holds even when $h(\mathbf{z})$ is not continuously differentiable. The identity can be easily extended to multivariate cases.

Multivariate Gaussian Variance-mean Mixture

We consider the following Gaussian variance-mean mixture with $v(w) > 0$.

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

We define $f_j(\mathbf{z}, w) := \mathbf{e}_j^T \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu} - u(w)\boldsymbol{\alpha}) \, h(\mathbf{z}) : \mathcal{R}^k \mapsto \mathcal{R}$ for $j \in [1,\dots,k]$. Since the conditional distribution, $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu} + u(w)\boldsymbol{\alpha}, v(w)\boldsymbol{\Sigma})$, is Gaussian, applying Lemma 1 and Theorem 2, we obtain the following identities for Gaussian variance-mean mixture.

Theorem 5: Gradient Identities for Gaussian Variance-mean Mixture

Let $h(\mathbf{z}) : \mathcal{R}^k \mapsto \mathcal{R}$ be a locally ACL function. If $\mathbb{E}_{q(z)} [|\nabla_{z_i} h(\mathbf{z})|] < \infty$ and $\mathbb{E}_{q(w,z)} [|u(w) \nabla_{z_i} h(\mathbf{z})|] < \infty$ for all i, and $\mathbb{E}_{q(w,z)} \left[\left| \nabla_{z_i} f_j(\mathbf{z}, w) \right| \right] < \infty$ for all i and j, we have the following identities.

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

Additionally, if $h(\mathbf{z})$ is continuously differentiable, its derivative $\nabla h(\mathbf{z})$ is locally ACL, and $\mathbb{E}_{q(w,z)} \left[\left| v(w) \nabla_{z_j} \nabla_{z_i} h(\mathbf{z}) \right| \right] < \infty$ for all i and j, we have a second-order identity.

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

Examples of new reparameterizable gradient identities for Gaussian mixtures:

Skew Gaussian distribution: $q(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) := \int \mathcal{N}(\mathbf{z}|\boldsymbol{\mu} + |w|\boldsymbol{\alpha}, \boldsymbol{\Sigma}) \mathcal{N}(w|0,1) dw$.

$$\nabla_\mu \mathbb{E}_{q(z)} [h(\mathbf{z})] = \mathbb{E}_{q(z)} [\nabla_z h(\mathbf{z})] , \quad \nabla_\Sigma \mathbb{E}_{q(z)} [h(\mathbf{z})] = \tfrac{1}{2} \mathbb{E}_{q(w,z)} \left[\nabla_z^2 h(\mathbf{z}) \right]$$
$$\nabla_\alpha \mathbb{E}_{q(z)} [h(\mathbf{z})] = \mathbb{E}_{q(w,z)} [|w| \nabla_z h(\mathbf{z})] = \mathbb{E}_{\mathcal{N}(t|\mu,\Sigma)} [u_1(\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) \nabla_t h(\mathbf{t})] + \mathbb{E}_{q(z)} [u_2(\mathbf{z},\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) \nabla_z h(\mathbf{z})]$$

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

Exponentially modified Gaussian distribution: $q(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) := \int \mathcal{N}(\mathbf{z}|\boldsymbol{\mu} + w\boldsymbol{\alpha}, \boldsymbol{\Sigma}) \mathrm{Exp}(w|1) dw$.

$$\nabla_\mu \mathbb{E}_{q(z)} [h(\mathbf{z})] = \mathbb{E}_{q(z)} [\nabla_z h(\mathbf{z})] , \quad \nabla_\Sigma \mathbb{E}_{q(z)} [h(\mathbf{z})] = \tfrac{1}{2} \mathbb{E}_{q(w,z)} \left[\nabla_z^2 h(\mathbf{z}) \right]$$
$$\nabla_\alpha \mathbb{E}_{q(z)} [h(\mathbf{z})] = \mathbb{E}_{q(w,z)} [w \nabla_z h(\mathbf{z})] = \mathbb{E}_{\mathcal{N}(t|\mu,\Sigma)} [u_1(\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) \nabla_t h(\mathbf{t})] + \mathbb{E}_{q(z)} [u_2(\mathbf{z},\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) \nabla_z h(\mathbf{z})]$$

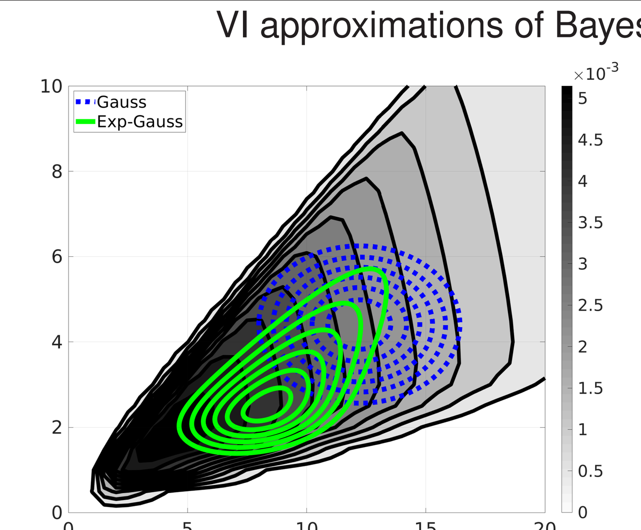
where $u_1(\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) = \frac{1}{\boldsymbol{\alpha}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}}$ and $u_2(\mathbf{z},\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) = \frac{(\mathbf{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}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

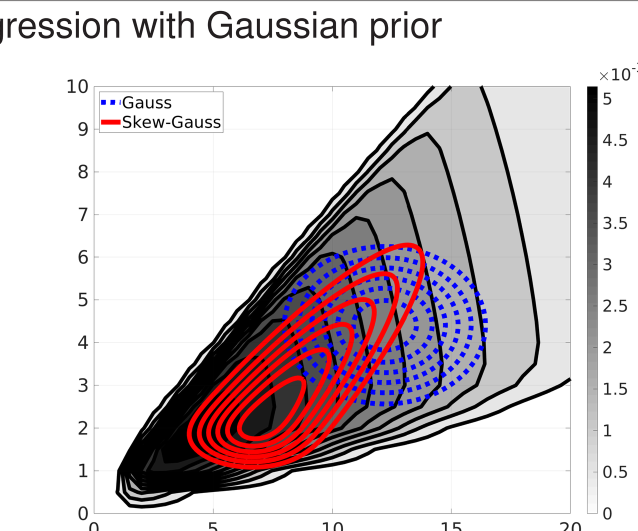
Symmetric normal inverse-Gaussian distribution: $q(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \int \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w^{-1}\boldsymbol{\Sigma}) \mathrm{IG}(w|1,1) dw$.
The modified Bessel function of the second kind is denoted by $\mathcal{K}_d(\cdot)$.

$$\nabla_\Sigma \mathbb{E}_{q(z)} [h(\mathbf{z})] = \tfrac{1}{2} \mathbb{E}_{q(w,z)} \left[w^{-1} \nabla_z^2 h(\mathbf{z}) \right] = \tfrac{1}{2} \mathbb{E}_{q(z)} \left[u_1(\mathbf{z},\boldsymbol{\mu},\boldsymbol{\Sigma}) \nabla_z^2 f_n(\mathbf{z}) \right]$$
$$\nabla_\mu \mathbb{E}_{q(z)} [h(\mathbf{z})] = \mathbb{E}_{q(z)} [\nabla_z h(\mathbf{z})]$$

where $t(\mathbf{z},\boldsymbol{\mu},\boldsymbol{\Sigma}) := \sqrt{(\mathbf{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu}) + 1}$ and $u_1(\mathbf{z},\boldsymbol{\mu},\boldsymbol{\Sigma}) := \frac{t(\mathbf{z},\boldsymbol{\mu},\boldsymbol{\Sigma}) \mathcal{K}_{\frac{k-3}{2}}(t(\mathbf{z},\boldsymbol{\mu},\boldsymbol{\Sigma}))}{\mathcal{K}_{\frac{k-1}{2}}(t(\mathbf{z},\boldsymbol{\mu},\boldsymbol{\Sigma}))}$.

VI approximations of Bayesian logistic regression with Gaussian prior





Conclusion: Our work enables machine learning applications of Stein's lemma in more general settings than before.