ABSTRACT

Variational Bayesian inference is an important machine-learning tool that finds application from statistics to robotics. The goal is to find an approximate probability density function (PDF) from a chosen family that is in some sense ‘closest’ to the full Bayesian posterior. Closeness is typically defined through the selection of an appropriate loss functional such as the Kullback-Leibler (KL) divergence. In this paper, we explore a new formulation of variational inference by exploiting the fact that the set of PDFs constitutes a Bayesian Hilbert space under careful definitions of vector addition, scalar multiplication and an inner product. We show that variational inference based on KL divergence then amounts to an iterative projection of the Bayesian posterior onto a subspace corresponding to the selected approximation family. In fact, the inner product chosen for the Bayesian Hilbert space suggests the definition of a new measure of the information contained in a PDF and in turn a new divergence is introduced. Each step in the iterative projection is equivalent to a local minimization of this divergence. We present an example Bayesian subspace based on exponentiated Hermite polynomials as well as work through the details of this general framework for the specific case of the multivariate Gaussian approximation family and show the equivalence to another Gaussian variational inference approach. We furthermore discuss the implications for systems that exhibit sparsity, which is handled naturally in Bayesian space.

Keywords Aitchison geometry · Bayesian Hilbert spaces · variational inference · Bayesian inference · compositional data · stochastic algebra

1 Introduction

In 1763, Richard Price published on behalf of his recently deceased friend, the Reverend Thomas Bayes, a paper that introduced what would become the atomic element of probabilistic inference: Bayes’ rule (Bayes, 1763). The paper though was widely ignored. About a decade later, the same rule was discovered by Pierre-Simon Laplace and, while Laplace laid its foundations, the theory of inference based on this rule became known as Bayesian inference. So confident was Laplace in the theory that he famously calculated the odds at 11,000 to 1 that the mass of Saturn as determined by a former student was correct to within 1%, 1,000,000-to-1 odds on the mass of Jupiter (Laplace, 1995, translated from 1825 French edition, pp. 46-47). (Based on the most recent available data, he would have collected on the bet on Saturn.) Bayesian inference has been used in a great variety of applications from Henri Poincaré’s defense of Captain Dreyfus to Alan Turing’s breaking of the Enigma code (McGrayne, 2011). In modern day, it provides the crucial framework for inference in such fields as statistics, decision theory, computational neuroscience, machine learning, computer vision, state estimation and robotics.

The objective common to all these applications is the determination of a posterior probability to test some hypothesis or to calculate some estimate based on prior information and observed measurements. However, it is not always possible to find the posterior exactly. Indeed, we must often resort to approximate techniques. One such technique, which will
occupy us here, is that of variational inference or variational Bayes (Bishop, 2006). In this variational approach, the goal is to find the probability density function that comes closest to the posterior as determined by minimizing a loss functional subject to the constraint that the distribution sought be drawn from a tractable class of densities, for example, where the posterior has to take the form of a Gaussian distribution. A common choice for the loss functional is the Kullback-Leibler divergence (Csiszar, 1975; Hinton and van Camp, 1993; Jordan et al., 1999; Bishop, 2006; Barber, 2012; Amari, 2016; Blei et al., 2017) although others such as Bregman (Adamcik, 2014; Painsky and Wornell, 2020), Wasserstein (Ambrogioni et al., 2018) and Renyi divergences (Li and Turner, 2016) have been used.

The field of variational inference based on the KL divergence is already well trodden although the research is hardly exhausted. The chosen class of densities from which the approximate posterior is to be shaped is key to variational inference. In the mean-field approximation, for example, the solution to the minimization of the divergence is constructed as a product of densities from a chosen family of admissible functions such as a Bayesian mixture of Gaussians (Bishop, 2006). Another possibility is using Bayesian mixtures of exponential families (Wainwright and Jordan, 2008; Amari, 2016). A number of algorithms by which to execute the minimization exist including the variational EM algorithm, natural gradient descent and Gaussian variational inference.

Jordan et al. (1999) observed that “there is not as yet a systematic algebra that allows particular variational transformations to be matched optimally to particular graphical models.” While this was written two decades ago and specifically about graphical models, the remark finds resonance in the present work.

Our aim is to introduce a kind of information algebra to variational inference that not only provides a convenient and effective framework for analysis but also reveals key relationships to past work. This algebra has its origins in the work of Aitchison (1982, 1986) on compositional data in statistics. Compositional data can be represented on a simplex as with probability distributions for a finite set of discrete events. The resulting Aitchison geometry or Aitchison simplex establishes a vector space, in which vector addition is a normalized multiplication (perturbation) and scalar multiplication is a normalized exponentiation (powering). With an appropriate inner product, the set of PDFs over a finite discrete space was formalized as a Hilbert space by Pawlowsky-Glahn and Egozcue (2001) and independently investigated by Barfoot (2002) and Barfoot and D’Eleuterio (2002) in their stochastic algebra. The extension to continuous variables was first published by Egozcue et al. (2006) and also studied by Barfoot and D’Eleuterio (2003) for the case of finite domains. The generalization to include probabilities and measures on the whole real line was made by van den Boogaart et al. (2010, 2014). These spaces are called Bayesian Hilbert spaces.

In such a space, Bayes’ venerated rule becomes

\[ p(x|z) = p(z|x) \oplus p(x) \]  

where \( \oplus \) indicates vector addition. (The normalization inherent in the operation accounts for the marginal \( p(z) \) automatically.) Each new measurement made to refine the posterior becomes one more term added to the sum. It is this linear feature of a Bayesian Hilbert space that makes the structure ideally suited to variational inference.

The set of Gaussians, in an appropriately extended sense, constitutes a subspace of the space of PDFs as do exponential families. An arbitrary PDF in one of these subspaces can be expressed in the simple and usual manner as a linear combination of a basis for the subspace. The problem of variational inference can thus be expressed as the minimization of a divergence over a set of Fourier coefficients.

The linear-algebraic structure of these spaces affords us a new perspective and provides new insight. We show, for example, that the solution to variational inference based on the KL divergence can be viewed as an iterative projection, in the Euclidean sense, onto a given subspace of densities. Indeed, this algorithm is essentially a Newton-like iteration scheme to solve for the minimum of the divergence, having a form identical to the natural-gradient-descent technique of Amari (1998). Moreover, using a subspace of Gaussians reproduces the recent results of Barfoot et al. (2020).

We also introduce a new information measure using a norm for the space of PDFs. This allows for a metric to be defined on the space, which can be interpreted as the distance between two PDFs. A new (symmetric and quadratic) divergence between PDFs can be based on the distance metric. It is notable that each step in our iterative-projection scheme is a local minimization of this divergence.

We shall begin with an overview of Bayesian Hilbert spaces in the next section. In §3, we discuss subspaces and bases, including exponentiated Hermite polynomials and Gaussian distributions. The variational inference problem for the KL divergence as viewed from the purchase of a Bayesian Hilbert space is considered in §4. The specific case of using a Gaussian subspace, that is, Gaussian variational inference, is treated in §5. Some numerical examples are provided in §6. We discuss the implications of sparsity in §7 and end with a few concluding remarks.
2 Bayesian Hilbert Spaces

Let us consider some domain $\mathcal{X}$ for our PDFs, e.g., $\mathbb{R}^N$; we shall refer to $x \in \mathcal{X}$ as the state. A PDF $p(x)$ assigns a nonnegative, finite value to each element of $\mathcal{X}$ such that

$$\int_{\mathcal{X}} p(x) \, dx = 1. \tag{2}$$

We define the set of all such bounded PDFs to be

$$\mathcal{P} = \left\{ p(x) \mid (\forall x \in \mathcal{X}) \ 0 \leq p(x) < C, \int_{\mathcal{X}} p(x) \, dx = 1 \right\}, \tag{3}$$

for some finite $C$. As a technicality, we must also compel each $p$ only to take on the value zero on a set of measure zero, but this will not be of major practical concern.

We define vector addition, $\oplus : \mathcal{P} \times \mathcal{P} \to \mathcal{P}$, between two elements $p_1, p_2 \in \mathcal{P}$ to be

$$p_1 \oplus p_2 = \frac{p_1(x)p_2(x)}{\int_{\mathcal{X}} p_1(x)p_2(x) \, dx}, \tag{4}$$

and scalar multiplication, $\cdot : \mathbb{R} \times \mathcal{P} \to \mathcal{P}$, of $p \in \mathcal{P}$ by $\alpha \in \mathbb{R}$ as

$$\alpha \cdot p = \frac{p^\alpha(x)}{\int_{\mathcal{X}} p^\alpha(x) \, dx}. \tag{5}$$

With these operations, $\mathcal{P}$ is established as a vector space, termed a Bayesian linear space, over the field $\mathbb{R}$ (van den Boogaart et al., 2010).

Vector subtraction is defined in the usual way,

$$p_1 \ominus p_2 = p_1 \oplus (-1) \cdot p_2. \tag{6}$$

We note that subtraction, or the inverse additive operation, is equivalent to the Radon-Nikodym derivative (van den Boogaart et al., 2010).

It will sometimes be convenient to avoid being explicit about the normalization constant for a PDF and so we define the normalization operator,

$$\downarrow p = \frac{p(x)}{\int_{\mathcal{X}} p(x) \, dx} \in \mathcal{P}. \tag{7}$$

Vector addition and scalar multiplication can then be simply expressed as

$$p_1 \oplus p_2 = \downarrow p_1(x)p_2(x), \quad \alpha \cdot p = \downarrow p^\alpha(x). \tag{8}$$

As mentioned above, Bayes’ rule can be rendered as $p(x|z) = p(z|x) \oplus p(x)$. The normalizing marginal $p(z)$ is invisible to the operation.

**Inner Product.** We endow the vector space with an inner product defined as

$$\langle p_1, p_2 \rangle = \frac{1}{2} \int_{\mathcal{X}} \int_{\mathcal{X}} \ln \left( \frac{p_1(x)}{p_1(y)} \right) \ln \left( \frac{p_2(x)}{p_2(y)} \right) \nu(x) \nu(y) \, dx \, dy, \tag{9}$$

where $\nu(\cdot)$ is a density function corresponding to an appropriate measure for $\mathcal{X}$.

If we take the measure $\nu$ to be also an element of $\mathcal{P}$, which we may think of as a reference PDF, then we can write the inner product in (9) as

$$\langle p_1, p_2 \rangle = \mathbb{E}_\nu \left[ \ln p_1 \ln p_2 \right] - \mathbb{E}_\nu \left[ \ln p_1 \right] \mathbb{E}_\nu \left[ \ln p_2 \right], \tag{10}$$

where $\mathbb{E}_\nu[\cdot]$ is the expectation with respect to $\nu$. In this work, we shall always take the measure to be a PDF; however, we shall refer to it as the measure to distinguish it from the other densities involved.

Let us further refine $\mathcal{P}$ to require the PDFs to be square-log integrable, that is,

$$\int_{\mathcal{X}} |\ln p|^2 \nu(x) \, dx < \infty. \tag{11}$$

Following van den Boogaart et al. (2014), then, we can claim that $\mathcal{P}$ with inner product (9) forms a separable Hilbert space. This is called a Bayesian Hilbert space. We shall sometimes briefly refer to it as a Bayesian space.
Information and Divergence. The norm of \( p \in \mathcal{P} \) can be taken as \( \| p \| = \langle p, p \rangle^{1/2} \). Accordingly, we can define the distance between two PDFs, \( p \) and \( q \), simply as \( d(p, q) = \| p \otimes q \| \), which induces a metric on the Bayesian space.

The norm of \( p \) can be used to express the information content of the PDF. In fact, we shall define

\[
I(p) = \frac{1}{2} \| p \|^2 = \frac{1}{2} \langle p, p \rangle
\]

as the information in \( p \). (The reason for the factor of \( \frac{1}{2} \) will become evident.) As an example, consider \( p = \mathcal{N}(\mu, \sigma^2) \) (over the domain \( \mathbb{R} \)) and measure \( \nu = \mathcal{N}(0, 1) \). The information is \( I(p) = (1 + 2\mu^2)/4\sigma^4 \). The smaller the variance the larger the information indicating that the PDF concentrates the random variable more tightly about its mean; that is, we know better where to expect the random variable so we may say that we have more information about it.

We shall furthermore find it useful to define a new divergence between two PDFs \( p \) and \( q \) as

\[
I(p \oplus q) = \frac{1}{2} \langle p \oplus q, p \oplus q \rangle.
\]

This is the information contained in the difference of \( p \) and \( q \). Unlike the Kullback-Leibler divergence, this divergence is symmetric in \( p \) and \( q \) and quadratic in Bayesian space. Clearly, \( p = q \) if and only if \( I(p \oplus q) = 0 \).

Stochastic Derivative. Accompanying this algebra is a functional calculus. Consider a PDF \( p(x|\theta) \in \mathcal{P} \) depending continuously on some parameter \( \theta \). We define the stochastic partial derivative of \( p \) with respect to \( \theta \) as (Barfoot and D’Eleuterio, 2003; Egozcue et al., 2013)

\[
\frac{\partial p}{\partial \theta} = \lim_{\lambda \to 0} \frac{1}{\lambda} \cdot \left( p(x|\theta + \lambda) \ominus p(x|\theta) \right).
\]

Note that the result of this operation remains an element in \( \mathcal{P} \). We can also define directional derivatives and a gradient operator but these will not be required here.

3 Subspaces and Bases

While \( \mathcal{P} \) is an infinite-dimensional space, it contains infinitely many finite-dimensional subspaces. We can in fact build a subspace \( \mathcal{Q} \) by taking the span of a set of \( M \) vectors \( B = \{ b_1, \ldots, b_M \} \), namely,

\[
\mathcal{Q} = \text{span} \{ b_1, \ldots, b_M \}.
\]

If we choose \( B \) to be linearly independent, it will form a basis for \( \mathcal{Q} \). We can accordingly write every vector \( q \) in \( \mathcal{Q} \) as a linear combination of \( B \), i.e.,

\[
q = \bigoplus_{m=1}^{M} \alpha_m \cdot b_m,
\]

where \( \alpha_m \in \mathbb{R} \) are unique.

As a shorthand, we will denote \( b = [b_1 \ b_2 \ \cdot\cdot\cdot \ b_M]^T \) as the basis. The inner products between all pairs of basis vectors forms the Gram matrix,

\[
\langle b, b \rangle = [\langle b_m, b_n \rangle],
\]

where \( (m, n) \) are the indices of the matrix. We furthermore have an orthonormal basis if \( \langle b_m, b_n \rangle = \delta_{mn} \), the Kronecker delta, in which case \( \langle b, b \rangle = 1 \), the identity matrix.

3.1 Projections

Given a subspace \( \mathcal{Q} \) of \( \mathcal{P} \) and \( p \in \mathcal{P} \), the PDF \( q^* \in \mathcal{Q} \) that minimizes the distance to, as well as the divergence (13) from, \( p \) is the projection of \( p \) onto \( \mathcal{Q} \), that is,

\[
q^* = \text{proj}_\mathcal{Q} p.
\]

As in Euclidean geometry, we can view \( p \) as being decomposed into a component \( p_\parallel \) lying in \( \mathcal{Q} \) and a component \( p_\perp \) perpendicular to it; therefore \( q^* = p_\parallel \) (see Figure 1).

The coordinates of \( q^* \) can be calculated as

\[
\alpha^* = \langle b, b \rangle^{-1} \langle b, p \rangle.
\]
We may also write the projection as an outer-product operation on $p$,
\[ q^*(x) = Q(x, x') \otimes p(x'), \tag{20} \]
where
\[ Q(x, x') = b(x) \langle b, b \rangle^{-1} b(x') \tag{21} \]
is the kernel of $Q$. (See also Appendix B.)

### 3.2 Hermite Basis on $\mathbb{R}$

As an example, consider the domain over which the PDFs of $P$ are defined to be $\mathbb{R}$. We can use the exponentiated Hermite polynomials as a basis for our infinite-dimensional $P$; in fact, they prove to be a natural choice. In one dimension, the first few probabilists’ Hermite polynomials are
\[ H_1(\xi) = \xi, \quad H_2(\xi) = \xi^2 - 1, \quad H_3(\xi) = \xi^3 - 3\xi, \quad H_4(\xi) = \xi^4 - 6\xi^2 + 3. \tag{22} \]

(We exclude $H_0(\xi) = 1$ as the resulting vector is the zero vector; however, it will need to be introduced when considering the domain $\mathbb{R}^N$ as explained in Appendix D.) Owing to the properties of the Hermite polynomials, namely, that
\[ \int_{-\infty}^{\infty} H_n(\xi) \nu(\xi) \, d\xi = 0, \quad \int_{-\infty}^{\infty} H_m(\xi) H_n(\xi) \nu(\xi) \, d\xi = n! \delta_{mn}, \quad m, n = 1, 2, 3, \ldots, \tag{23} \]
where $\nu(\xi) = \mathcal{N}(0, 1)$ is the standard normal density, we can construct an orthonormal basis for $P$ following Egozcue et al. (2006). Accordingly,
\[ E_\nu[H_n] = 0, \quad E_\nu[H_m H_n] = n! \delta_{mn}, \quad m, n = 1, 2, 3, \ldots, \tag{24} \]

Our basis functions are
\[ h_n(\xi) = \exp(-\eta_n(\xi)), \quad \eta_n(\xi) = \frac{1}{\sqrt{n!}} H_n(\xi). \tag{25} \]

Orthogonality follows as
\[ \langle h_m, h_n \rangle = E_\nu[\eta_m \eta_n] - E_\nu[\eta_m] E_\nu[\eta_n] = \frac{1}{\sqrt{m!n!}} \int_{-\infty}^{\infty} H_m(\xi) H_n(\xi) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \xi^2\right) \, d\xi = \delta_{mn}. \tag{26} \]

An arbitrary PDF $p \in P$ can be expanded elegantly in terms of this Hermite basis. However, we first need two lemmata, resting on the recursive definition of Hermite polynomials; these are

**Lemma 1.** For the standard normal measure, $\nu \sim \mathcal{N}(0, 1)$,
\[ E_\nu[H_{n+1}(\xi)f(\xi)] = E_\nu \left[ H_n(\xi) \frac{\partial f(\xi)}{\partial \xi} \right]. \tag{27} \]

where $f(\xi)$ is a differentiable function and is such that the expectations exist.

**Lemma 2.** For the standard normal measure, $\nu \sim \mathcal{N}(0, 1)$,
\[ E_\nu[H_n(\xi)f(\xi)] = E_\nu \left[ \frac{\partial^n f(\xi)}{\partial \xi^n} \right]. \tag{28} \]

where $f(\xi)$ is an $n$-fold differentiable function and is such that the expectations exist.
These may be regarded as generalizations of Stein’s lemma (Stein, 1981). The proofs are in Appendix C.

Now consider any \( p \in \mathcal{P} \) expressed as \( p(\xi) = \downarrow \exp(-\phi(\xi)) \). The coordinates are given by

\[
\alpha_n = \langle h_n, p \rangle = \frac{1}{\sqrt{n!}} \mathbb{E}_\nu \left[ \frac{\partial^n \phi(\xi)}{\partial \xi^n} \right] \quad (29)
\]

and hence

\[
p(\xi) = \bigoplus_{n=1}^{\infty} \alpha_n \cdot h_n(\xi) = \downarrow \exp \left( -\sum_{n=1}^{\infty} \frac{1}{n!} \mathbb{E}_\nu \left[ \frac{\partial^n \phi(\xi)}{\partial \xi^n} \right] H_n(\xi) \right) \quad (30)
\]

We can account for measures other than the standard normal density, say \( \nu \sim \mathcal{N}(\mu, \sigma^2) \), by the well known reparameterization ‘trick’,

\[
x = \mu + \sigma \xi,
\]

which leads to

\[
p(x) = \bigoplus_{n=1}^{\infty} \alpha_n \cdot h_n \left( \frac{x - \mu}{\sigma} \right) = \downarrow \exp \left( -\sum_{n=1}^{\infty} \frac{\sigma^n}{n!} \mathbb{E}_\nu \left[ \frac{\partial^n \phi(x)}{\partial x^n} \right] H_n \left( \frac{x - \mu}{\sigma} \right) \right). \quad (32)
\]

It is instructive to rewrite this expression by replacing \(-\phi\) with \( \ln p \) giving

\[
p(x) = \downarrow \exp \left( \sum_{n=1}^{\infty} \frac{\sigma^n}{n!} \mathbb{E}_\nu \left[ \frac{\partial^n \ln p(x)}{\partial x^n} \right] H_n \left( \frac{x - \mu}{\sigma} \right) \right). \quad (33)
\]

This is a Taylor-like expansion of \( p \) pivoting on a given mean \( \mu \) and standard deviation \( \sigma \).

Any subset of the basis functions \( \{h_1, h_2, \ldots\} \) establishes a subspace of \( \mathcal{P} \); however, as far as such subspaces are concerned, it would be natural to choose an \( M \)-dimensional subspace \( \mathcal{H} \) spanned by the first \( M \) basis functions. As the basis is orthonormal, the Gram matrix is \( \langle h, h \rangle = 1 \).

The Hermite functions can also be used to generate a basis for \( \mathcal{P} \) on the domain \( \mathbb{R}^N \) (see Appendix D).

### 3.3 Numerical Example With the Hermite Basis

Let us consider a simple one-dimensional, nonlinear estimation problem as a numerical example motivated by the type of inverse-distance nonlinearity found in a stereo camera model. This same experiment (with the same parameter settings) was used as a running example by Barfoot (2017, §4). We assume that our true state is drawn from a Gaussian prior:

\[
x \sim \mathcal{N}(\mu_p, \sigma_p^2).
\]

We then generate a measurement according to

\[
z = \frac{fb}{x} + n, \quad n \sim \mathcal{N}(0, \sigma_n^2),
\]

where \( n \) is measurement noise. The numerical values of the parameters used were

\[
\mu_p = 20 \text{ [m]}, \quad \sigma_p^2 = 9 \text{ [m]²},
\]

\[
f = 400 \text{ [pixel]}, \quad b = 0.1 \text{ [m]}, \quad \sigma_n^2 = 0.09 \text{ [pixel]²}.
\]

The true posterior is given by

\[
p(x|z) = \downarrow \exp(-\phi(x)), \quad \phi(x) = \frac{1}{2} \frac{(x - \mu_p)^2}{\sigma_p^2_{\text{prior}}} + \frac{1}{2} \frac{(z - \frac{fb}{x})^2}{\sigma_n^2_{\text{measurement}}}.
\]

We seek to find \( q(x) \), in a subspace spanned by a Hermite basis, closest to \( p(x|z) \). This problem can also be viewed as the correction step of the Bayes filter (Jazwinski, 1970): Start from a prior and correct it based on the latest (nonlinear) measurement.

Figure 2 provides an example of projection using (29) with an increasing number of Hermite basis functions. The measure used was \( \nu(x) = \mathcal{N}(\mu, 4) \). We see that even with a fixed measure, the approximation gets better as the number of basis functions increases to the point that it is difficult to distinguish visually between the posterior and the approximation. The bottom panel shows that divergence, \( I(p \ast q) \), essentially decreases exponentially fast with more basis functions.
Variational Inference as Iterative Projection in a Bayesian Hilbert Space

Figure 2: An example of projection of a posterior onto a finite basis with increasing number of basis functions. The top panel qualitatively shows that adding more basis functions brings the approximation closer to the posterior. The bottom shows the same quantitatively where $I(p \ominus q)$ decreases exponentially fast with more basis functions.

3.4 Indefinite-Gaussian Subspace of $\mathcal{P}$

We now turn to a subspace involving Gaussian PDFs, which are quintessentially important to statistics and estimation theory. Gaussians, as traditionally defined with a positive-definite covariance matrix, do not in themselves form a subspace. We need to expand the set to include covariance matrices that are sign-indefinite. Let us accordingly define an $N$-dimensional indefinite-Gaussian distribution as

$$p(x) = \mathcal{N}(\mu, \Sigma) = 1/(2\pi)^{N/2} |\Sigma|^{1/2} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right),$$

(38)

which has mean, $\mu$, and symmetric covariance, $\Sigma$. The set of all indefinite Gaussians is

$$\mathcal{G} = \{ p(x) = \mathcal{N}(\mu, \Sigma) \mid x, \mu \in \mathbb{R}^N, \Sigma \in \mathbb{R}^{N \times N} \}.$$  

(39)

It is easy to show that $\mathcal{G}$ is in fact a subspace of $\mathcal{P}$ as the zero vector is contained therein ($\Sigma^{-1} = \mathbf{0}$, allowing that $\Sigma \to \infty$) and the set is closed under vector addition and scalar multiplication.

To establish $\mathcal{G}$ as a Bayesian Hilbert space, we must have an appropriate measure, $\nu$. In our case, we choose the measure to also be a Gaussian, $\nu = \mathcal{N}(\mu, \Sigma) \in \mathcal{G}$. We may thus declare $\mathcal{G}$ to be a Bayesian Hilbert space for a measure $\nu \in \mathcal{G}$ and note that Gaussian PDFs are square-log integrable satisfying (11).

Several possibilities exist to parameterize Gaussians (Barfoot, 2020). There are $\frac{1}{2}N(N + 3)$ unique elements contained in the mean and the symmetric covariance matrix on $\mathbb{R}^N$; hence the dimension of $\mathcal{G}$ is $\frac{1}{2}N(N + 3)$. We shall construct our basis on a positive-definite choice of covariance $\Sigma$ that we can decompose in Cholesky fashion, i.e., $\Sigma = LL^T$.

Now consider

$$\gamma_1(x) = L^{-1}(x - \mu), \quad \gamma_2(x) = \sqrt{\frac{1}{2} \text{D}^T \text{D}} \text{vech} \left( L^{-1}(x - \mu)(x - \mu)^T L^{-T} \right),$$

(40)

wherein $\text{vech} (\cdot)$ is the half-vectorization of its matrix argument and $\text{D}$ is the associated duplication matrix (see Appendix A). Note that $\gamma_1$ is an $N \times 1$ column and $\gamma_2$ is an $\frac{1}{2}N(N + 1) \times 1$ column. With a little abuse of notation, we set the basis functions as

$$g(x) = \begin{bmatrix} g_1(x) \\ g_2(x) \end{bmatrix} = \exp \left( - \begin{bmatrix} \gamma_1(x) \\ \gamma_2(x) \end{bmatrix} \right);$$

(41)

that is, the exponential is applied elementwise. We claim that $g(x)$ is a basis for $\mathcal{G}$. 

7
It is instructive to show that \( g(x) \) spans \( \mathcal{G} \) as well as serving as the proof that it is a basis. Consider again the reparameterization ‘trick’ given by
\[
x = \mu + L\xi
\]  
with \( \xi \sim \mathcal{N}(0, 1) \). This renders (41) as
\[
g(\xi) = \begin{bmatrix} g_1(\xi) \\ g_2(\xi) \end{bmatrix} = \downarrow \exp \left( - \begin{bmatrix} \gamma_1(\xi) \\ \gamma_2(\xi) \end{bmatrix} \right) = \downarrow \exp \left( - \begin{bmatrix} \xi \\ \sqrt{\frac{1}{2}D^T D \text{vech} \xi^T} \end{bmatrix} \right).
\]  
A linear combination of the basis functions can be written as
\[
p(\xi) = \downarrow \exp \left( -\alpha_1^T \gamma_1(\xi) - \alpha_2^T \gamma_2(\xi) \right).
\]  
Now
\[
\alpha_1^T \gamma_1 = \alpha_1^T \xi,
\]  
Also, we can in general express the second set of coordinates as
\[
\alpha_2 = \sqrt{\frac{1}{2}D^T D \text{vech} S}
\]  
for some symmetric \( S \) that can easily be reconstructed from \( \alpha_2 \). Hence
\[
\alpha_2^T \gamma_2 = \frac{1}{2} \text{vech} S^T D^T D \text{vech} \xi^T = \frac{1}{2} \text{vec} S \text{vec} \xi^T
\]  
given the identities \( \text{vec} A = D^\dagger \text{vec} A \) and \( D D^\dagger \text{vec} A = \text{vec} A \), where \( D^\dagger \) is the Moore-Penrose inverse of \( D \) (Appendix A). Moreover, the identity \( \text{vec} A^T \text{vec} B = \text{tr} AB \) leads to
\[
\alpha_2^T \gamma_2 = \frac{1}{2} \text{tr} \left( S \xi \xi^T \right) = \frac{1}{2} \xi^T S \xi.
\]  
Then
\[
p(x) = \downarrow \exp \left( -\alpha_1^T \xi - \frac{1}{2} \xi^T S \xi \right) = \downarrow \exp \left( -\frac{1}{2} \left( \xi + S^{-1}\alpha_1 \right)^T S \left( \xi + S^{-1}\alpha_1 \right) \right) = \downarrow \exp \left( -\frac{1}{2} \left( x - (\mu - LS^{-1}\alpha_1) \right)^T L^{-T} S L^{-1} \left( x - (\mu - LS^{-1}\alpha_1) \right) \right).
\]  
This can represent any Gaussian distribution, where the mean is \( \mu - LS^{-1}\alpha_1 \) and the covariance \( LS^{-1}L^T \). Thus \( g \) spans \( \mathcal{G} \). Furthermore, as the dimension of \( \mathcal{G} \) is \( \frac{1}{2} N(N + 3) \), the number of functions in \( g \), \( g \) is a basis for \( \mathcal{G} \).

This basis is, in addition, orthonormal as can be shown in a straightforward fashion by using the reparameterized form \( g(\xi) \) and recognizing that the entries in \( \gamma_1(\xi) \) are \( \xi_i \) and those in \( \gamma_2(\xi) \) are either \( \xi_i \xi_j (i \neq j) \) or \( \xi_i^2/\sqrt{2} \). Hence, \( \langle g, g \rangle = 1 \).

It can be shown that
\[
\alpha_1 = \langle g_1, p \rangle = L^T \mathbb{E}_\nu \left[ \frac{\partial \phi(x)}{\partial x^T} \right], \quad \alpha_2 = \langle g_2, p \rangle = \sqrt{\frac{1}{2}D^T D \text{vech} \left( L^T \mathbb{E}_\nu \left[ \frac{\partial^2 \phi(x)}{\partial x^T \partial x} \right] L \right)}
\]  
are the coordinates for \( p(x) = \downarrow \exp(-\phi(x)) \in \mathcal{G} \). Consult Appendix E, where the more general case that the projection of any PDF in \( \mathcal{P} \) onto \( \mathcal{G} \) is shown to be given also by (50). Another rendering of (49) is
\[
p(x) = \downarrow \exp \left( -\frac{1}{2} \left( x - (\mu - LS^{-1}\alpha_1) \right)^T L^{-T} S L^{-1} \left( x - (\mu - LS^{-1}\alpha_1) \right) \right).
\]  
which also expresses the projection of a PDF in \( \mathcal{P} \) onto \( \mathcal{G} \). We include the derivation of the information \( I \) of a Gaussian distribution in Appendix F.

### 4 Variational Bayesian Inference

We shall now address the problem of variational Bayesian inference using the algebraic tools of the Bayesian space. But we begin by reviewing the *Fisher information matrix* and showing that with respect to the coordinates used in a given subspace it is simply the Gram matrix of the chosen basis.
4.1 Fisher Information Matrix

Let \( q(x|\theta) \in Q \), a finite-dimensional subspace of \( P \) with basis \( B \), depending on some parameter \( \theta \). The Fisher information on \( \theta \) with respect to the measure \( \nu \) is defined to be the covariance of the score (Fisher, 1922), i.e.,

\[
I_\theta = E_\nu \left[ \left( \frac{\partial \ln q}{\partial \theta} - E_\nu \left[ \frac{\partial \ln q}{\partial \theta} \right] \right)^2 \right] = E_\nu \left[ \left( \frac{\partial \ln q}{\partial \theta} \right)^2 \right] - \left( E_\nu \left[ \frac{\partial \ln q}{\partial \theta} \right] \right)^2. \tag{52}
\]

While our Fisher information may appear slightly unfamiliar, by taking the measure to be the density \( \nu = q \) then \( E_\nu[\partial \ln q/\partial \theta] = 0 \) and we have the traditional version. We purposely delay setting \( \nu = q \) to show the connection to Bayesian space.

Take \( q \) to be expressed as a linear combination of the basis functions \( b_n \), that is,

\[
q(x|\theta) = \bigoplus_n \alpha_n(\theta) \cdot b_n. \tag{53}
\]

The score is

\[
\frac{\partial \ln q}{\partial \theta} = \frac{1}{q} \frac{\partial q}{\partial \theta}. \tag{54}
\]

As \( q = \prod_n b_n^\alpha / \int \prod_n b_n^{\alpha_n} dx \),

\[
\frac{\partial q}{\partial \theta} = \frac{\partial}{\partial \theta} \left( \prod_n b_n^{\alpha_n} / \int \prod_n b_n^{\alpha_n} dx \right) = \sum_n \left( \frac{\partial \alpha_n}{\partial \theta} \right) \ln b_n \prod_n b_n^{\alpha_n} / \int \prod_n b_n^{\alpha_n} dx \left( \prod_n b_n^{\alpha_n} / \int \prod_n b_n^{\alpha_n} dx \right) - \sum_n \left( \frac{\partial \alpha_n}{\partial \theta} \right) \ln b_n \prod_n b_n^{\alpha_n} / \int \prod_n b_n^{\alpha_n} dx \cdot \frac{\partial q}{\partial \theta} = q \sum_n \left( \frac{\partial \alpha_n}{\partial \theta} \right) \left( \ln b_n - E_\nu[\ln b_n] \right). \tag{55}
\]

Hence

\[
\frac{\partial \ln q}{\partial \theta} = \sum_n \left( \frac{\partial \alpha_n}{\partial \theta} \right) \left( \ln b_n - E_\nu[\ln b_n] \right). \tag{56}
\]

(We note in passing that \( \partial \ln q/\partial \theta = \sum_k (\partial \alpha_m/\partial \theta) \) clr \( b_m \), where clr \( b_m \) is the centered log-ratio transformation (van den Boogaart et al., 2014).)

Substituting (56) into (52) produces

\[
I_\theta = \sum_n \sum_m \left( \frac{\partial \alpha_m}{\partial \theta} \right) \left( \frac{\partial \alpha_n}{\partial \theta} \right) \left( E_\nu[\ln b_m \ln b_n] - E_\nu[\ln b_n]E_\nu[\ln b_m] \right) = \left( \frac{\partial \alpha}{\partial \theta} \right)^T (\mathbf{b}, \mathbf{b}) \left( \frac{\partial \alpha}{\partial \theta} \right). \tag{57}
\]

The traditional Fisher information uses \( q \) as the measure and we will indicate that explicitly with a subscript on the inner product, e.g., \( (\mathbf{b}, \mathbf{b})_q \).

We mention for interest that the stochastic derivative of \( q(x|\theta) \) with respect to \( \theta \) is

\[
\frac{\partial q}{\partial \theta} = \bigoplus_m \left( \frac{\partial \alpha_m}{\partial \theta} \right) \cdot b_m(x) \tag{58}
\]

and so

\[
I_\theta = \left( \frac{\partial \alpha}{\partial \theta} \right)^T (\mathbf{b}, \mathbf{b}) \left( \frac{\partial \alpha}{\partial \theta} \right) = \left( \bigoplus_m \left( \frac{\partial \alpha_m}{\partial \theta} \right) \cdot b_m, \bigoplus_n \left( \frac{\partial \alpha_n}{\partial \theta} \right) \cdot b_n \right) = \left( \frac{\partial q}{\partial \theta}, \frac{\partial q}{\partial \theta} \right), \tag{59}
\]

which makes the inner-product expression of the Fisher information coordinate-free.

For multiple parameters, \( \theta_1, \theta_2 \ldots \theta_K \), the \( (m, n) \) entry in the Fisher information matrix (FIM) is

\[
I_{\theta, mn} = E_\nu \left[ \left( \frac{\partial \ln q}{\partial \theta_m} - E_\nu \left[ \frac{\partial \ln q}{\partial \theta_m} \right] \right) \left( \frac{\partial \ln q}{\partial \theta_n} - E_\nu \left[ \frac{\partial \ln q}{\partial \theta_n} \right] \right) \right] = E_\nu \left[ \frac{\partial \ln q}{\partial \theta_m} \frac{\partial \ln q}{\partial \theta_n} \right] - E_\nu \left[ \frac{\partial \ln q}{\partial \theta_m} \right] E_\nu \left[ \frac{\partial \ln q}{\partial \theta_n} \right]. \tag{60}
\]
leading to

$$I_{\theta} = \left( \frac{\partial \alpha}{\partial \theta} \right)^T (b, b) \left( \frac{\partial \alpha}{\partial \theta} \right).$$  \hfill (61)

We shall be particularly interested in the FIM with respect to the coordinates for a given basis, that is, when $\theta = \alpha$. In this case, the FIM is simply the Gram matrix,

$$I_{\alpha} = (b, b).$$  \hfill (62)

When $q$ is used as the measure, we shall write $I_{\alpha} = (b, b)|_q$.

### 4.2 Variation on the Kullback-Leibler Divergence

In variational Bayesian inference, we seek to find an approximation, $q$, from some family of distributions constituting a subspace $Q$, to the true Bayesian posterior $p \in P$. In general,

$$Q \subseteq P,$$

where equality implies that $q = p$ will match the posterior exactly. But $P$ is infinite-dimensional and, in practice, $Q \subset P$ is a finite-dimensional subspace.

There are many possible divergences that can be defined to characterize the ‘closeness’ of $q$ to $p$ including the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951), Bregman divergence (Bregman, 1967), Wasserstein divergence/Earth mover’s distance (Monge, 1781) and Rényi divergence (Rényi, 1961). We shall focus on the KL divergence, which is defined as

$$KL(q||p) = -\int_X q(x) \ln \left( \frac{p(x|z)}{q(x)} \right) dx = -E_q[\ln p - \ln q].$$  \hfill (64)

Sometimes the reverse of this is used: $KL(p||q)$. Note, we show the divergence with respect to the posterior, $p(x|z)$, but in practice during the calculations we use that $p(x|z) = p(x, z)/p(z) = \downarrow p(x, z)$ since the joint likelihood is easy to construct and then the $p(z)$ can be dropped for it does not result in a KL term that depends on $x$. We will generically use $p$ in what follows to keep the notation clean.

**KL Gradient.** We assume a basis $B = \{b_1, b_2 \cdots b_M\}$ for $Q$ and we write $q$ as

$$q = \bigoplus_{m=1}^M \alpha_m \cdot b_m.$$  \hfill (65)

We desire to minimize the KL divergence with respect to the coordinates $\alpha_m$. The gradient of $KL(q||p)$ can be computed as follows:

$$\frac{\partial KL}{\partial \alpha_n} = -\int_X \left( \frac{\partial q}{\partial \alpha_n} (\ln p - \ln q) - q \frac{\partial \ln q}{\partial \alpha_n} \right) dx.$$  \hfill (66)

Recalling (55) and (56), this reduces to

$$\frac{\partial KL}{\partial \alpha_n} = -E_q[\ln b_n (\ln p - \ln q)] + E_q[\ln b_n]E_q[\ln p - \ln q] = -\langle b_n, p \oplus q \rangle_q,$$  \hfill (67)

or, collecting these in matrix form,

$$\frac{\partial KL}{\partial \alpha^T} = -\langle b, p \oplus q \rangle_q.$$  \hfill (68)

The necessary condition for a minimum of the KL divergence is that the gradient is zero. Newton’s method suggests the manner in which we might iteratively solve for the optimal distribution. Following the established procedure, the iteration for the coordinates is given by

$$\alpha^{(i+1)} = \alpha^{(i)} + H^{(i)}^{-1} \langle b, p \oplus q^{(i)} \rangle_{q^{(i)}},$$  \hfill (69)

where $H$ is the Hessian of the KL divergence.
KL Hessian. The \((m, n)\) entry of the Hessian is

\[
\frac{\partial^2 \text{KL}}{\partial \alpha_m \partial \alpha_n} = -\frac{\partial}{\partial \alpha_m} \langle b_n, p \otimes q \rangle_q.
\]  

(70)

This differentiation must take into account the effect of the ‘measure’ \(q\). The product rule applies here and we can break down the differentiation as

\[
\frac{\partial^2 \text{KL}}{\partial \alpha_m \partial \alpha_n} = -\left( \frac{\partial}{\partial \alpha_m} \langle b_n, p \otimes q \rangle_q \right) - \langle b_n, p \otimes q \rangle_{\partial q/\partial \alpha_m},
\]  

(71)

the first term of which is to be read as the derivative of the inner product holding the measure fixed and the second of which deals with the derivative of the measure while holding the arguments of inner product fixed. The first term is

\[
\left( \frac{\partial}{\partial \alpha_m} \langle b_n, p \otimes q \rangle_q \right) = \left( \frac{\partial}{\partial \alpha_m} \langle b_n, p \rangle - \frac{\partial}{\partial \alpha_m} \sum_k \alpha_k \langle b_n, b_k \rangle \right)_q = -\langle b_n, b_m \rangle_q = -\langle b_m, b_n \rangle_q.
\]  

(72)

As shown in Appendix G, the second becomes

\[
\langle b_n, p \otimes q \rangle_{\partial q/\partial \alpha_m} = \left( b_n, \frac{\partial \ln q}{\partial \alpha_m} \cdot (p \otimes q) \right)_q - E_q[\ln p - \ln q] \langle b_m, b_n \rangle_q.
\]  

(73)

We advise that the coefficient \(\partial \ln q/\partial \alpha_m\) of \(p \otimes q\) is in fact a function of the state and as such cannot be transferred to the other argument of the inner product as would be possible for a scalar in the field \(\mathbb{R}\). We also recognize the factor of the last term as \(\text{KL}(q||p)\). Therefore, substituting (72) and (73) into (71) yields

\[
\frac{\partial^2 \text{KL}}{\partial \alpha_m \partial \alpha_n} = (1 - \text{KL}(q||p)) \langle b_m, b_n \rangle_q - \left( b_n, \frac{\partial \ln q}{\partial \alpha_m} \cdot (p \otimes q) \right)_q.
\]  

(74)

We observe that the second term on the right-hand side is symmetric in the indices as the substitution of (56) will attest. In matrix form, the Hessian is

\[
H = \frac{\partial^2 \text{KL}}{\partial \alpha^T \partial \alpha} = (1 - \text{KL}(q||p)) I_{\alpha} - \left( b, \frac{\partial \ln q}{\partial \alpha} \cdot (p \otimes q) \right)_q.
\]  

(75)

Newton’s method (69) can now be implemented. But the Hessian bears a closer look.

The Hessian can also be explicitly written as

\[
\frac{\partial^2 \text{KL}}{\partial \alpha_m \partial \alpha_n} = \langle b_m, b_n \rangle_q - E_q[\ln b_m \ln b_n (\ln p - \ln q)] + E_q[\ln b_m \ln b_n E_q[\ln p - \ln q]] + E_q[\ln b_n E_q[\ln b_m (\ln p - \ln q)] - 2E_q[\ln b_m] E_q[\ln b_n] E_q[\ln p - \ln q],
\]  

(76)

the terms of which can be collected as

\[
\frac{\partial^2 \text{KL}}{\partial \alpha_m \partial \alpha_n} = \langle b_m, b_n \rangle_q + (-b_{mn} + E_q[\ln b_n] \cdot b_m + E_q[\ln b_m] \cdot b_n, p \otimes q)_q.
\]  

(77)

where \(b_{mn} = \downarrow \exp(\ln b_m \ln b_n)\). The symmetry in the Hessian is plainly evident in this version.

Iterative Projection. In the vicinity of the optimal distribution, with a sufficiently large subspace \(Q\), we may expect \(p \otimes q\) to be small in the sense that \(\ln p - \ln q\) is small almost everywhere. This makes all the terms in the Hessian of first order except \(I_{\alpha}\), which is of zeroth order. The gradient (68) is also of first order. Thus to keep Newton’s descent to this order, we may approximate the Hessian as \(H \approx I_{\alpha}\) and the iterative procedure (69) becomes simply

\[
\alpha^{(i+1)} = \alpha^{(i)} + I_{\alpha}^{-1} \left( b, p \otimes q^{(i)} \right)_q^{(i)}.
\]  

(78)

However, as \(q^{(i)} = \bigcup_m \alpha_m^{(i)} \cdot b_m\),

\[
\left( b, p \otimes q^{(i)} \right)_q^{(i)} = \langle b, p \rangle_{q^{(i)}} - \langle b, b \rangle_{q^{(i)}} \alpha^{(i)} = \langle b, p \rangle_{q^{(i)}} - I_{\alpha}^{(i)} \alpha^{(i)}.
\]  

(79)
Figure 3: Iterative projection onto a sequence of Bayesian Hilbert spaces, \((Q, q^{(i)})\).

Hence (78) becomes
\[
\alpha^{(i+1)} = I_{\alpha}^{-1}(b, p)_{q^{(i)}}. \tag{80}
\]
The iterative update to \(q\) is
\[
q^{(i+1)} = \bigoplus_{m} \alpha^{(i+1)} m \cdot b_m.
\]
which describes a projection onto \(Q\) using measure \(q^{(i)}\). That is, the procedure can be viewed as an iterative projection,
\[
q^{(i+1)} = \text{proj}_{(Q, q^{(i)})} p. \tag{82}
\]
Figure 3 depicts the scheme. The procedure is essentially the application of Newton’s method on the KL divergence with the Hessian approximated as the FIM. This is precisely the approximation made in natural gradient descent (Amari, 1998). In our Bayesian space, the operating point of the Newton step becomes the measure for the inner product. This highlights a key aspect of using the algebra associated with a Bayesian space. It recognizes the dual role of \(q\): On the one hand it is the approximating PDF and on the other it serves as a measure that weights the difference between the approximation and the approximated.

Convergence of iterative projection is guaranteed if the Hessian is positive-definite. Provided that the subspace is large enough, we can expect convergence when we begin in a neighborhood of optimal \(q\) where the first-order terms in the Hessian are sufficiently small.

It is notable that each step of the iterative projection is equivalent to the local minimization of the divergence \(I(p \ominus q)\) with the measure fixed at \(q^{(i)}\) because
\[
I(p \ominus (q^{(i)} \oplus \delta q)) = I(p \ominus q^{(i)}) + \delta \alpha^T \left( \frac{\partial I}{\partial \alpha^T} \right)_{q^{(i)}} + \frac{1}{2} \delta \alpha^T \left( \frac{\partial^2 I}{\partial \alpha^T \partial \alpha} \right)_{q^{(i)}} \delta \alpha, \tag{83}
\]
where \(\delta \alpha = \alpha^{(i+1)} - \alpha^{(i)}\) and
\[
\left( \frac{\partial I}{\partial \alpha^T} \right)_{q^{(i)}} = -\left( b, p \ominus q^{(i)} \right)_{q^{(i)}}, \quad \left( \frac{\partial^2 I}{\partial \alpha^T \partial \alpha} \right)_{q^{(i)}} = \langle b, b \rangle_{q^{(i)}} = I_{\alpha}^{(i)}, \tag{84}
\]
which are identical to the linearized forms for the KL divergence.

Throughout this section we have assumed that the basis \(B\) remains constant across iterations, but this need not be the case. We may also choose to update the basis along with the measure to maintain, for example, orthonormality. This is explored in the next section on Gaussian iterative projection.

5 Iterative Gaussian Projection

Let us investigate a little more closely iterative projection using Gaussian PDFs, given their importance in statistics and estimation theory.

5.1 Projections

As mentioned at the end of the last section, we do not have to maintain the same basis from step to step as long as each basis spans the same subspace. This is a particularly useful maneuver when using the subspace \(G\) of indefinite
Gaussians. Denote the mean and variance of \( q(i) \in \mathcal{G} \) as \( \mu(i) \) and \( \Sigma(i) \) and let the basis \( g(i) \) be defined as in (40) and (41). Note that this basis is orthonormal with respect to \( q(i) \). As such, \( I_{\alpha}^{(i)} = \langle g(i), g(i) \rangle_{q(i)} = 1 \). Imagine the PDF to be approximated is expressed as \( p = \downarrow \exp(-\phi(x)) \in \mathcal{P} \). The coordinates resulting from the next projection are given by (50), namely,

\[
\alpha_1^{(i+1)} = \langle g_1^{(i)}, p \rangle = L^{(i)} T E_{q(i)} \left[ \frac{\partial \phi(x)}{\partial x^T} \right],
\]

\[
\alpha_2^{(i+1)} = \langle g_2^{(i)}, p \rangle = \sqrt{\frac{1}{2}} D^T \text{vech} \left( L^{(i)} T E_{q(i)} \left[ \frac{\partial^2 \phi(x)}{\partial x^T \partial x} \right] L^{(i)} \right),
\]

where \( L^{(i)} \) issues from the Cholesky decomposition of \( \Sigma^{(i)} \).

The new iteration is

\[
q^{(i+1)} = \text{proj}_{p} \left( \downarrow \exp \left( -\alpha_1^{(i+1)} \gamma_1^{(i)} - \alpha_2^{(i+1)} \gamma_2^{(i)} \right) \right).
\]

Using (51), this becomes

\[
q^{(i+1)} = \downarrow \exp \left( -\frac{1}{2} (x - \mu^{(i)})^T \Sigma^{(i+1)}^{-1} (x - \mu^{(i)}) \right).
\]

Herein

\[
\Sigma^{(i+1)}^{-1} = E_{q(i)} \left[ \frac{\partial^2 \phi(x)}{\partial x^T \partial x} \right],
\]

\[
\Sigma^{(i+1)}^{-1} \delta \mu = -E_{q(i)} \left[ \frac{\partial \phi(x)}{\partial x^T} \right],
\]

\[
\mu^{(i+1)} = \mu^{(i)} + \delta \mu,
\]

give the updates from \( q(i) \) to \( q^{(i+1)} \). In this way, the same Gaussian variational inference approach is presented by Barfoot et al. (2020). We have arrived at the same variational updates but have done so from the framework of a Bayesian Hilbert space, where it becomes abundantly clear that the minimization algorithm is in fact a slightly simplified version of Newton’s method. This also provides the connection back to the classic Gaussian filtering and smoothing algorithms as discussed by Barfoot et al. (2020).

### 5.2 Convex Sets

The indefinite-Gaussian subspace does not guarantee that the covariance matrix of the approximation is positive-definite for every PDF, \( p = \downarrow \exp(-\phi(x)) \). However, as this covariance, or more precisely its inverse, is given by

\[
\Sigma^{-1} = E_{q} \left[ \frac{\partial^2 \phi(x)}{\partial x^T \partial x} \right],
\]

positive-definiteness can be ensured if the Hessian of \( \phi(x) \) is positive-definite. Equivalently, if \( \phi(x) \) is a convex function in \( x \), that is, if \( p(x) \) is logarithmically concave, then the covariance of the projection on \( \mathcal{G} \) will be positive-definite.

The same can in fact be said about projection \( p \) onto the Hermite subspace \( \mathcal{H} \) provided that at least the first two Hermite polynomials are used in the basis (and combinations thereof in the multivariate case). The indefinite-Gaussian subspace is also a subspace of \( \mathcal{H} \) so constructed.

This conjures the picture in Figure 4. Let the subset \( \mathcal{P}^+ \subset \mathcal{P} \) include all logarithmically concave PDFs or all PDFs with a negative convex log-likelihood, i.e.,

\[
\mathcal{P}^+ = \left\{ \text{p(x)} = \downarrow \exp(-\phi(x)) \in \mathcal{P} \mid \phi(x) \text{ convex in } x \in \mathcal{X} \right\}.
\]
Variational Inference as Iterative Projection in a Bayesian Hilbert Space

Figure 4: Relationship of subsets of PDFs with positive-definite variances. Specifically, \( \mathcal{G}^+ \subset \mathcal{G} \), \( \mathcal{H}^+ \subset \mathcal{H} \), and \( \mathcal{P}^+ \subset \mathcal{P} \).

We can similarly define convex subsets for \( \mathcal{H} \),

\[
\mathcal{H}^+ = \left\{ p(x) = \downarrow \exp(-\phi(x)) \in \mathcal{H} \mid \phi(x) \text{ convex in } x \in \mathcal{X} \right\},
\]

and \( \mathcal{G} \),

\[
\mathcal{G}^+ = \left\{ p(x) = \downarrow \exp(-\phi(x)) \in \mathcal{G} \mid \phi(x) \text{ convex in } x \in \mathcal{X} \right\}.
\]

But we emphasize that these are subsets not subspaces. However, \( \mathcal{P}^+ \) can be projected onto \( \mathcal{H}^+ \subset \mathcal{H} \), which in turn can be projected onto \( \mathcal{G}^+ \subset \mathcal{G} \).

Unfortunately, starting with \( p \in \mathcal{P}^+ \) is a very strict condition that can be difficult to meet in practice. Fortunately, this is not the only way that we can end up with \( q \in \mathcal{G}^+ \). One possibility is to start with \( q \) in a local basin of convergence where the expected curvature remains positive. Another is to make an input correction by adding a regularizer term to \( p \in \mathcal{P} \) to cause it to move into \( \mathcal{P}^+ \) before projection. On the output end, if \( q \) does not end up in \( \mathcal{G}^+ \) it can be rejected in favour of the closest \( q \in \mathcal{G}^+ \) to \( p \); in other words, one can add a constraint to the variational objective. In the Maximum A Posteriori (MAP) setting, the last two ideas are tied to Levenberg-Marquardt regularization.

6 Numerical Examples

We return to the Hermite basis to provide examples of iterative projection on 2- and 4-dimensional subspaces using again the posterior of §3.3. In all examples, the expectations were computed with generic numerical integration although, as discussed by Barfoot et al. (2020), there are several other options including Gaussian quadrature.

We consider two different examples. In the first, shown in Figure 5, we hold the Hermite basis functions fixed and take the full estimate \( q^{(i)} \) from the previous iteration to be the measure. With respect to the changing (and in the 4-basis-function case non-Gaussian) measure, the fixed basis is not orthogonal. This then necessitates calculation of the Gram matrix at each iteration to implement the projection. We initialized both the 2- and 4-basis-function estimates to the prior, which corresponds to the first panel in Figure 5. The next three panels show subsequent iterations of the estimates. The last panel shows the KL divergence between the estimates and the true posterior for 10 iterations. We see that both estimates converged in a few iterations with the 4-basis-function version performing better at each iteration and becoming visibly indistinguishable from the true posterior after four iterations.

In the second example, shown in Figure 6, we chose to use a Gaussian measure, even for the 4-basis-function estimate. This was motivated by the fact that the Hermite basis is orthogonal with respect to a Gaussian measure. Using an orthogonal basis brings along with it a simpler form of the projection since the Gram matrix is the identity matrix. As the measure changes from one iteration to the next, we then have to update the basis to retain the desired orthogonality. This can be accomplished by again using the reparameterization ‘trick’ to adjust the basis to be orthogonal with respect to the current Gaussian measure. For the 4-basis-function estimate, the Gaussian measure was calculated by further projecting the current estimate onto the Gaussian subspace, a trivial matter of retaining only the first two coordinates corresponding to \( H_1 \) and \( H_2 \). The first panel of Figure 6 again depicts both estimates initialized to the prior. The next three panels show the iterative scheme progressing. The last panel provides the KL divergence between the estimates and the true posterior. The results are similar to the previous example. Interestingly, the 4-basis-function estimate was only slightly worse in terms of its final KL than the first example despite the fact that it is constrained by its use of a Gaussian measure.
Variational Inference as Iterative Projection in a Bayesian Hilbert Space

Figure 5: Example of iterative projection onto 2- and 4-dimensional subspaces spanned by Hermite polynomials, where the measure is taken to be the estimate $q^{(i)}$ at the previous iteration and the basis held fixed. Both the 2- and 4-basis-function estimates were initialized to the prior (first panel) and then iteratively updated (next three panels). The last panel shows the KL divergence between the estimates and the true posterior for 10 iterations. We see that the estimates converged in a few iterations with the 4-basis-function estimate matching the posterior better at every iteration.
Figure 6: Same setup as Figure 5, except that the measure was Gaussian for both estimates. For the 4-basis-function estimate, this involved projecting the current estimate onto the Gaussian subspace. The basis functions were reorthogonalized at each iteration as described in §5, accounting for the small difference in the 2-basis-function case from the previous example. Again, we see both estimates converged well. The 4-basis-function estimate converged to a slightly worse value than when the full estimate was used as the measure rather than its Gaussian projection.
We note that the second example exhibited a slightly slower convergence than the first. This is understandable in the
4-basis-function case as the algorithm is slightly compromised by essentially truncating the measure to a Gaussian. As
for the 2-basis-function case, while the two examples differ only in the reorthogonalization of the basis functions, it
must be borne in mind that the use of different bases for even the same subspace can result in a different progression
of the projections. At convergence, there is still a very slight difference in the KL-divergence values but this is an artefact
of the integration tolerance operating on different variables (ξ instead of x).

7 Exploiting Sparsity

One of the major advantages of thinking of $P$ as a vector space with the definition of vector addition $\oplus$ is that
Bayesian inference in general can be viewed as the addition of vectors. Consider the posterior $p(x|z)$ where $z$ are some
measurements. Bayes’ rule states that

$$p(x|z) = \frac{p(z|x)p(x)}{p(z)} = p(z|x) \oplus p(x), \quad (94)$$

where $p(x)$ is a prior, $p(z|x)$ is a measurement factor and, as mentioned earlier, we needn’t introduce the normalization
constant $p(z)$ explicitly when writing the posterior as a vector addition in Bayesian space.

7.1 Multiple Measurements

If we have several measurements that are statistically independent, then this can be factored as

$$p(x|z) = p(x) \oplus K \bigoplus_{k=1}^{K} p(z_k|x_k), \quad (95)$$

where $x_k = P_k x$ is a subset of the variables in $x$, $P_k$ is a projection matrix, and $z_k$ is the $k$th measurement. This
expresses sparsity in the state description and in the measurements. To keep the notation economical, we shall simply
write

$$p = \bigoplus_{k=0}^{K} p_k, \quad (96)$$

where $p$ is the posterior and the $p_k$ comprise the prior and the measurements, corresponding to statistically independent
data. In other words, the factorization becomes a summation in the Bayesian space $P$.

Now consider our projective approach to inference. As usual, given a subspace $Q \subset P$, the optimal estimate to (96) is
given by

$$q^* = \text{proj}_Q p = \text{proj}_Q \bigoplus_{k=0}^{K} p_k = \bigoplus_{k=0}^{K} \text{proj}_Q p_k, \quad (97)$$

That is, the projection of the sum is the sum of the projections. Each individual projection can be done separately
because we are in a linear space. This is of enormous practical advantage because it means that we do not need all of $Q$
to represent each projection.

We can see this more clearly by defining $P_k \subset P$ as the subspace corresponding to the variables $x_k$. Then

$$p = \bigoplus_{k=0}^{K} p_k \in P_0 \oplus P_1 \oplus \cdots \oplus P_K \subseteq P. \quad (98)$$

In other words, $p$ is contained in the direct sum of the subspaces $P_k$. Each constituent part $p_k$ may be confined to a
smaller subspace of $P$, depending on the variable dependencies in each term.

If we wish to project $p_k \in P_k$ onto $Q$ it will suffice to consider the projection on just $Q_k = P_k \cap Q$, i.e.,

$$\text{proj}_Q p_k = \text{proj}_{Q_k} p_k. \quad (99)$$

The subspace $Q_k$ may, and ideally would, be smaller than $Q$. We may refer to $Q_k$ as the marginal subspace of $Q$ with
respect to the subset of variables $x_k$.

Therefore, the optimal estimate will be given by

$$q^* = \text{proj}_Q p = \bigoplus_{k=0}^{K} \text{proj}_{Q_k} p_k. \quad (100)$$

17
This means that we can project the PDF associated with each measurement onto a smaller subspace and simply add up the estimates, lifting the overall estimate up into a potentially much larger space. The decomposition and reconstitution is illustrated in Figure 7. Just as with the total posterior, we may describe \( q^* \) as being an element of a direct sum of the individual subspaces of \( Q \), i.e.,

\[
q^* \in Q_0 \oplus Q_1 \oplus \cdots \oplus Q_K \subseteq Q.
\]

The subspace sum may be substantially smaller than \( Q \) but again it will depend on the variable dependencies of each term.

This is the key result that allows most practical inference frameworks to function in a tractable way. Depending on the chosen basis for \( Q \), many of the coordinates can potentially be zero and thus it will not be necessary to waste effort computing them or space storing them.

### 7.2 Gaussian Inference

The effect of sparsity as it applies to iterative Gaussian inference is of particular interest. Let us consider the decomposition of a posterior \( p \) in accordance to the foregoing; that is,

\[
p = \downarrow \exp(-\phi(x)) = \downarrow \exp \left( - \sum_{k=0}^{K} \phi_k(x_k) \right) = \bigoplus_{k=0}^{K} \exp(-\phi_k(x_k)) = \bigoplus_{k=0}^{K} p_k,
\]

where \( \phi_k(x_k) \) is the \( k \)th (negative log) factor expression and \( x_k = P_k \mathbf{x} \).

As in (51), we may express the variational estimate as

\[
q^{(i+1)} = \text{proj}_{(Q, q^{(i)})} p = \downarrow \exp \left( -(\mathbf{x} - \mathbf{\mu}^{(i)})^T \mathbb{E}_{q^{(i)}} \left[ \frac{\partial \phi(x)}{\partial \mathbf{x}^T} \right] - \frac{1}{2} (\mathbf{x} - \mathbf{\mu}^{(i)})^T \mathbb{E}_{q^{(i)}} \left[ \frac{\partial^2 \phi(x)}{\partial \mathbf{x}^T \partial \mathbf{x}} \right] (\mathbf{x} - \mathbf{\mu}^{(i)}) \right),
\]
using the measure \( q^{(i)} = \mathcal{N}(\mu^{(i)}, \Sigma^{(i)}) \). To take advantage of sparsity, we need to have it reflected in the expectations herein. The first one leads to

\[
\mathbb{E}_{q^{(i)}} \left[ \frac{\partial \phi(x)}{\partial x^T} \right] = \sum_{k=0}^{K} \mathbb{E}_{q^{(i)}} \left[ \frac{\partial \phi_k(x_k)}{\partial x^T} \right] = \sum_{k=0}^{K} \left( \frac{\partial x_k}{\partial x} \right)^T \mathbb{E}_{q^{(i)}} \left[ \frac{\partial \phi_k(x_k)}{\partial x^T} \right] = \sum_{k=0}^{K} P_k^T \mathbb{E}_{q^{(i)}} \left[ \frac{\partial \phi_k(x_k)}{\partial x^T} \right] \left( \frac{\partial x_k}{\partial x} \right)^T \] (104)

given that \( x_k = P_k x \). For each factor \( k \), then, we are able to shift the differentiation from \( x \) to \( x_k \). We draw attention to the last equality, where the expectation simplifies to using \( q^{(i)}_k = q^{(i)}(x_k) \), the marginal of the measure for just the variables in factor \( k \). In a similar fashion,

\[
\mathbb{E}_{q^{(i)}} \left[ \frac{\partial^2 \phi(x)}{\partial x^T \partial x} \right] = \sum_{k=0}^{K} P_k^T \mathbb{E}_{q^{(i)}} \left[ \frac{\partial^2 \phi_k(x_k)}{\partial x^T \partial x} \right] P_k
\] (105)

accounts for the second expectation in (103).

The implication of the factorization is that each factor, identified by \( \phi_k(x_k) \), is projected onto \( G_k \), the marginal subspace associated with variables \( x_k \). The results can then be recombined for the full variational estimate as

\[
q^{(i+1)} = \text{proj}_{\{G_k,q^{(i)}_k\}} p = \bigoplus_{k=0}^{K} \text{proj}_{G_k,q^{(i)}_k} p_k = \bigoplus_{k=0}^{K} q^{(i+1)}_k.
\] (106)

The individual projections of \( p_k = \downarrow \exp \left( -\phi_k(x_k) \right) \) onto \( \{G_k,q^{(i)}_k\} \) are

\[
q^{(i+1)}_k = \text{proj}_{\{G_k,q^{(i)}_k\}} p_k = \downarrow \exp \left( -\frac{1}{2}(x_k - \mu^{(i+1)}_k)^T \Sigma^{(i+1)-1}_k (x_k - \mu^{(i+1)}_k) \right)
\]

\[
= \downarrow \exp \left( -\frac{1}{2}(x - P_k^T \mu^{(i+1)}_k)^T \left( P_k^T \Sigma^{(i+1)-1}_k P_k \right) (x - P_k^T \mu^{(i+1)}_k) \right),
\] (107)

where

\[
\mu^{(i+1)}_k = \mu^{(i)}_k - \Sigma^{(i+1)-1}_k \mathbb{E}_{q^{(i)}_k} \left[ \frac{\partial \phi_k(x_k)}{\partial x_k} \right], \quad \Sigma^{(i+1)-1}_k = \mathbb{E}_{q^{(i)}_k} \left[ \frac{\partial^2 \phi_k(x_k)}{\partial x_k^T \partial x_k} \right].
\] (108)

It is straightforward to show that the vector sum of \( q_k \) from (107) reproduces (103). (Note that \( P_k P_k^T = I \) as \( P_k \) is a projection matrix and \( P_k^T \) the corresponding dilation.)

As explained in detail by Barfoot et al. (2020), it would be too expensive for practical problems to construct first \( \Sigma^{(i)} \) and then extract the required blocks for the marginals, \( q^{(i)}_k = \mathcal{N}(\mu^{(i)}_k, \Sigma^{(i)}_k) = \mathcal{N}(P_k \mu^{(i)}_k, P_k \Sigma^{(i)}_k P_k^T) \). We see from the above development that we actually only require the blocks of \( \Sigma^{(i)}_k \) corresponding to the nonzero blocks of its inverse and the method of Takahashi et al. (1973) can be used to extract the required blocks efficiently. Barfoot et al. (2020) provide numerical experiments showing the efficacy of this approach.

Moreover, if we choose to project onto a larger Hermite basis, we can still continue to use a Gaussian measure since the Hermite basis functions we defined are orthonormal with respect to this measure; this is what was done in the second numerical example above. This means that even with a larger approximation family, we can continue to exploit the method of Takahashi et al. (1973) to extract only the required marginals.

8 Concluding Remarks

Our principal goal in this work has been to provide a new perspective on the problem of variational inference. This new vantage point is afforded by considering probability density functions as elements in a Bayesian Hilbert space, where vector addition is a normalized multiplication (perturbation) that accounts for Bayes’ rule and scalar multiplication is a normalized exponentiation (powering). Gaussians and, more generally, exponential families, which are often used in variational inference, constitute subspaces. We thus have at our disposal all the familiar instruments of linear algebra.

The use of the Kullback-Leibler divergence \( \text{KL}(q||p) \) in variational inference to find the best approximation \( q \) to a given posterior \( p \) is widespread. In most approaches, the canvas on which the minimization is carried out is a set, usually
convex, or a manifold of admissible functions (Csiszár, 1975; Csiszár and Tusnády, 1984; Amari, 1995; Adamčík, 2014; Amari, 2016). ‘Projections’ of \( p \) onto the set or manifold are \textit{ipso facto} the PDF \( q \) that minimizes the divergence. However, in Bayesian space, we may interpret projections as standard linear-algebraic projections, reminding us of a Euclidean world.

We take particular note of the information geometry of Csiszár and Amari. They along with their colleagues (Csiszár and Tusnády, 1984; Amari et al., 1992) separately developed the \textit{em} algorithm—not to be confused with the EM (expectation-maximization) algorithm although the two are in many cases equivalent—to solve the generalized variational problem, which involves a dual minimization of \( q \) over its manifold and \( p \) over its own. (The minimum is therefore the minimum ‘distance’ between manifolds.) The \textit{e}-step of the algorithm is performed by making the manifold ‘flat,’ i.e., linear, as a result of using an exponential family of densities. This flattening is equivalent to thinking in terms of a Bayesian space as we have done here. Indeed, as we have shown, the natural-gradient-descent algorithm of Amari (1998) can be explained using this framework as a Newton-like iterative projection.

Based on the inner product of our Bayesian space, we have suggested a new information measure. It is proportional to the squared norm of a probability distribution, which can be used to establish a (symmetric and quadratic) divergence between two PDFs. The connection to the KL divergence is worthwhile mentioning. Each step in the iterative-projection algorithm presented here for variational inference based on the KL divergence amounts to a local minimization of our Bayesian-space divergence.

The linear structure of a Bayesian space furthermore allows us to treat sparsity in measurement data very neatly as the vector sum of the measurements, each of which can be expressed as an element in a subspace restricted to the local variables dictated by the sparsity of the problem, for example, as in the simultaneous-localization-and-mapping (SLAM) problem in robotics (Barfoot et al., 2020). The mean-field approximation in variational inference can be handled in much the same way in this framework. The factorization of a distribution with respect to a desired family of distributions would again be rendered as a vector sum of PDFs.

In his fictional autobiography, \textit{Zen and the Art of Motorcycle Maintenance}, Robert M. Pirsig notes that “One geometry cannot be more true than another; it can only be more convenient.” The same can be said of algebra. Whether one takes a geometric or algebraic tack in analyzing a problem, it can be agreed that different perspectives offer different views and given a particular problem or even a particular class of problem one tack may sometimes be more convenient than others. We hope the perspective presented here on variational inference using a Bayesian Hilbert space of probability density functions offers not only convenience in some respects but insight and a degree of elegance as well.

**References**

Adamčík, M., “The Information Geometry of Bregman Divergences and Some Applications in Multi-Expert Reasoning,” \textit{Entropy}, 16:6338–6381, 2014.

Aitchison, J., “The statistical analysis of compositional data (with discussion),” \textit{Journal of the Royal Statistical Society, Series B}, 44:139–177, 1982.

Amari, S.-I., “The EM Algorithm and Information Geometry in Neural Network Learning,” \textit{Neural Computation}, 7:13–18, 1995.

Amari, S.-I., “Natural gradient works efficiently in learning,” \textit{Neural computation}, 10(2):251–276, 1998.

Amari, S.-I., \textit{Information Geometry and Its Applications}, Springer, Japan, 2016.

Amari, S.-I., Kurata, K., and Nagaoka, H., “Information Geometry of Boltzmann Machines,” \textit{IEEE Transactions on Neural Networks}, 3(2):260–271, 1992.

Ambrogioni, L., Güçlü, U., Güçlütürk, Y., Hinne, M., Maris, E., and van Gerven, M. A. J., “Wasserstein Variational Inference,” in \textit{32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada}, 2018.

Barber, D., \textit{Bayesian Reasoning and Machine Learning}, Cambridge University Press, 2012.

Barfoot, T. D., “Stochastic Decentralized Systems,” Ph.D. Thesis, University of Toronto, 2002.

Barfoot, T. D., \textit{State Estimation for Robotics}, Cambridge University Press, 2017.

Barfoot, T. D., “Multivariate Gaussian Variational Inference by Natural Gradient Descent,” Technical report, Autonomous Space Robotics Lab, University of Toronto, 2020, arXiv:2001.10025 [stat.ML].

Barfoot, T. D. and D’Eleuterio, G. M. T., “An Algebra for the Control of Stochastic Systems: Exercises in Linear Algebra,” in \textit{Proceedings of the 5th International Conference on Dynamics and Control of Structures in Space (DCSS)}, Cambridge, England, 2002.
Barfoot, T. D. and D’Eleuterio, G. M. T., “Stochastic Algebra for Continuous Variables,” Technical report, University of Toronto Institute for Aerospace Studies, 2003.

Barfoot, T. D., Forbes, J. R., and Yoon, D. J., “Exactly Sparse Gaussian Variational Inference with Application to Derivative-Free Batch Nonlinear State Estimation,” International Journal of Robotics Research, to appear, 2020, arXiv:2001.10025 [stat.ML].

Bayes, T., “Essay towards solving a problem in the doctrine of chances,” Philosophical Transactions of the Royal Society of London, 53:370–418, 1763.

Bishop, C. M., Pattern Recognition and Machine Learning, Springer, 2006.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D., “Variational Inference: A Review for Statisticians,” Journal of the American Statistical Association, 112(518):859–877, 2017.

Bregman, L. M., “The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming,” USSR Computational Mathematics and Mathematical Physics, 7(3):200–217, 1967.

Csiszár, I., “I-Divergence Geometry of Probability Distributions and Minimization Problems,” The Annals of Probability, 3(1):146–158, 1975.

Csiszár, I. and Tusnády, G., “Information Geometry and Alternating Minimization Procedures,” in Statistics and Decisions, Supplement Issue No. 1, R. Oldenberg, 1984.

Egozcue, J., Pawlowsky-Glahn, V., Tolosana-Delgado, R., Ortega, M. I., and van den Boogaart, K. G., “Bayes spaces: use of improper distributions and exponential families.” RACSAM: Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales. Serie A. Matemáticas, 107:475–486, 2013.

Egozcue, J. J., Diaz-Barrero, J. L., and Pawlowsky-Glahn, V., “Hilbert Space of Probability Density Functions Based on Aitchison Geometry,” Acta Mathematica Sinica, 22:1175–1182, 2006.

Fisher, R. A., “On the mathematical foundations of theoretical statistics,” Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 222(594-604):309–368, 1922.

Hinton, G. E. and van Camp, D., “Keeping Neural Networks Simple by Minimizing the Description Length of the Weights,” in Sixth ACM Conference on Computational Learning Theory, Santa Cruz, California, 1993.

Jazwinski, A. H., Stochastic Processes and Filtering Theory, Academic, New York, 1970.

Jordan, M. I., Ghahramani, Z., Jaakkola, T., and Saul, L. K., “An Introduction to Variational Methods for Graphical Models,” Machine Learning, 37:183–233, 1999.

Kullback, S. and Leibler, R. A., “On information and sufficiency,” The annals of mathematical statistics, 22(1):79–86, 1951.

Laplace, P.-S., Philosophical Essay on Probabilities, Springer, 1995, translated by Andrew I. Dale from Fifth French Edition, 1825.

Li, Y. and Turner, R. E., “Rényi Divergence Variational Inference,” in 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 2016.

Magnus, J. R. and Neudecker, H., “The elimination matrix: some lemmas and applications,” SIAM Journal on Algebraic Discrete Methods, 1(4):422–449, 1980.

Magnus, J. R. and Neudecker, H., Matrix differential calculus with applications in statistics and econometrics, John Wiley & Sons, 2019.

Manton, J. H. and Amblard, P.-O., “A Primer on Reproducing Kernel Hilbert Spaces,” Technical report, The University of Melbourne and CNRS, 2015, arXiv:1408.0952v2 [math.HO].

McGrayne, S. B., The Theory That Would Not Die: How Bayes’ Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy, Yale University Press, New Haven, Connecticut, 2011.

Monge, G., “Mémoire sur la théorie des déblais et des remblais,” Histoire de l’Académie Royale des Sciences de Paris, 1781.

Painsky, A. and Wornell, G. G., “Bregman Divergence Bounds and Universality Properties of the Logarithmic Loss,” Technical report, Department of Industrial Engineering, Tel Aviv University, 2020, arXiv:1810.07014v2 [cs.IT].

Pawlowsky-Glahn, V. and Egozcue, J. J., “Geometric Approach to Statistical Analysis on the Simplex,” Stochastic Environmental Research and Risk Assessment, 15:384–398, 2001.
A Kronecker Product, vec and vech Operators, and Duplication Matrices

There are several identities, of which we have made use, involving the Kronecker product $\otimes$ and the vectorization operator vec($\cdot$) that stacks the columns of a matrix:

\[
\begin{align*}
\text{vec}(a) & \equiv a \\
\text{vec}(ab^T) & \equiv b \otimes a \\
\text{vec}(ABC) & \equiv (C^T \otimes A) \text{vec}(B) \\
\text{vec}(A)^T \text{vec}(B) & \equiv \text{tr}(AB) \\
(A \otimes B)(C \otimes D) & \equiv (AC) \otimes (BD) \\
(A \otimes B)^{-1} & \equiv A^{-1} \otimes B^{-1} \\
(A \otimes B)^T & \equiv A^T \otimes B^T. 
\end{align*}
\]

(109)

It is worth noting that $\otimes$ and vec($\cdot$) are linear operators.

As we will be working with (symmetric) covariance matrices when discussing Gaussians, we would like to be able to represent them parsimoniously in terms of only their unique variables. Following Magnus and Neudecker (2019, §18), we introduce the half-vectorization operator vech($\cdot$) that stacks up the elements in a column matrix, excluding all the elements above the main diagonal. The duplication matrix $D$ allows us to recover a full symmetric matrix from its unique parts:

\[
\text{vec}(A) = D \text{vech}(A) \quad \text{(symmetric } A). 
\]

(110)

It is helpful to consider a simple $2 \times 2$ example:

\[
A = \begin{bmatrix} a & b \\
     b & c \end{bmatrix}, \quad \text{vec}(A) = \begin{bmatrix} a \\
     b \\
     b \\
     c \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 & 0 \\
     0 & 1 & 0 \\
     0 & 1 & 0 \\
     0 & 0 & 1 \end{bmatrix}, \quad \text{vech}(A) = \begin{bmatrix} a \\
     b \\
     c \end{bmatrix}. 
\]

(111)

The Moore-Penrose pseudoinverse of $D$ will be denoted $D^\dagger$ and is given by

\[
D^\dagger = (D^T D)^{-1} D^T. 
\]

(112)

We can then use $D^\dagger$ to convert the vectorization of a matrix into its half-vectorization:

\[
\text{vech}(A) = D^\dagger \text{vec}(A) \quad \text{(symmetric } A). 
\]

(113)

For our $2 \times 2$ example we have

\[
D^\dagger = \begin{bmatrix} 1 & 0 & 0 & 0 \\
     0 & \frac{1}{2} & \frac{1}{2} & 0 \\
     0 & \frac{1}{2} & \frac{1}{2} & 0 \\
     0 & 0 & 0 & 1 \end{bmatrix}. 
\]

(114)
Useful identities involving $D$ are then

\[ D^\dagger D = 1 \]
\[ D^\dagger T^T = D D^\dagger \]
\[ DD^\dagger \text{vec}(A) = \text{vec}(A) \quad \text{(symmetric A)} \]
\[ DD^\dagger (A \otimes A) D = (A \otimes A) D \quad \text{(any A)} \]

which can be found in Magnus and Neudecker (1980).

**B Outer Products**

The outer product $\Phi : \mathcal{P} \to \mathcal{P}$ of two vectors $b = b(x), c = c(x') \in \mathcal{P}$, denoted $\Phi(x, x') = b(x)\langle c(x') \rangle$ or briefly $\Phi = b\langle c, \rangle$, is defined by its operation on arbitrary $d = d(x') \in \mathcal{P}$ as

\[ \Phi(x, x') \otimes d(x') = b(x)\langle c(x') \rangle \otimes d(x') = b(x) \cdot (c, d) = (c, d) \cdot b(x). \quad (116) \]

Thus, dropping the functional dependence,

\[ \langle a, \Phi \rangle \langle d \rangle = \langle a, b \rangle \langle c, d \rangle \quad (117) \]

for arbitrary $a \in \mathcal{P}$. More generally,

\[ \Phi = \bigoplus_{i=1}^{M} \bigoplus_{j=1}^{N} \phi_{ij} \cdot b_i \langle c_j, \rangle \quad (118) \]

where $b_i, c_j \in \mathcal{P}$ and $\phi_{ij} \in \mathbb{R}$, so that

\[ \Phi \otimes d = \bigoplus_{i=1}^{M} \sum_{j=1}^{N} \phi_{ij} \langle c_j, d \rangle \cdot b_i \quad (119) \]

and

\[ \langle a, \Phi \rangle \langle d \rangle = \sum_{i=1}^{M} \sum_{j=1}^{N} \phi_{ij} \langle a, b_i \rangle \langle c_j, d \rangle. \quad (120) \]

Defining the matrix $\Phi = [\phi_{ij}] \in \mathbb{R}^{M \times N}$ and

\[ b(x) = \begin{bmatrix} b_1(x) \\ b_2(x) \\ \vdots \\ b_M(x) \end{bmatrix}, \quad c(x) = \begin{bmatrix} c_1(x) \\ c_2(x) \\ \vdots \\ c_N(x) \end{bmatrix}, \quad (121) \]

we may abbreviate (118) to

\[ \Phi(x, x') = b(x) \langle c(x') \rangle \quad (122) \]

and hence $\langle a, \Phi \rangle \langle d \rangle = \langle a, b \rangle \Phi(c, d)$, where $\langle a, b \rangle$ is interpreted as a row and $\langle c, d \rangle$ as a column.

Given an orthonormal basis $\{b_1, b_2 \cdots b_M\}$ for a subspace $\mathcal{S} \subset \mathcal{P}$,

\[ Q = \bigoplus_{m=1}^{M} b_m \langle b_m \rangle b_m \equiv b \quad (123) \]

is the kernel of $\mathcal{S}$ and thus, for any $s \in \mathcal{S}$, $Q \otimes s = s$ (Manton and Amblard, 2015). For an nonorthonormal basis,

\[ Q = \bigoplus_{m=1}^{M} \bigoplus_{n=1}^{M} \kappa_{mn} \cdot b_m \langle b_n \rangle b_n \equiv b \langle b, b \rangle^{-1} b, \quad (124) \]

where $\kappa_{mn}$ is the $(m, n)$ entry in $\langle b, b \rangle^{-1}$.
C  Proofs of Lemma 1 and Lemma 2

Proof of Lemma 1. The \( n = 0 \) case,
\[
E_\nu [\xi f] = E_\nu \left[ \frac{\partial f}{\partial \xi} \right],
\]
(125)
is immediately true by Stein’s lemma (Stein, 1981). For general case \( n \),
\[
E_\nu [H_{n+1} f] = E_\nu \left[ \left( \xi H_n - \frac{\partial H_n}{\partial \xi} \right) f \right] = E_\nu \left[ \frac{\partial}{\partial \xi} (H_n f) - \frac{\partial H_n}{\partial \xi} f \right] = E_\nu \left[ \frac{\partial H_n}{\partial \xi} f + H_n \frac{\partial f}{\partial \xi} - \frac{\partial H_n}{\partial \xi} f \right] = E_\nu \left[ H_n \frac{\partial f}{\partial \xi} \right],
\]
(126)
where we have used the recurrence relation,
\[
H_{n+1} = \xi H_n - \frac{\partial H_n}{\partial \xi},
\]
(127)
for the Hermite polynomials.

Proof of Lemma 2. Repeatedly applying Lemma 1,
\[
E_\nu [H_n f] = E_\nu [H_{n-1} \frac{\partial f}{\partial \xi}] = E_\nu [H_{n-2} \frac{\partial^2 f}{\partial \xi^2}] = \cdots = E_\nu [H_1 \frac{\partial^{n-1} f}{\partial \xi^{n-1}}] = E_\nu [H_0 \frac{\partial^n f}{\partial \xi^n}] = E_\nu \left[ \frac{\partial^n f}{\partial \xi^n} \right],
\]
(128)
yields the desired result.

D  Hermite Basis on \( \mathbb{R}^N \)

We can extend the results of §3.2 to create a Hermite basis for \( \mathcal{P} \) on \( \mathbb{R}^N \). Let
\[
\eta(\xi) = \frac{1}{\sqrt{n!}} \begin{bmatrix} H_0(\xi) \\ H_1(\xi) \\ \vdots \\ H_M(\xi) \end{bmatrix},
\]
(129)
Note that we have reintroduced \( H_0(\xi) \) because the basis will be created by all possible combinatorial \( N \)-products of these functions, one for each variable in \( \xi \in \mathbb{R}^N \). However, we will have to exclude the combination made up of only \( H_0 \) because once again this function gives the zero vector of \( \mathcal{P} \). We may express this operation as a Kronecker product, i.e.,
\[
\eta(\xi) = C (\eta(\xi_1) \otimes \eta(\xi_2) \otimes \cdots \otimes \eta(\xi_N)),
\]
(130)
where \( C = [0 \ 1] \) contains zero in the first column followed by the identity matrix; this removes the offending function. Observe that \( C C^T = 1 \). The basis is then
\[
h(\xi) = \xi \exp(-\eta(\xi)).
\]
(131)
The total number of basis functions is \( (M+1)^N - 1 \).

This set of basis functions retains its orthonormality because
\[
\langle h(\xi), h(\xi) \rangle = E_\nu \left[ (C \eta(\xi_1) \otimes \cdots \otimes \eta(\xi_N)) (C^T \eta(\xi_1) \otimes \cdots \otimes \eta(\xi_N)^T) \right] = C E_\nu \left[ \eta(\xi_1) \eta(\xi_1)^T \otimes \cdots \otimes \eta(\xi_N) \eta(\xi_N)^T \right] C^T
\]
(132)
by a property of the Kronecker product (Appendix A). Now
\[
E_\nu \left[ \eta(\xi_1) \eta(\xi_1)^T \otimes \cdots \otimes \eta(\xi_N) \eta(\xi_N)^T \right] = \int_{-\infty}^{\infty} \eta(\xi_1) \eta(\xi_1)^T \nu(\xi_1) d\xi_1 \times \cdots \times \int_{-\infty}^{\infty} \eta(\xi_N) \eta(\xi_N)^T \nu(\xi_N) d\xi_N = 1_{(M+1)^N \times (M+1)^N},
\]
(133)
wherein each of the integrals expresses the orthonormality of the Hermite functions and results in an \( (M+1) \times (M+1) \) identity matrix. Hence
\[
\langle h(\xi), h(\xi) \rangle = C 1_{(M+1)^N \times (M+1)^N} C^T = 1_{((M+1)^N - 1) \times ((M+1)^N - 1)}.
\]
(134)
To determine the coordinates of an arbitrary PDF \( p \in \mathcal{H} \), we shall require the multivariate version of Lemma 2:
Lemma 3. For the standard normal measure, $\nu \sim \mathcal{N}(0, 1)$,
\[
\mathbb{E}_\nu [H_{n_1}(\xi_1)H_{n_2}(\xi_2) \cdots H_{n_N}(\xi_N) f(\xi)] = \mathbb{E}_\nu \left[ \frac{\partial^{n_1+n_2+\cdots+n_N} f(\xi)}{\partial \xi_1^{n_1}\partial \xi_2^{n_2} \cdots \partial \xi_N^{n_N}} \right], \quad n_k = 1, 2 \cdots M.
\] (135)

where $f : \mathbb{R}^N \rightarrow \mathbb{R}$ is $n_k$-fold differentiable in $\xi_k$ and is such that the expectations exist.

The proof relies on the use of Lemma 2 for each individual partial derivative; for example, with respect to the variable $\xi_1$,
\[
\mathbb{E}_\nu [H_{n_1}(\xi_1)H_{n_2}(\xi_2) \cdots H_{n_N}(\xi_N) f(\xi)] = \mathbb{E}_\nu \left[ H_{n_1-1}(\xi_1)H_{n_2}(\xi_2) \cdots H_{n_N}(\xi_N) \frac{\partial f(\xi)}{\partial \xi_1} \right].
\] (136)

The product $H_{n_1}(\xi_2) \cdots H_{n_N}(\xi_N)$ has no dependence on $\xi_1$ and can therefore be treated as a constant. Doing the same for all the other variables and for the indicated number of times leads to the stated result.

We can streamline the notation by defining
\[
\partial \xi = \begin{bmatrix} \frac{\partial}{\partial \xi} & \frac{\partial^2}{\partial \xi^2} & \cdots & \frac{\partial^M}{\partial \xi^M} \end{bmatrix}
\] (137)

and, as above,
\[
\partial \xi = C(\partial \xi_1 \otimes \partial \xi_2 \otimes \cdots \otimes \partial \xi_N).
\] (138)

Using the measure $\nu = \mathcal{N}(0, 1)$, then,
\[
\alpha = \langle h(\xi), p(\xi) \rangle = \mathbb{E}_\nu [\partial \xi p(\xi)]
\] (139)

are the coordinates of $p(\xi) \in \mathcal{H}$.

E Coordinates of Multivariate Gaussian Projection

Let $p(x) = \downarrow \exp(-\phi(x)) \in \mathcal{P}$. Projecting onto $\mathcal{G}$, the coordinates associated with basis functions $g_1$ are
\[
\alpha_1 = \langle g_1, p \rangle = \mathbb{E}_\nu [\gamma_1(x) \phi(x)] - \mathbb{E}_\nu [\gamma_1(x)] \mathbb{E}_\nu [\phi(x)]
\]
\[
= \mathbb{E}_\nu \left[ \mathbf{L}^{-1}(x-\mu) \phi(x) \right] - \mathbb{E}_\nu \left[ \mathbf{L}^{-1}(x-\mu) \right] \mathbb{E}_\nu [\phi(x)]
\] (140)

\[
= \mathbf{L}^{-1} \mathbf{\Sigma} \mathbb{E}_\nu \left[ \frac{\partial \phi(x)}{\partial x^T} \right]
\]
\[
= \mathbf{L}^T \mathbb{E}_\nu \left[ \frac{\partial \phi(x)}{\partial x^T} \right],
\]

where we have employed Stein’s lemma (Stein, 1981) to go from the third line to the fourth. Taking the inner product of these coefficients with the associated basis functions we have
\[
\alpha_1^T \gamma_1(x) = \mathbb{E}_\nu \left[ \frac{\partial \phi(x)}{\partial x^T} \right]^T \mathbf{L}^{-1}(x-\mu) = \mathbb{E}_\nu \left[ \frac{\partial \phi(x)}{\partial x^T} \right]^T (x-\mu).
\] (141)

The coordinates associated with basis functions $g_2$ are
\[
\alpha_2 = \langle g_2, p \rangle
\]
\[
= \mathbb{E}_\nu [\gamma_2(x) \phi(x)] - \mathbb{E}_\nu [\gamma_2(x)] \mathbb{E}_\nu [\phi(x)]
\]
\[
= \mathbb{E}_\nu \left[ \sqrt{\frac{1}{2} D^T \text{Dvec} \left( \mathbf{L}^{-1}(x-\mu)(x-\mu)^T \mathbf{L}^{-T} \right) \phi(x)} \right]
\]
\[
- \mathbb{E}_\nu \left[ \sqrt{\frac{1}{2} D^T \text{Dvec} \left( \mathbf{L}^{-1}(x-\mu)(x-\mu)^T \mathbf{L}^{-T} \right) \phi(x)} \right] \mathbb{E}_\nu [\phi(x)]
\] (142)

\[
= \sqrt{\frac{1}{2} D^T \text{Dvec} \left( \mathbf{L}^{-1} \mathbf{\Sigma} \mathbb{E}_\nu \left[ \frac{\partial^2 \phi(x)}{\partial x^T \partial x} \right] \mathbf{\Sigma} \mathbf{L}^{-T} \right) L^{-T}}
\]
\[
= \sqrt{\frac{1}{2} D^T \text{Dvec} \left( \mathbf{L}^{-1} \mathbf{\Sigma} \mathbb{E}_\nu \left[ \frac{\partial^2 \phi(x)}{\partial x^T \partial x} \right] \mathbf{\Sigma} \mathbf{L}^{-T} \right) L^{-T}}
\]
where we have again used Stein’s lemma to go from the fourth line to the fifth, this time a double application. Taking
the inner product of these coefficients with the associated basis functions we have
\[
\alpha_2^T \gamma_2(x) = \frac{1}{2} \text{vec} \left( L^T \mathbb{E}_\nu \left[ \frac{\partial^2 \phi(x)}{\partial x^T \partial x} \right] L \right)^T D^T D \text{vec} \left( L^{-1}(x - \mu)(x - \mu)^T L^{-T} \right)
\]
\[
= \frac{1}{2} \text{vec} \left( L^T \mathbb{E}_\nu \left[ \frac{\partial^2 \phi(x)}{\partial x^T \partial x} \right] L \right)^T D^T D \text{vec} \left( L^{-1}(x - \mu)(x - \mu)^T L^{-T} \right)
\]
\[
= \frac{1}{2} \text{vec} \left( L^T \mathbb{E}_\nu \left[ \frac{\partial^2 \phi(x)}{\partial x^T \partial x} \right] L \right)^T D^T D \text{vec} \left( L^{-1}(x - \mu)(x - \mu)^T L^{-T} \right)
\]
\[
= \frac{1}{2} \text{vec} \left( L^T \mathbb{E}_\nu \left[ \frac{\partial^2 \phi(x)}{\partial x^T \partial x} \right] L \right)^T \text{vec} \left( L^{-1}(x - \mu)(x - \mu)^T L^{-T} \right)
\]
\[
= \frac{1}{2} \text{tr} \left( L^T \mathbb{E}_\nu \left[ \frac{\partial^2 \phi(x)}{\partial x^T \partial x} \right] L L^{-1}(x - \mu)(x - \mu)^T L^{-T} \right)
\]
\[
= \frac{1}{2} \text{tr} \left( (x - \mu)^T L^{-T} L^T \mathbb{E}_\nu \left[ \frac{\partial^2 \phi(x)}{\partial x^T \partial x} \right] L L^{-1}(x - \mu) \right)
\]
\[
= \frac{1}{2} (x - \mu)^T \mathbb{E}_\nu \left[ \frac{\partial^2 \phi(x)}{\partial x^T \partial x} \right] (x - \mu).
\]
Combining these we have
\[
\text{proj } p = \downarrow \exp \left( -(x - \mu)^T \mathbb{E}_\nu \left[ \frac{\partial \phi(x)}{\partial x} \right] - \frac{1}{2} (x - \mu)^T \mathbb{E}_\nu \left[ \frac{\partial^2 \phi(x)}{\partial x^T \partial x} \right] (x - \mu) \right)
\]
for the projected PDF in terms of its Gaussian basis.

### F Gaussian Information

We calculate here the information \( I \) contained in a multivariate Gaussian distribution, \( g(x) = \mathcal{N}(\mu', \Sigma') \in \mathcal{G}^+ \). We have
\[
g(x) = \downarrow \exp (-\phi(x)) \tag{145}
\]

with
\[
\phi(x) = \frac{1}{2} (x - \mu')^T \Sigma'^{-1} (x - \mu'). \tag{146}
\]
The measure is taken as \( \nu = \mathcal{N}(\mu, \Sigma) \).

Using our orthonormal basis for \( \mathcal{G} \), the information in \( g \) is
\[
I(g) = \frac{1}{2} \|g\|^2 = \frac{1}{2} \langle g, g \rangle = \frac{1}{2} (\alpha_1^T \alpha_1 + \alpha_2^T \alpha_2), \tag{147}
\]
where \( \alpha_1 \) and \( \alpha_2 \) are the coordinates. As
\[
\mathbb{E}_\nu \left[ \frac{\partial \phi(x)}{\partial x} \right] = \Sigma'^{-1} (\mu - \mu'), \quad \mathbb{E}_\nu \left[ \frac{\partial^2 \phi(x)}{\partial x^T \partial x} \right] = \Sigma'^{-1}, \tag{148}
\]
these coordinates are, by (50),
\[
\alpha_1 = L^T \Sigma'^{-1} (\mu - \mu'), \quad \alpha_2 = \sqrt{\frac{1}{2} D^T D} \text{vec} \left( L^T \Sigma'^{-1} L \right). \tag{149}
\]
Hence, from (147),
\[
I(g) = \frac{1}{2} \left( (\mu - \mu')^T \Sigma'^{-1} \Sigma \Sigma'^{-1} (\mu - \mu') + \frac{1}{2} \text{tr} \Sigma'^{-1} \Sigma \Sigma'^{-1} \Sigma \right), \tag{150}
\]
26
where the second term is a result of the fourth identity in (109) and the third in (115). It will, however, be instructive to rewrite the terms as

\[
(\mu - \mu')^T \Sigma^{-1} \Sigma' \Sigma^{-1} (\mu - \mu') = \mu^T \Sigma^{-1} \Sigma' \Sigma^{-1} \mu - 2 \mu^T \Sigma^{-1} \Sigma (\mu^T \otimes 1) \text{vec} \Sigma^{-1} + \left( \text{vec} \Sigma^{-1} \right)^T (\mu \otimes 1) \Sigma (\mu^T \otimes 1) \text{vec} \Sigma^{-1}
\]

\[
\text{tr} \Sigma^{-1} \Sigma' \Sigma^{-1} \Sigma = \left( \text{vec} \Sigma^{-1} \right)^T (\Sigma \otimes \Sigma) \text{vec} \Sigma^{-1},
\]

with the help of the third and fourth identities in (109). Now (150) can be neatly expressed as

\[
I(q) = \frac{1}{2} \left[ \Sigma^{-1} \mu' \right]^T \left[ \Sigma - (\mu^T \otimes 1) \frac{1}{2} (\Sigma \otimes \Sigma) + (\mu^T \otimes 1) \mu^T \otimes 1 \right] \left[ \text{vec} \Sigma^{-1} \mu' \right].
\]

This is the information contained in the Gaussian \( \mathcal{N}(\mu', \Sigma') \) although it is conditioned by the choice of measure \( \mathcal{N}(\mu, \Sigma) \) used to define the inner product. Note that as \( \Sigma^{-1} \) tends to zero, indicating a broadening of the distribution, the information also goes to zero; that is, we lose information of what we can expect in drawing from the distribution. The expression (153) can also be interpreted as simply writing the information using a different basis associated with the so-called natural parameters of a Gaussian (Barfoot, 2020).

### G Derivation of Equation (73): Derivative of the Measure in the Inner Product

We consider the inner product

\[
\langle p, q \rangle_\nu = \mathbb{E}_\nu[\ln p \ln q] - \mathbb{E}_\nu[\ln p] \mathbb{E}_\nu[\ln q]
\]

with

\[
\nu = \bigoplus_m \alpha_m \cdot b_m.
\]

We emphasize that here \( p \) and \( q \) are held fixed. The partial derivative with respect to \( \alpha_m \) is

\[
\frac{\partial}{\partial \alpha_m} \langle p, q \rangle_\nu = \frac{\partial}{\partial \alpha_m} \mathbb{E}_\nu[\ln p \ln q] - \left( \frac{\partial}{\partial \alpha_m} \mathbb{E}_\nu[\ln p] \right) \mathbb{E}_\nu[\ln q] - \mathbb{E}_\nu[\ln p] \left( \frac{\partial}{\partial \alpha_m} \mathbb{E}_\nu[\ln q] \right).
\]

In general,

\[
\frac{\partial}{\partial \alpha_m} \mathbb{E}_\nu[\ln r] = \int \frac{\partial \nu}{\partial \alpha_m} \ln r dx = \int \frac{\partial \ln \nu}{\partial \alpha_m} \ln r dx = \mathbb{E}_\nu \left[ \frac{\partial \ln \nu}{\partial \alpha_m} \ln r \right].
\]

(This quantity may in fact alternatively be written as \( \langle b_m, r \rangle_\nu \)). The last two derivatives in (156) are accounted for; as for the first, replacing \( \ln r \) with \( \ln p \ln q \) above, gives

\[
\frac{\partial}{\partial \alpha_m} \mathbb{E}_\nu[\ln p \ln q] = \mathbb{E}_\nu \left[ \frac{\partial \ln \nu}{\partial \alpha_m} \ln p \ln q \right].
\]

Thus

\[
\frac{\partial}{\partial \alpha_m} \langle p, q \rangle_\nu = \mathbb{E}_\nu \left[ \frac{\partial \ln \nu}{\partial \alpha_m} \ln p \ln q \right] - \mathbb{E}_\nu \left[ \frac{\partial \ln \nu}{\partial \alpha_m} \ln p \right] \mathbb{E}_\nu[\ln q] - \mathbb{E}_\nu[\ln p] \mathbb{E}_\nu \left[ \frac{\partial \ln \nu}{\partial \alpha_m} \ln q \right].
\]

Now we may rewrite this as

\[
\frac{\partial}{\partial \alpha_m} \langle p, q \rangle_\nu = \mathbb{E}_\nu \left[ \ln p \ln q \frac{\partial \ln \nu}{\partial \alpha_m} \right] - \mathbb{E}_\nu[\ln p] \mathbb{E}_\nu \left[ \ln q \frac{\partial \ln \nu}{\partial \alpha_m} \right] - \mathbb{E}_\nu \left[ \frac{\partial \ln \nu}{\partial \alpha_m} \ln p \right] \mathbb{E}_\nu[\ln q].
\]

We recognize that \( q^{\partial \ln \nu/\partial \alpha_m} \) in not a PDF; however, the self-normalizing feature of the inner product allows us to write

\[
\mathbb{E}_\nu \left[ \ln p \ln q^{\partial \ln \nu/\partial \alpha_m} \right] - \mathbb{E}_\nu[\ln p] \mathbb{E}_\nu \left[ \ln q^{\partial \ln \nu/\partial \alpha_m} \right] = \langle p, q^{\partial \ln \nu/\partial \alpha_m} \rangle_\nu = \langle p, \frac{\partial \ln \nu}{\partial \alpha_m} \cdot q \rangle_\nu.
\]

For the last term in (160), we use (56) yielding

\[
\mathbb{E}_\nu \left[ \frac{\partial \ln \nu}{\partial \alpha_m} \ln p \right] \mathbb{E}_\nu[\ln q] = (\mathbb{E}_\nu[\ln b_n \ln p] - \mathbb{E}_\nu[\ln b_n] \mathbb{E}_\nu[\ln p]) \mathbb{E}_\nu[\ln q] = \mathbb{E}_\nu[\ln q] \langle b_n, p \rangle_\nu.
\]

Finally then

\[
\frac{\partial}{\partial \alpha_m} \langle p, q \rangle_\nu = \langle p, \frac{\partial \ln \nu}{\partial \alpha_m} \cdot q \rangle_\nu - \mathbb{E}_\nu[\ln q] \langle b_n, p \rangle_\nu.
\]
As the inner product is symmetric in its arguments, this is also
\[
\frac{\partial}{\partial \alpha_n} \langle p, q \rangle_\nu = \frac{\partial}{\partial \alpha_n} \langle q, p \rangle_\nu = \left\langle q, \frac{\partial \ln \nu}{\partial \alpha_n} \cdot p \right\rangle_\nu - \mathbb{E}_\nu [\ln p] \langle b_n, q \rangle_\nu.
\] (164)

There is a caveat, however, in that we cannot transfer \(\partial \nu / \partial \alpha_n\) as the coefficient of \(p\) to that of \(q\); this is because the coefficient is a function of the domain variables of the PDFs. That transformation, though, may be expressed as
\[
\left\langle q, \frac{\partial \ln \nu}{\partial \alpha_n} \cdot p \right\rangle_\nu = \left\langle p, \frac{\partial \ln \nu}{\partial \alpha_n} \cdot q \right\rangle_\nu + \mathbb{E}_\nu [\ln p] \langle b_n, q \rangle_\nu - \mathbb{E}_\nu [\ln q] \langle b_n, p \rangle_\nu.
\] (165)

We have used the shorthand \(\langle p, q \rangle_{\partial \nu / \partial \alpha_n}\) to denote the derivative in (163) as in (73).