The Kullback–Leibler Divergence Between Lattice Gaussian Distributions

Article *in* Journal of the Indian Institute of Science · January 2022

DOI: 10.1007/s41745-021-00279-5

1 author:

Frank Nielsen
Sony Computer Science Laboratories, Inc.
574 PUBLICATIONS **8,553** CITATIONS

J. Indian Inst. Sci.
A Multidisciplinary Reviews Journal
ISSN: 0970-4140 Coden-JIISAD

REVIEW ARTICLE

The Kullback–Leibler Divergence Between Lattice Gaussian Distributions

Frank Nielsen[*]

Abstract | A lattice Gaussian distribution of given mean and covariance matrix is a discrete distribution supported on a lattice maximizing Shannon's entropy under these mean and covariance constraints. Lattice Gaussian distributions find applications in cryptography and in machine learning. The set of Gaussian distributions on a given lattice can be handled as a discrete exponential family whose partition function is related to the Riemann theta function. In this paper, we first report a formula for the Kullback–Leibler divergence between two lattice Gaussian distributions and then show how to efficiently approximate it numerically either via Rényi's α-divergences or via the projective γ-divergences. We illustrate how to use the Kullback-Leibler divergence to calculate the Chernoff information on the dually flat structure of the manifold of lattice Gaussian distributions.

Keywords: *Lattice Gaussian distribution, Discrete exponential family, Riemann theta function, Statistical divergence, Information geometry*

1 Introduction

It is well-known that the multivariate Gaussian distributions $N(\mu, \Sigma)$ are continuous distributions with support $\mathbb{R}^d$ which maximize Shannon's differential entropy under mean μ and covariance matrix Σ constraints[1]. Similarly, the d-variate lattice Gaussian distribution $N_\Lambda(\mu, \Sigma)$ can be defined as the distribution supported on a lattice[2] $\Lambda = \Lambda(L) = L\mathbb{Z}^d = \{Lz : z \in \mathbb{Z}^d\}$ which maximizes Shannon's entropy for the prescribed mean μ and covariance matrix Σ, where $\mathbb{Z}^d$ denotes the d-dimensional integer lattice and $L = [l_1 \mid \ldots \mid l_d]$ is a lattice basis of d column vectors l_i's arranged in the lattice basis matrix Λ. We consider full-rank lattices satisfying $\det(L) \neq 0$, and the lattice $\mathbb{Z}^d$ is called the d-dimensional integer lattice. Two lattices $\Lambda(L)$ and $\Lambda(L')$ are equal if and only if there exists a unimodular matrix U (i.e., a square matrix with integer entries and determinant ± 1) such that $L = L'U$. For example, the second lattice $L' = \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix}$ of Fig. 2 is equal to the integer lattice $\mathbb{Z}^2$ because $L' = U \times L$ with unimodular matrix $U = \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} (\det(U) = -1)$.

The set $\{N_\Lambda(\mu, \Sigma)\}$ of lattice Gaussian distributions form a discrete exponential family[1] (Chapter 2) since they are maximum entropy distributions. The probability mass function (pmf) $p_{\mu,\Sigma}$ is written canonically as

$$p_{\mu,\Sigma}(l) = \frac{1}{P_\Lambda(\mu, \Sigma)} \exp\left(\left\langle \zeta(\mu, \Sigma), s(l) \right\rangle\right), \quad (l \in \Lambda)$$

where $\zeta(\mu, \Sigma)$ denotes the natural parameter corresponding to the ordinary parameterization $\lambda = (\mu, \Sigma)$, $s(x) = (x, xx^\top)$ are the sufficient statistics, and $\langle \zeta, \zeta' \rangle$ is the following compound vector-matrix inner product between $\zeta = (\zeta_1, \zeta_2)$ and $\zeta' = (\zeta_1', \zeta_2')$ (with $\zeta_1, \zeta_1' \in \mathbb{R}^d$ and $\zeta_2, \zeta_2' \in \mathcal{P}_d$, the open cone of positive-definite matrices):

$$\langle \zeta, \zeta' \rangle := \zeta_1^\top \zeta_1' + \mathrm{tr}(\zeta_2^\top \zeta_2').$$

The term $P_\Lambda(\mu, \Sigma) = \sum_{l \in \Lambda} \exp\left(\left\langle \zeta(\mu, \Sigma), s(l) \right\rangle\right)$ in the denominator of the pmf is called the

[1] *Fundamental Research Laboratory, Sony Computer Science Laboratories Inc., 3-14-13 Higashi Gotanda, Tokyo 141-0022, Shingagawa-Ku, Japan. *Frank.Nielsen@acm.org*

partition function. The discrete natural exponential family $\mathcal{G}_\Lambda = \{N_\Lambda(\zeta)\}$ is a regular and minimal exponential family[3] of order $D = \frac{d(d+3)}{2}$: That is, the natural parameter space $\mathbb{R}^d \times \mathcal{P}_d$ is open and the $D+1$ functions $1, x_1, \ldots, x_i, \ldots, x_d, x_1^2, \ldots, x_i x_j, \ldots, x_d^2$ are linearly independent[A4]. When $\Lambda = \mathbb{Z}^d$, the lattice Gaussian distributions are called the discrete Gaussian distributions or the discrete normal distributions[5] with pmf written as

$$p_\zeta(z) = \frac{1}{P_{\mathbb{Z}^d}(\zeta)} \exp\left(\langle \zeta, s(z) \rangle\right), \quad (z \in \mathbb{Z}^d).$$

The canonical decomposition of exponential families is not unique: For example, we can choose $s_\alpha(x) = \alpha s(x)$ (for any non-zero scalar α) and adjust accordingly $\zeta_\alpha = \frac{1}{\alpha}\zeta$ so that the inner product remains invariant: $\langle s(x), \zeta \rangle = \langle s_\alpha(x), \zeta_\alpha \rangle$. To link the partition function of lattice Gaussian distributions to the Riemann Theta function[6], we consider the following sufficient statistic $t(x) = \left(2\pi x, -\pi x x^\top\right)$ with corresponding natural parameter ξ. Notice that the natural parameter ξ cannot be expressed easily as a function of (μ, Σ). Thus the lattice Gaussian distributions are mathematically different to handle than the continuous Gaussian distributions where conversions between ordinary to natural parameters are available in closed-form (see Appendix B of[7]). Using the $(\xi, t(x))$ parameterization for the discrete exponential family of lattice distributions $N_{\mathbb{Z}}^d(\xi)$, we get the following pmf decomposition:

$$p_\xi(z) = \frac{1}{Z_{\mathbb{Z}^d}(\xi)} \exp\left(2\pi\left(-\frac{1}{2} z^\top \xi_2 z + z^\top \xi_1\right)\right), (z \in \mathbb{Z}^d), \tag{1}$$

where the partition function is

$$Z_{\mathbb{Z}^d}(\xi) = \sum_{z \in \mathbb{Z}^d} \exp\left(2\pi\left(-\frac{1}{2} z^\top \xi_2 z + z^\top \xi_1\right)\right).$$

Although the partition function is not available as a bounded-size formula of ξ because it is an infinite summation over the elements of $\mathbb{Z}^d$, it can nevertheless be conveniently expressed as

$$Z_{\mathbb{Z}^d}(\xi) = \theta(-i\xi_1, i\xi_2), \tag{2}$$

where the complex-valued Riemann theta function[6] is the holomorphic function defined by its Fourier series as follows:

$$\theta : \mathbb{C}^d \times \mathcal{H}_d \to \mathbb{C}$$
$$\theta(\omega, \Omega) := \sum_{z \in \mathbb{Z}^d} \exp\left(2\pi i\left(\frac{1}{2} z^\top \Omega z + z^\top \omega\right)\right),$$

where $\mathcal{H}_d$ denotes the Siegel upper space of symmetric complex matrices with positive-definite imaginary parts[8]. A symmetric complex matrix with positive-definite imaginary part is called a Riemann matrix[9, 10]. Thus the Siegel upper space is the set of all Riemann matrices. Agostini and Améndola[5] consider complex-valued discrete Gaussian distributions (Definition 2.3 of[5]) by relaxing the parameter ξ to belong to $(\mathbb{C}^d \times \mathcal{H}_d) \backslash \Theta_d$, where $\Theta_d := \{(\omega, \Omega) \in \mathbb{C}^d \times \mathcal{H}_d : \theta(\omega, \Omega) = 0\}$ is the so-called universal theta divisor[11]. They report in Proposition 3 of[5] of the equivalence of complex-valued discrete Gaussian distributions, and deduce from that proposition that the natural parameter space is a quotient space for complex-valued discrete Gaussian distributions. In this paper, we consider real-valued lattice Gaussian distributions and Proposition 3 of[5] proves that the natural parameter space is $\Xi = \mathbb{R}^d \times \mathcal{P}_d$.

In practice, there exists efficient techniques to approximate the Riemann theta function[12] with many available software packages that implement them in various programming languages (e.g., package `Theta.jl` in Julia[13]). We refer to[14] (Chapter 22) for a basic introduction of approximation techniques of the Riemann theta functions, to[15] for approximations of the Jacobi's theta function for the univariate case $d = 1$, and to[12] (Theorem 2) for guaranteed ϵ-approximations using lattice points falling inside ellipsoids to approximate the infinite Riemann Theta sums.

Let $p_\xi(z) = \frac{\tilde{p}_\xi(z)}{Z_{\mathbb{Z}^d}(\xi)}$ where $\tilde{p}_\xi(z)$ denotes the unnormalized pmf:

$$\tilde{p}_\xi(z) = \exp\left(2\pi\left(-\frac{1}{2} z^\top \xi_2 z + z^\top \xi_1\right)\right). \tag{3}$$

Figure 1 displays the plots of two 1D discrete normal unnormalized pmfs and two 2D discrete unnormalized pmfs.

The univariate discrete normal distribution was historically first mentioned by Lisman and Van Zuylen[16] (1972), and later studied by Kemp[17] (1997). The relationship of its partition function $Z_{\mathbb{Z}}$ with the Jacobi theta function was pointed out by Szablowski[18] (2001), and later

[A] Definition: n univariate functions $f_1(x), \ldots, f_n(x)$ are said to be linearly dependent if there exists n constants $c_1, \ldots, c_n$, not all zero, such that $\sum_{i=1}^n c_i f_i(x) = 0$ for some x belonging to an interval $I \subset \mathbb{R}$. Otherwise, the functions are said linearly independent.

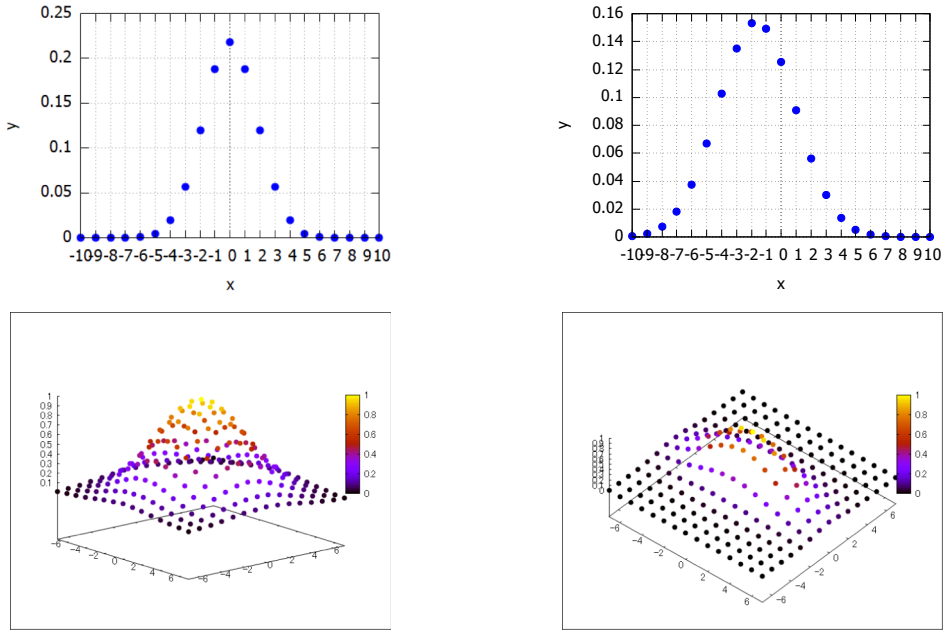

Figure 1: Plot of unnormalized discrete normal distributions: Top: p_ξ on the 1D integer lattice $\mathbb{Z}$ clipped at $[-10, 10]$ for $\xi = (0, 0.3)$ (left) and $\xi = (0.25, 0.15)$ (right). Notice that when $\xi_1 \in \mathbb{Z}$, the discrete normal is symmetric (left) but not for $\xi_1 \notin \mathbb{Z}$ (right). Bottom: $\tilde{p}_\xi$ on the 2D integer lattice $\mathbb{Z}^2$ clipped at $[-7, 7] \times [-7, 7]$: (Left) $\xi_1 = (0, 0)$ and $\xi_2 = \mathrm{diag}\left(\frac{1}{10}, \frac{1}{10}\right)$, (right) $\xi_1 = (0, 0)$ and $\xi_2 = \mathrm{diag}\left(\frac{1}{10}, \frac{1}{2}\right)$.

fully extended to Riemann multivariate theta function by Agostini and Améndola[5] (2019). Furthermore, Agostini and Améndola[5] extended the discrete Gaussian distributions to complex-valued pmfs by allowing the parameter $\xi_1 \in \mathbb{C}^d$ and the parameter ξ_2 to range in the Siegel upper space $\mathcal{H}_d$ instead of the positive-define matrix cone $\mathcal{P}_d$. Doing so allowed them to use the quasi-periodicity properties of the Riemann theta functions to deduce corresponding properties for the complex-valued Gaussian distributions. Namely, the Riemann theta function enjoys the following quasiperiodicity property:

- periodicity in ω with integer periods:
- $$\theta(\omega + u, \Omega) = \theta(\omega, \Omega),$$

  and
- quasi-periodicity:
- $$\theta(\omega + \Omega v, \Omega) = \exp\left(-2\pi i\left(\frac{1}{2} v^\top \Omega v + v^\top \omega\right)\right) \theta(\omega, \Omega),$$

  for any $u, v \in \mathbb{Z}^d$.

A statistical model $\mathcal{P} := \{p_\xi : \xi \in \Xi\}$ is said identifiable[4] when the mapping $\xi \mapsto p_\xi$ is one-to-one. When $\xi \in \mathbb{C}^d \times \mathcal{H}_d$, the discrete Gaussian model is not identifiable (i.e., there can exist two parameters ξ and ξ' such that $p_\xi(z) = p_{\xi'}(z)$). But the real-valued discrete Gaussian model is identifiable when $\xi \in \mathbb{R}^d \times \mathcal{P}_d$.

A key property of Gaussian distributions is that the family is invariant under the action of affine automorphisms of $\mathbb{R}^d$. Similarly, the family of discrete Gaussian distributions is invariant under the action of affine automorphisms of $\mathbb{Z}^d$ (Proposition 3.5[5]):

$$\forall \alpha \in \mathrm{GL}(d, \mathbb{Z}), \quad \alpha X_\xi = X_{\alpha^{-\top}\xi_1, \alpha^{-\top}\xi_2 \alpha^{-1}}.$$

The parity property of discrete Gaussians follows (Remark 3.7[5]):

$$X_{-\xi_1, \xi_2} \sim -X_\xi.$$

The discrete normal distributions play an important role as the counterpart of the normal distributions in robust implementations on finite-precision arithmetic computers of algorithms designed for cryptography[19] and

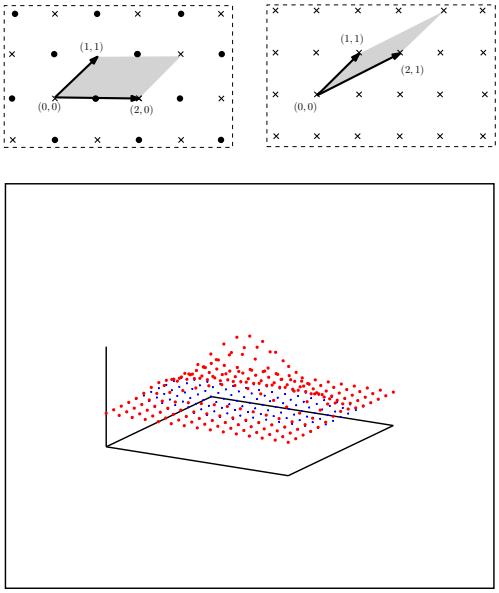

Figure 2: Top: Two examples of lattices with their basis defining fundamental parallelepipeds: the left lattice is a subset of $\mathbb{Z}^2$ while the second lattice coincides with $\mathbb{Z}^2$ although $L \neq I_2$, the 2×2 identity matrix. Bottom: Lattice Gaussian $N_\Lambda(\xi)$ with $\Lambda = L\mathbb{Z}^2$ obtained for $L = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, and $\xi_1 = (0,0)$ and $\xi_2 = \mathrm{diag}(0.1, 0.5)$. The lattice points are displayed in blue and the unnormalized pmf values at the lattice points are shown in red.

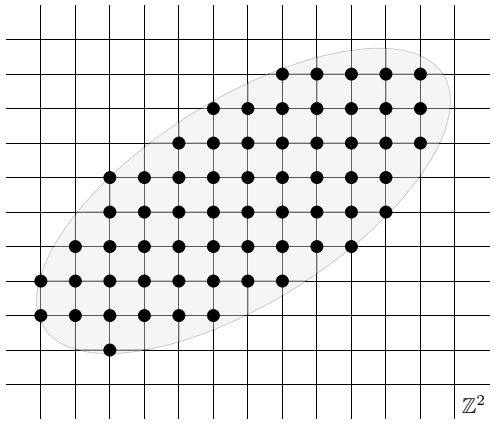

Figure 3: Approximating the function $\theta_{\mathbb{Z}^d}(\xi)$ by summing on the integer lattice points falling inside an ellipsoid E_ξ: $\theta(\xi) \simeq \tilde{\theta}(\xi; E_\xi \cap \mathbb{Z}^2)$.

differential privacy[20–22]. Recently, the discrete normal distributions have also been elicited as marginalized distributions in a particular type of Boltzmann machines with continuous visible states and discrete hidden states[23] in machine learning.

Let us extend the definition of discrete Gaussian distributions to arbitrary full-rank lattice Gaussian distributions:

Definition 1 (Lattice Gaussian distribution) A lattice Gaussian random variable $X \sim N_\Lambda(\xi)$ has the following pmf:

$$p_\xi(l) = \frac{1}{\theta_\Lambda(\xi)} \exp\left(2\pi\left(-\frac{1}{2}l^\top \xi_2 l + l^\top \xi_1\right)\right), \quad l \in \Lambda,$$

where the partition function

$$\theta_\Lambda(\xi) := \sum_{l \in \Lambda} \exp\left(2\pi\left(-\frac{1}{2}l^\top \xi_2 l + l^\top \xi_1\right)\right),$$

is related to the Riemann theta function by:

$$\theta_\Lambda(\xi) = \theta(-iL^\top \xi_1, iL^\top \xi_2 L), \tag{4}$$

with $i \in \mathbb{C}$ such that $i^2 = -1$.

Figure 2 displays some examples of lattices and plots a 2D lattice Gaussian distribution. In the remainder, we consider ξ to belong to the parameter space $\Xi = \mathbb{R}^d \times \mathcal{P}_d$. Hence, all pmfs $p_\xi(l)$ are real-valued. Notice that by definition, we have $\theta_\Lambda(\xi) = \sum_{l \in \Lambda} \tilde{p}_\xi(l)$.

In practice, we can approximate efficiently the Riemann theta function by replacing the infinite summation by a finite summation over a selected region R of integer lattice points[12,13]:

$$\tilde{\theta}(\xi; R) = \sum_{x \in R} \exp\left(2\pi\left(-\frac{1}{2}x^\top \xi_2 x + x^\top \xi_1\right)\right).$$

When $R = \mathbb{Z}_d$, we have $\tilde{\theta}(\xi; R) = \theta(\xi)$. For example, the method proposed in[12] to approximate the theta function consists in choosing the integer lattice points falling inside an ellipsoid E_ξ for calculating $\tilde{\theta}(\xi; R_\xi)$ with $R_\xi = E_\xi \cap \mathbb{Z}^d$. A theta ellipsoid with its integer lattice points R_ξ is illustrated in Fig. 3.

The paper is organized as follows: First, we consider the maximum likelihood estimator for the lattice Gaussian distributions in Sect. 2. This let us introduce the dual moment parameterization of exponential families, and describe methods to perform numerically parameter conversions when handling lattice Gaussian distributions. Then we report a formula for the cross-entropy and the Kullback-Leibler divergence (KLD) between two lattice Gaussian distributions in Sect. 3 (Proposition 1 and Proposition 2). In practice, we show how to approximate the KLD in Sect. 4 using asymptotic limits of either the Rényi α-divergences[24] (Proposition 4) or the γ-divergences[25] (Proposition 7). Finally, we

illustrate the use of the KLD for calculating the Chernoff information between two lattice Gaussian distributions in Sect. 5. Chernoff information characterizes the best error exponent in Bayesian hypothesis testing[26]. The Chernoff information can also be used in information fusion tasks (e.g., fusion of mixtures of lattice Gaussian distributions following the work described in Julier[27] for merging Gaussian mixture models).

2 Maximum Likelihood Estimator and Parameter Conversions

2.1 Parameterizations of the Lattice Gaussian Distributions

The log-normalizer $F_\Lambda(\xi) := \log \theta_\Lambda(\xi)$ is a real-analytic convex function[3] which is also called the cumulant function since the cumulant generating function $K_{X_\xi}(u)$ (cgf, see[1], page 30) of $X_\xi \sim p_\xi$ is

$$\begin{aligned} K_{X_\xi}(u) :&= \log\left(E[\exp(u^\top t(x))]\right) \\ &= F_\Lambda(\xi+u) - F_\Lambda(\xi). \end{aligned}$$

The geometric moments of the sufficient statistics can be retrieved from the partial derivatives of the cgf[28]. In particular, we have $\eta := E_{p_\xi}[t(x)] = \nabla F_\Lambda(\xi)$. There exists a bijection between the set of natural parameters ξ and the set of moment parameters $\eta(\xi) = E_{p_\xi}[t(x)] = \nabla F_\Lambda(\xi)$ for regular minimal exponential families. Moreover, the log-normalizer $F_\Lambda(\xi)$ is a Legendre-type function and hence its convex conjugate function $F^*_\Lambda(\eta)$ obtained from the Legendre-Fenchel transformation is also of Legendre-type[29]:

$$F^*_\Lambda(\eta) := \langle \xi, \eta\rangle - F_\Lambda(\xi),$$

with $\xi = \nabla F^*_\Lambda(\eta)$. Thus a pmf of a lattice Gaussian family can either be parameterized by the ordinary mean-covariance parameter $\lambda = (\mu, \Sigma)$, the natural parameter ξ, or its corresponding dual moment parameter η.

By definition of p_ξ as a maximum entropy distribution with mean μ and covariance matrix Σ, we have $E_{p_\xi}[x] = \mu$ and $\mathrm{Cov}_{p_\xi}[x] = E_{p_\xi}[(x-\mu)(x-\mu)^\top] = \Sigma$. Therefore we can express the moment parameter using the ordinary parameters as

$$\begin{aligned} \eta_1 &= E_{p_\xi}[2\pi x] = 2\pi\,\mu, \\ \eta_2 &= E_{p_\xi}[-\pi x x^\top] = -\pi\,(\Sigma + \mu\mu^\top). \end{aligned}$$

Notice that we can also use other alternative parameterizations like $\tau = (a, B)$ for $a \in \mathbb{R}^d$, $B \in \mathcal{P}_d$ with the following pmf[30]:

$$p_\tau(l) \propto \exp\left(-\frac{1}{2}(x-a)^\top B^{-1}(x-a)\right).$$

In that case, observe that $E_{p_\tau}[x] \neq \mu$ and $\mathrm{Cov}_{p_\tau}[x] \neq \Sigma$ for $\tau = (\mu, \Sigma)$.

2.2 Maximum Likelihood Estimator

Let $v_1, \ldots, v_m$ be a set of m identically and independently distributed variables sampled from p_ξ. The estimating equation for the maximum likelihood estimator (mle, see[1], page 135) is

$$\hat{\eta} = \frac{1}{m}\sum_{i=1}^m t(v_i).$$

Thus we get the following estimating equation when considering the lattice Gaussian family:

$$\begin{aligned} \hat{\eta}_1 &= \frac{2\pi}{m}\sum_{i=1}^n x_i = 2\pi\,\hat{\mu}, \\ \hat{\eta}_2 &= -\frac{\pi}{m}\sum_{i=1}^n x_i x_i^\top = -\pi\,(\hat{\Sigma} + \hat{\mu}\hat{\mu}^\top). \end{aligned}$$

To get the corresponding natural parameter of $\hat{\eta}$, we shall use the equivariance property of the mle:

$$\hat{\xi} = \nabla F^*_\Lambda(\hat{\eta}).$$

Since $F^*_\Lambda(\eta)$ is not available in closed-form, we need to numerically approximate it as $\tilde{\xi} \approx F^*_\Lambda(\eta)$ by solving the following gradient system:

$$\eta = \nabla F_\Lambda(\xi).$$

2.3 Converting Moment Parameters to Natural Parameters and Vice Versa

Given the moment parameter η, we may approximate $\xi = \nabla F^*(\eta)$ by solving the gradient system $\eta = \nabla F(\xi)$. For example, one way to solve the gradient system is by using the technique described in[31] (with corresponding Matlab® code publicly available in[32]) that we summarize as follows:

First, let us choose the following canonical parameterization of the densities of an exponential family of order D with cumulant function $F(\psi)$ and sufficient statistics $t_i(x)$'s:

$$p_\psi(x) := \exp\left(-\sum_{i=0}^D \psi_i t_i(x)\right).$$

That is, $\psi_0 = F(\psi)$ and $\psi_i = -\xi_i$ for $i \in \{1, \ldots, D\}$. The parameter ψ is an augmented

natural parameter which includes the log-normalizer in its first coefficient ψ_0.

Let $K_i(\psi) := E_{p_\theta}[t_i(x)] = \eta_i$ denote the set of $D+1$ non-linear equations for $i \in \{0,\ldots,D\}$ (with $t_0(x)=1$ and $\eta_0=1$). The method of[31] converts iteratively p_η to p_ψ using an approximate Newton's method for solving the system of equations $K_i(\psi)=\eta_i$. We initialize $\psi_i^{(0)}$ for $i \in \{1,\ldots,D\}$ and calculate numerically $\psi_0^{(0)} = F(\psi^{(0)})$.

At iteration t with current estimate $\psi^{(t)}$, we use the following first-order Taylor approximation:

$$K_i(\psi) \approx K_i(\psi^{(t)}) + (\psi - \psi^{(t)})^\top \nabla K_i(\psi^{(t)}).$$

Let $H(\psi)$ denote the $(D+1)\times(D+1)$ matrix:

$$H(\psi) := \left[\frac{\partial}{\partial \psi_j} K_i(\psi)\right]_{ij}.$$

We have

$$H_{ij}(\psi) = H_{ji}(\psi) = -E_{p_\psi}[t_i(x)t_j(x)]. \quad (5)$$

We update as follows:

$$\psi^{(t+1)} = \psi^{(t)} + H^{-1}(\psi^{(t)}) \begin{bmatrix} \eta_0 - K_0(\psi^{(t)}) \\ \vdots \\ \eta_D - K_D(\psi^{(t)}) \end{bmatrix}. \quad (6)$$

We implemented this method for d-variate discrete normal distributions with $D = \frac{d(d+3)}{2}$ and $t_1(x)=x_1,\ldots,t_d(x)=x_d, t_{d+1}(x) = -\frac{1}{2}x_1x_1$, $t_{d+2}(x) = -\frac{1}{2}x_1x_2,\ldots,t_D(x) = -\frac{1}{2}x_dx_d$. We approximated $H_{ij} = -E_{p_\psi}[t_i(x)t_j(x)]$ of Eq. 5 using the theta ellipsoid integer lattice points.

By definition, the d-variate standard discrete normal distribution $N_{\mathbb{Z}^d}(0,I)$ has zero mean and identity covariance matrix: Its corresponding natural parameters ξ_{std} can be approximated numerically as $\xi_{\mathrm{std}} \simeq (0, 0.1591549 \times I)$[5], where I denotes the $d\times d$ identity matrix.

Reciprocally, given a natural parameter ξ, we may estimate its corresponding dual moment parameter $\eta = \nabla F(\xi) = E_{p_\xi}[t(x)]$. We can either use derivatives of $F(\xi)$ (e.g., the derivatives of the Riemann theta function[13]), or estimate the parameter as $\tilde{\eta} = \frac{1}{m}\sum_{i=1}^m t(v_i)$ where $v_1,\ldots,v_m$ are m i.i.d variates sampled from p_ξ. Sampling from multivariate discrete normal distributions is studied in[23,33].

3 A Formula for the Kullback–Leibler Divergence Between Lattice Gaussian Distributions

The Kullback–Leibler divergence (KLD, Eq. 2.26 of[26], page 19) between two pmfs $r(x)$ and $s(x)$ with support $\mathcal{X}$ is defined by

$$D_{\mathrm{KL}}[r:s] := \sum_{x\in\mathcal{X}} r(x)\log\frac{r(x)}{s(x)}.$$

The KLD is also called the relative entropy because it can be expressed as the difference between the cross-entropy and the entropy:

$$D_{\mathrm{KL}}[r:s] = H[r:s] - H[r],$$

where $H[r:s]$ denotes the cross-entropy

$$H[r:s] := -\sum_{x\in\mathcal{X}} r(x)\log s(x),$$

and $H[r]=H[r:r]$ is Shannon's entropy.

It was shown in[34] that the cross-entropy between two densities p_ξ and $p_{\xi'}$ of an exponential family can be expressed as

$$H[p_\xi : p_{\xi'}] = F_\Lambda(\xi') - \langle \xi', \eta\rangle.$$

In particular, when $\xi'=\xi$, we get

$$H[p_\xi : p_\xi] = H[p_\xi] = F_\Lambda(\xi) - \langle \xi,\eta\rangle = -F_\Lambda^*(\eta).$$

Hence, we get the following propositions:

Proposition 1 *The cross-entropy between two discrete normal distributions $p_\xi \sim N_\Lambda(\mu,\Sigma)$ and $p_{\xi'} \sim N_\Lambda(\mu',\Sigma')$ is*

$$H[p_\xi : p_{\xi'}] = \log\theta_\Lambda(\xi') - 2\pi\mu^\top\xi_1' + \pi\,\mathrm{tr}\left(\xi_2'(\Sigma+\mu\mu^\top)\right). \quad (7)$$

Proposition 1 generalizes Proposition 4.4 of[5].

Proposition 2 *The Kullback–Leibler divergence between two lattice Gaussian distributions $p_\xi \sim N_\Lambda(\mu,\Sigma)$ and $p'_\xi \sim N_\Lambda(\mu',\Sigma')$ is:*

$$D_{\mathrm{KL}}[p_\xi : p_{\xi'}] = \log\left(\frac{\theta_\Lambda(\xi')}{\theta_\Lambda(\xi)}\right) - 2\pi\mu^\top(\xi_1'-\xi_1) + \pi\,\mathrm{tr}\left((\xi_2'-\xi_2)(\Sigma+\mu\mu^\top)\right). \quad (8)$$

Notice that we use the mixed (λ,ξ)-parameterizations in the above formula. We now consider two fast approximation techniques that bypass the need of the λ-parameterization.

4 Approximating and Estimating the Kullback–Leibler Divergence

4.1 Approximating the Kullback–Leibler Divergence via Rényi α-Divergences

The Rényi α-divergence[24] between pmf $r(x)$ and pmf $s(x)$ on support $\mathcal{X}$ is defined for any positive real $\alpha \neq 1$ by

$$\begin{aligned} D_\alpha[r:s] :=& \frac{1}{\alpha-1}\log\left(\sum_{x\in\mathcal{X}} r(x)^\alpha s(x)^{1-\alpha}\right), \\ =& \frac{1}{\alpha-1}\log\left(E_s\left[\left(\frac{r(s)}{s(x)}\right)^\alpha\right]\right), \\ & (\alpha>0, \alpha\neq 1). \end{aligned}$$

When both pmfs are from the same discrete exponential family with log-normalizer $F(\xi)$, the Rényi α-divergence amounts to a α-skewed Jensen divergence[35, 36] between the corresponding natural parameters:

$$D_\alpha[p_\xi : p_{\xi'}] = \frac{1}{1-\alpha} J_{F,\alpha}(\xi:\xi'),$$

where

$$\begin{aligned} J_{F,\alpha}(\xi:\xi') :=& \alpha F(\xi) + (1-\alpha)F(\xi') \\ & - F(\alpha\xi + (1-\alpha)\xi'). \end{aligned}$$

Indeed, let

$$I_{\alpha,\beta}[r:s] = \sum_{x\in\mathcal{X}} r(x)^\alpha s(x)^\beta, \quad \alpha,\beta\in\mathbb{R}.$$

Then we have the following lemma:

Proposition 3 *For two pmfs p_ξ and $p_{\xi'}$ of a discrete exponential family with log-normalizer $F(\xi)$ with $\alpha\xi+\beta\xi'\in\Xi$, we have*

$$\begin{aligned} I_{\alpha,\beta}[p_\xi : p_{\xi'}] =& \exp\big(F(\alpha\xi+\beta\xi') \\ & -(\alpha F(\xi)+\beta F(\xi'))\big). \end{aligned}$$

Proof We have

$$\begin{aligned} I_{\alpha,\beta}[p_\xi : p_{\xi'}] =& \sum_{x\in\mathcal{X}} \exp(\langle t(x), \alpha\xi\rangle - \alpha F(\xi)) \\ & \exp(\langle t(x), \beta\xi'\rangle - \beta F(\xi')), \\ =& e^{F(\alpha\xi+\beta\xi')-(\alpha F(\xi)+\beta F(\xi'))} \\ & \underbrace{\sum_{x\in\mathcal{X}} e^{\langle t(x),\alpha\xi+\beta\xi'\rangle - F(\alpha\xi+\beta\xi')}}_{=1}, \end{aligned}$$

since $\sum_{x\in\mathcal{X}} p_{\alpha\xi+\beta\xi'}(x) = 1$ when $\alpha\xi+\beta\xi'\in\Xi$. □

Thus we get the following proposition:

Proposition 4 *The Rényi α-divergence between two Gaussian lattice distributions p_ξ and $p_{\xi'}$ for $\alpha>0$ and $\alpha\neq 1$ is*

$$\begin{aligned} D_\alpha[p_\xi : p_{\xi'}] =& \frac{1}{1-\alpha}\left(\alpha\log\frac{\theta_\Lambda(\xi)}{\theta_\Lambda(\alpha\xi+(1-\alpha)\xi')}\right. \\ & \left.+(1-\alpha)\log\frac{\theta_\Lambda(\xi')}{\theta_\Lambda(\alpha\xi+(1-\alpha)\xi')}\right). \end{aligned} \tag{9}$$

Proof We have

$$\begin{aligned} D_\alpha[p_\xi : p_{\xi'}] =& \frac{1}{1-\alpha}\big(\alpha\log\theta_\Lambda(\xi) + (1-\alpha) \\ & \log\theta(\xi') - \log\theta_\Lambda(\alpha\xi+(1-\alpha)\xi')\big). \end{aligned}$$

Plugging $\log\theta_\Lambda(\alpha\xi+(1-\alpha)\xi') = (\alpha+1-\alpha)\log\theta_\Lambda(\alpha\xi+(1-\alpha)\xi')$ in the right-hand-side equation yields the result. Notice that we can also express equivalently the Rényi divergences as

$$D_\alpha[p_\xi : p_{\xi'}] = \frac{1}{1-\alpha}\log\frac{\theta_\Lambda(\xi)^\alpha\theta_\Lambda(\xi')^{1-\alpha}}{\theta_\Lambda(\alpha\xi+(1-\alpha)\xi')}.$$

□

When $\alpha=\frac{1}{2}$, Rényi α-divergence amounts to twice the symmetric Bhattacharyya divergence[37]: $D_{\frac{1}{2}}[r:s] = 2\,D_{\text{Bhattacharyya}}[r,s]$ with:

$$D_{\text{Bhattacharyya}}[r,s] := -\log\left(\sum_{x\in\mathcal{X}}\sqrt{r(x)s(x)}\right).$$

The Bhattacharyya divergence is the negative logarithm of the Bhattacharyya coefficient:

$$\rho_{\text{Bhattacharyya}}[r,s] := \sum_{x\in\mathcal{X}}\sqrt{r(x)s(x)},$$

and the squared Hellinger divergence is related to the Bhattacharyya coefficient as follows:

$$\begin{aligned} D^2_{\text{Hellinger}}[r,s] =& \frac{1}{2}\sum_{x\in\mathcal{X}}(\sqrt{r(x)}-\sqrt{s(x)})^2 \\ =& 1-\rho_{\text{Bhattacharyya}}[r,s]. \end{aligned}$$

Thus we get the following proposition:

Proposition 5 *The squared Hellinger distance between two lattice Gaussian distributions p_ξ and $p_{\xi'}$ is*

$$D^2_{\text{Hellinger}}[p_\xi, p_{\xi'}] = 1 - \frac{\theta_\Lambda\left(\frac{\xi+\xi'}{2}\right)}{\sqrt{\theta_\Lambda(\xi)\theta_\Lambda(\xi')}}.$$

Now, the Rényi α-divergences tend asymptotically to the KLD when $\alpha\to 1$. Hence in practice, we can approximate the KLD between

two lattice Gaussians distributions by the Rényi α-divergence for $\alpha = 1 - \epsilon$ for sufficient small value of $\epsilon \neq 0$.

Proposition 6 *The Kullback–Leibler divergence between two lattice Gaussian distributions p_ξ and $p_{\xi'}$ can be approximated by the Rényi α-divergence for $\alpha = 1 - \epsilon$ and $\epsilon \neq 0$ close to 0:*

$$D_{\mathrm{KL}}[p_\xi : p_{\xi'}] \simeq D_{1-\epsilon}[p_\xi : p_{\xi'}] = \frac{1}{\epsilon} J_{F_\Lambda,1-\epsilon}(\xi : \xi') = \frac{1}{\epsilon} \log \frac{\theta_\Lambda(\xi)^{1-\epsilon}\theta_\Lambda(\xi')^\epsilon}{\theta_\Lambda((1-\epsilon)\xi + \epsilon\xi')}$$

Let $\Delta_\epsilon(\xi : \xi') := \left|D_{\mathrm{KL}}[p_\xi : p_{\xi'}] - D_{1-\epsilon}[p_\xi : p_{\xi'}]\right|$ denote the absolute value of the error of the approximation. Since $D_{\mathrm{KL}}[p_\xi : p_{\xi'}] = B_{F_\Lambda}(\xi' : \xi)$ and $D_{1-\epsilon}[p_\xi : p_{\xi'}] = \frac{1}{\epsilon} J_{F_\Lambda,1-\epsilon}(\xi : \xi')$, we have:

$$\begin{aligned}\Delta_\epsilon(\xi : \xi') &= \left|\log \frac{\theta_\Lambda(\xi')}{\theta_\Lambda(\xi)} - \frac{1}{\theta_\Lambda(\xi)}\left\langle \xi' - \xi, \nabla\theta_\Lambda(\xi)\right\rangle \right.\\ &\quad \left. -\frac{1}{\epsilon}\log \frac{\theta_\Lambda(\xi)^{1-\epsilon}\theta_\Lambda(\xi')^\epsilon}{\theta_\Lambda((1-\epsilon)\xi + \epsilon\xi')}\right|,\\ &= \left|\frac{1}{\epsilon}\log \frac{\theta_\Lambda((1-\epsilon)\xi + \epsilon\xi')}{\theta_\Lambda(\xi)}\right.\\ &\quad \left. -\left\langle \xi' - \xi, \nabla\theta_\Lambda(\xi)\right\rangle\right|.\end{aligned}$$

In practice, the error $\Delta_\epsilon(\xi : \xi')$ is also related to the numerical errors incurred when calculating approximations of Theta functions[13].

Since Rényi α-divergences are non-decreasing with α[24], we obtained both lower and upper bounds of the KLD.

4.2 Approximating the Kullback–Leibler Divergence via Projective γ-Divergences

The γ-divergences[25, 38] between two pmfs $p(x)$ and $q(x)$ defined over the support $\mathcal{X}$ for a real $\gamma > 1$ is defined by:

$$\vec{D}_\gamma[p : q] := \frac{1}{\gamma(\gamma-1)} \log\left(\frac{\left(\sum_{x\in\mathcal{X}} p^\gamma(x)\right)\left(\sum_{x\in\mathcal{X}} q^\gamma(x)\right)^{\gamma-1}}{\left(\sum_{x\in\mathcal{X}} p(x)q^{\gamma-1}(x)\right)^\gamma}\right),$$

$(\gamma > 1)$.

The γ-divergences are projective divergences, i.e., they satisfy the following identity:

$$\vec{D}_\gamma[p : p'] = \vec{D}_\gamma[\lambda p : \lambda' p'], \quad (\forall \lambda, \lambda' > 0).$$

We use the vector notation $\vec{D}$ to indicate that this divergence is projective. Thus let us rewrite $p(x) = \frac{\tilde{p}(x)}{Z_p}$ and $q(x) = \frac{\tilde{q}(x)}{Z_q}$ where $\tilde{p}(x)$ and $\tilde{q}(x)$ are unnormalized pmfs, and Z_p and Z_q their respective normalizers. Then we have

$$\vec{D}_\gamma[p : p'] = \vec{D}_\gamma[\tilde{p} : \tilde{p}'].$$

Let us define

$$I_\gamma[p : q] := \sum_{x\in\mathcal{X}} p(x)q(x)^{\gamma-1}.$$

Then the γ-divergence can be written as:

$$\begin{aligned}\vec{D}_\gamma[p : q] &= \vec{D}_\gamma[\tilde{p} : \tilde{q}]\\ &= \frac{1}{\gamma(\gamma-1)} \log\left(\frac{I_\gamma[\tilde{p} : \tilde{p}]\, I_\gamma[\tilde{q} : \tilde{q}]^{\gamma-1}}{I_\gamma[\tilde{p} : \tilde{q}]^\gamma}\right).\end{aligned}$$

Consider $p = p_\xi$ and $q = p_{\xi'}$ two pmfs belonging to the lattice Gaussian exponential family, and let

$$\tilde{I}_\gamma\left(\xi : \xi'\right) = I_\gamma\left[\tilde{p}_\xi : \tilde{p}_{\xi'}\right].$$

Provided that $\xi + (\gamma - 1)\xi' \in \Xi$, we have following the proof of Proposition 3 that

$$\begin{aligned}\tilde{I}_\gamma\left(\xi : \xi'\right) &= \sum_{l\in\Lambda} \tilde{p}_\xi(l)\tilde{p}_{\xi'}(l)^{\gamma-1},\\ &= \sum_{l\in\Lambda} \exp(\left\langle \xi + (\gamma-1)\xi', t(x)\right\rangle),\\ &= \exp(F_\Lambda(\xi + (\gamma-1)\xi')) \underbrace{\sum_{l\in\Lambda} p_{\xi+(\gamma-1)\xi'}(l)}_{=1},\\ &= \exp(F_\Lambda(\xi + (\gamma-1)\xi')),\end{aligned}$$

where $F_\Lambda(\xi) = \log\theta_\Lambda(\xi)$ denotes the cumulant function of the Gaussian distributions on lattice Λ. That is, we have

$$\tilde{I}_\gamma\left(\xi : \xi'\right) = \theta_\Lambda(\xi + (\gamma-1)\xi'),$$

and therefore, we can express the γ-divergences as

$$\vec{D}_\gamma[p_\xi : p_{\xi'}] = \frac{1}{\gamma(\gamma-1)} \log\left(\frac{\theta_\Lambda(\gamma\xi)\,\theta_\Lambda(\gamma\xi')^{\gamma-1}}{\theta_\Lambda(\xi + (\gamma-1)\xi')^\gamma}\right). \tag{10}$$

Notice that the exact values of the infinite summations $\tilde{I}_\gamma\left(\xi : \xi'\right)$ depend on the Riemann theta function.

Now, the γ-divergences tend asymptotically to the Kullback–Leibler divergence between normalized densities when $\gamma \to 1$[25, 38]: $\lim_{\gamma\to 1} \vec{D}_\gamma[\tilde{p} : \tilde{q}] = D_{\mathrm{KL}}\left[\frac{\tilde{p}}{Z_p} : \frac{\tilde{q}}{Z_q}\right]$. Let us notice that the KLD is not a projective divergence, and that for small enough $\gamma > 1$, we have $\xi + (\gamma-1)\xi'$ always falling inside the natural parameter space Ξ. Moreover, we can approximate the infinite summation using a finite region of integer lattice points $R_{\xi,\xi'}$:

Table 1: Summary of statistical divergences with corresponding formula for lattice Gaussian distributions with partition function $\theta_\Lambda(\xi)$. Ordinary parameterization $\lambda(\xi) = (\mu = E_{p_\xi}[X], \Sigma = \mathrm{Cov}_{p_\xi}[X])$ for $X \sim N_\Lambda(\xi)$.

Divergence	Definition
	Closed-form formula for lattice Gaussians
Kullback–Leibler divergence	$D_{\mathrm{KL}}[p_\xi : p_{\xi'}] = \sum_{l\in\Lambda} p_\xi(l) \log \frac{p_\xi(l)}{p_{\xi'}(l)}$
	$D_{\mathrm{KL}}[p_\xi : p_{\xi'}] = \log\left(\frac{\theta_\Lambda(\xi')}{\theta_\Lambda(\xi)}\right)$
	$-2\pi\mu^\top(\xi_1' - \xi_1) + \pi\,\mathrm{tr}\big((\xi_2' - \xi_2)(\Sigma + \mu\mu^\top)\big)$
squared Hellinger divergence	$D^2_{\mathrm{Hellinger}}[p_\xi : p_{\xi'}] = \frac{1}{2}\sum_{l\in\Lambda}(\sqrt{p_\xi(l)} - \sqrt{p_{\xi'}(l)})^2$
	$D^2_{\mathrm{Hellinger}}[p_\xi : p_{\xi'}] = 1 - \frac{\theta_\Lambda\left(\frac{\xi+\xi'}{2}\right)}{\sqrt{\theta_\Lambda(\xi)\theta_\Lambda(\xi')}}$
Rényi α-divergence	$D_\alpha[p_\xi : p_{\xi'}] = \frac{1}{\alpha-1}\log\left(\sum_{l\in\Lambda} p_\xi(l)^\alpha p_{\xi'}(l)^{1-\alpha}\right)$
$(\alpha > 0, \alpha \neq 1)$	$D_\alpha[p_\xi : p_{\xi'}] = \frac{\alpha}{1-\alpha}\log\frac{\theta_\Lambda(\xi)}{\theta_\Lambda(\alpha\xi+(1-\alpha)\xi')} + \log\frac{\theta_\Lambda(\xi')}{\theta_\Lambda(\alpha\xi+(1-\alpha)\xi')}$
	$\lim_{\alpha\to 1} D_\alpha[p_\xi : p_{\xi'}] = D_{\mathrm{KL}}[p_\xi : p_{\xi'}]$
γ-divergence	$\vec{D}_\gamma[p_\xi : p_{\xi'}] = \frac{1}{\gamma(\gamma-1)}\log\left(\frac{\left(\sum_{l\in\Lambda} p_\xi^\gamma(x)\right)\left(\sum_{l\in\Lambda} p_{\xi'}^\gamma(l)\right)^{\gamma-1}}{\left(\sum_{l\in\Lambda} p_\xi(l) p_{\xi'}^{\gamma-1}(l)\right)^\gamma}\right)$
$(\gamma > 1)$	$\vec{D}_\gamma[p_\xi : p_{\xi'}] = \frac{1}{\gamma(\gamma-1)}\log\left(\frac{\theta_\Lambda(\gamma\xi)\theta_\Lambda(\gamma\xi')^{\gamma-1}}{\theta_\Lambda(\xi+(\gamma-1)\xi')^\gamma}\right)$
	$\lim_{\gamma\to 1}\vec{D}_\gamma[p_\xi : p_{\xi'}] = D_{\mathrm{KL}}[p_\xi : p_{\xi'}]$
Hölder divergence	$\vec{D}^H_{\alpha,\gamma}[r : s] := \left\lvert\log\left(\frac{\sum_{x\in\mathcal{X}} r(x)^{\gamma/\alpha} s(x)^{\gamma/\beta}}{\left(\sum_{x\in\mathcal{X}} r(x)^\gamma\right)^{1/\alpha}\left(\sum_{x\in\mathcal{X}} s(x)^\gamma\right)^{1/\beta}}\right)\right\rvert$
$(\gamma > 0, \frac{1}{\alpha} + \frac{1}{\beta} = 1)$	$\vec{D}^H_{\alpha,\gamma}[p_\xi : p_{\xi'}] = \left\lvert\log\frac{\theta_\Lambda(\gamma\xi)^{\frac{1}{\alpha}}\theta_\Lambda(\gamma\xi')^{\frac{1}{\beta}}}{\theta_\Lambda(\frac{\gamma}{\alpha}\xi + \frac{\gamma}{\beta}\xi')}\right\rvert$
Cauchy–Schwarz divergence	$\vec{D}_{\mathrm{CS}}[r : s] := -\log\frac{\sum_{x\in\mathcal{X}} r(x)s(x)}{\sqrt{\left(\sum_{x\in\mathcal{X}} r^2(x)\right)\left(\sum_{x\in\mathcal{X}} s^2(x)\right)}}$
(Hölder with $\alpha = \beta = \gamma = 2$)	$\vec{D}_{\mathrm{CS}}[p_\xi : p_{\xi'}] = \log\frac{\sqrt{\theta_\Lambda(2\xi)\theta_\Lambda(2\xi')}}{\theta_\Lambda(\xi+\xi')}$

$$\tilde{I}_{\gamma,R_{\xi,\xi'}}(\xi : \xi') := \sum_{x\in R_{\xi,\xi'}} \tilde{p}_\xi\, \tilde{p}_{\xi'}(x)^\gamma.$$

For example, we can use the theta ellipsoids[12] E_ξ and $E_{\xi'}$ used to approximate $\theta(\xi)$ and $\theta(\xi')$, respectively (Fig. 3): We choose $R_{\xi,\xi'} = (E_\xi \cup E_{\xi'}) \cap \mathbb{Z}^d$. In practice, this approximation of the I_γ summations scales well in high dimensions. Overall, we get our approximation of the KLD between two lattice Gaussian distributions summarized in the following proposition:

Proposition 7 *The Kullback–Leibler divergence between two lattice Gaussian distributions p_ξ and $p_{\xi'}$ can be approximated:*

$$D_{\mathrm{KL}}[p_\xi : p_{\xi'}] \approx \vec{D}_\gamma[p_\xi : p_{\xi'}] = \frac{1}{\gamma(\gamma-1)}\log\left(\frac{\tilde{I}_{\gamma,R_\xi}(\xi : \xi)\,\tilde{I}_{\gamma,R_{\xi'}}(\xi' : \xi')^{\gamma-1}}{\tilde{I}_{\gamma,R_{\xi,\xi'}}(\xi : \xi')^\gamma}\right), \tag{11}$$

for $\gamma > 1$ close to 1 (say, $\gamma = 1 + 10^{-5}$), where R_ξ and $R_{\xi'}$ denote the integer lattice points falling inside the theta ellipsoids E_ξ and $E_{\xi'}$ used to approximate the theta functions[12] $\theta_\Lambda(\xi)$ and $\theta_\Lambda(\xi')$, respectively.

Table 1 summarizes the various closed-formula obtained for the statistical divergences between lattice Gaussian distributions considered in this paper.

Other statistical divergences like the projective Hölder divergences[39] between lattice Gaussian distributions can be obtained similarly in closed-form:

$$\vec{D}^{\mathrm{H}}_{\alpha,\gamma}[r : s] : = \left\lvert\log\left(\frac{\sum_{x\in\mathcal{X}} r(x)^{\gamma/\alpha} s(x)^{\gamma/\beta}}{\left(\sum_{x\in\mathcal{X}} r(x)^\gamma\right)^{1/\alpha}\left(\sum_{x\in\mathcal{X}} s(x)^\gamma\right)^{1/\beta}}\right)\right\rvert, \quad \left(\gamma > 0, \frac{1}{\alpha} + \frac{1}{\beta} = 1\right),$$

for $\alpha, \gamma > 0$. The Hölder divergences include the Cauchy–Schwarz divergence[40] for $\gamma = \alpha = \beta = 2$:

$$\vec{D}_{\mathrm{CS}}[r : s] := -\log \frac{\sum_{x\in\mathcal{X}} r(x)s(x)}{\sqrt{(\sum_{x\in\mathcal{X}} r^2(x))\,(\sum_{x\in\mathcal{X}} s^2(x))}}.$$

Since the natural parameter space Ξ is a cone[39], we get:

$$\vec{D}^{\mathrm{H}}_{\alpha,\gamma}[p_\xi : p_{\xi'}] = \left| \log \frac{\theta_\Lambda(\gamma\xi)^{\frac{1}{\alpha}}\theta_\Lambda(\gamma\xi')^{\frac{1}{\beta}}}{\theta_\Lambda(\frac{\gamma}{\alpha}\xi + \frac{\gamma}{\beta}\xi')} \right|.$$

Thus we get the following closed-form for the Cauchy–Schwarz divergence between two lattice Gaussian distributions:

$$\vec{D}_{\mathrm{CS}}[p_\xi : p_{\xi'}] = \log \frac{\sqrt{\theta_\Lambda(2\xi)\theta_\Lambda(2\xi')}}{\theta_\Lambda(\xi + \xi')}.$$

5 Bayesian Hypothesis Testing and Chernoff Information

We conclude with an application of the KLD in statistics which highlights the information-geometric structure of the exponential family[41]. Chernoff information characterizes the error exponent in Bayesian hypothesis testing (see Section 11.9 of[26], page 384) and is also widely used in information fusion[27]. The Chernoff information between two pmfs $r(x)$ and $s(x)$ is defined by

$$D_{\mathrm{Chernoff}}[r, s] := -\min_{\alpha\in[0,1]} \log\left(\sum_{x\in\mathcal{X}} r^\alpha(x)s^{1-\alpha}(x)\right).$$

Let α^* denotes the best exponent: $\alpha^* = \arg\min_{\alpha\in[0,1]} \sum_{x\in\mathcal{X}} r^\alpha(x)s^{1-\alpha}(x)$. When $r(x) = p_\xi(x)$ and $s(x) = p_{\xi'}(x)$ are pmfs of a discrete exponential family with cumulant function $F(\xi)$, we have (Theorem 1 of[42]):

$$D_{\mathrm{Chernoff}}[p_\xi, p_{\xi'}] = B_F(\xi : \xi^*) = B_F(\xi' : \xi^*),$$

where $\xi^* := \alpha^*\xi + (1-\alpha)\xi'$, and $B_F(\xi : \xi')$ is the Bregman divergence[43]:

$$B_F(\xi' : \xi) := F(\xi') - F(\xi) - \left\langle \xi' - \xi, \nabla F(\xi) \right\rangle.$$

Thus calculating the Chernoff information amounts to first find the best value α^* and second to compute $D_{\mathrm{KL}}[p_{\xi^*} : p_\xi]$ or equivalently $D_{\mathrm{KL}}[p_{\xi^*} : p_{\xi'}]$.

By modeling the exponential family as a manifold $\mathcal{G}_\Lambda = \{p_\xi \ : \ \xi \in \Xi\}$ equipped with the Fisher information metric[41, 44], we can characterize geometrically the exact α^* (Theorem 2 of[42]) from the unique intersection of an exponential geodesic $\gamma^e_{\xi,\xi'}$ with a mixture bisector $\mathrm{Bi}_m(\xi, \xi')$ where

$$\gamma^e_{\xi,\xi'} := \{p_{\lambda\xi+(1-\lambda)\xi'} \propto p_\xi^\lambda p_{\xi'}^{1-\lambda} \ : \ \lambda \in (0, 1)\},$$
$$\mathrm{Bi}_m(\xi, \xi') := \{p_\omega \in M \ : \ D_{\mathrm{KL}}[p_\omega : p_\xi] = D_{\mathrm{KL}}[p_\omega : p_{\xi'}]\}.$$

Thus we have

$$p_{\xi^*} = \gamma^e_{\xi,\xi'} \cap \mathrm{Bi}_m(\xi, \xi').$$

This geometric characterization yields a fast numerical approximation technique to obtain α^* within a prescribed machine precision error[42]. Since the lattice Gaussian distributions form an exponential family with a dually flat structure[41] (also called a Hessian structure[44]), we can apply the above technique derived from information geometry to calculate numerically the Chernoff information between two lattice Gaussian distributions. More precisely, the geodesics γ^e are called e-geodesics and are defined with respect to the exponential connection ∇^e: The γ^e-geodesics are visualized as straight line segments in the natural parameter coordinate system. The mixture bisector Bi_m is an autoparallel submanifold with respect to ∇^m, the mixture connection[41]: The mixture bisectors are visualized as straight line segments in the moment parameter coordinate system[45].

As a final remark, let us state that knowing that the KLD between two lattice Gaussian distributions amounts to a Bregman divergence is also helpful for a number of tasks like clustering[46]: For example, the left-sided KLD centroid of n lattice Gaussian distributions $p_{\xi_1}, \ldots, p_{\xi_n}$ amounts to a right-sided Bregman centroid which is always the center of mass of the natural parameters[47]:

$$\begin{aligned}\xi^* &= \arg\min_\xi \sum_{i=1}^n \frac{1}{n} D_{\mathrm{KL}}[p_\xi : p_{\xi_i}] \\ &= \arg\min_\xi \sum_{i=1}^n \frac{1}{n} B_F(\xi_i : \xi), \\ \Rightarrow \xi^* &= \frac{1}{n}\sum_{i=1}^n \xi_i.\end{aligned}$$

6 Conclusion and Discussion

In this paper, we have considered the family of real-valued lattice Gaussian distributions (Definition 1) as a discrete exponential family defined on a lattice support. We reported in Sect. 2.3 a Newton's method to convert numerically a moment parameter to its corresponding natural parameter. We then give formula to calculate the cross-entropy (Proposition 1), the

Kullback–Leibler divergence (Proposition 2), and the Rényi α-divergences (Proposition 4) between two lattice Gaussian distributions. Furthermore, we showed how to approximate numerically the Kullback-Leibler divergence between two lattice Gaussian distributions using either the Rényi α-divergences (Proposition 6) or the γ-divergences (Proposition 7). Finally, in Sect. 5, we consider the exponential family manifold structure[41] of the lattice Gaussian family, and show how to compute the Chernoff information which characterizes the best error exponent in Bayesian hypothesis testing[26].

We leave for future work the analysis of the approximation errors and corresponding time complexities for computing the various statistical divergences reported in Table 1 when we approximate the Riemann Theta functions $\theta_\Lambda(\xi)$ and $\theta_\Lambda(\xi')$ with error $\epsilon > 0$ (say, using the guaranteed approximation error of Theorem 2 of[9]). In practice, approximating Riemann Theta functions scale up to dimensions 50 to 60. In high dimensions, one may also consider parsimonious models[48] for the matrix parameter ξ_2 of lattice Gaussian distributions (which otherwise shall require a quadratic number of coefficients with the dimension to define). Another open question is how to choose the finite subsets of lattice points R_ξ, $R_{\xi'}$ and $R_{\xi,\xi'}$ so that we get an upper bound $\epsilon > 0$ of the approximation error of the Kullback-Leibler divergence $D_{\mathrm{KL}}[p_\xi : p_{\xi'}]$. What is the best precision-computation tradeoff one can achieve to guarantee an ϵ-approximation of the Kullback-Leibler divergence?

Publisher's Note

Acknowledgements

We thank the reviewers for the constructive and helpful suggestions on this paper.

Received: 8 October 2021 Accepted: 12 December 2021
Published online: 23 January 2022

References

1. Keener RW (2010) Theoretical statistics: topics for a core course. Springer, Heidelberg
2. Grätzer G (2011) Lattice theory: foundation. Springer, Heidelberg
3. Barndorff-Nielsen O (2014) Information and Exponential Families in Statistical Theory. Wiley, New Jersey
4. Calin O, Udrişte C (2014) Geometric modeling in probability and statistics. Springer, Heidelberg, Germany
5. Agostini D, Améndola C (2019) Discrete Gaussian distributions via theta functions. SIAM J Appl Algebra Geometry 3(1):1–30
6. Olver FW, Lozier DW, Boisvert RF, Clark CW (2010) NIST Handbook of Mathematical Functions. Cambridge University Press, Cambridge
7. Nielsen F (2020) An elementary introduction to information geometry. Entropy 22(10):1100
8. Siegel CL (2014) Symplectic Geometry. Elsevier, Amsterdam
9. Deconinck B, Van Hoeij M (2001) Computing Riemann matrices of algebraic curves. Physica D 152:28–46
10. Frauendiener J, Jaber C, Klein C (2019) Efficient computation of multidimensional theta functions. J Geometry Phys 141:147–158
11. Mumford, D., Musili, C.: Tata Lectures on Theta I. Birkhäuser, Boston, USA (2007). With the collaboration of C. Musili, M. Nori, E. Previato, and M. Stillman
12. Deconinck B, Heil M, Bobenko A, Van Hoeij M, Schmies M (2004) Computing Riemann theta functions. Math Comput 73(247):1417–1442
13. Agostini D, Chua L (2021) Computing theta functions with Julia. J Softw Algebra Geometry 11(1):41–51
14. Osborne, A.R.: Nonlinear ocean wave and the inverse scattering transform. In: Scattering, pp. 637–666. Elsevier, The Netherlands (2002)
15. Labrande H (2018) Computing Jacobi's theta in quasi-linear time. Math Comput 87(311):1479–1508
16. Lisman J, Van Zuylen M (1972) Note on the generation of most probable frequency distributions. Statistica Neerlandica 26(1):19–23
17. Kemp AW (1997) Characterizations of a discrete normal distribution. J Stat Planning Inference 63(2):223–229
18. Szabłowski PJ (2001) Discrete normal distribution and its relationship with Jacobi theta functions. Stat Prob Lett 52(3):289–299
19. Budroni, A., Semaev, I.: New Public-Key Crypto-System EHT. arXiv preprint arXiv:2103.01147 (2021)
20. Wang, L., Jia, R., Song, D.: D2P-Fed: Differentially private federated learning with efficient communication. arXiv preprint arXiv:2006.13039 (2020)
21. Canonne, C.L., Kamath, G., Steinke, T.: The discrete Gaussian for differential privacy. arXiv preprint arXiv:2004.00010 (2020)
22. Canonne, C.L., Kamath, G., Steinke, T.: The discrete gaussian for differential privacy. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, Virtual (2020)

23. Carrazza S, Krefl D (2020) Sampling the Riemann-Theta Boltzmann machine. Comput Phys Commun 256:107464
24. Van Erven T, Harremos P (2014) Rényi divergence and Kullback–Leibler divergence. IEEE Trans Inform Theory 60(7):3797–3820
25. Fujisawa H, Eguchi S (2008) Robust parameter estimation with a small bias against heavy contamination. J Multivariate Anal 99(9):2053–2081
26. Cover TM (1999) Elements of Information Theory. Wiley, New Jersey
27. Julier SJ An empirical study into the use of Chernoff information for robust, distributed fusion of Gaussian mixture models. In: 9th International conference on information Fusion, pp. 1–8 (2006). IEEE
28. Pistone G, Wynn HP Finitely generated cumulants. Statistica Sinica, 1029–1052 (1999)
29. Rockafellar RT (2015) Convex Analysis. Princeton University Press, Princeton
30. Navarro J, Ruiz J (2005) A note on the discrete normal distribution. Adv Appl Stat 5(2):229–245
31. Zellner A, Highfield RA (1988) Calculation of maximum entropy distributions and approximation of marginal-posterior distributions. J Econ 37(2):195–209
32. Mohammad-Djafari A A Matlab program to calculate the maximum entropy distributions. In: Maximum Entropy and Bayesian Methods, pp. 221–233. Springer, Heidelberg (1992)
33. George AJ, Kashyap N An MCMC Method to Sample from Lattice Distributions. arXiv:2101.06453 (2021)
34. Nielsen F, Nock R Entropies and cross-entropies of exponential families. In: 2010 IEEE International Conference on Image Processing, pp. 3621–3624 (2010). IEEE
35. Zhang J (2004) Divergence function, duality, and convex analysis. Neural Comput 16(1):159–195
36. Nielsen F, Boltz S (2011) The Burbea-Rao and Bhattacharyya centroids. IEEE Trans Inform Theory 57(8):5455–5466
37. Bhattacharyya A On a measure of divergence between two multinomial populations. Sankhyā: the indian journal of statistics, 401–406 (1946)
38. Cichocki A, Amari S-i Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. Entropy 12(6), 1532–1568 (2010)
39. Nielsen F, Sun K, Marchand-Maillet S (2017) On Hölder projective divergences. Entropy 19(3):122
40. Jenssen R, Principe JC, Erdogmus D, Eltoft T (2006) The Cauchy-Schwarz divergence and Parzen windowing: Connections to graph theory and Mercer kernels. J Franklin Inst 343(6):614–629
41. Amari S-i Information Geometry and Its Applications vol. 194. Springer, Heidelberg (2016)
42. Nielsen F (2013) An information-geometric characterization of Chernoff information. IEEE Signal Process Lett 20(3):269–272
43. Bregman LM (1967) The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comput Math Math Phys 7(3):200–217
44. Shima H (2007) The Geometry of Hessian Structures. World Scientific, Singapore
45. Boissonnat J-D, Nielsen F, Nock R (2010) Bregman Voronoi diagrams. Discrete Comput Geometry 44(2):281–307
46. Garcia V, Nielsen F (2010) Simplification and hierarchical representations of mixtures of exponential families. Signal Process 90(12):3197–3212
47. Banerjee A, Merugu S, Dhillon IS, Ghosh J Clustering with Bregman divergences. Journal of machine learning research **6**(10) (2005)
48. McNicholas PD, Murphy TB (2008) Parsimonious Gaussian mixture models. Stat Comput 18(3):285–296

Frank Nielsen was awarded his PhD on adaptive computational geometry (1996) from INRIA/University of Cote d'Azur (France). He is a fellow of Sony Computer Science Laboratories Inc. (Sony CSL, Tokyo) where he currently conducts research on the fundamentals and practice of geometric machine learning and artificial intelligence. He taught at Ecole Polytechnique (France) visual computing and high-performance computing for data science, and currently is on the board of the following journals: Information Geometry (Springer), Entropy (MDPI), and IEEE Transactions on Information Theory. Since 2013, Frank Nielsen co-organizes with Frédéric Barbaresco the biannual conference Geometric Science of Information (GSI).