On the Kullback-Leibler divergence between discrete normal distributions

— KLD between lattice Gaussian distributions —

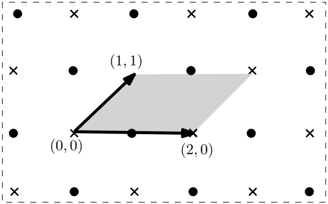
https://arxiv.org/abs/2109.14920

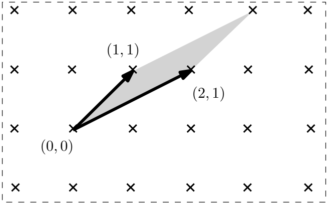
Frank Nielsen
Sony Computer Science Laboratories Inc.
Tokyo, Japan

https://FrankNielsen.github.io/

October 2021

Lattices: Integer lattice $\mathbb{Z}^d$ and full-rank lattices





- lattice basis of d column vectors, arranged in a matrix $L = [l_1 \mid \ldots \mid l_d]$
- lattice $\Lambda = \Lambda(L) = L\mathbb{Z}^d := \{Lz \ : \ z \in \mathbb{Z}^d\}$
- full rank lattice: $\det(L) \neq 0$
- two lattices $\Lambda(L)$ and $\Lambda(L')$ coincide iff. $L = L'U$ for a unimodular matrix U (= square matrix with integer entries and determinant ± 1)

$$L' = \left[\begin{array}{c|c} 1 & 2 \\ 1 & 1 \end{array}\right] = U \times L, L = I, \quad U = \left[\begin{array}{c|c} 1 & 2 \\ 1 & 1 \end{array}\right], \quad \det(U) = -1$$

- integer lattice $\mathbb{Z}^d$: one-hot basis vectors $e_1, \ldots, e_d$.
  $L = I_{d,d} = [e_1 \mid \ldots \mid e_d] = \mathrm{diag}(1, \ldots, 1)$

Lattice Gaussian distributions

- Lattice Gaussian random variable $X_\xi \sim N_\Lambda(\xi)$ has pmf:

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

- Partition function expressed using the Riemann theta function $\theta(\omega, \Omega)$:

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

- Riemann theta: holomorphic function [26], converging Fourier series:

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

$\mathcal{H}_d$ = Siegel upper space of symmetric complex matrices with positive-definite imaginary parts [27].

Discrete normal distributions

Studied in [1]

- Probability mass function:

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

- Partition function $Z_\mathbb{Z}(\xi) = \theta_R(-i\xi_1, i\xi_2)$
- Proposition 3.5 [1]:

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

- Parity (Remark 3.7 [1]):

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

- But marginals of discrete Gaussians are not discrete Gaussians

Lattice Gaussians: A Discrete exponential family

- Lattice Gaussian distributions form a discrete (minimal regular) exponential family $\mathcal{G}_\Lambda = \{p_\xi \ : \ \xi \in \Xi\}$:

$$p_\xi(l) = \exp\left(\langle t(x), \xi\rangle - F_\Lambda(\xi)\right), \quad F_\Lambda(\xi) := \log\theta_\Lambda(\xi)$$

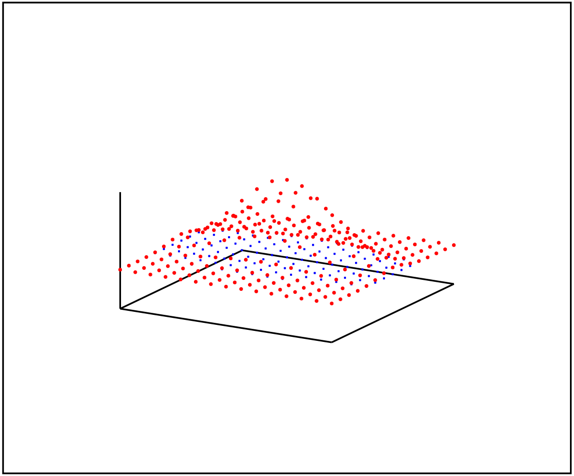
- Natural parameter space: $\Xi = \mathbb{R}^d \times \mathcal{P}_d$ with $\mathcal{P}_d$ the open cone of positive-definite matrices. Exp. fam. of order $D = \frac{d(d+3)}{2}$
- Sufficient statistics: $t(x) = \left(2\pi x, -\pi x x^\top\right)$
- Compound vector-matrix inner product:

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

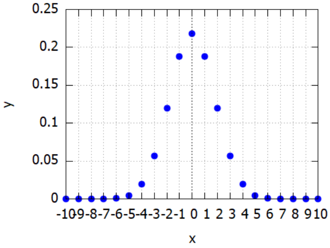
- Thus lattice Gaussians are maximum entropy distributions on the lattice support Λ: $\max_p \in H(p)$ such that $E[t(x)] = \eta$. Constraint $E[t(x)] = \eta$ is equivalent to the two constraints $\mu = E_p[X]$ and $\Sigma = \mathrm{Cov}_p[X] = E_p[(X - E_p[X])(X - E_p[X])^\top]$
- Prior work: Lisman and Van Zuylen [19] (1972), Kemp [18] (1997), partition function with Jacobi theta function by Szablowski [28] (2001), Riemann multivariate theta and complex-valued pmf with $\Xi = \mathbb{C}^d \times \mathcal{H}_d$ by Agostini and Améndola [1] (2019)

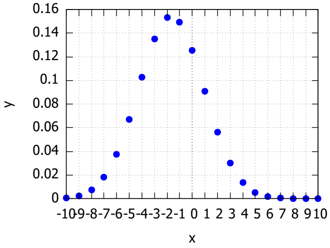
Lattice Gaussian distribution $N_\Lambda(\xi)$

- Lattice: $\Lambda = L\mathbb{Z}^2$ with $L = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ $(\det(L) = 1)$
- Natural parameter $\xi = (\xi_1, \xi_2)$: $\xi_1 = (0, 0)$ and $\xi_2 = \mathrm{diag}(0.1, 0.5)$



Discrete normal distributions $N_{\mathbb{Z}}(\xi)$





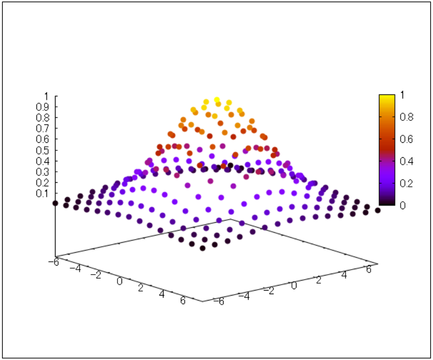
- Symmetric distribution (left) or not (right) depending on the parameters
- Periodicity of Riemann θ with integer periods $u \in \mathbb{Z}^d$:

$$\theta(\omega + u, \Omega) = \theta(\omega, \Omega)$$

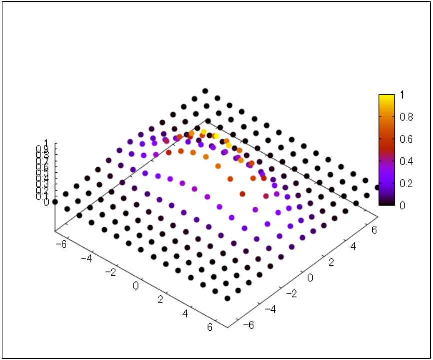
- yields $X_{(a,Bu)} \sim X_{(a,B)} + u$ for any $u \in \mathbb{Z}^d$:

$$\Pr(X_{(a,Bu)} = l) = \Pr(X_{(a,B)} = l - u)$$

Discrete normal distributions: Bivariate examples



$\xi_1 = (0,0)$ and $\xi_2 = \mathrm{diag}\left(\frac{1}{10}, \frac{1}{10}\right)$



$\xi_1 = (0,0)$ and $\xi_2 = \mathrm{diag}\left(\frac{1}{10}, \frac{1}{2}\right)$

2D integer lattice $\Lambda = \mathbb{Z}^2$

Some applications of lattice Gaussian distributions

- Applications:
  - Lattice-based cryptography [6]
  - Machine learning:
    - Differential privacy [30, 7]
    - Boltzmann machines with continuous visible states and discrete hidden states [8]
- Sampling discrete Gaussian distributions:
  - in 1D [7]
  - using simple rejection sampling [8]
  - using Markov chain Monte Carlo [15]

MLE and dual moment parameterization η

- A set $\{v_1, \ldots, v_m\}$ of m IID variates sampled from p_ξ.
- Estimating equation for the maximum likelihood estimator (MLE):

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

- Equivariance property of the MLE yields $\hat{\xi} = \nabla F_\Lambda^*(\hat{\eta})$
- MLE of lattice Gaussians using ordinary parameterization $\lambda = (\mu, \Sigma)$

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

- Dual moment/expectation parameterization of exponential families: $\eta = \nabla F(\theta) = E[t(X)]$ and $\theta = \nabla F^*(\eta)$ with Legendre-Fenchel convex conjugate

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

with $\xi = \nabla F_\Lambda^*(\eta)$.

Converting moment η to natural ξ parameters

- Solve a concave maximization program given η:

$$F^*(\eta) := \sup_{\xi} L_\eta(\xi) := \langle \xi, \eta \rangle - F(\xi)$$

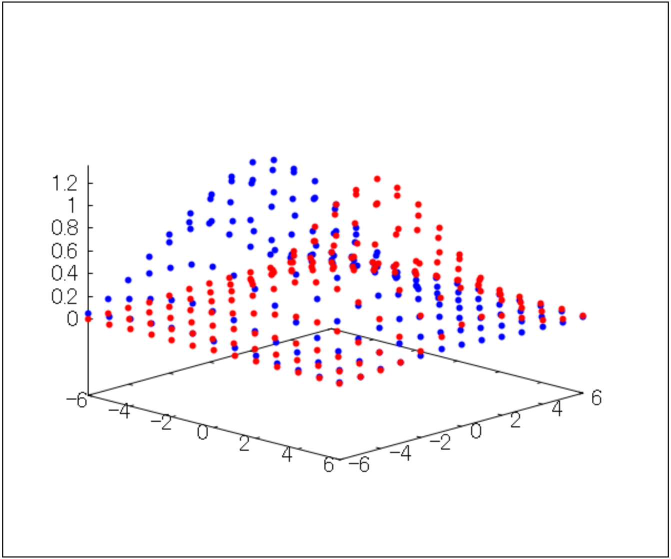
  Concave maximization $\nabla^2 L_\eta(\xi) = -\nabla^2 F(\xi)$ or equivalently convex minimization $-L_\eta(\xi)$. Unique optimal solution $\boxed{\xi = \nabla F^*(\eta)}$. Then get $F^*(\eta) = \langle \eta, \nabla F^*(\eta) \rangle - F(\nabla F^*(\eta))$ (negentropy)

- Solve iteratively [31, 20]: $p_\psi(x) := \exp\left(-\sum_{i=0}^{D} \psi_i t_i(x)\right)$, with $\psi_i = -\xi_i$, and $\psi_0 = F(\psi)$. Let $K_i(\psi) := E_{p_\xi}[t_i(x)] = \eta_i$ and $\eta_0 = 1$.
- Update iteratively:

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

- $H_{ij}(\psi) = H_{ji}(\psi) = -E_{p_\psi}[t_i(x) t_j(x)]$ (need to be approximated)

Statistical divergences



How to measure the dissimilarity between bivariate discrete normal distributions?

Cross-entropy and Kullback-Leibler divergence

- Kullback-Leibler divergence [10] between two pmfs $r(x)$ and $s(x)$ with support $\mathcal{X}$:
$$D_{\mathrm{KL}}[r:s] := \sum_{x\in\mathcal{X}} r(x)\log\frac{r(x)}{s(x)}.$$
- KLD also called relative entropy $D_{\mathrm{KL}}[r:s] = H[r:s] - H[r]$ with $H[r:s] := -\sum_{x\in\mathcal{X}} r(x)\log s(x)$ and $H[r] = H[r:r]$ is Shannon's entropy
- Cross-entropy between two densities p_ξ and $p_{\xi'}$ of an exponential family [23]:
$$H[p_\xi : p_{\xi'}] = F_\Lambda(\xi') - \langle \xi', \eta\rangle.$$
- When $\xi' = \xi$, get from Fenchel-Young's inequality:
$$H[p_\xi : p_\xi] = H[p_\xi] = F_\Lambda(\xi) - \langle \xi, \eta\rangle = -F_\Lambda^*(\eta).$$

$\Rightarrow$ The convex conjugate is Shannon's negentropy [23] (convex)

Cross-entropy and Kullback-Leibler divergence

Proposition

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

$$\boxed{H[p_\xi : p_{\xi'}] = \log \theta_\Lambda(\xi') - 2\pi \mu^\top \xi'_1 + \pi \,\mathrm{tr}\left(\xi'_2(\Sigma + \mu\mu^\top)\right)}$$

Proposition

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:*

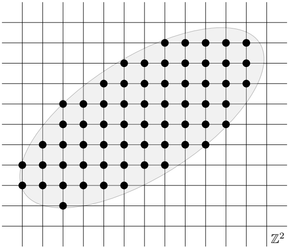
$$\boxed{D_{\mathrm{KL}}[p_\xi : p_{\xi'}] = \log\left(\frac{\theta_\Lambda(\xi')}{\theta_\Lambda(\xi)}\right) - 2\pi\mu^\top(\xi'_1 - \xi_1) + \pi \,\mathrm{tr}\left((\xi'_2 - \xi_2)(\Sigma + \mu\mu^\top)\right)}$$

Note that $\mu := E_{p_\xi}[X]$ and $\Sigma := \mathrm{Cov}_{p_\xi}[X] = E_{p_\xi}[(X-\mu)(X-\mu)^\top]$ need to be computed from p_ξ or ξ be calculated from $\lambda = (\mu, \Sigma)$

Computing expectations with approximations of θ

- In practice, computing expectations like means or covariance matrices require approximating Riemann θ function [12, 13]
- Replace infinite sums by finite sums on integer lattice points $R_\xi := E_\xi \cap \mathbb{Z}^d$, where E_ξ is theta ellipsoid (with $\tilde{\theta}(\xi; \mathbb{Z}_d) = \theta(\xi)$):

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



Approximating $\theta_{\mathbb{Z}^d}(\xi)$ by finite sum of $\tilde{p}_\xi$ on the integer lattice points R_ξ falling inside an ellipsoid E_ξ

KLD via Rényi α-divergences

- Rényi α-divergence [29]:

$$D_\alpha[r:s] \quad := \quad \frac{1}{\alpha-1}\log\left(\sum_{x\in\mathcal{X}} r(x)^\alpha s(x)^{1-\alpha}\right)$$

- 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

$$\begin{aligned} I_{\alpha,\beta}[p_\xi : p_{\xi'}] \quad &:= \quad \sum_{l\in\Lambda} p_\xi(l)^\alpha p_{\xi'}(l)^\beta \\ &= \quad \exp\left(F(\alpha\xi+\beta\xi') - (\alpha F(\xi)+\beta F(\xi'))\right) \end{aligned}$$

Proposition

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

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

Bhattacharyya and Hellinger divergences

- Bhattacharyya divergence:

$$D_{\mathrm{Bhattacharyya}}[r, s] := -\log\left(\sum_{x\in\mathcal{X}} \sqrt{r(x)s(x)}\right) = \frac{1}{2} D_{\frac{1}{2}}[r : s]$$

- Bhattacharyya coefficient $\rho_{\mathrm{Bhattacharyya}}[r, s] := \sum_{x\in\mathcal{X}} \sqrt{r(x)s(x)}$,
- Squared Hellinger divergence ($D_{\mathrm{Hellinger}}$ is a metric distance):

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

Proposition

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

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

Approximating the KLD via Rényi α-divergences

Proposition

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

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

- Rényi α-divergences are non-decreasing with α [29]: obtain both lower and upper bounds of the KLD.
- When $\alpha \to 1$, $J_{F,\alpha}(\xi : \xi') \to B_F(\xi' : \xi)$ and $D_\alpha[p_\xi : p'_\xi] \to D_{\mathrm{KL}}[p_\xi : p'_\xi]]$ (see [22])

Approximating KLD via γ-divergences

- γ-divergences proposed for robust estimation [14, 9] ($\gamma > 1$):

$$D_\gamma[p:q] := \frac{1}{\gamma(\gamma-1)} \log\left(\frac{\left(\sum_{x\in\mathcal{X}} p^\gamma(x)\right)\left(\sum_{x\in\mathcal{X}} q^\gamma(x)\right)^{\gamma-1}}{\left(\sum_{x\in\mathcal{X}} p(x) q^{\gamma-1}(x)\right)^\gamma} \right), \quad (\gamma > 0)$$

- γ-divergences are projective divergences: Let $p(x) = \frac{\tilde{p}(x)}{Z_p}$ and $q(x) = \frac{\tilde{q}(x)}{Z_q}$. Then we have:

$$D_\gamma[p:p'] = D_\gamma[\tilde{p}:\tilde{p}'].$$

- Let $I_\gamma[p:q] := \sum_{x\in\mathcal{X}} p(x) q(x)^{\gamma-1}$. γ-divergence can be written as

$$D_\gamma[p:q] = D_\gamma[\tilde{p}:\tilde{q}] = \frac{1}{\gamma(\gamma-1)} \log\left(\frac{I_\gamma[\tilde{p}:\tilde{p}]\, I_\gamma[\tilde{q}:\tilde{q}]^{\gamma-1}}{I_\gamma[\tilde{p}:\tilde{q}]^\gamma} \right).$$

Approximating KLD via γ-divergences

- Let $\tilde{I}_\gamma(\xi : \xi') := I_\gamma\left[\tilde{p}_\xi : \tilde{p}_{\xi'}\right]$.
- Notice that $\tilde{I}_\gamma$ depends on θ [24]:

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

- Thus we have

$$\boxed{D_\gamma[p_\xi : p_{\xi'}] = \frac{1}{\gamma(\gamma-1)} \log\left(\frac{\theta_\Lambda(\gamma\xi)\,\theta_\Lambda(\gamma\xi')^{\gamma-1}}{\theta_\Lambda(\xi + (\gamma-1)\xi')^\gamma}\right)}$$

- Approximate $\tilde{I}_\gamma(\xi : \xi')$ using theta ellipsoids (finite sums)

$$\tilde{I}_{\gamma, R_{\xi,\xi'}}(\xi : \xi') := \sum_{x \in R_\xi \cup R_{\xi'}} \tilde{p}_\xi\, \tilde{p}_{\xi'}(x)^{\gamma-1} \approx \theta_\Lambda(\xi + (\gamma-1)\xi')$$

Approximating KLD via γ-divergences

- When $\gamma \to 1$, $D_\gamma[\tilde{p} : \tilde{q}] = D_{\mathrm{KL}}\left[\frac{\tilde{p}}{Z_p} : \frac{\tilde{q}}{Z_q}\right]$
- γ-divergences are projective but not the KLD which is homogeneous of degree 1: $D_{\mathrm{KL}}[\lambda p : \lambda q] = \lambda D_{\mathrm{KL}}[p : q]$

Proposition

The Kullback-Leibler divergence between two lattice Gaussian distributions p_ξ and $p_{\xi'}$ can be efficiently approximated:

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

for $\gamma > 0$ close to 1 (say, $\gamma = 1 + 10^{-5}$), where $R_\xi = E_\xi \cap \mathbb{Z}^d$ and $R_{\xi'} = E_{\xi'} \cap \mathbb{Z}^d$ denote the integer lattice points falling inside the theta ellipsoids E_ξ and $E_{\xi'}$ used to approximate the theta functions $\theta_\Lambda(\xi)$ and $\theta_\Lambda(\xi')$, respectively.

Hölder and Cauchy-Schwarz divergences

- projective Hölder divergence [25], $\alpha > 0, \gamma > 0$, $\frac{1}{\alpha} + \frac{1}{\beta} = 1$:

$$D^{\text{Hölder}}_{\alpha,\gamma}[r : s] := \left| \log \left(\frac{\sum_{x \in \mathcal{X}} r(x)^{\gamma/\alpha} s(x)^{\gamma/\beta}}{\left(\sum_{x \in \mathcal{X}} r(x)^{\gamma}\right)^{1/\alpha} \left(\sum_{x \in \mathcal{X}} s(x)^{\gamma}\right)^{1/\beta}} \right) \right|$$

- generalize the Cauchy-Schwarz divergence [16] for $\alpha = \gamma = \beta = 2$:

$$D_{\mathrm{CS}}[r : s] := -\log \frac{\sum_{x \in \mathcal{X}} r(x) s(x)}{\sqrt{\left(\sum_{x \in \mathcal{X}} r^2(x)\right) \left(\sum_{x \in \mathcal{X}} s^2(x)\right)}}.$$

- Closed-form formula between lattice Gaussian distributions:

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

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

Bayesian hypothesis testing: Chernoff information

- Chernoff information between pmfs $r(x)$ and $s(x)$:

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

- best exponent α^*: $\alpha^* = \arg\min_{\alpha\in[0,1]} \sum_{x\in\mathcal{X}} r^{\alpha}(x)s^{1-\alpha}(x)$.
- Theorem: Chernoff information for pmfs of a discrete exponential family amounts to a Bregman divergence [21]:

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

  where $\xi^* := \alpha^*\xi + (1-\alpha^*)\xi'$
- Bregman divergence [5]: $B_F(\xi':\xi) := F(\xi') - F(\xi) - \langle \xi'-\xi, \nabla F(\xi)\rangle$
- Chernoff information can also used in information fusion tasks [17]

Chernoff information: Lattice Gaussian manifold

- $\mathcal{G}_\Lambda = \{p_\xi \ : \ \xi \in \Xi\}$ equipped with the Fisher information metric [3] $g_F(\xi) = \nabla^2 F_\lambda(\xi)$ (Hessian metric) yields dually flat structure $(\mathcal{G}_\Lambda, g_F, \nabla^e, \nabla^m)$ with dual e-connection ∇^e and m-connection ∇^m
- Define exponential geodesic (wrt ∇^e connection) and mixture bisector (wrt ∇^m connection):

$$
\begin{aligned}
\gamma^e_{\xi,\xi'} &:= \{p_{\lambda\xi+(1-\lambda)\xi'} \propto p_\xi^\lambda p_{\xi'}^{1-\lambda} \ : \ \lambda \in (0,1)\} \\
\mathrm{Bi}_m(\xi,\xi') &:= \{p_\omega \in \mathcal{G}_\Lambda \ : \ D_{\mathrm{KL}}[p_\omega : p_\xi] = D_{\mathrm{KL}}[p_\omega : p_{\xi'}]\}
\end{aligned}
$$

- Chernoff point is characterized by

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

- Bisection search [21] on $\alpha \in (0,1)$ to get α^* from $\xi^* := \alpha^*\xi + (1-\alpha^*)\xi'$

Clustering lattice Gaussian distributions

- Use the property that the KLD between two lattice Gaussian distributions amounts to a Bregman divergence for various tasks.
- For example, clustering of lattice Gaussian distributions [4, 11] (say, for mixture simplification):

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

Summary: KLD between lattice Gaussians

- Lattice Gaussian distributions form a discrete exponential family with cumulant function related to Riemann theta function
- Maximum likelihood estimator in closed-form for $\hat{\eta}$. Convert iteratively to get the corresponding natural parameter $\hat{\xi}$
- Kullback-Leibler divergence in closed form using the mixed parameterizations ξ and $\lambda = (\mu, \Sigma)$ (or moment parameter η)
- Kullback-Leibler divergence using natural parameters ξ approximated using Rényi α-divergences for $\alpha \simeq 1$
- Kullback-Leibler divergence using natural parameters ξ approximated using projective γ-divergences for $\gamma \simeq 0$ ($\gamma > 0$)
- Chernoff information amounts to KLD once the optimal exponent α^* is found. Information geometry yields simple efficient algorithm on the dually flat manifold of lattice Gaussian distributions
- Many available packages for calculating Riemann θ function and its derivatives [13, 2]

Summary of closed-form formula

Divergence	definition/closed-form formula for lattice Gaussians
Kullback-Leibler divergence	$D_{\mathrm{KL}}[p_\xi : p_{\xi'}] = \sum_{l\in\Lambda} p_\xi(l) \log \frac{p_\xi(l)}{p_{\xi'}(l)}$
	$D_{\mathrm{KL}}[p_\xi : p_{\xi'}] = \log\left(\frac{\theta_\Lambda(\xi')}{\theta_\Lambda(\xi)}\right) - 2\pi\mu^\top(\xi'_1 - \xi_1) + \pi\,\mathrm{tr}\left((\xi'_2 - \xi_2)(\Sigma + \mu\mu^\top)\right)$
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	$D_\gamma[p_\xi : p_{\xi'}] = \frac{1}{\gamma(\gamma-1)}\log\left(\frac{\left(\sum_{l\in\Lambda} p_\xi^\gamma(x)\right)\left(\sum_{l\in\Lambda} p_{\xi'}^\gamma(l)\right)^{\gamma-1}}{\left(\sum_{l\in\Lambda} p_\xi(l) p_{\xi'}^{\gamma-1}(l)\right)^\gamma}\right)$
$(\gamma > 1)$	$D_\gamma[p_\xi : p_{\xi'}] = \frac{1}{\gamma(\gamma-1)}\log\left(\frac{\theta_\Lambda(\gamma\xi)\,\theta_\Lambda(\gamma\xi')^{\gamma-1}}{\theta_\Lambda(\xi+(\gamma-1)\xi')^\gamma}\right)$
	$\lim_{\gamma\to 1} D_\gamma[p_\xi : p_{\xi'}] = D_{\mathrm{KL}}[p_\xi : p_{\xi'}]$
Hölder divergence	$D^{\text{Hölder}}_{\alpha,\gamma}[r : s] := \left\lvert \log\left(\frac{\sum_{x\in\mathcal{X}} r(x)^{\gamma/\alpha} s(x)^{\gamma/\beta}}{\left(\sum_{x\in\mathcal{X}} r(x)^\gamma\right)^{1/\alpha}\left(\sum_{x\in\mathcal{X}} s(x)^\gamma\right)^{1/\beta}}\right)\right\rvert$
$(\gamma > 0,\ \frac{1}{\alpha} + \frac{1}{\beta} = 1)$	$D^{\text{Hölder}}_{\alpha,\gamma}[p_\xi : p_{\xi'}] = \left\lvert \log\frac{\theta_\Lambda(\gamma\xi)^{\frac{1}{\alpha}}\,\theta_\Lambda(\gamma\xi')^{\frac{1}{\beta}}}{\theta_\Lambda(\frac{\gamma}{\alpha}\xi + \frac{\gamma}{\beta}\xi')}\right\rvert$
Cauchy-Schwarz divergence	$D_{\mathrm{CS}}[r : s] := -\log\frac{\sum_{x\in\mathcal{X}} r(x)s(x)}{\sqrt{\left(\sum_{x\in\mathcal{X}} r^2(x)\right)\left(\sum_{x\in\mathcal{X}} s^2(x)\right)}}$
(Hölder with $\alpha = \beta = \gamma = 2$)	$D_{\mathrm{CS}}[p_\xi : p_{\xi'}] = \log\frac{\sqrt{\theta_\Lambda(2\xi)\theta_\Lambda(2\xi')}}{\theta_\Lambda(\xi+\xi')}$

//Partition function θ_Λ related to Riemann theta function θ_R (with $i^2 = -1$):

$$\theta_\Lambda(\xi) = \theta_R(-il^\top\xi_1, il^\top\xi_2 l),\ \theta_R(z,\Omega) = \sum \exp\left(2\pi i\left(\frac{1}{2}l^\top\Omega l + l^\top z\right)\right)$$

References I

[1] Daniele Agostini and Carlos Améndola.
Discrete Gaussian distributions via theta functions.
SIAM Journal on Applied Algebra and Geometry, 3(1):1–30, 2019.

[2] Daniele Agostini and Lynn Chua.
Computing theta functions with Julia.
Journal of Software for Algebra and Geometry, 11(1):41–51, 2021.

[3] Shun-ichi Amari.
Information geometry and its applications, volume 194.
Springer, Heidelberg, 2016.

[4] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh.
Clustering with Bregman divergences.
Journal of machine learning research, 6(10), 2005.

[5] Lev M Bregman.
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.

[6] Alessandro Budroni and Igor Semaev.
New Public-Key Crypto-System EHT.
arXiv preprint arXiv:2103.01147, 2021.

[7] Clément L Canonne, Gautam Kamath, and Thomas Steinke.
The discrete Gaussian for differential privacy.
arXiv preprint arXiv:2004.00010, 2020.

References II

[8] Stefano Carrazza and Daniel Krefl.
Sampling the Riemann-Theta Boltzmann machine.
Computer Physics Communications, 256:107464, 2020.

[9] Andrzej Cichocki and Shun-ichi Amari.
Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities.
Entropy, 12(6):1532–1568, 2010.

[10] Thomas M Cover.
Elements of information theory.
John Wiley & Sons, New Jersey, 1999.

[11] Jason V. Davis and Inderjit Dhillon.
Differential entropic clustering of multivariate gaussians.
Advances in Neural Information Processing Systems, 19:337, 2007.

[12] Bernard Deconinck, Matthias Heil, Alexander Bobenko, Mark Van Hoeij, and Marcus Schmies.
Computing Riemann theta functions.
Mathematics of Computation, 73(247):1417–1442, 2004.

[13] Jörg Frauendiener, Carine Jaber, and Christian Klein.
Efficient computation of multidimensional theta functions.
Journal of Geometry and Physics, 141:147–158, 2019.

[14] Hironori Fujisawa and Shinto Eguchi.
Robust parameter estimation with a small bias against heavy contamination.
Journal of Multivariate Analysis, 99(9):2053–2081, 2008.

References III

[15] Anand Jerry George and Navin Kashyap.
An MCMC Method to Sample from Lattice Distributions.
arXiv:2101.06453, 2021.

[16] Robert Jenssen, Jose C Principe, Deniz Erdogmus, and Torbjørn Eltoft.
The Cauchy–Schwarz divergence and Parzen windowing: Connections to graph theory and Mercer kernels.
Journal of the Franklin Institute, 343(6):614–629, 2006.

[17] Simon J Julier.
An empirical study into the use of Chernoff information for robust, distributed fusion of Gaussian mixture models.
In *9th International Conference on Information Fusion*, pages 1–8. IEEE, 2006.

[18] Adrienne W Kemp.
Characterizations of a discrete normal distribution.
Journal of Statistical Planning and Inference, 63(2):223–229, 1997.

[19] JHC Lisman and MCA Van Zuylen.
Note on the generation of most probable frequency distributions.
Statistica Neerlandica, 26(1):19–23, 1972.

[20] Ali Mohammad-Djafari.
A Matlab program to calculate the maximum entropy distributions.
In *Maximum entropy and Bayesian methods*, pages 221–233. Springer, Heidelberg, 1992.

[21] Frank Nielsen.
An information-geometric characterization of Chernoff information.
IEEE Signal Processing Letters, 20(3):269–272, 2013.

References IV

[22] Frank Nielsen and Sylvain Boltz.
The Burbea-Rao and Bhattacharyya centroids.
IEEE Transactions on Information Theory, 57(8):5455–5466, 2011.

[23] Frank Nielsen and Richard Nock.
Entropies and cross-entropies of exponential families.
In *2010 IEEE International Conference on Image Processing*, pages 3621–3624. IEEE, 2010.

[24] Frank Nielsen and Richard Nock.
Patch matching with polynomial exponential families and projective divergences.
In *International Conference on Similarity Search and Applications*, pages 109–116. Springer, 2016.

[25] Frank Nielsen, Ke Sun, and Stéphane Marchand-Maillet.
On Hölder projective divergences.
Entropy, 19(3):122, 2017.

[26] Frank WJ Olver, Daniel W Lozier, Ronald F Boisvert, and Charles W Clark.
NIST Handbook of mathematical functions.
Cambridge university press, Cambridge, 2010.

[27] Carl Ludwig Siegel.
Symplectic geometry.
Elsevier, Amsterdam, 2014.

[28] Paweł J Szabłowski.
Discrete normal distribution and its relationship with Jacobi theta functions.
Statistics & probability letters, 52(3):289–299, 2001.

References V

[29] Tim Van Erven and Peter Harremos.
Rényi divergence and Kullback-Leibler divergence.
IEEE Transactions on Information Theory, 60(7):3797–3820, 2014.

[30] Lun Wang, Ruoxi Jia, and Dawn Song.
D2P-Fed: Differentially private federated learning with efficient communication.
arXiv preprint arXiv:2006.13039, 2020.

[31] Arnold Zellner and Richard A Highfield.
Calculation of maximum entropy distributions and approximation of marginalposterior distributions.
Journal of Econometrics, 37(2):195–209, 1988.