Regularized Information Geometric and Optimal Transport distances between covariance operators and Gaussian processes

Hà Quang Minh

RIKEN Center for Advanced Intelligence Project, Tokyo

Outline

Generalization of the following geometrical structures for Gaussian measures in $\mathbb{R}^n$

- Riemannian distances
- Divergences
- Optimal transport distances
- Connections and unifying formulations

to infinite-dimensional Gaussian measures and Gaussian processes

Motivations for studying geometrical structures of Gaussian measures and SPD matrices

- Central role in statistics, probability, machine learning
- Numerous practical applications
  - Brain imaging (Arsigny et al 2005, Dryden et al 2009, Qiu et al 2015, Zhou et al 2016, Ning 2018)
  - Computer vision: object detection (Tuzel et al 2008, Tosato et al 2013), image retrieval (Cherian et al 2013), visual recognition (Jayasumana et al 2015), person re-identification (Devyatkov et al 2018, Matsukawa 2019)
  - Radar signal processing: Barbaresco (2013), Formont et al 2013, Braca et al 2018, Aubry et al 2018
  - Brain Computer Interfaces (BCI) Li et al 2011, Barachant et al 2013, Uehara et al 2015, Congedo et al 2017, Rodrigues et al 2018, Yair et al 2019
  - Many more applications and references...

Fisher-Rao metric - Affine-invariant Riemannian metric

$\mathrm{Sym}^{++}(n)$ = set of $n \times n$ SPD matrices

- Multivariate zero-mean Gaussian densities on $\mathbb{R}^n \Longleftrightarrow \mathrm{Sym}^{++}(n)$

$$\mathcal{S} = \left\{ P(x;\theta) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma(\theta))}} \exp\left(-\frac{1}{2}x^T\Sigma(\theta)^{-1}x\right), \theta \in \Theta \right\}$$

$$\Theta = \left\{ \theta = [\theta^1, \ldots, \theta^k], k = \frac{n(n+1)}{2} : \Sigma(\theta) \in \mathrm{Sym}^{++}(n) \right\}$$

- Fisher information matrix

$$g_{ij}(\theta) = \int_{\mathbb{R}^n} \frac{\partial \ln P(x;\theta)}{\partial \theta^i} \frac{\partial \ln P(x;\theta)}{\partial \theta^j} P(x;\theta)dx$$

- This defines a Riemannian metric on $\mathcal{S}$, so-called *Fisher-Rao metric*, or *Fisher information metric*

Fisher-Rao metric - Affine-invariant Riemannian metric

- Fisher-Rao metric: central element in Information Geometry (Amari 1985, Amari & Nagaoka 2000, Amari 2016)
- Explicit expression for Fisher-Rao metric on $\mathcal{S}$

$$g_{ij}(\theta) = \frac{1}{2}\mathrm{tr}\left[\Sigma^{-1}\left(\frac{\partial}{\partial\theta^i}\Sigma\right)\Sigma^{-1}\left(\frac{\partial}{\partial\theta^j}\Sigma\right)\right]$$

- Corresponds to the affine-invariant Riemannian metric on $\mathrm{Sym}^{++}(n)$

$$\begin{aligned}\langle A, B\rangle_\Sigma &= \frac{1}{2}\langle \Sigma^{-1/2}A\Sigma^{-1/2}, \Sigma^{-1/2}B\Sigma^{-1/2}\rangle_F \\ &= \frac{1}{2}\mathrm{tr}(\Sigma^{-1}A\Sigma^{-1}B), \quad A, B \in \mathrm{Sym}(n), \Sigma \in \mathrm{Sym}^{++}(n)\end{aligned}$$

Geometry of SPD Matrices - Riemannian manifold viewpoint

- Affine-invariant Riemannian metric (e.g. Pennec et al 2006, Bhatia 2007)
- Unique geodesic joining $A, B \in \mathrm{Sym}^{++}(n)$

$$\gamma_{AB}(t) = A^{1/2}(A^{-1/2}BA^{-1/2})^t A^{1/2}$$
$$\gamma_{AB}(0) = A, \quad \gamma_{AB}(1) = B$$

- Riemannian (geodesic) distance

$$d_{\mathrm{aiE}}(A, B) = ||\log(A^{-1/2}BA^{-1/2})||_F$$

- Rich mathematical structures
- Corresponds to the Fisher-Rao distance between zero-mean Gaussian measures on $\mathbb{R}^n$

Geometry of SPD Matrices - Riemannian manifold viewpoint

Log-Euclidean metric (Arsigny et al 2007)

- Unique geodesic joining $A, B \in \mathrm{Sym}^{++}(n)$

$$\gamma_{AB}(t) = \exp[(1-t)\log(A) + t\log(B)], \quad \gamma(0) = A, \gamma(1) = B$$

- Riemannian (geodesic) distance

$$d_{\mathrm{logE}}(A, B) = ||\log(A) - \log(B)||_F$$

- Faster to compute than affine-invariant distance
- Lead to positive definite kernels, e.g. Gaussian kernel

$$K(A, B) = \exp(-||\log(A) - \log(B)||^2/\sigma^2)$$

Infinite-dimensional generalization of Riemannian distances

- Substantially different from the finite-dimensional formulation
- Problems: for A an SPD matrix

$$A = U\mathrm{diag}(\lambda_1, \ldots, \lambda_n)U^T,$$
$$\log(A) = U\mathrm{diag}(\log\lambda_1, \ldots, \log\lambda_n)U^T$$

If A is a strictly positive Hilbert-Schmidt operator

1. Eigenvalues $\lambda_k \to 0$ as $k \to \infty$
2. $\frac{1}{\lambda_k} \to \infty$ and $\log(\lambda_k) \to -\infty$
3. A^{-1} is unbounded
4. $\log(A)$ is unbounded

Infinite-dimensional generalization of $\mathrm{Sym}^{++}(n)$

- Generalizing the Log-Euclidean distance

$$d_{\mathrm{logE}}(A, B) = ||\log(A) - \log(B)||_F, \quad A, B \in \mathrm{Sym}^{++}(n)$$

  to the setting where A, B are self-adjoint, positive Hilbert-Schmidt operators on a separable Hilbert space $\mathcal{H}$

- Two issues to consider
  1. Generalization of the principal matrix $\log$ function
  2. Generalization of the Frobenius inner product and norm (the Hilbert-Schmidt norm is not sufficient)

First problem: unboundedness of $\log(A)$ since $\lim_{k\to\infty} \lambda_k = 0$

$$\log(A) = \sum_{k=1}^{\infty} \log(\lambda_k)(\mathbf{u}_k \otimes \mathbf{u}_k), \quad \lim_{k\to\infty} \log(\lambda_k) = -\infty$$

Note: $\mathbf{u}_k \otimes \mathbf{u}_k$ is the generalization of the product $\mathbf{u}_k\mathbf{u}_k^T$ in $\mathbb{R}^n$

Resolution: Regularization with $\gamma \in \mathbb{R}, \gamma > 0$

$$\log(A + \gamma I) = \sum_{k=1}^{\infty} \log(\lambda_k + \gamma)(\mathbf{u}_k \otimes \mathbf{u}_k),$$

$$\lim_{k\to\infty} \log(\lambda_k + \gamma) = \log(\gamma)$$

so $\log(A + \gamma I)$ is bounded

Infinite-dimensional generalization of $\mathrm{Sym}^{++}(n)$

Consider the generalization

$$||\log(A) - \log(B)||_F \to ||\log(A + \gamma I) - \log(B + \nu I)||_{\mathrm{HS}}$$

Second problem: The identity operator I is not Hilbert-Schmidt:

$$||I||_{\mathrm{HS}} = \mathrm{tr}(I) = \infty$$

For $\gamma \neq 1$

$$||\log(A + \gamma I)||^2_{\mathrm{HS}} = \sum_{k=1}^{\infty} [\log(\lambda_k + \gamma)]^2 = \infty$$

For $A = B = 0, \gamma \neq \nu$

$$d(\gamma I, \nu I) = ||\log(\gamma/\nu) I||_{\mathrm{HS}} = |\log(\gamma/\nu)|\ ||I||_{\mathrm{HS}} = \infty$$

Infinite-dimensional generalization of $\mathrm{Sym}^{++}(n)$

- Second problem: The identity I is not Hilbert-Schmidt
- Resolution: Extended Hilbert-Schmidt norm (Larotonda, *Differential Geometry and Its Applications*, 2007)

$$||A + \gamma I||^2_{\mathrm{HS_X}} = ||A||^2_{\mathrm{HS}} + \gamma^2$$

- Extended Hilbert-Schmidt inner product

$$\langle A + \gamma I, B + \nu I \rangle_{\mathrm{HS_X}} = \langle A, B \rangle_{\mathrm{HS}} + \gamma\nu$$

i.e. the scalar operators γI are orthogonal to the Hilbert-Schmidt operators

$$||A + \gamma I||^2_{\mathrm{HS_X}} = ||A||^2_{\mathrm{HS}} + \gamma^2, \quad ||I||_{\mathrm{HS_X}} = 1$$

Geometry of positive definite Hilbert-Schmidt operators

- Larotonda (*Differential Geometry and Its Applications* 2007): generalization of the manifold $\mathrm{Sym}^{++}(n)$ of SPD matrices to the infinite-dimensional Hilbert manifold

$$\Sigma(\mathcal{H}) = \{A + \gamma I > 0 : A^* = A, A \in \mathrm{HS}(\mathcal{H}), \gamma \in \mathbb{R}\}$$

- Hilbert-Schmidt operators on the Hilbert space $\mathcal{H}$

$$\mathrm{HS}(\mathcal{H}) = \{A : ||A||^2_{\mathrm{HS}} < \infty\}$$

- A self-adjoint $||A||^2_{\mathrm{HS}} = \sum_{k=1}^{\infty} \lambda_k^2$
- Generalization of the affine-invariant Riemannian metric

Affine-invariant Riemannian metric - Infinite-dimensional generalization

- Larotonda (*Differential Geometry and Its Applications* 2007)
- Tangent space $T_P(\Sigma(\mathcal{H})) \cong \mathrm{HS}_X(\mathcal{H}) \cap \mathrm{Sym}(\mathcal{H}) \quad \forall P \in \Sigma(\mathcal{H})$
- Riemannian metric: For $P \in \Sigma(\mathcal{H})$

$$
\begin{aligned}
&\langle (A+\gamma I), (B+\nu I)\rangle_P \\
&= \langle P^{-1/2}(A+\gamma I)P^{-1/2}, P^{-1/2}(B+\nu I)P^{-1/2}\rangle_{\mathrm{HS_X}}
\end{aligned}
$$

- Riemannian (geodesic) distance

$$
\begin{aligned}
&d_{\mathrm{aiHS}}[(A+\gamma I), (B+\nu I)] \\
&= ||\log[(B+\nu I)^{-1/2}(A+\gamma I)(B+\nu I)^{-1/2}]||_{\mathrm{HS_X}}
\end{aligned}
$$

- Related work: Lawson and Lim (PNAS, 2013), Pálfia (Advances in Math., 2016): means of positive operators

Log-Hilbert-Schmidt distance

Generalizing Log-Euclidean distance $d_{\mathrm{logE}}(A, B) = ||\log(A) - \log(B)||$

- Log-Hilbert-Schmidt distance (H.Q.Minh et al 2014)

$$d_{\mathrm{logHS}}[(A + \gamma I), (B + \nu I)] = ||\log(A + \gamma I) - \log(B + \nu I)||_{\mathrm{HS_X}}$$

- Log-Hilbert-Schmidt inner product

$$\langle (A + \gamma I), (B + \nu I)\rangle_{\mathrm{logHS}} = \langle \log(A + \gamma I), \log(B + \nu I)\rangle_{\mathrm{HS_X}}$$

- All quantities are guaranteed to be finite

Computation of distances and divergences - RKHS methodology

1. Distances/divergences between RKHS covariance operators
2. Distances/divergences between Gaussian processes and covariance operators of stochastic processes in general
3. Both involve RKHS methodology

Reproducing kernel Hilbert space (RKHS) setting

- K = positive definite kernels on $\mathcal{X} \times \mathcal{X}$
- $\mathcal{H}_K$ = corresponding RKHS (reproducing kernel Hilbert space)
- Positive definite kernel K on $\mathcal{X} \times \mathcal{X}$ induces canonical feature map $\Phi : \mathcal{X} \to \mathcal{H}_K$, $K_x : \mathcal{X} \to \mathbb{R}$, $K_x(t) = K(x,t)$

$$\Phi(x) = K_x \in \mathcal{H}_K, \quad \mathcal{H}_K \text{ = feature space}$$

$$\langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}_K} = \langle K_x, K_y \rangle_{\mathcal{H}_K} = K(x,y)$$

- Assume ρ = Borel probability distribution on $\mathcal{X}$, with

$$\int_{\mathcal{X}} ||\Phi(x)||^2_{\mathcal{H}_K} d\rho(x) = \int_{\mathcal{X}} K(x,x) d\rho(x) < \infty$$

RKHS mean vector and covariance operator

- $\mathbf{X} = [x_1, \ldots, x_m]$ = data matrix randomly sampled from $\mathcal{X}$ according to ρ, with m observations
- Informally, Φ gives an infinite feature matrix in the feature space $\mathcal{H}_K$, of size $\dim(\mathcal{H}_K) \times m$

$$\Phi(\mathbf{X}) = [\Phi(x_1), \ldots, \Phi(x_m)]$$

- Formally, $\Phi(\mathbf{X}) : \mathbb{R}^m \to \mathcal{H}_K$ is the bounded linear operator

$$\Phi(\mathbf{X})w = \sum_{i=1}^{m} w_i \Phi(x_i), \quad w \in \mathbb{R}^m$$

RKHS mean vector and covariance operator

- Theoretical RKHS mean

$$\mu_\Phi = \int_{\mathcal{X}} \Phi(x) d\rho(x) \in \mathcal{H}_K$$

- Empirical RKHS mean

$$\mu_{\Phi(\mathbf{X})} = \frac{1}{m}\sum_{i=1}^{m} \Phi(x_i) = \frac{1}{m}\Phi(\mathbf{X})\mathbf{1}_m \in \mathcal{H}_K$$

- Linear kernel $K(x, y) = \langle x, y \rangle$ on $\mathbb{R}^d$: $\mu_{\mathbf{X}} = \frac{1}{m}\sum_{i=1}^{m} x_i$

RKHS mean vector and covariance operator

- Theoretical covariance operator $C_\Phi : \mathcal{H}_K \to \mathcal{H}_K$

$$C_\Phi = \int_{\mathcal{X}} \Phi(x) \otimes \Phi(x) d\rho(x) - \mu_\Phi \otimes \mu_\Phi$$

- Empirical covariance operator $C_{\Phi(\mathbf{x})} : \mathcal{H}_K \to \mathcal{H}_K$

$$C_{\Phi(\mathbf{X})} = \frac{1}{m} \sum_{i=1}^{m} \Phi(x_i) \otimes \Phi(x_i) - \mu_{\Phi(\mathbf{X})} \otimes \mu_{\Phi(\mathbf{X})}$$
$$= \frac{1}{m} \Phi(\mathbf{X}) J_m \Phi(\mathbf{X})^*$$

$J_m = I_m - \frac{1}{m}\mathbf{1}_m \mathbf{1}_m^T$ = centering matrix

- Linear kernel $K(x,y) = \langle x, y \rangle$ on $\mathbb{R}^d$: $C_{\mathbf{X}} = \frac{1}{m}\mathbf{X} J_m \mathbf{X}^T$ (sample covariance matrix)

Log-Hilbert-Schmidt distance between RKHS covariance operators

The distance

$$
\begin{aligned}
& d_{\text{logHS}}[(C_{\Phi(\mathbf{X})} + \gamma I_{\mathcal{H}_K}), (C_{\Phi(\mathbf{Y})} + \nu I_{\mathcal{H}_K})] \\
&= d_{\text{logHS}}\left[\left(\frac{1}{m}\Phi(\mathbf{X})J_m\Phi(\mathbf{X})^* + \gamma I_{\mathcal{H}_K}\right), \left(\frac{1}{m}\Phi(\mathbf{Y})J_m\Phi(\mathbf{Y})^* + \nu I_{\mathcal{H}_K}\right)\right]
\end{aligned}
$$

has a closed form in terms of $m \times m$ Gram matrices

$$
\begin{aligned}
K[\mathbf{X}] &= \Phi(\mathbf{X})^*\Phi(\mathbf{X}), (K[\mathbf{X}])_{ij} = K(x_i, x_j), \\
K[\mathbf{Y}] &= \Phi(\mathbf{Y})^*\Phi(\mathbf{Y}), (K[\mathbf{Y}])_{ij} = K(y_i, y_j), \\
K[\mathbf{X}, \mathbf{Y}] &= \Phi(\mathbf{X})^*\Phi(\mathbf{Y}), (K[\mathbf{X}, \mathbf{Y}])_{ij} = K(x_i, y_j) \\
K[\mathbf{Y}, \mathbf{X}] &= \Phi(\mathbf{Y})^*\Phi(\mathbf{X}), (K[\mathbf{Y}, \mathbf{x}])_{ij} = K(y_i, x_j)
\end{aligned}
$$

Example: Log-Hilbert-Schmidt distance between RKHS covariance operators

Theorem (H.Q.M. et al - NIPS 2014)

Assume that $\dim(\mathcal{H}_K) = \infty$*. Let* $\gamma > 0$*,* $\nu > 0$*. The Log-Hilbert-Schmidt distance between* $(C_{\Phi(\mathbf{X})} + \gamma I_{\mathcal{H}_K})$ *and* $(C_{\Phi(\mathbf{Y})} + \nu I_{\mathcal{H}_K})$ *is*

$$d^2_{\text{logHS}}[(C_{\Phi(\mathbf{X})} + \gamma I_{\mathcal{H}_K}), (C_{\Phi(\mathbf{Y})} + \nu I_{\mathcal{H}_K})] = \text{tr}[\log(I_{N_A} + \Sigma_A)]^2 + \text{tr}[\log(I_{N_B} + \Sigma_B)]^2 - 2C_{AB} + (\log\gamma - \log\nu)^2$$

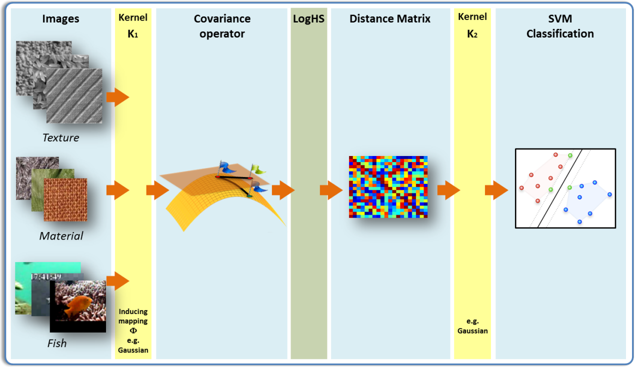
$$\frac{1}{\gamma m} J_m K[\mathbf{X}] J_m = U_A \Sigma_A U_A^T, \quad \frac{1}{\nu m} J_m K[\mathbf{Y}] J_m = U_B \Sigma_B U_B^T,$$

$$A^* B = \frac{1}{\sqrt{\gamma\nu} m} J_m K[\mathbf{X}, \mathbf{Y}] J_m,$$

$$C_{AB} = \mathbf{1}_{N_A}^T \log(I_{N_A} + \Sigma_A)\Sigma_A^{-1}(U_A^T A^* B U_B \circ U_A^T A^* B U_B)\Sigma_B^{-1}\log(I_{N_B} + \Sigma_B)\mathbf{1}_{N_B}$$

Log-Hilbert-Schmidt distance between RKHS covariance operators

Two-layer kernel machine for image classification
Distances are expressed in terms of kernel Gram matrices



[1] H.Q.M., San Biagio, Murino. Log-Hilbert-Schmidt metric between positive definite operators on Hilbert spaces, NIPS 2014

[2] H.Q.M., San Biagio, Bazzani, Murino. Approximate Log-Hilbert-Schmidt distances between covariance operators for image classification, CVPR 2016

Outline

- Riemannian distances
- Divergences
  - Finite-dimensional setting
  - Infinite-dimensional generalizations
- Optimal transport distances
- Distances/divergences between Gaussian processes

Log-Determinant divergences - finite-dimensional setting

Convex cone viewpoint of $\mathrm{Sym}^{++}(n)$

- Alpha Log-Determinant divergences (Chebbi and Moakher, 2012)

$$d^{\alpha}_{\mathrm{logdet}}(A,B) = \frac{4}{1-\alpha^2}\log\frac{\det(\frac{1-\alpha}{2}A + \frac{1+\alpha}{2}B)}{\det(A)^{\frac{1-\alpha}{2}}\det(B)^{\frac{1+\alpha}{2}}}, \quad -1 < \alpha < 1$$

- Limiting cases

$$d^{1}_{\mathrm{logdet}}(A,B) = \lim_{\alpha\to 1} d^{\alpha}_{\mathrm{logdet}}(A,B) = \mathrm{tr}(B^{-1}A - I) - \log\det(B^{-1}A)$$

$$d^{-1}_{\mathrm{logdet}}(A,B) = \lim_{\alpha\to -1} d^{\alpha}_{\mathrm{logdet}}(A,B) = \mathrm{tr}(A^{-1}B - I) - \log\det(A^{-1}B)$$

- Correspond to Rényi and Kullback-Leibler (KL) divergences between zero-mean Gaussian measures on $\mathbb{R}^n$

Infinite-dimensional generalizations

- Trace class operators $\mathrm{Tr}(\mathcal{H}) = \{A : ||A||_{\mathrm{tr}} = \mathrm{tr}|A| < \infty\}$
- A self-adjoint, compact, $||A||_{\mathrm{tr}} = \sum_{k=1}^{\infty} |\lambda_k|$, $\mathrm{tr}(A) = \sum_{k=1}^{\infty} \lambda_k$
- Covariance operators are trace class operators
- For a strictly positive compact operator A on a Hilbert space $\mathcal{H}$, with eigenvalues $\{\lambda_k\}_{k=1}^{\infty}$

$$\log\det(A) = \mathrm{tr}\log(A) = \sum_{k=1}^{\infty} \log(\lambda_k) = -\infty$$

- Need to define properly consider the set of operators A and extend the functions tr and $\det$

Infinite-Dimensional Log-Determinant Divergences

(H.Q.M. Linear Algebra and App. 2017)

- tr_X: extended trace class operators and extended trace

$$\mathrm{tr}_X(A + \gamma I) = \mathrm{tr}(A) + \gamma \quad \mathrm{tr}(I) = \infty$$

- $\det$: Fredholm determinant

$$\det(A + I) = \prod_{k=1}^{\infty}(1 + \lambda_k) = \exp[\mathrm{tr}\log(A + I)]$$

- $\det_X$: extended Fredholm determinant

$$\det_X(A + \gamma I) = \gamma \det[(A/\gamma) + I] = \exp[\mathrm{tr}_X \log(A + \gamma I)]$$

Infinite-Dimensional Log-Determinant Divergences

Alpha Log-Determinant divergences (H.Q.M., Linear Algebra and App., 2017)

$$d^{\alpha}_{\mathrm{logdet}}[(A+\gamma I),(B+\gamma I)] \qquad \text{(simplified version)}, -1 < \alpha < 1$$
$$= \frac{4}{1-\alpha^2}\log\left[\frac{\mathrm{det_X}\left(\frac{1-\alpha}{2}(A+\gamma I)+\frac{1+\alpha}{2}(B+\gamma I)\right)}{\mathrm{det_X}(A+\gamma I)^{\frac{1-\alpha}{2}}\mathrm{det_X}(B+\gamma I)^{\frac{1+\alpha}{2}}}\right],$$

- $A+\gamma I > 0, B+\gamma I > 0$: A, B = trace class operators
- $\mathrm{det_X}$: extended Fredholm determinant
- Closed form formulas in RKHS setting
- Related work in RKHS setting: Zhou & Chellapa (PAMI 2006), Harandi et al (CVPR 2014) (valid in finite-dimension)

Infinite-Dimensional Log-Determinant Divergences

H.Q.Minh (2020). Regularized divergences between covariance operators and Gaussian measures on Hilbert spaces, Journal of Theoretical Probability

Alpha Log-Determinant divergences $\Longleftrightarrow$ Rény divergences

$\alpha = 1 \Longleftrightarrow$ Kullback-Leibler (KL) divergence

- On $\mathbb{R}^d$, two Gaussian densities μ_0, μ_1 are always equivalent, $\mu_0 \sim \mu_1$ (have the same support)
- Feldman-Hajek Theorem: on $\mathcal{H}$, $\dim(\mathcal{H}) = \infty$, two Gaussian measures μ_0, μ_1 are either equivalent or mutually singular, $\mu_0 \perp \mu_1$ (have disjoint support)

$$\mu_0 \perp \mu_1 \Rightarrow \mathrm{KL}(\mu_0 || \mu_1) = \infty$$

Infinite-Dimensional Log-Determinant Divergences

Theorem (H.Q.M. 2020)

Consider two equivalent Gaussian measures $\mathcal{N}(m, C), \mathcal{N}(m, C_0)$ on a Hilbert space $\mathcal{H}$. Let S be a self-adjoint Hilbert-Schmidt operator on $\mathcal{H}$ such that $C = C_0^{1/2}(I - S)C_0^{1/2}$, then

$$\lim_{\gamma \to 0} d^1_{\mathrm{logdet}}[(C + \gamma I), (C_0 + \gamma I)] = 2D_{\mathrm{KL}}(\mathcal{N}(m, C) || \mathcal{N}(m, C_0))$$
$$= -\log \det_2(I - S)$$

$\det_2$ = Hilbert-Carleman determinant

$$\det_2(I + A) = \det[(I + A)\exp(-A)], \quad A \in \mathrm{HS}(\mathcal{H})$$

Infinite-Dimensional Log-Determinant Divergences

H.Q.Minh. Alpha-Beta Log-Determinant divergences between positive definite Hilbert-Schmidt operators (Geometric Science of Information 2017, Information Geometry 2019, Positivity 2020)

- General formulation encompassing
  - Alpha Log-Determinant divergences
  - Affine-invariant Riemannian distance
- Employs extended Hilbert-Carleman determinant
- These divergences all induce the Affine-invariant Riemannian metric
- Closed form formulas in RKHS setting

Outline

- Riemannian distances
- Divergences
- Optimal transport distances and related geometrical structures
- Entropic regularization of optimal transport
  1. Finite-dimensional Gaussian setting
  2. Infinite-dimensional Gaussian setting
- Distances/divergences between Gaussian processes

Optimal Transport distances between probability measures

- (X, d) = complete separable metric space (e.g. $X = \mathbb{R}^n$)
- $c : X \times X \to \mathbb{R}_{\geq 0}$ = lower semi-continuous *cost function* (e.g. $c(x, y) = ||x - y||^2$ for $X = \mathbb{R}^n$)
- $\mathcal{P}(X)$ = set of all probability measures on X.
- The *optimal transport (OT)* problem between two probability measures $\nu_0, \nu_1 \in \mathcal{P}(X)$ is (Villani 2009, 2016)

$$\mathrm{OT}(\nu_0, \nu_1) = \min_{\gamma \in \mathrm{Joint}(\nu_0, \nu_1)} \mathbb{E}_\gamma[c] = \min_{\gamma \in \mathrm{Joint}(\nu_0, \nu_1)} \int_{X \times X} c(x, y) d\gamma(x, y)$$

- $\mathrm{Joint}(\nu_0, \nu_1)$ is the set of joint probabilities with marginals ν_0 and ν_1

Optimal transport distances

- $\mathcal{P}_p(X)$ = set of all probability measures μ on X of finite moment of order p, $1 \leq p < \infty$, i.e.

$$\int_X d^p(x_0, x) d\mu(x) < \infty \text{ for some (any) } x_0 \in X.$$

- *p-Wasserstein distance* W_p between ν_0 and ν_1

$$W_p(\nu_0, \nu_1) = \mathrm{OT}_{d^p}(\nu_0, \nu_1)^{\frac{1}{p}}.$$

- This distance defines a metric on $\mathcal{P}_p(X)$
- Takes into account the geometry of the underlying space X (via the distance d)

Gaussian setting - Bures-Wasserstein distance

- For two multivariate Gaussian distributions $\mu_i = \mathcal{N}(m_i, C_i)$, $i = 0, 1$, on $\mathbb{R}^n$, with the square cost function

$$W_2^2(\mu_0, \mu_1) = \min_{\gamma \in \text{Joint}(\mu_0, \mu_1)} \int_{\mathbb{R}^n \times \mathbb{R}^n} ||x - y||^2 d\gamma(x, y)$$

- $W_2(\mu_0, \mu_1)$ admits the following closed form (Dowson & Landau 1982, Olkin & Pukelsheim 1982, Givens & Shortt 1984)

$$W_2^2(\mu_0, \mu_1) = \|m_0 - m_1\|^2 + \text{tr}(C_0) + \text{tr}(C_1) - 2\text{tr}\left(C_1^{1/2} C_0 C_1^{1/2}\right)^{1/2}.$$

- Bures-Wasserstein distance between SPD matrices: $m_0 = m_1$,

$$d_{\text{BW}}^2(C_0, C_1) = \text{tr}(C_0) + \text{tr}(C_1) - 2\text{tr}\left(C_1^{1/2} C_0 C_1^{1/2}\right)^{1/2}.$$

Wasserstein Riemannian geometry of SPD matrices

- Riemannian metric: For each $P \in \mathrm{Sym}^{++}(n)$ and each pair $Y, Z \in T_P(\mathrm{Sym}^{++}(n)) \cong \mathrm{Sym}(n)$ (Takatsu 2011)

$$\langle Y, Z\rangle_P = \mathrm{tr}[\mathcal{L}_P(Y)P\mathcal{L}_P(Z)]$$

  where $\mathcal{L}_P(Y) = X \in \mathrm{Sym}(n)$ is the unique solution of

$$\text{Lyapunov equation} \quad XP + PX = Y$$

- Riemannian distance (Bures-Wasserstein distance) is the length of the geodesic

$$\gamma(t) = (1-t)^2A + t^2B + t(1-t)[(AB)^{1/2} + (BA)^{1/2}]$$

Infinite-dimensional Wasserstein distance

- $\mathcal{L}^2$-Wasserstein distance between two Gaussian measures $\mu_i = \mathcal{N}(m_i, C_i)$, $i = 0, 1$, on an infinite-dimensional Hilbert space $\mathcal{H}$ (Gelbrich 1990)

$$W_2^2(\mu_0, \mu_1) = ||m_0 - m_1||^2 + \mathrm{tr}[C_0 + C_1 - 2(C_0^{1/2} C_1 C_0^{1/2})^{1/2}]^{1/2}$$

- Same expression as in the finite-dimensional case
- Some recent work on Gaussian processes: Mallasto and Feragen (NIPS2017), Masarotto, Panaretos and Zemel (Sankhya 2019)
- Bures-Wasserstein distance between two covariance operators

$$d_{\mathrm{BW}}(A, B) = (\mathrm{tr}[A + B - 2(A^{1/2} B A^{1/2})^{1/2}])^{1/2}$$

- Valid for singular covariance operators
- Not Fréchet differentiable

Entropic regularization of optimal transport

- Exact optimal transport distances generally computationally demanding
- Exact Wasserstein distance $\mathcal{W}_p$ can have bad sample complexity (worst case exponentially $O(n^{-1/d})$ in $\mathbb{R}^d$ (Dudley 1969, Weed,Bach 2019)
- Entropic regularization (Cuturi 2013)

$$\mathrm{OT}_c^{\epsilon}(\mu,\nu) = \min_{\gamma \in \mathrm{Joint}(\mu,\nu)} \{\mathbb{E}_{\gamma}[c] + \epsilon \mathrm{KL}(\gamma || \mu \otimes \nu)\},$$

- $\mathrm{KL}(\nu||\mu)$ = Kullback-Leibler divergence between ν and μ
- Equivalent to the classical Schrödinger Bridge Problem (Schrödinger 1931)
- Optimization problem can be solved efficiently using Sinkhorn algorithm

Entropic regularization of optimal transport

- Biased: $\mathrm{OT}^{\epsilon}_{d^p}(\mu, \mu) \neq 0$ (neither a distance nor divergence)
- *Sinkhorn divergence* (Genevay et al 2018, Feydy et al 2019)

$$S^{\epsilon}_p(\mu, \nu) = \mathrm{OT}^{\epsilon}_{d^p}(\mu, \nu) - \frac{1}{2}(\mathrm{OT}^{\epsilon}_{d^p}(\mu, \mu) + \mathrm{OT}^{\epsilon}_{d^p}(\nu, \nu)).$$

- Much research interest recently,e.g. Sommerfeld 2017, Ripani 2017, Mena, Niles-Weed (2019), Gerolin et al 2019
- Some recent applications: learning generative models (Genevay et al 2018), Sinkhorn autoencoders (Patrini et al 2019), density functional theory in chemistry (Gerolin et al 2019)

Entropic regularization - Gaussian setting

For $\mu_i = \mathcal{N}(m_i, C_i)$, $i = 0, 1$, on $\mathbb{R}^n$,

$$\mathrm{OT}^{\epsilon}_{d^2}(\mu_0, \mu_1) = \min_{\gamma \in \mathrm{Joint}(\mu_0, \mu_1)} \left\{ \mathbb{E}_{\gamma} ||x - y||^2 + \epsilon \mathrm{KL}(\gamma || \mu_0 \otimes \mu_1) \right\}$$

- Janati, Muzellec, Peyré, and M. Cuturi (2020), Mallasto, Gerolin, Minh (2020), del Barrio, Loubes (2020)
- Mutual information $\mathrm{KL}(\gamma || \mu_0 \otimes \mu_1) = H(\mu_0) + H(\mu_1) - H(\gamma)$
- $H(X) = -\int_{\mathbb{R}^n} \log[f_X(x)] f_X(x) dx$ is the *differential entropy*
- Maximum Entropy property of Gaussian densities: if X has mean zero and covariance matrix C, then

$$H(X) \leq \frac{1}{2} \log[(2\pi e)^n \det(C)], \text{with equality if and only if } X \sim \mathcal{N}(0, C).$$

Entropic regularization - Gaussian setting

For $\mu_i = \mathcal{N}(m_i, C_i)$, $i = 0, 1$, on $\mathbb{R}^n$,

$$(*)\ \mathrm{OT}^{\epsilon}_{d^2}(\mu_0, \mu_1) = \min_{\gamma \in \mathrm{Joint}(\mu_0,\mu_1)} \left\{ \mathbb{E}_{\gamma}||x - y||^2 + \epsilon \mathrm{KL}(\gamma || \mu_0 \otimes \mu_1) \right\}$$

- Maximum entropy of Gaussian densities: $\mathrm{KL}(\gamma||\mu_0 \otimes \mu_1)$ is minimum if and only if γ is a joint Gaussian density of μ_0 and μ_1
- A minimizer γ of $(*)$ must be a joint Gaussian density
  $\gamma = \mathcal{N}\left(\begin{pmatrix} m_0 \\ m_1 \end{pmatrix}, \Gamma\right)$, $\Gamma = \begin{pmatrix} C_0 & C \\ C^T & C_1 \end{pmatrix}$, C = *cross-covariance matrix*
  $\mathrm{KL}(\gamma||\mu_0 \otimes \mu_1) = \frac{1}{2}\log\left(\frac{\det(C_0)\det(C_1)}{\det(\Gamma)}\right)$

Entropic regularization - Gaussian setting

For $\mu_i = \mathcal{N}(m_i, C_i)$, $i = 0, 1$, both $\mathrm{OT}^{\epsilon}_{d^2}(\mu_0, \mu_1)$ and $\mathrm{S}^{\epsilon}_2(\mu_0, \mu_1)$ admit closed form formulas. Let $N^{\epsilon}_{ij} = I + \left(I + \frac{16}{\epsilon^2} C_i^{\frac{1}{2}} C_j C_i^{\frac{1}{2}}\right)^{\frac{1}{2}}$, $i, j = 0, 1$, then

$$\mathrm{OT}^{\epsilon}_{d^2}(\mu_0, \mu_1) = \|m_0 - m_1\|^2 + \mathrm{Tr}(C_0) + \mathrm{Tr}(C_1) - \frac{\epsilon}{2}\left[\mathrm{Tr}(N^{\epsilon}_{01}) - \log\det(N^{\epsilon}_{01}) + n\log 2 - 2n\right],$$

$$S^{\epsilon}_2(\mu_0, \mu_1) = \|m_0 - m_1\|^2_2 + \frac{\epsilon}{4}\left(\mathrm{Tr}\left(N^{\epsilon}_{00} - 2N^{\epsilon}_{01} + N^{\epsilon}_{11}\right) + \log\left(\frac{\det^2(N^{\epsilon}_{01})}{\det(N^{\epsilon}_{00})\det(N^{\epsilon}_{11})}\right)\right).$$

The unique minimizer γ (optimal transport plan) is a joint Gaussian measure of μ_0 and μ_1.

From finite to infinite-dimensional setting - entropic regularization

- The entropy $H(X) = \frac{1}{2}\log[(2\pi e)^d \det(C)]$ does **not** generalize to infinite dimension ($\det(C) = \prod_{k=1}^{\infty}\lambda_k$, $\lim_{k\to\infty}\lambda_k = 0$)
- For two Gaussian measures $\mathcal{N}(m_i, C_i)$, $i = 0, 1$ and their joint Gaussian measure γ,

$$\gamma = \mathcal{N}\left(\begin{pmatrix} m_0 \\ m_1 \end{pmatrix}, \Gamma\right), \Gamma = \begin{pmatrix} C_0 & C \\ C^* & C_1 \end{pmatrix}, C = \textit{cross-covariance operator}$$

- The right hand side of the following expression is **not** well-defined

$$\mathrm{KL}(\gamma || \mu_0 \otimes \mu_1) = \frac{1}{2}\log\left(\frac{\det(C_0)\det(C_1)}{\det(\Gamma)}\right)$$

- However, the mutual information $\mathrm{KL}(\gamma || \mu_0 \otimes \mu_1)$ is well-defined (can be infinite)

Theorem (**Minimum Mutual Information of Joint Gaussian Measures**)

Let $\mathcal{H}_1, \mathcal{H}_2$ *be two separable Hilbert spaces. Let* $\mu_X = \mathcal{N}(m_X, C_X) \in \mathrm{Gauss}(\mathcal{H}_1)$, $\mu_Y = \mathcal{N}(m_Y, C_Y) \in \mathrm{Gauss}(\mathcal{H}_2)$, $\ker(C_X) = \ker(C_Y) = \{0\}$. *Let* $\gamma \in \mathrm{Joint}(\mu_X, \mu_Y)$, $\gamma_0 \in \mathrm{Gauss}(\mu_X, \mu_Y)$, γ_0 *is equivalent to* $\mu_X \otimes \mu_Y$. *Assume that* γ *and* γ_0 *have the same covariance operator* Γ *and* $\mu_X \otimes \mu_Y$ *has covariance operator* Γ_0. *Then*

$$\mathrm{KL}(\gamma || \mu_X \otimes \mu_Y) \geq \mathrm{KL}(\gamma_0 || \mu_X \otimes \mu_Y) = -\frac{1}{2} \log\det(I - V^* V).$$

Equality happens if and only if $\gamma = \gamma_0$. *Here* V *is the unique bounded linear operator satisfying* $V \in \mathrm{HS}(\mathcal{H}_2, \mathcal{H}_1)$, $||V|| < 1$, *such that*

$$\Gamma = \Gamma_0^{1/2} \begin{pmatrix} I & V \\ V^* & I \end{pmatrix} \Gamma_0^{1/2}$$

Entropic regularization - Gaussian measures on Hilbert space

For $\mu_i = \mathcal{N}(m_i, C_i)$, $i = 0, 1$, on $\mathcal{H}$,

$$(**)\ \mathrm{OT}^{\epsilon}_{d^2}(\mu_0, \mu_1) = \min_{\gamma \in \mathrm{Joint}(\mu_0,\mu_1)} \left\{ \mathbb{E}_{\gamma}||x - y||^2 + \epsilon \mathrm{KL}(\gamma || \mu_0 \otimes \mu_1) \right\}$$

- A minimizer γ must satisfy $\gamma \in \mathrm{Gauss}(\mu_0, \mu_1), \gamma \sim \mu_0 \otimes \mu_1$
- **Direct solution**: problem $(**)$ is equivalent to

$$\mathrm{OT}^{\epsilon}_{d^2}(\mu_0, \mu_1) = ||m_0 - m_1||^2 + \mathrm{tr}(C_0) + \mathrm{tr}(C_1) - \max_{V \in \mathrm{HS}(\mathcal{H}), ||V|| < 1} \left\{ 2\mathrm{tr}(VC_1^{1/2}C_0^{1/2}) + \frac{\epsilon}{2} \log\det(I - V^*V) \right\}$$

- Infinite-dimensional optimization problem but can be solved for V

Entropic regularization - Gaussian measures on Hilbert space

Solution via Schrödinger system

$$(**)\ \mathrm{OT}^{\epsilon}_{d^2}(\mu_0, \mu_1) = \min_{\gamma \in \mathrm{Joint}(\mu_0,\mu_1)} \left\{ \mathbb{E}_{\gamma}||x-y||^2 + \epsilon \mathrm{KL}(\gamma || \mu_0 \otimes \mu_1) \right\}$$

- Since $\gamma \in \mathrm{Gauss}(\mu_0, \mu_1), \gamma \sim \mu_0 \otimes \mu_1$, we solve for the Radon-Nikodym density

$$\frac{d\gamma}{d(\mu_0 \otimes \mu_1)}(x,y) = \alpha^{\epsilon}(x)\beta^{\epsilon}(y)k(x,y)$$

  for $k(x,y) = \exp(-\frac{||x-y||^2}{\epsilon})$
- $\alpha^{\epsilon}, \beta^{\epsilon}$ obtained via solving the Schrödinger system

$$\alpha^{\epsilon}(x)\mathbb{E}_{\mu_1}[\beta^{\epsilon}(y)k(x,y)] = 1,$$
$$\beta^{\epsilon}(y)\mathbb{E}_{\mu_0}[\alpha^{\epsilon}(x)k(x,y)] = 1.$$

Entropic regularization - Gaussian measures on Hilbert space

Theorem (Sinkhorn divergence between Gaussian measures on Hilbert space)

Let $\mu_0 = \mathcal{N}(m_0, C_0)$ *and* $\mu_1 = \mathcal{N}(m_1, C_1)$. *For each fixed* $\epsilon > 0$,

$$\mathrm{S}_2^{\epsilon}(\mu_0, \mu_1) = ||m_0 - m_1||^2 + \frac{\epsilon}{4}\mathrm{tr}\,[M_{00}^{\epsilon} - 2M_{01}^{\epsilon} + M_{11}^{\epsilon}] + \frac{\epsilon}{4}\log\left[\frac{\det\left(I + \frac{1}{2}M_{01}^{\epsilon}\right)^2}{\det\left(I + \frac{1}{2}M_{00}^{\epsilon}\right)\det\left(I + \frac{1}{2}M_{11}^{\epsilon}\right)}\right].$$

Here $M_{ij}^{\epsilon} = -I + \left(I + \frac{16}{\epsilon^2}C_i^{1/2}C_jC_i^{1/2}\right)^{1/2}$ *is a trace class operator,* det *is the* *infinite-dimensional Fredholm determinant*,

$\lim_{\epsilon \to 0} S_2^{\epsilon}(\mu_0, \mu_1) = W_2^2(\mu_0, \mu_1)$, $\quad \lim_{\epsilon \to \infty} S_2^{\epsilon}(\mu_0, \mu_1) = ||m_0 - m_1||^2$

[1] H.Q.M. Entropic regularization of Wasserstein distance between infinite-dimensional Gaussian measures and Gaussian processes, Journal of Theoretical Probability, 2022

RKHS Gaussian measures

- ρ_1, ρ_2 = Borel probability measures on $\mathcal{X} \to$ Gaussian measures $\mathcal{N}(\mu_{\Phi,\rho_i}, C_{\Phi,\rho_i})$, $i = 1, 2$ on RKHS $\mathcal{H}_K$
- Sinkhorn divergence $\mathrm{S}_2^{\epsilon}[\mathcal{N}(\mu_{\Phi,\rho_1}, C_{\Phi,\rho_1}), \mathcal{N}(\mu_{\Phi,\rho_2}, C_{\Phi,\rho_2})]$ is well-defined with closed form formula
- $\mathbf{X} = (x_i)_{i=1}^m, \mathbf{Y} = (y_j)_{j=1}^n$ = independently sampled from $(\mathcal{X}, \rho_1)$, $(\mathcal{X}, \rho_2)$
- Empirical Sinkhorn divergence

$$\mathrm{S}_2^{\epsilon}[\mathcal{N}(\mu_{\Phi(\mathbf{X})}, C_{\Phi(\mathbf{X})}), \mathcal{N}(\mu_{\Phi(\mathbf{Y})}, C_{\Phi(\mathbf{Y})})]$$

  has closed form formula in terms of Gram matrices

Kernel Gaussian-Sinkhorn divergence as a semi-metric between Borel probability measures

Theorem

Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *be a characteristic kernel. Then, for* $0 \leq \epsilon \leq \infty$,

$$\mathrm{S}_2^{\epsilon}[\mathcal{N}(\mu_{\Phi,\rho_1}, C_{\Phi,\rho_1}), \mathcal{N}(\mu_{\Phi,\rho_2}, C_{\Phi,\rho_2})] = \mathrm{S}_2^{\epsilon}[\mathcal{N}(\mu_{\Phi,\rho_2}, C_{\Phi,\rho_2}), \mathcal{N}(\mu_{\Phi,\rho_1}, C_{\Phi,\rho_1})],$$
$$\mathrm{S}_2^{\epsilon}[\mathcal{N}(\mu_{\Phi,\rho_1}, C_{\Phi,\rho_1}), \mathcal{N}(\mu_{\Phi,\rho_2}, C_{\Phi,\rho_2})] \geq 0,$$
$$\mathrm{S}_2^{\epsilon}[\mathcal{N}(\mu_{\Phi,\rho_1}, C_{\Phi,\rho_1}), \mathcal{N}(\mu_{\Phi,\rho_2}, C_{\Phi,\rho_2})] = 0 \iff \rho_1 = \rho_2 \;\; \forall \rho_1, \rho_2 \in \mathcal{P}(\mathcal{X}).$$

Examples of characteristic kernels ($\rho \to \mu_{\Phi,\rho}$ is injective (Fukumizu et al NIPS2007)): Gaussian kernel $K(x,y) = \exp(-\frac{||x-y||^2}{\sigma^2})$, $\sigma \neq 0$, $\mathcal{X} = \mathbb{R}^d$; Laplacian kernel $K(x,y) = \exp(-a||x-y||)$, $a > 0, \mathcal{X} = \mathbb{R}^d$

For $\rho_1 = \mathcal{N}(\mu_{\Phi(\mathbf{X})}, C_{\Phi(\mathbf{X})})$, $\rho_2 = \mathcal{N}(\mu_{\Phi(\mathbf{Y})}, C_{\Phi(\mathbf{Y})})$

$$
\begin{aligned}
S_2^{\epsilon}(\rho_1,\rho_2) &= \frac{1}{m^2}\mathbf{1}_m^T K[\mathbf{X}]\mathbf{1}_m + \frac{1}{n^2}\mathbf{1}_n^T K[\mathbf{Y}]\mathbf{1}_n - \frac{2}{mn}\mathbf{1}_m^T K[\mathbf{X},\mathbf{Y}]\mathbf{1}_n \\
&+ \frac{\epsilon}{4}\mathrm{tr}\left[-I + \left(I + \frac{16}{\epsilon^2 m^2}(J_m K[\mathbf{X}] J_m)^2\right)^{1/2}\right] \\
&+ \frac{\epsilon}{4}\mathrm{tr}\left[-I + \left(I + \frac{16}{\epsilon^2 n^2}(J_n K[\mathbf{Y}] J_n)^2\right)^{1/2}\right] \\
&- \frac{\epsilon}{2}\mathrm{tr}\left[-I + \left(I + \frac{16}{\epsilon^2 mn}J_m K[\mathbf{X},\mathbf{Y}] J_n K[\mathbf{Y},\mathbf{X}] J_m\right)^{1/2}\right] \\
&+ \frac{\epsilon}{2}\log\det\left(\frac{1}{2}I + \frac{1}{2}\left(I + \frac{16}{\epsilon^2 mn}J_m K[\mathbf{X},\mathbf{Y}] J_n K[\mathbf{Y},\mathbf{X}] J_m\right)^{1/2}\right) \\
&- \frac{\epsilon}{4}\log\det\left(\frac{1}{2}I + \frac{1}{2}\left(I + \frac{16}{\epsilon^2 m^2}(J_m K[\mathbf{X}] J_m)^2\right)^{1/2}\right) \\
&- \frac{\epsilon}{4}\log\det\left(\frac{1}{2}I + \frac{1}{2}\left(I + \frac{16}{\epsilon^2 n^2}(J_n K[\mathbf{Y}] J_n)^2\right)^{1/2}\right).
\end{aligned}
$$

Limiting behavior

As $\epsilon \to \infty$ and $\epsilon \to 0$

$$\lim_{\epsilon\to\infty} S_2^{\epsilon}(\mu_0, \mu_1) = ||\mu_{\Phi(\mathbf{X})} - \mu_{\Phi(\mathbf{Y})}||^2_{\mathcal{H}_K}$$
$$= \frac{1}{m^2}\mathbf{1}_m^T K[\mathbf{X}]\mathbf{1}_m + \frac{1}{n^2}\mathbf{1}_n^T K[\mathbf{Y}]\mathbf{1}_n - \frac{2}{mn}\mathbf{1}_m^T K[\mathbf{X},\mathbf{Y}]\mathbf{1}_n.$$

Empirical squared Kernel MMD distance (Gretton et al 2006)

$$\lim_{\epsilon\to 0} S_2^{\epsilon}(\mu_0, \mu_1) = \frac{1}{m^2}\mathbf{1}_m^T K[\mathbf{X}]\mathbf{1}_m + \frac{1}{n^2}\mathbf{1}_n^T K[\mathbf{Y}]\mathbf{1}_n - \frac{2}{mn}\mathbf{1}_m^T K[\mathbf{X},\mathbf{Y}]\mathbf{1}_n$$
$$+ \frac{1}{m}\mathrm{tr}(K[\mathbf{X}]J_m) + \frac{1}{n}\mathrm{tr}(K[\mathbf{Y}]J_n)$$
$$- \frac{2}{\sqrt{mn}}\mathrm{tr}[J_m K[\mathbf{X},\mathbf{Y}]J_n K[\mathbf{Y},\mathbf{X}]J_m]^{1/2}.$$

Kernel Wasserstein Distance (Zhang et al PAMI 2020, H.Q.M GSI 2019, Linear Algebra and Its Applications 2022)

Outline

- Riemannian distances
- Divergences
- Optimal transport distances and related geometrical structures
- Entropic regularization of optimal transport
    1. Finite-dimensional Gaussian setting
    2. Infinite-dimensional Gaussian setting
- Distances/divergences between Gaussian processes

Distances/divergences between Gaussian processes

- T = compact metric space (in general σ-compact metric space)
- ν = nondegenerate Borel probability measure on T
- Gaussian process $\xi = (\xi_t)_{t \in T} = (\xi(\omega, t))_{t \in T}$ on a probability space $(\Omega, \mathcal{F}, P)$ with mean function $\mu(t)$ and covariance function $K(s,t)$

$$\mu(t) = \mathbb{E}\xi(t), \quad K(s,t) = \mathbb{E}[(\xi(s) - \mu(s))(\xi(t) - \mu(t))], \quad s, t \in T$$

- For each finite set $\mathbf{X} = (x_j)_{j=1}^m$ in T, $(\xi(., x_j))_{j=1}^m$ is a random vector distributed according to the Gaussian measure $\mathcal{N}(\mu[\mathbf{X}], K[\mathbf{X}])$ in $\mathbb{R}^m$, $\mu[\mathbf{X}] = (\mu(x_j))_{j=1}^m$, $(K[\mathbf{X}])_{ij} = K(x_i, x_j)$

Distances/divergences between Gaussian processes

$$\text{Assume } \int_T (\mu(t))^2 d\nu(t) < \infty, \quad \int_T K(t,t) d\nu(t) < \infty$$

- The sample paths of ξ are in $\mathcal{L}^2(T, \nu)$ almost surely
- If $\dim(\mathcal{H}_K) = \infty$, the sample paths are **outside** $\mathcal{H}_K$ almost surely
- There is a one-to-one correspondence[1] between measurable Gaussian process $\mathrm{GP}(\mu, K) \Longleftrightarrow \mathcal{N}(\mu, C_K)$ (Gaussian measure) on $\mathcal{H} = \mathcal{L}^2(T, \nu)$

$$(C_K f)(s) = \int_T K(s,t) f(t) d\nu(t)$$

[1] Rajput and Cambanis. Gaussian Processes and Gaussian Measures, 1972

Distances/divergences between Gaussian processes

- Any distance/divergence function D between Gaussian measures on $\mathcal{H} = \mathcal{L}^2(T, \nu)$ induces a distance/divergence function D_{GP} between Gaussian processes with paths in $\mathcal{L}^2(T, \nu)$
- Given $\xi^i = \mathrm{GP}(\mu_i, K^i)$

$$D_{\mathrm{GP}}(\xi^1, \xi^2) = D(\mathcal{N}(\mu_1, C_{K^1}), \mathcal{N}(\mu_2, C_{K^2}))$$

- Subsequently, assume $\mu_1 = \mu_2 = 0$
- Related work: Panaretos, Kraus, and Maddocks (2010), Horváth and Kokoszka (2012), Fremdt, Steinebach, Horváth, and Kokoszka (2013) (Hilbert-Schmidt distance), Pigoli, Aston, Dryden, and Secchi (2014), Mallasto and Feragen 2017, Masarotto, Panaretos and Zemel (2019) (Wasserstein distance), Matthews et al (AISTATS 2016), Sun et al (ICLR 2019) (KL divergence, functional Bayes NN)

Distances/divergences between Gaussian processes

$\xi^i \sim \mathrm{GP}(0, K^i)$, $i = 1, 2$

- Log-Hilbert-Schmidt distance, $\gamma \in \mathbb{R}, \gamma > 0$ fixed

$$D^{\gamma}_{\mathrm{logHS}}(\xi^1, \xi^2) = D^{\gamma}_{\mathrm{logHS}}[\mathcal{N}(0, C_{K^1}), \mathcal{N}(0, C_{K^2})]$$
$$= ||\log(\gamma I + C_{K^1}) - \log(\gamma I + C_{K^2})||_{\mathrm{HS}_X}$$

- Affine-invariant Riemannian distance, $\gamma \in \mathbb{R}, \gamma > 0$ fixed

$$D^{\gamma}_{\mathrm{aiHS}}(\xi^1, \xi^2) = D^{\gamma}_{\mathrm{aiHS}}[\mathcal{N}(0, C_{K^1}), \mathcal{N}(0, C_{K^2})]$$
$$= ||\log[(\gamma I + C_{K^1})^{-1/2}(\gamma I + C_{K^2})(\gamma I + C_{K^1})^{-1/2}]||_{\mathrm{HS}_X}$$

- Wasserstein distance/Sinkhorn divergence, $\epsilon > 0$ fixed

$$W_2(\xi^1, \xi^2) = W_2[\mathcal{N}(0, C_{K^1}), \mathcal{N}(0, C_{K^2})],$$
$$\mathrm{S}^{\epsilon}_2(\xi^1, \xi^2) = \mathrm{S}^{\epsilon}_2[\mathcal{N}(0, C_{K^1}), \mathcal{N}(0, C_{K^2})]$$

Estimation of distances/divergences

- $\mathbf{X} = (x_j)_{j=1}^M \in T^m$
- $(K^i[\mathbf{X}])_{jk} = K^i(x_j, x_k) = \mathbb{E}[\xi^i(\omega, x_j)\xi(\omega, x_k)]$, $1 \leq j, k \leq m$
- $(\xi^i(., x_j))_{j=1}^m \sim \mathcal{N}(0, K^i[\mathbf{X}])$ in $\mathbb{R}^m$
- We can estimate the infinite-dimensional formula in $\mathcal{L}^2(T, \nu)$

$$D\,[\mathcal{N}(0, C_{K^1}), \mathcal{N}(0, C_{K^2})]$$

by the finite-dimensional formula in $\mathbb{R}^m$

$$D\left[\mathcal{N}\left(0, \frac{1}{m}K^1[\mathbf{X}]\right), \mathcal{N}\left(0, \frac{1}{m}K^2[\mathbf{X}]\right)\right]$$

where $D = ||\ ||_{\mathrm{HS}}, D^{\gamma}_{\mathrm{logHS}}, D^{\gamma}_{\mathrm{aiHS}}, \mathrm{S}^{\epsilon}_2, W_2$.

RKHS covariance and cross-covariance operators

$\mathcal{H}_{K^i}$ = reproducing kernel Hilbert space (RKHS) associated with K^i

$$R_{K^i} : \mathcal{L}^2(T,\nu) \to \mathcal{H}_{K^i}, \quad R_{K^i}f(x) = \int_T K^i(x,t)f(t)d\nu(t),$$

$$C_{K^i} = R_{K^i}^* R_{K^i} : \mathcal{L}^2(T,\nu) \to \mathcal{L}^2(T,\nu)$$

with $R_{K^i}^* : \mathcal{H}_{K^i} \to \mathcal{L}^2(T,\nu)$ = inclusion operator

- **RKHS cross-covariance operators**

$$R_{ij} = R_{K^i} R_{K^j}^* : \mathcal{H}_{K^j} \to \mathcal{H}_{K^i},$$

$$R_{ij} = \int_T (K_t^i \otimes K_t^j) d\nu(t), \quad R_{ij}f = \int_T K_t^i \langle f, K_t^j \rangle_{\mathcal{H}_{K^j}} d\nu(t), \quad i,j = 1,2,$$

$$R_{ij}f(x) = \int_T K_t^i(x) f(t) d\nu(t) = \int_T K^i(x,t)f(t)d\nu(t), \quad f \in \mathcal{H}_{K^j},$$

- **RKHS covariance operators** $L_{K^i} = R_{ii} = R_{K^i} R_{K^i}^* : \mathcal{H}_{K^i} \to \mathcal{H}_{K^i}$

RKHS covariance and cross-covariance operators

- $L_{K^i} = R_{ii} = R_{K^i} R^*_{K^i} : \mathcal{H}_{K^i} \to \mathcal{H}_{K^i}$ have the same nonzero eigenvalues as $C_{K^i} = R^*_{K^i} R_{K^i} : \mathcal{L}^2(T, \nu) \to \mathcal{L}^2(T, \nu)$, so have the same trace, same $|| \ ||_{\mathrm{HS}}$
- They are the same when restricted to $\mathcal{H}_{K^i} \subset \mathcal{L}^2(T, \nu)$
- Both appear extensively in learning theory with kernel methods, e.g. Cucker and Smale (2000), Smale and Zhou (2007), Rosasco, Belkin, and De Vito (2010)
- C_{K^i} and L_{K^i} are generally **not interchangeable**
- $D[\mathcal{N}(0, C_{K^1}), \mathcal{N}(0, C_{K^2})]$ is well-defined
- $D[\mathcal{N}(0, L_{K^1}), \mathcal{N}(0, L_{K^2})]$ is generally **not** well-defined

RKHS covariance and cross-covariance operators

- Empirical version given $\mathbf{X} = (x_j)_{j=1}^m \in T^m$

$$R_{ij,\mathbf{x}} : \mathcal{H}_{K^j} \to \mathcal{H}_{K^i},$$

$$R_{ij,\mathbf{x}} = \frac{1}{m}\sum_{k=1}^{m}(K^i_{x_k} \otimes K^j_{x_k}) : \mathcal{H}_{K^j} \to \mathcal{H}_{K^i},$$

$$R_{ij,\mathbf{x}}f = \frac{1}{m}\sum_{k=1}^{m} K^i_{x_k}\langle f, K^j_{x_k}\rangle_{\mathcal{H}_{K^j}} = \frac{1}{m}\sum_{k=1}^{m} f(x_k)K^i_{x_k}, \; f \in \mathcal{H}_{K^j},$$

- $L_{K^i,\mathbf{x}} = R_{ii,\mathbf{x}} : \mathcal{H}_{K^i} \to \mathcal{H}_{K^i}$ has the same nonzero eigenvalues as $\frac{1}{m}K^i[\mathbf{X}] : \mathbb{R}^m \to \mathbb{R}^m$

RKHS covariance and cross-covariance operators

Assume $\sup_{x \in T} K^i(x,x) \leq \kappa_i^2$

Proposition (Convergence of RKHS empirical covariance and cross-covariance operators)

$||R_{ij}||_{\mathrm{HS}(\mathcal{H}_{K^j}, \mathcal{H}_{K^i})} \leq \kappa_i \kappa_j$, $||R_{ij,\mathbf{X}}||_{\mathrm{HS}(\mathcal{H}_{K^j}, \mathcal{H}_{K^i})} \leq \kappa_i \kappa_j$, $i,j = 1,2$, $\forall \mathbf{X} \in T^m$. *Let* $\mathbf{X} = (x_i)_{i=1}^m$ *be independently sampled from* (T, ν). $\forall 0 < \delta < 1$, *with probability at least* $1 - \delta$,

$$||R_{ij,\mathbf{X}} - R_{ij}||_{\mathrm{HS}(\mathcal{H}_{K^j}, \mathcal{H}_{K^i})} \leq \kappa_i \kappa_j \left[\frac{2 \log \frac{2}{\delta}}{m} + \sqrt{\frac{2 \log \frac{2}{\delta}}{m}} \right]$$

Convergence in Hilbert-Schmidt norm

Compare with 2-Wasserstein distance (weak convergence)

$$\lim_{n\to\infty} W_2[\mathcal{N}(0, A_n), \mathcal{N}(0, A)] = 0 \Longleftrightarrow \lim_{n\to\infty} ||A_n - A||_{\mathrm{tr}} = 0$$

Theorem (Convergence in Sinkhorn divergence)

Let $\{A_N\}_{N\in\mathbb{N}}, A \in \mathrm{Sym}^+(\mathcal{H}) \cap \mathrm{Tr}(\mathcal{H})$*. Then*

$$\mathrm{S}^{\epsilon}_{d^2}[\mathcal{N}(0, A_N), \mathcal{N}(0, A)] \leq \frac{3}{\epsilon}[||A_N||_{\mathrm{HS}} + ||A||_{\mathrm{HS}}]||A_N - A||_{\mathrm{HS}}.$$

In particular, $\lim_{N\to\infty} ||A_N - A||_{\mathrm{HS}} = 0 \Rightarrow \lim_{N\to\infty} \mathrm{S}^{\epsilon}_{2}[\mathcal{N}(0, A_N), \mathcal{N}(0, A)] = 0$

Consequence: We can apply laws of large numbers for Hilbert space-valued random variables to obtain sample complexity

Estimation of Sinkhorn divergence

$$G(A) = \mathrm{tr}[M(A)] - \log\det\left(I + \frac{1}{2}M(A)\right), \text{where } M(A) = -I + (I + c^2 A)^{1/2}$$

With this definition, with $c = \frac{4}{\epsilon}$,

Proposition (RKHS covariance and cross-covariance operator representation for Sinkhorn divergence)

Let $\mathbf{X} = (x_i)_{i=1}^m \in T^m$. *Then*

$$\mathrm{S}_2^{\epsilon}[\mathcal{N}(0, C_{K^1}), \mathcal{N}(0, C_{K^2})] = \frac{1}{c}\left[G(L_{K^1}^2) + G(L_{K^2}^2) - 2G(R_{12}^* R_{12})\right]$$

$$\mathrm{S}_2^{\epsilon}\left[\mathcal{N}\left(0, \frac{1}{m}K^1[\mathbf{X}]\right), \mathcal{N}\left(0, \frac{1}{m}K^2[\mathbf{X}]\right)\right] = \frac{1}{c}\left[G(L_{K^1,\mathbf{X}}^2) + G(L_{K^2,\mathbf{X}}^2) - 2G(R_{12,\mathbf{X}}^* R_{12,\mathbf{X}})\right]$$

Estimation of Sinkhorn divergence

Assume $\sup_{x \in T} K^i(x, x) \leq \kappa_i^2$

Theorem (Estimation of Sinkhorn divergence between Gaussian processes from finite covariance matrices - bounded kernels)

Let $\mathbf{X} = (x_i)_{i=1}^m$ *be independently sampled from* (T, ν)*. For any* $0 < \delta < 1$*, with probability at least* $1 - \delta$*,*

$$\left| \mathrm{S}_2^{\epsilon}\left[\mathcal{N}\left(0, \frac{1}{m}K^1[\mathbf{X}]\right), \mathcal{N}\left(0, \frac{1}{m}K^2[\mathbf{X}]\right) - \mathrm{S}_2^{\epsilon}[\mathcal{N}(0, C_{K^1}), \mathcal{N}(0, C_{K^2})]\right]\right|$$

$$\leq \frac{6}{\epsilon}(\kappa_1^2 + \kappa_2^2)^2 \left[\frac{2 \log \frac{6}{\delta}}{m} + \sqrt{\frac{2 \log \frac{6}{\delta}}{m}}\right]$$

The convergence is dimension-independent

Estimation of Log-Hilbert-Schmidt distance

Theorem (**Estimation of Log-Hilbert-Schmidt distance from finite covariance matrices**)

Let $\gamma \in \mathbb{R}, \gamma > 0$ *be fixed. Let* $\mathbf{X} = (x_i)_{i=1}^m$ *be independently sampled from* (T, ν)*.* $\forall 0 < \delta < 1$*, with probability at least* $1 - \delta$*,*

$$\left| \left\| \log\left(\gamma I + \frac{1}{m} K^1[\mathbf{X}]\right) - \log\left(\gamma I + \frac{1}{m} K^2[\mathbf{X}]\right) \right\|_F^2 \right.$$
$$\left. - || \log(\gamma I + C_{K^1}) - \log(\gamma I + C_{K^2}) ||^2_{\mathrm{HS}(\mathcal{L}^2(T,\nu))} \right|$$
$$\leq \frac{2(\kappa_1^4 + \kappa_2^4)}{\gamma^2} \left(\frac{2 \log \frac{6}{\delta}}{m} + \sqrt{\frac{2 \log \frac{6}{\delta}}{m}} \right)$$
$$+ \frac{2\kappa_1^2 \kappa_2^2}{\gamma^2} \left(1 + \frac{\kappa_1^2 + \kappa_2^2}{2\gamma}\right) \left(\frac{2 \log \frac{24}{\delta}}{m} + \sqrt{\frac{2 \log \frac{24}{\delta}}{m}} \right).$$

Estimation of affine-invariant Riemannian distance

Theorem (Estimation of affine-invariant Riemannian distance from finite covariance matrices)

Let $\gamma \in \mathbb{R}, \gamma > 0$ be fixed. Let $\mathbf{X} = (x_j)_{j=1}^m$ be independently sampled from (T, ν). For any $0 < \delta < 1$, with probability at least $1 - \delta$,

$$\Bigg| \left\| \log\left[\left(\gamma I + \frac{1}{m}K^1[\mathbf{X}]\right)^{-1/2}\left(\gamma I + \frac{1}{m}K^2[\mathbf{X}]\right)\left(\gamma I + \frac{1}{m}K^1[\mathbf{X}]\right)^{-1/2}\right] \right\|_F^2$$

$$- || \log[(\gamma I + C_{K^1})^{-1/2}(\gamma I + C_{K^2})(\gamma I + C_{K^1})^{-1/2}]||^2_{\mathrm{HS}(\mathcal{L}^2(T,\nu))} \Bigg|$$

$$\leq \frac{1}{\gamma^2}\left(1 + \frac{\kappa_1^2}{\gamma}\right)^3 \left[(\kappa_1 + \kappa_2)^2 + \frac{\kappa_1^2\kappa_2^2}{\gamma}\right]\left(\kappa_1 + \kappa_2 + \frac{\kappa_1\kappa_2}{\gamma}\right)^2 \left[\frac{2\log\frac{6}{\delta}}{m} + \sqrt{\frac{2\log\frac{6}{\delta}}{m}}\right]$$

The convergence is dimension-independent

Estimation of 2-Wasserstein distance

Theorem (Estimation of 2-Wasserstein distance from finite covariance matrices)

Let $\mathbf{X} = (x_i)_{i=1}^m$ *be independently sampled from* (T, ν). *Assume further that* $\dim(\mathcal{H}_{K^2}) < \infty$. $\forall 0 < \delta < 1$, *with probability at least* $1 - \delta$,

$$\left| W_2^2\left[\mathcal{N}\left(0, \frac{1}{m}K^1[\mathbf{X}]\right), \mathcal{N}\left(0, \frac{1}{m}K^2[\mathbf{X}]\right)\right] - W_2^2[\mathcal{N}(0, C_{K^1}), \mathcal{N}(0, C_{K^2})]\right|$$

$$\leq (\kappa_1^2 + \kappa_2^2)\left[\frac{2\log\frac{6}{\delta}}{m} + \sqrt{\frac{2\log\frac{6}{\delta}}{m}}\right]$$

$$+ 2\sqrt{2}\kappa_1\kappa_2\sqrt{\dim(\mathcal{H}_{K^2})}\sqrt{\frac{2\log\frac{6}{\delta}}{m} + \sqrt{\frac{2\log\frac{6}{\delta}}{m}}}$$

Estimation of distances from finite samples

- The finite covariance matrix $K[\mathbf{X}]$ is generally unknown
- $K[\mathbf{X}]$ needs to estimated from finite samples
- $\xi \sim \mathrm{GP}(0, K)$ defined on probability space $(\Omega, \mathcal{F}, P)$
- Assume $\mathbf{W} = (\omega_i)_{i=1}^N$, corresponding to N sample paths $\xi_i(x) = \xi(\omega_i, x)$
- On a set $\mathbf{X} = (x_i)_{i=1}^m \in T^m$, this gives the $m \times N$ data matrix

$$\mathbf{Z} = \begin{pmatrix} \xi(\omega_1, x_1), \ldots, \xi(\omega_N, x_1), \\ \cdots \\ \xi(\omega_1, x_m), \ldots, \xi(\omega_N, x_m) \end{pmatrix} = [\mathbf{z}(\omega_1), \ldots \mathbf{z}(\omega_N)] \in \mathbb{R}^{m \times N}$$

Here $\mathbf{z}(\omega) = (\xi(\omega, x_i))_{i=1}^m$

Estimation of distances from finite samples

Given the $m \times N$ data matrix

$$\mathbf{Z} = \begin{pmatrix} \xi(\omega_1, x_1), \ldots, \xi(\omega_N, x_1), \\ \cdots \\ \xi(\omega_1, x_m), \ldots, \xi(\omega_N, x_m) \end{pmatrix} = [\mathbf{z}(\omega_1), \ldots \mathbf{z}(\omega_N)] \in \mathbb{R}^{m \times N}$$

Since $(K[\mathbf{X}])_{ij} = \mathbb{E}[\xi(\omega, x_i)\xi(\omega, x_j)]$,

$$K[\mathbf{X}] = \mathbb{E}[\mathbf{z}(\omega)\mathbf{z}(\omega)^T] = \int_\Omega \mathbf{z}(\omega)\mathbf{z}(\omega)^T dP(\omega)$$

The empirical version of $K[\mathbf{X}]$, using the random sample $\mathbf{W} = (\omega_i)_{i=1}^N$,

$$\hat{K}_{\mathbf{W}}[\mathbf{X}] = \frac{1}{N}\sum_{i=1}^N \mathbf{z}(\omega_i)\mathbf{z}(\omega_i)^T = \frac{1}{N}\mathbf{Z}\mathbf{Z}^T$$

Estimation of Sinkhorn divergence

Theorem (Estimation of Sinkhorn divergence between Gaussian processes from finite samples - bounded kernels)

Let $\mathbf{X} = (x_i)_{i=1}^m$ *be independently sampled from* (T, ν). *Let* $\mathbf{W}^1 = (\omega_j^1)_{j=1}^N$, $\mathbf{W}^2 = (\omega_j^2)_{j=1}^N$ *be independently sampled from* (Ω_1, P_1) *and* (Ω_2, P_2), *respectively.* $\forall 0 < \delta < 1$, *with probability at least* $1 - \delta$,

$$\left| \mathrm{S}_2^{\epsilon}\left[\mathcal{N}\left(0, \frac{1}{m}\hat{K}^1_{\mathbf{W}^1}[\mathbf{X}]\right), \mathcal{N}\left(0, \frac{1}{m}\hat{K}^2_{\mathbf{W}^2}[\mathbf{X}]\right)\right] - \mathrm{S}_2^{\epsilon}[\mathcal{N}(0, C_{K^1}), \mathcal{N}(0, C_{K^2})] \right|$$

$$\leq \frac{6}{\epsilon}(\kappa_1^2 + \kappa_2^2)^2 \left[\frac{2\log\frac{12}{\delta}}{m} + \sqrt{\frac{2\log\frac{12}{\delta}}{m}}\right]$$

$$+ \frac{24\sqrt{3}}{\epsilon\delta}\left[\left(1 + \frac{8}{\delta}\right)\kappa_1^4 + \left(3 + \frac{16}{\delta}\right)\kappa_1^2\kappa_2^2 + \kappa_2^4\right]\frac{1}{\sqrt{N}}$$

Here the probability is with respect to the space $(T, \nu)^m \times (\Omega_1, P_1)^N \times (\Omega_2, P_2)^N$

Numerical experiments

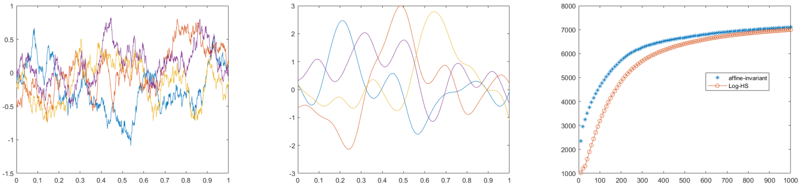


Figure: Samples of the centered Gaussian processes $\mathcal{N}(0, K^1)$, $\mathcal{N}(0, K^2)$ on $T = [0, 1]$ and approximations of squared distances between them. Left: $K^1(x, y) = \exp(-a||x - y||)$, $a = 1$. Right: $K^2(x, y) = \exp(-||x - y||^2/\sigma^2)$, $\sigma = 0.1$. Here the number of sample paths is $N = 10, 20, \ldots, 1000$, and $\gamma = 10^{-7}$

Numerical experiments

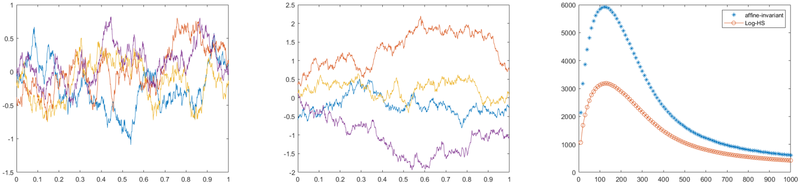


Figure: Samples of the centered Gaussian processes $\mathcal{N}(0, K^1)$, $\mathcal{N}(0, K^2)$ on $T = [0, 1]$ and approximations of squared distances between them. Left: $K^1(x, y) = \exp(-a||x - y||)$, $a = 1$. Right: $K^2(x, y) = \exp(-a||x - y||)$, $a = 1.2$. Here the number of sample paths is $N = 10, 20, \ldots, 1000$, and $\gamma = 10^{-7}$

Numerical experiments

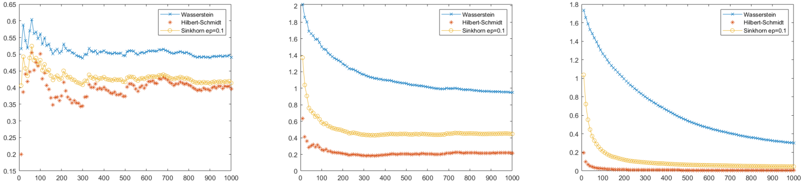


Figure: Approximate divergences/squared distances between the previous Gaussian processes on $T = [0,1]^d \subset \mathbb{R}^d$. Left: $d = 1$. Middle: $d = 5$. Right: $d = 50$. The estimation is obtained using N realizations of each process. Here $N = 10, 20, \ldots, 1000$.

Summary

- Generalization of Riemannian distances between Gaussian measures from $\mathbb{R}^n$ to the Hilbert space setting
- Affine-invariant Riemannian and Log-Euclidean distances, Log-Determinant divergences: regularization is theoretically necessary
- Wasserstein distance: entropic regularization leads to favorable theoretical properties
- Hilbert-Schmidt convergence leads to dimension-independent sample complexities
- Many more theoretical results can be obtained
- **Upcoming: Kullback-Leibler (KL) and Rényi divergences between Gaussian processes**
- Future work: beyond Gaussian process setting

References

- H.Q.Minh, M. San Biagio, V. Murino. Log-Hilbert-Schmidt metric between positive definite operators on Hilbert spaces, NIPS 2014
- H.Q.Minh. Affine-invariant Riemannian distance between infinite-dimensional covariance operators, *Geometric Science of Information*, 2015
- H.Q.Minh, M. San Biagio, L. Bazzani, V. Murino. Approximate Log-Hilbert-Schmidt distances between covariance operators for image classification, CVPR 2016
- H.Q.Minh and V.Murino. From covariance matrices to covariance operators: Data representation from finite to infinite-dimensional settings, in *Algorithmic Advances in Riemannian Geometry and Applications*, 2016.
- H.Q.Minh and V. Murino. Covariances in Computer Vision and Machine Learning, Morgan & Claypool Publishers *Synthesis Lectures on Computer Vision*, 2017.

References

- H.Q.Minh. Infinite-dimensional Log-Determinant divergences between positive definite trace class operators, *Linear Algebra and its Applications*, 2017
- H.Q. Minh. Alpha-Beta Log-Determinant divergences between positive definite trace class operators, *Information Geometry*, 2019
- H.Q.Minh. Infinite-dimensional Log-Determinant divergences between positive definite Hilbert-Schmidt operators, *Positivity*, 2020
- H.Q. Minh. Regularized divergences between covariance operators and Gaussian measures on Hilbert spaces, *Journal of Theoretical Probability*, 2020

References

- H.Q.Minh. Entropic regularization of Wasserstein distance between infinite-dimensional Gaussian measures and Gaussian processes, Journal of Theoretical Probability, 2022
- H.Q.Minh. Convergence and finite sample approximations of entropic regularized Wasserstein distances in Gaussian and RKHS settings, 2021 (arXiv:2101.01429)
- H.Q.Minh. Finite sample approximations of exact and entropic Wasserstein distances between covariance operators and Gaussian processes, *SIAM/ASA Journal on Uncertainty Quantification*, 2022

Thank you!