Information Theory with Kernel Methods

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France

Münster, March 2024

Measuring "distance" between probability distributions

- **Common sub-task in many areas of data science**
  - Model fitting
  - Independence or homogeneity tests
  - Quantifying loss of information or uncertainty
  - Independent component analysis
  - Mean field analysis of neural networks

Measuring "distance" between probability distributions

- **Common sub-task in many areas of data science**
  - Model fitting
  - Independence or homogeneity tests
  - Quantifying loss of information or uncertainty
  - Independent component analysis
  - Mean field analysis of neural networks
- **Main difficulties**
  - Beyond discrete random variables and Gaussians
  - Non-linear dependencies
  - Need to be estimated from data
  - Physical / statistical meaning

Classical comparison frameworks

- **Information theory** (Cover and Thomas, 1999)
  - Kullback-Leibler divergence for finite set $\mathcal{X}$

$$D(p\|q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

Classical comparison frameworks

- **Information theory** (Cover and Thomas, 1999)
  - Kullback-Leibler divergence for finite set $\mathcal{X}$

$$D(p\|q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

  - Invariance properties and strong physical interpretation
  - Link with probabilistic inference
  - Hard to estimate beyond small discrete and Gaussian distributions

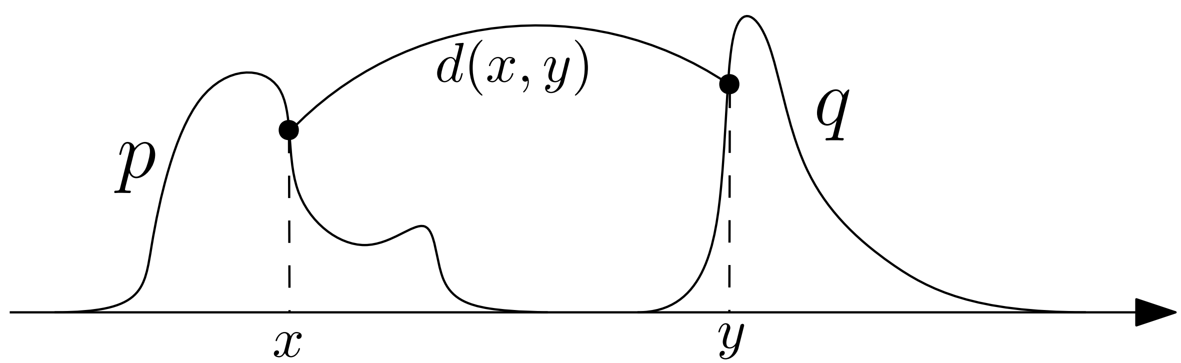
Classical comparison frameworks

- **Information theory** (Cover and Thomas, 1999)
  - Kullback-Leibler divergence for finite set $\mathcal{X}$

$$D(p\|q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

  - Invariance properties and strong physical interpretation
  - Link with probabilistic inference
  - Hard to estimate beyond small discrete and Gaussian distributions

- **Optimal transport** (Peyré and Cuturi, 2019)
  - Physical interpretation through base distance d



Studying probability distributions through moments

- **Moments of feature map $\varphi : \mathcal{X} \to \mathcal{H}$ Hilbert space (or $\mathbb{R}^d$)**
  - Probability distributions p on $\mathcal{X}$
  - Mean element: $\mu_p = \int_{\mathcal{X}} \varphi(x) dp(x)$

Studying probability distributions through moments

- **Moments of feature map $\varphi : \mathcal{X} \to \mathcal{H}$ Hilbert space (or $\mathbb{R}^d$)**
  - Probability distributions p on $\mathcal{X}$
  - Mean element: $\mu_p = \displaystyle\int_{\mathcal{X}} \varphi(x)dp(x)$

- **Full characterization if $\mathcal{H}$ large enough**
  - See Sriperumbudur et al. (2010); Micchelli et al. (2006)
  - Natural metric: $(p, q) \mapsto \|\mu_p - \mu_q\|$
  - Easy to estimate with convergence rates $\propto 1/\sqrt{n}$
  - Only the kernel $k(x, y) = \langle \varphi(x), \varphi(y) \rangle$ is needed

Studying probability distributions through moments

- **Moments of feature map $\varphi : \mathcal{X} \to \mathcal{H}$ Hilbert space (or $\mathbb{R}^d$)**
  - Probability distributions p on $\mathcal{X}$
  - Mean element: $\mu_p = \int_{\mathcal{X}} \varphi(x) dp(x)$

- **Full characterization if $\mathcal{H}$ large enough**
  - See Sriperumbudur et al. (2010); Micchelli et al. (2006)
  - Natural metric: $(p, q) \mapsto \|\mu_p - \mu_q\|$
  - Easy to estimate with convergence rates $\propto 1/\sqrt{n}$
  - Only the kernel $k(x, y) = \langle \varphi(x), \varphi(y) \rangle$ is needed

$$\|\hat{\mu}_p - \hat{\mu}_q\|^2 = \Big\| \frac{1}{n} \sum_{i=1}^{n} \varphi(x_i) - \frac{1}{m} \sum_{j=1}^{m} \varphi(y_j) \Big\|^2$$

Studying probability distributions through moments

- **Moments of feature map $\varphi : \mathcal{X} \to \mathcal{H}$ Hilbert space (or $\mathbb{R}^d$)**
  - Probability distributions p on $\mathcal{X}$
  - Mean element: $\mu_p = \int_{\mathcal{X}} \varphi(x) dp(x)$

- **Full characterization if $\mathcal{H}$ large enough**
  - See Sriperumbudur et al. (2010); Micchelli et al. (2006)
  - Natural metric: $(p, q) \mapsto \|\mu_p - \mu_q\|$
  - Easy to estimate with convergence rates $\propto 1/\sqrt{n}$
  - Only the kernel $k(x, y) = \langle \varphi(x), \varphi(y) \rangle$ is needed

- **Many applications** (see, e.g., Muandet et al., 2017)
  - Model fitting, independence tests, GANs, gradient flows, etc.

Studying probability distributions through moments

- **Moments of feature map $\varphi : \mathcal{X} \to \mathcal{H}$ Hilbert space (or $\mathbb{R}^d$)**
  - Probability distributions p on $\mathcal{X}$
  - Mean element: $\mu_p = \int_{\mathcal{X}} \varphi(x) dp(x)$

- **Full characterization if $\mathcal{H}$ large enough**
  - See Sriperumbudur et al. (2010); Micchelli et al. (2006)
  - Natural metric: $(p, q) \mapsto \|\mu_p - \mu_q\|$
  - Easy to estimate with convergence rates $\propto 1/\sqrt{n}$
  - Only the kernel $k(x, y) = \langle \varphi(x), \varphi(y) \rangle$ is needed

- **Many applications** (see, e.g., Muandet et al., 2017)
  - Model fitting, independence tests, GANs, gradient flows, etc.

- **Any link with information-theoretic quantities?**

From mean element to covariance operator

- **Covariance operator / matrix** $\Sigma_p = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$
  - Self-adjoint / symmetric / Hermitian, positive-semidefinite

From mean element to covariance operator

- **Covariance operator / matrix** $\Sigma_p = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$
  - Self-adjoint / symmetric / Hermitian, positive-semidefinite

- **Main tool: Quantum entropies**
  - Von Neumann entropy: $\mathrm{tr}\left[\Sigma_p \log \Sigma_p\right]$
  - Relative entropy: $\mathrm{tr}\left[\Sigma_p(\log \Sigma_p - \log \Sigma_q) - \Sigma_p + \Sigma_q\right]$

From mean element to covariance operator

- **Covariance operator / matrix** $\Sigma_p = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$
  - Self-adjoint / symmetric / Hermitian, positive-semidefinite

- **Main tool: Quantum entropies**
  - Von Neumann entropy: $\operatorname{tr}\big[\Sigma_p \log \Sigma_p\big]$
  - Relative entropy: $\operatorname{tr}\big[\Sigma_p(\log \Sigma_p - \log \Sigma_q) - \Sigma_p + \Sigma_q\big]$

- **Many properties** (`https://arxiv.org/abs/2202.08545`)
  - Clear relationships with regular information theory
  - Estimation in $1/\sqrt{n}$
  - Use in multivariate modelling
  - Variational inference

- **Related work**: Giraldo et al. (2014); Minh (2021)

Covariance operators $\Sigma_p = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$

- **Assumptions**
  - $(x, y) \mapsto k(x, y)$ positive definite kernel on $\mathcal{X} \times \mathcal{X}$
  - $\mathcal{X}$ compact, and $\forall x \in \mathcal{X}$, $k(x, x) \leqslant 1$

Covariance operators $\Sigma_p = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$

- **Assumptions**
  - $(x, y) \mapsto k(x, y)$ positive definite kernel on $\mathcal{X} \times \mathcal{X}$
  - $\mathcal{X}$ compact, and $\forall x \in \mathcal{X}$, $k(x, x) \leqslant 1$
  - There exists a feature map $\varphi : \mathcal{X} \to \mathcal{H}$ Hilbert space such that

$$\forall x, y \in \mathcal{X}, \quad k(x, y) = \langle \varphi(x), \varphi(y) \rangle$$

  - Space of linear functions in φ, that is, $f(x) = \langle f, \varphi(x) \rangle$ for $f \in \mathcal{H}$

Covariance operators $\Sigma_p = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$

- **Assumptions**
  - $(x,y) \mapsto k(x,y)$ positive definite kernel on $\mathcal{X} \times \mathcal{X}$
  - $\mathcal{X}$ compact, and $\forall x \in \mathcal{X}$, $k(x,x) \leqslant 1$
  - There exists a feature map $\varphi : \mathcal{X} \to \mathcal{H}$ Hilbert space such that

$$\forall x, y \in \mathcal{X}, \quad k(x,y) = \langle \varphi(x), \varphi(y) \rangle$$

  - Space of linear functions in φ, that is, $f(x) = \langle f, \varphi(x) \rangle$ for $f \in \mathcal{H}$
  - Universal kernel (Steinwart, 2001): dense in the set of continuous functions with uniform norm

Covariance operators $\Sigma_p = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$

- **Assumptions**
  - $(x,y) \mapsto k(x,y)$ positive definite kernel on $\mathcal{X} \times \mathcal{X}$
  - $\mathcal{X}$ compact, and $\forall x \in \mathcal{X}$, $k(x,x) \leqslant 1$
  - There exists a feature map $\varphi : \mathcal{X} \to \mathcal{H}$ Hilbert space such that

$$\forall x, y \in \mathcal{X}, \quad k(x,y) = \langle \varphi(x), \varphi(y) \rangle$$

  - Space of linear functions in φ, that is, $f(x) = \langle f, \varphi(x) \rangle$ for $f \in \mathcal{H}$
  - Universal kernel (Steinwart, 2001): dense in the set of continuous functions with uniform norm

- **Classical example for $\mathcal{X} \subset \mathbb{R}^d$:** $k(x,y) = \exp(-\|x-y\|_2^2/\sigma^2)$
  - Infinitely differentiable functions

Covariance operators $\Sigma_p = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$

- **Torus** $\mathcal{X} = [0,1]^d$
  - $k(x,y) = q(x-y)$, q 1-periodic, with positive Fourier series $\hat{q}$
  - Corresponds to $\varphi(x)_\omega = \hat{q}(\omega)^{1/2} e^{2i\pi\omega^\top x}$, $\omega \in \mathbb{Z}^d$

$$k(x,y) = \langle \varphi(x), \varphi(y) \rangle = \sum_{\omega \in \mathbb{Z}^d} \hat{q}(\omega) e^{2i\pi\omega^\top (x-y)} = q(x-y)$$

Covariance operators $\Sigma_p = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$

- **Torus** $\mathcal{X} = [0,1]^d$
  - $k(x,y) = q(x-y)$, q 1-periodic, with positive Fourier series $\hat{q}$
  - Corresponds to $\varphi(x)_\omega = \hat{q}(\omega)^{1/2} e^{2i\pi\omega^\top x}$, $\omega \in \mathbb{Z}^d$
  - Link to characteristic functions

$$(\Sigma_p)_{\omega\omega'} = \hat{q}(\omega)^{1/2}\hat{q}(\omega')^{1/2} \cdot \mathbb{E}\big[e^{2i\pi(\omega-\omega')^\top x}\big]$$

  - Example: $\hat{q}(\omega) \propto \exp(-\sigma\|\omega\|_1)$

Covariance operators $\Sigma_p = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$

- **Torus** $\mathcal{X} = [0,1]^d$
  - $k(x,y) = q(x-y)$, q 1-periodic, with positive Fourier series $\hat{q}$
  - Corresponds to $\varphi(x)_\omega = \hat{q}(\omega)^{1/2} e^{2i\pi\omega^\top x}$, $\omega \in \mathbb{Z}^d$
  - Link to characteristic functions

$$(\Sigma_p)_{\omega\omega'} = \hat{q}(\omega)^{1/2}\hat{q}(\omega')^{1/2} \cdot \mathbb{E}\big[e^{2i\pi(\omega-\omega')^\top x}\big]$$

  - Example: $\hat{q}(\omega) \propto \exp(-\sigma\|\omega\|_1)$

- **Finite sets** $\mathcal{X} = \{1,\dots,m\}$
  - "One-hot" encoding ($\forall i,\ \varphi(x)_i = 1_{x=i}$) leads to $\Sigma_p = \mathrm{Diag}(p)$
  - $\mathcal{X} = \{-1,1\}^d$, with $\varphi(x)$ composed of monomials

- **Beyond!**

Properties of covariance operators

$$\Sigma_p = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$$

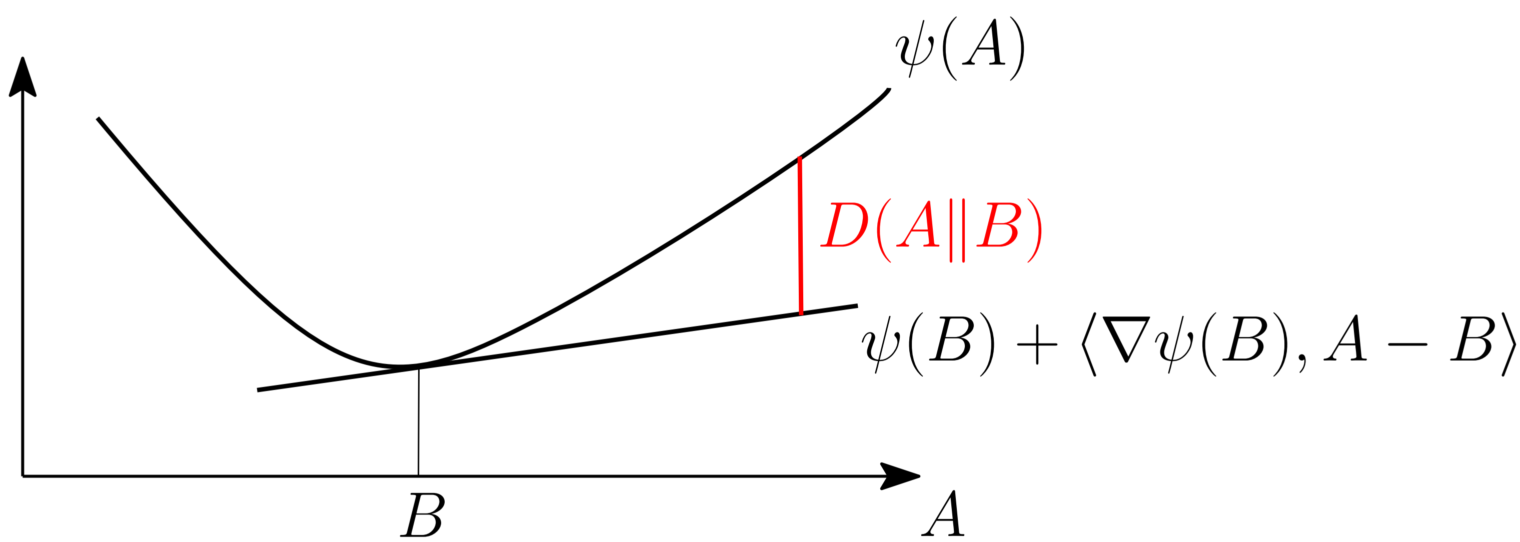
- **Characterization of probability distributions**
  - Σ_p is positive semi-definite, with trace less than one
  - Sequence of positive eigenvalues tending to zero
  - The mapping $p \mapsto \Sigma_p$ is injective
- **Similar to the mean element** $\mu_p = \int_{\mathcal{X}} \varphi(x) dp(x)$

Quantum entropies

- **Negative entropy** (von Neumann, 1932): $\mathrm{tr}\left[A \log A\right] = \sum_{\lambda \in \Lambda(A)} \lambda \log \lambda$
  - $\Lambda(A)$ set of eigenvalues of A

Quantum entropies

- **Negative entropy** (von Neumann, 1932): $\operatorname{tr}\left[A \log A\right] = \sum_{\lambda \in \Lambda(A)} \lambda \log \lambda$
  - $\Lambda(A)$ set of eigenvalues of A
- **Relative entropy**: $D(A \| B) = \operatorname{tr}[A(\log A - \log B) - A + B]$
  - Kullback-Leibler divergence
  - Bregman divergence $\psi(A) - \psi(B) - \langle \nabla \psi(B), A - B \rangle$ for $\psi(A) = \operatorname{tr}\left[A \log A\right]$



Quantum entropies

- **Negative entropy** (von Neumann, 1932): $\operatorname{tr}\big[A\log A\big] = \sum_{\lambda\in\Lambda(A)} \lambda\log\lambda$
  - $\Lambda(A)$ set of eigenvalues of A

- **Relative entropy**: $D(A\|B) = \operatorname{tr}[A(\log A - \log B) - A + B]$
  - Kullback-Leibler divergence
  - Bregman divergence $\psi(A) - \psi(B) - \langle\nabla\psi(B), A - B\rangle$ for $\psi(A) = \operatorname{tr}\big[A\log A\big]$

- **Properties** (Petz, 1986; Ruskai, 2007; Wilde, 2013)
  - $D(A\|B) \geqslant 0$ with equality if and only if $A = B$
  - $(A, B) \mapsto D(A\|B)$ jointly convex in A and B
  - Applications to matrix concentration inequalities (Tropp, 2015)
  - Used in optimization (Chandrasekaran and Shah, 2017)

Kernel relative entropy (Bach, 2022a)

- **Definition**: $D(\Sigma_p \| \Sigma_q) = \mathrm{tr}\left[\Sigma_p(\log \Sigma_p - \log \Sigma_q) - \Sigma_p + \Sigma_q\right]$
  - Σ_p and Σ_q covariance operators

Kernel relative entropy (Bach, 2022a)

- **Definition**: $D(\Sigma_p \| \Sigma_q) = \mathrm{tr}\left[\Sigma_p(\log \Sigma_p - \log \Sigma_q) - \Sigma_p + \Sigma_q\right]$
  - Σ_p and Σ_q covariance operators

- **Properties**
  - Finite if $\left\|\frac{dp}{dq}\right\|_\infty$ finite
  - Always non-negative, with equality if and only $p = q$
  - Jointly convex in (p, q)

Kernel relative entropy (Bach, 2022a)

- **Definition**: $D(\Sigma_p \| \Sigma_q) = \mathrm{tr}\left[\Sigma_p(\log \Sigma_p - \log \Sigma_q) - \Sigma_p + \Sigma_q\right]$
  - Σ_p and Σ_q covariance operators

- **Properties**
  - Finite if $\left\|\frac{dp}{dq}\right\|_\infty$ finite
  - Always non-negative, with equality if and only $p = q$
  - Jointly convex in (p, q)

- **Extension to non-relative entropy**
  - See Bach (2022a)

Kernel relative entropy (Bach, 2022a)

- **Definition**: $D(\Sigma_p \| \Sigma_q) = \mathrm{tr}\left[\Sigma_p(\log \Sigma_p - \log \Sigma_q) - \Sigma_p + \Sigma_q\right]$
  - Σ_p and Σ_q covariance operators

- **Properties**
  - Finite if $\left\|\frac{dp}{dq}\right\|_\infty$ finite
  - Always non-negative, with equality if and only $p = q$
  - Jointly convex in (p, q)

- **Extension to non-relative entropy**
  - See Bach (2022a)

- **Not all properties of Shannon relative entropy will be satisfied**
  - For axiomatic definition of entropy, see Csiszár (2008)

Finite sets with orthonormal embeddings

- **Finite set $\mathcal{X}$**
  - Orthonormal embeddings $\langle \varphi(x), \varphi(y) \rangle = 1_{x=y}$
  - All covariance operators jointly diagonalizable with probability mass values as eigenvalues

Finite sets with orthonormal embeddings

- **Finite set $\mathcal{X}$**
  - Orthonormal embeddings $\langle \varphi(x), \varphi(y) \rangle = 1_{x=y}$
  - All covariance operators jointly diagonalizable with probability mass values as eigenvalues
- **Recovering regular relative entropy exactly**

$$D(\Sigma_p \| \Sigma_q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = D(p \| q)$$

  - Beyond finite sets?

Lower bound on Shannon relative entropy

- **Using Jensen's inequality and** $\forall x \in \mathcal{X}$**,** $\|\varphi(x)\|^2 = 1$

$$D(\Sigma_p \| \Sigma_q) = D\Big(\int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x) \Big\| \int_{\mathcal{X}} \frac{dq}{dp}(x) \varphi(x)\varphi(x)^* dp(x) \Big)$$

Lower bound on Shannon relative entropy

- **Using Jensen's inequality and** $\forall x \in \mathcal{X}$**,** $\|\varphi(x)\|^2 = 1$

$$
\begin{aligned}
D(\Sigma_p \| \Sigma_q) &= D\Big(\int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x) \Big\| \int_{\mathcal{X}} \frac{dq}{dp}(x)\varphi(x)\varphi(x)^* dp(x) \Big) \\
&\leqslant \int_{\mathcal{X}} D\Big(\varphi(x)\varphi(x)^* \Big\| \frac{dq}{dp}(x)\varphi(x)\varphi(x)^* \Big) dp(x)
\end{aligned}
$$

Lower bound on Shannon relative entropy

- **Using Jensen's inequality and** $\forall x \in \mathcal{X}$**,** $\|\varphi(x)\|^2 = 1$

$$
\begin{aligned}
D(\Sigma_p \| \Sigma_q) &= D\Big(\int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x) \Big\| \int_{\mathcal{X}} \frac{dq}{dp}(x)\varphi(x)\varphi(x)^* dp(x) \Big) \\
&\leqslant \int_{\mathcal{X}} D\Big(\varphi(x)\varphi(x)^* \Big\| \frac{dq}{dp}(x)\varphi(x)\varphi(x)^* \Big) dp(x)
\end{aligned}
$$

Lower bound on Shannon relative entropy

- **Using Jensen's inequality and** $\forall x \in \mathcal{X}$, $\|\varphi(x)\|^2 = 1$

$$
\begin{aligned}
D(\Sigma_p \| \Sigma_q) &= D\Big(\int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x) \Big\| \int_{\mathcal{X}} \frac{dq}{dp}(x)\varphi(x)\varphi(x)^* dp(x) \Big) \\
&\leqslant \int_{\mathcal{X}} D\Big(\varphi(x)\varphi(x)^* \Big\| \frac{dq}{dp}(x)\varphi(x)\varphi(x)^* \Big) dp(x) \\
&= \int_{\mathcal{X}} \|\varphi(x)\|^2 \log \frac{\|\varphi(x)\|^2}{\|\varphi(x)\|^2 \frac{dq}{dp}(x)} dp(x)
\end{aligned}
$$

Lower bound on Shannon relative entropy

- **Using Jensen's inequality and** $\forall x \in \mathcal{X}$, $\|\varphi(x)\|^2 = 1$

$$
\begin{aligned}
D(\Sigma_p \| \Sigma_q) &= D\Big(\int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x) \Big\| \int_{\mathcal{X}} \frac{dq}{dp}(x)\varphi(x)\varphi(x)^* dp(x) \Big) \\
&\leqslant \int_{\mathcal{X}} D\Big(\varphi(x)\varphi(x)^* \Big\| \frac{dq}{dp}(x)\varphi(x)\varphi(x)^* \Big) dp(x) \\
&= \int_{\mathcal{X}} \|\varphi(x)\|^2 \log \frac{\|\varphi(x)\|^2}{\|\varphi(x)\|^2 \frac{dq}{dp}(x)} dp(x) \\
&\leqslant \int_{\mathcal{X}} \log\Big(\frac{dp}{dq}(x) \Big) dp(x) = D(p\|q)
\end{aligned}
$$

Lower bound on Shannon relative entropy

- **Using Jensen's inequality and** $\forall x \in \mathcal{X}$, $\|\varphi(x)\|^2 = 1$

$$
\begin{aligned}
D(\Sigma_p \| \Sigma_q) &= D\Big(\int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x) \Big\| \int_{\mathcal{X}} \frac{dq}{dp}(x)\varphi(x)\varphi(x)^* dp(x) \Big) \\
&\leqslant \int_{\mathcal{X}} D\Big(\varphi(x)\varphi(x)^* \Big\| \frac{dq}{dp}(x)\varphi(x)\varphi(x)^* \Big) dp(x) \\
&= \int_{\mathcal{X}} \|\varphi(x)\|^2 \log \frac{\|\varphi(x)\|^2}{\|\varphi(x)\|^2 \frac{dq}{dp}(x)} dp(x) \\
&\leqslant \int_{\mathcal{X}} \log\Big(\frac{dp}{dq}(x)\Big) dp(x) = D(p\|q)
\end{aligned}
$$

- **How tight?**

Small-width asymptotics for metric spaces

- **Approximation bound**: assuming that p, q have strictly positive Lipschitz-continuous densities

$$0 \leqslant \ D(p\|q) - D(\Sigma_p\|\Sigma_q) \ \leqslant E(p,q) \times \Delta(k)$$

  - $\Delta(k)$ characterizes lack of orthonormality of embedding φ
  - Explicit constant $E(p,q)$, see Bach (2022a)
  - Proof based on quantum information theory

Proof

- **Quantum measurement** (with $\Sigma = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* d\tau(x)$)
  - Define for all $y \in \mathcal{X}$, operator $D(y) = \Sigma^{-1/2}\big(\varphi(y)\varphi(y)^*\big)\Sigma^{-1/2}$
  - Positive self-adjoint operators such that $\displaystyle\int_{\mathcal{X}} D(y) d\tau(y) = I$

Proof

- **Quantum measurement** (with $\Sigma = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* d\tau(x)$)
  - Define for all $y \in \mathcal{X}$, operator $D(y) = \Sigma^{-1/2}\big(\varphi(y)\varphi(y)^*\big)\Sigma^{-1/2}$
  - Positive self-adjoint operators such that $\displaystyle\int_{\mathcal{X}} D(y)d\tau(y) = I$
  - Measurement $\mathrm{tr}[D(y)\Sigma_p] = \tilde{p}(y)$, with

$$\tilde{p}(y) = \int_{\mathcal{X}} \mathrm{tr}\left[\Sigma^{-1/2}\big(\varphi(y)\varphi(y)^*\big)\Sigma^{-1/2}\varphi(x)\varphi(x)^*\right]dp(x) = \int_{\mathcal{X}} h(x,y)dp(x)$$

where $h(x,y) = \langle \varphi(x), \Sigma^{-1/2}\varphi(y)\rangle^2$, and $\displaystyle\int_{\mathcal{X}} h(x,y)d\tau(x) = 1$

Proof

- **Quantum measurement** (with $\Sigma = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* d\tau(x)$)
  - Define for all $y \in \mathcal{X}$, operator $D(y) = \Sigma^{-1/2}\big(\varphi(y)\varphi(y)^*\big)\Sigma^{-1/2}$
  - Positive self-adjoint operators such that $\displaystyle\int_{\mathcal{X}} D(y)d\tau(y) = I$
  - Measurement $\mathrm{tr}[D(y)\Sigma_p] = \tilde{p}(y)$, with

$$\tilde{p}(y) = \int_{\mathcal{X}} \mathrm{tr}\left[\Sigma^{-1/2}\big(\varphi(y)\varphi(y)^*\big)\Sigma^{-1/2}\varphi(x)\varphi(x)^*\right] dp(x) = \int_{\mathcal{X}} h(x,y)dp(x)$$

where $h(x,y) = \langle \varphi(x), \Sigma^{-1/2}\varphi(y)\rangle^2$, and $\displaystyle\int_{\mathcal{X}} h(x,y)d\tau(x) = 1$

- **Monotonicity of quantum measurements**: $D(\tilde{p}\|\tilde{q}) \leqslant D(\Sigma_p\|\Sigma_q)$

- **“Sandwich”**: $D(\tilde{p}\|\tilde{q}) \leqslant D(\Sigma_p\|\Sigma_q) \leqslant D(p\|q)$

Small-width asymptotics for metric spaces

- **Approximation bound**: assuming that p, q have strictly positive Lipschitz-continuous densities

$$0 \leqslant \ D(p\|q) - D(\Sigma_p\|\Sigma_q) \ \leqslant E(p,q) \times \Delta(k)$$

  - $\Delta(k)$ characterizes lack of orthonormality of embedding φ
  - Explicit constant $E(p,q)$, see Bach (2022a)
  - Proof based on quantum information theory

- **Consequences on the d-dimensional torus**
  - With $\hat{q}(\omega) \propto \exp(-\sigma\|\omega\|_1)$, we have $D(p\|q) - D(\Sigma_p\|\Sigma_q) = O(\sigma^2)$
  - Corresponds to $k(x,y)$ being a function of $\frac{1}{\sigma}(x-y)$

Estimation from finite sample - I

- **Canonical problem**: estimate $D(\Sigma_p \| \Sigma_q)$ from n i.i.d. samples of p
  - With $D(\Sigma_p \| \Sigma_q) = \operatorname{tr}\left[\Sigma_p \log \Sigma_p - \Sigma_p \log \Sigma_q - \Sigma_p + \Sigma_q\right]$

Estimation from finite sample - I

- **Canonical problem**: estimate $D(\Sigma_p \| \Sigma_q)$ from n i.i.d. samples of p
  - With $D(\Sigma_p \| \Sigma_q) = \operatorname{tr}\left[\Sigma_p \log \Sigma_p - \Sigma_p \log \Sigma_q - \Sigma_p + \Sigma_q\right]$
  - Natural estimator of $\operatorname{tr}\left[\Sigma_p \log \Sigma_p\right]$ is $\operatorname{tr}\left[\hat{\Sigma}_p \log \hat{\Sigma}_p\right]$, with

$$\hat{\Sigma}_p = \frac{1}{n} \sum_{i=1}^{n} \varphi(x_i) \varphi(x_i)^*$$

Estimation from finite sample - I

- **Canonical problem**: estimate $D(\Sigma_p \| \Sigma_q)$ from n i.i.d. samples of p
  - With $D(\Sigma_p \| \Sigma_q) = \operatorname{tr}\left[\Sigma_p \log \Sigma_p - \Sigma_p \log \Sigma_q - \Sigma_p + \Sigma_q\right]$
  - Natural estimator of $\operatorname{tr}\left[\Sigma_p \log \Sigma_p\right]$ is $\operatorname{tr}\left[\hat{\Sigma}_p \log \hat{\Sigma}_p\right]$, with

$$\hat{\Sigma}_p = \frac{1}{n} \sum_{i=1}^{n} \varphi(x_i)\varphi(x_i)^*$$

- **Proposition**: $\operatorname{tr}\left[\hat{\Sigma}_p \log \hat{\Sigma}_p\right] = \operatorname{tr}\left[\frac{1}{n} K \log\left(\frac{1}{n} K\right)\right]$
  - with $K \in \mathbb{R}^{n \times n}$ the kernel matrix defined as $K_{ij} = k(x_i, x_j)$
  - Running time complexity: from $O(n^3)$ to $O(nm^2)$ (Boutsidis et al., 2009; Rudi et al., 2015)
  - Applicable to other divergences (Giraldo et al., 2014; Minh, 2021)

Estimation from finite sample - II

- **Statistical performance**
  - Let $c = \int_0^{+\infty} \sup_{x \in \mathcal{X}} \langle \varphi(x), (\Sigma + \lambda I)^{-1} \varphi(x) \rangle^2 d\lambda$
  - Assume $\frac{dp}{dq}(x) \geqslant \alpha$

$$\mathbb{E}\Big[\big|\operatorname{tr}\big[\hat{\Sigma}_p \log \hat{\Sigma}_p\big] - \operatorname{tr}\big[\Sigma_p \log \Sigma_p\big]\big|\Big] \leqslant 34 \cdot \frac{\sqrt{c}}{\sqrt{n}} + \frac{1 + c(8 \log n)^2}{n\alpha} + \frac{17 \log n}{\sqrt{n}}$$

  - No need to regularize

Estimation from finite sample - II

- **Statistical performance**
  - Let $c = \displaystyle\int_0^{+\infty} \sup_{x \in \mathcal{X}} \langle \varphi(x), (\Sigma + \lambda I)^{-1} \varphi(x) \rangle^2 d\lambda$
  - Assume $\dfrac{dp}{dq}(x) \geqslant \alpha$

$$\mathbb{E}\Big[\big| \operatorname{tr}\big[\hat{\Sigma}_p \log \hat{\Sigma}_p\big] - \operatorname{tr}\big[\Sigma_p \log \Sigma_p\big]\big|\Big] \leqslant 34 \cdot \frac{\sqrt{c}}{\sqrt{n}} + \frac{1 + c(8 \log n)^2}{n\alpha} + \frac{17 \log n}{\sqrt{n}}$$

  - No need to regularize
  - Proof technique: $A \log A = A \log(A + \nu I) - \displaystyle\int_0^{\nu} A(A + \lambda I)^{-1} d\lambda$

Estimation from finite sample - II

- **Statistical performance**
  - Let $c = \int_0^{+\infty} \sup_{x \in \mathcal{X}} \langle \varphi(x), (\Sigma + \lambda I)^{-1} \varphi(x) \rangle^2 d\lambda$
  - Assume $\frac{dp}{dq}(x) \geqslant \alpha$

$$\mathbb{E}\Big[\Big| \operatorname{tr}\big[\hat{\Sigma}_p \log \hat{\Sigma}_p\big] - \operatorname{tr}\big[\Sigma_p \log \Sigma_p\big]\Big|\Big] \leqslant 34 \cdot \frac{\sqrt{c}}{\sqrt{n}} + \frac{1 + c(8 \log n)^2}{n\alpha} + \frac{17 \log n}{\sqrt{n}}$$

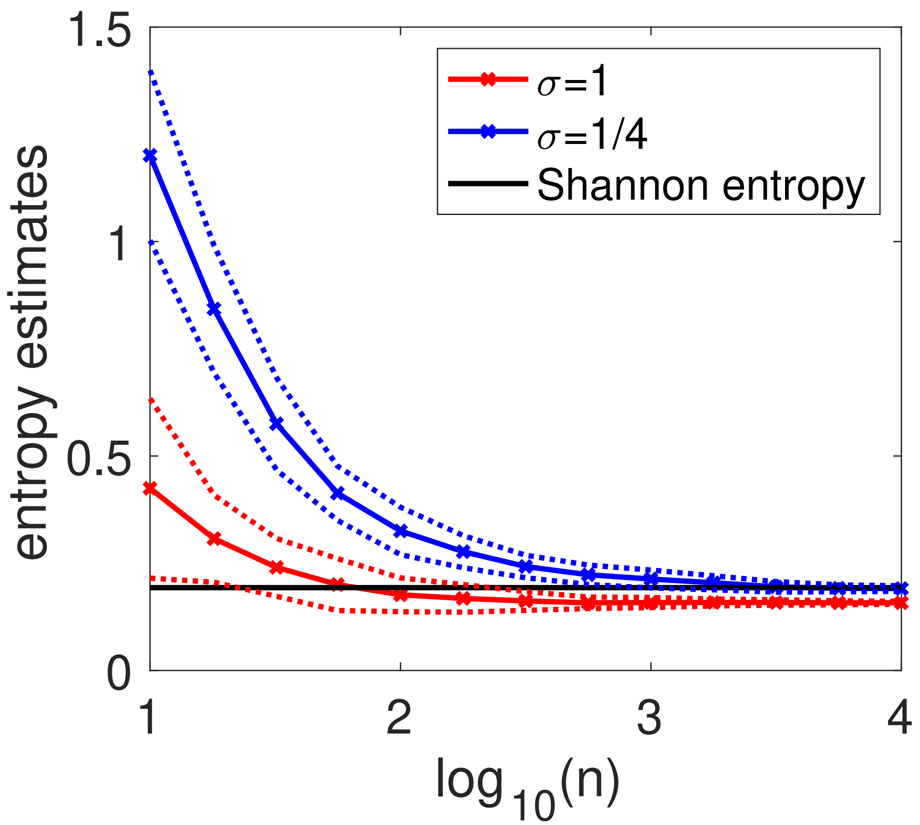
  - No need to regularize

- **Torus**: $c \propto \sigma^{-d} \Rightarrow$ estimation rate proportional to $\sigma^{-d/2}/\sqrt{n}$
  - Entropy estimation in $n^{-2/(d+4)}$
  - NB: optimal rate equal to $n^{-4/(d+4)}$ (Han et al., 2020)

- **Extension:** estimating $D(\Sigma_p \| \Sigma_q)$ from samples of p *and* q

Estimation from finite sample - III

- **Negative entropy estimation**
  - From i.i.d. samples with 20 replications, $d = 1$
  - Two values of the kernel bandwidth σ, as n increases



- NB: Faster estimation from oracles $\int_{\mathcal{X}} k(x,y)k(x,z)dp(x)$

Log-partition functions and variational inference

- **Log-partition function**: given $f : \mathcal{X} \to \mathbb{R}$ and a distribution q on $\mathcal{X}$

$$\log \int_{\mathcal{X}} e^{f(x)} dq(x) = \sup_{p \text{ probability}} \int_{\mathcal{X}} f(x) dp(x) - D(p\|q)$$

  – Used within variational inference (Wainwright and Jordan, 2008)
  – Duality between maximum entropy and maximum likelihood

Log-partition functions and variational inference

- **Log-partition function**: given $f : \mathcal{X} \to \mathbb{R}$ and a distribution q on $\mathcal{X}$

$$\log \int_{\mathcal{X}} e^{f(x)} dq(x) = \sup_{p \text{ probability}} \int_{\mathcal{X}} f(x) dp(x) - D(p\|q)$$

  – Used within variational inference (Wainwright and Jordan, 2008)

- **Upper-bound** (assuming unit norm features)

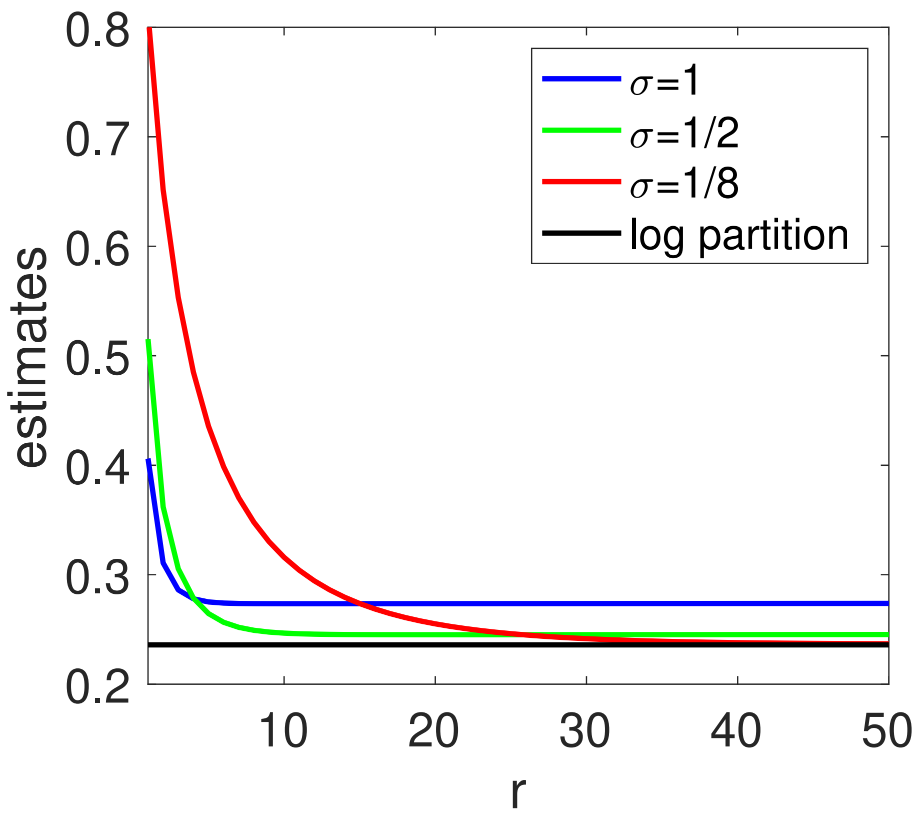
$$b(f) = \sup_{p \text{ measure}} \int_{\mathcal{X}} f(x) dp(x) - D(\Sigma_p \| \Sigma_q)$$

  – If $f(x) = \langle \varphi(x), H\varphi(x) \rangle$, $\; b(f) = \sup_{p \text{ measure}} \operatorname{tr}[H\Sigma_p] - D(\Sigma_p \| \Sigma_q)$

  – Computable by semi-definite programming

Log-partition functions and variational inference

- **Simple example**
  - $\mathcal{X} = [0,1]$, $f(x) = \cos(2\pi x)$, with $\log(\int_0^1 e^{f(x)} dx) \approx 0.2359$
  - $\hat{\varphi}(x)_\omega = \hat{q}(\omega) e^{2i\pi\omega x}$, for $\omega \in \{-r, \dots, r\}$



Relationship with optimization

- **Adding a temperature** (regular entropy and partition function):

$$\varepsilon \log \int_{\mathcal{X}} e^{\frac{1}{\varepsilon} f(x)} dq(x) = \sup_{p \text{ probability}} \int_{\mathcal{X}} f(x) dp(x) - \varepsilon D(p\|q)$$

  - When $\varepsilon \to 0$, converges to $\displaystyle \sup_{p \text{ probability}} \int_{\mathcal{X}} f(x) dp(x) = \sup_{x \in \mathcal{X}} f(x)$
  - What about for kernel entropies?

Relationship with optimization

- **Adding a temperature** (regular entropy and partition function):

$$\varepsilon \log \int_{\mathcal{X}} e^{\frac{1}{\varepsilon} f(x)} dq(x) = \sup_{p \text{ probability}} \int_{\mathcal{X}} f(x) dp(x) - \varepsilon D(p \| q)$$

  - When $\varepsilon \to 0$, converges to $\displaystyle \sup_{p \text{ probability}} \int_{\mathcal{X}} f(x) dp(x) = \sup_{x \in \mathcal{X}} f(x)$
  - What about for kernel entropies?

- **"Sum-of-squares" optimization of** $f(x) = \langle \varphi(x), F\varphi(x) \rangle$

$$\max_{\int_{\mathcal{X}} dp(x) = 1} \operatorname{tr} \Big[F \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x) \Big] \text{ such that } \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x) \succcurlyeq 0$$

  - Kernel sums-of-squares (Rudi, Marteau-Ferey, and Bach, 2020)
  - Extends polynomial formulations (Lasserre, 2001; Parrilo, 2003)

Extensions

- f-**divergences**: $D(p\|q) = \int_{\mathcal{X}} f\left(\frac{dp}{dq}(x)\right) dq(x)$
  - Need f operator convex (KL, squared Hellinger, Pearson, χ^2)
  - All properties are preserved

Extensions

- f-**divergences**: $D(p\|q) = \int_{\mathcal{X}} f\left(\frac{dp}{dq}(x)\right) dq(x)$
  - Need f operator convex (KL, squared Hellinger, Pearson, χ^2)
  - All properties are preserved

- **Other notions of quantum divergences** (Matsumoto, 2015)

$$\operatorname{tr}\left[A \log(B^{-1/2} A B^{-1/2})\right] \geqslant \operatorname{tr}\left[A(\log A - \log B)\right]$$

Extensions

- f-**divergences**: $D(p\|q) = \int_{\mathcal{X}} f\Big(\frac{dp}{dq}(x)\Big)dq(x)$
  - Need f operator convex (KL, squared Hellinger, Pearson, χ^2)
  - All properties are preserved

- **Other notions of quantum divergences** (Matsumoto, 2015)

$$\mathrm{tr}\left[A\log(B^{-1/2}AB^{-1/2})\right] \geqslant \mathrm{tr}\left[A(\log A - \log B)\right]$$

- **Optimal lower-bound**

$$\inf_{p,q \text{ probability measures}} D(p\|q) \quad \text{such that} \quad \Sigma_p = A \text{ and } \Sigma_q = B$$

  - Tractable sum-of-squares relaxations
  - See `https://arxiv.org/abs/2206.13285` for details

Discussion

- **Is this just a Gaussian assumption in feature space?**
  - No, as this would lead to (up to constants)

$$\frac{1}{2}\,\mathrm{tr}[\Sigma_p \Sigma_q^{-1}] - \frac{1}{2}\log\det[\Sigma_p \Sigma_q^{-1}]$$

Discussion

- **Is this just a Gaussian assumption in feature space?**
  - No, as this would lead to (up to constants)

$$\frac{1}{2}\,\mathrm{tr}[\Sigma_p \Sigma_q^{-1}] - \frac{1}{2}\log\det[\Sigma_p \Sigma_q^{-1}]$$

- **Any links with quantum mechanics / information theory?**
  - Balian (1992, 2014); Wilde (2013)
  - We consider only a subclass of density matrices

$$\Sigma_p = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$$

Discussion

- **Is this just a Gaussian assumption in feature space?**
  - No, as this would lead to (up to constants)

$$\frac{1}{2}\operatorname{tr}[\Sigma_p \Sigma_q^{-1}] - \frac{1}{2}\log\det[\Sigma_p \Sigma_q^{-1}]$$

- **Any links with quantum mechanics / information theory?**
  - Balian (1992, 2014); Wilde (2013)
  - We consider only a subclass of density matrices

$$\Sigma_p = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$$

- **Any links with quantum computing?**

Conclusion

- **Information theory with kernel methods**
  - Quantum entropies applied to covariance operators
  - Precise relationships with Shannon entropies
  - Estimation with no optimization
  - Applications to variational inference

Conclusion

- **Information theory with kernel methods**
  - Quantum entropies applied to covariance operators
  - Precise relationships with Shannon entropies
  - Estimation with no optimization
  - Applications to variational inference

- **Extensions / applications**
  - Large-scale algorithms (Bach, 2022b)
  - Structured objects beyond finite sets and $\mathbb{R}^d$
  - Differential privacy (Domingo-Enrich and Mroueh, 2022)
  - Variational inference beyond Gaussian or discrete variables

References

Francis Bach. Information theory with kernel methods. Technical Report 2202.08545, arXiv, 2022a.

Francis Bach. Sum-of-squares relaxations for information theory and variational inference. Technical Report 2206.13285, arXiv, 2022b.

Roger Balian. *Physique Statistique*. Ecole Polytechnique, 1992.

Roger Balian. The entropy-based quantum metric. *Entropy*, 16(7):3878–3888, 2014.

Christos Boutsidis, Michael W. Mahoney, and Petros Drineas. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the Symposium on Discrete algorithms*, pages 968–977, 2009.

Venkat Chandrasekaran and Parikshit Shah. Relative entropy optimization and its applications. *Mathematical Programming*, 161(1):1–32, 2017.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1999.

Imre Csiszár. Axiomatic characterizations of information measures. *Entropy*, 10(3):261–273, 2008.

Carles Domingo-Enrich and Youssef Mroueh. Auditing differential privacy in high dimensions with the kernel quantum Renyi divergence. *arXiv preprint arXiv:2205.13941*, 2022.

Luis Gonzalo Sanchez Giraldo, Murali Rao, and Jose C. Principe. Measures of entropy from data using infinitely divisible kernels. *IEEE Transactions on Information Theory*, 61(1):535–548, 2014.

Yanjun Han, Jiantao Jiao, Tsachy Weissman, and Yihong Wu. Optimal rates of entropy estimation over Lipschitz balls. *The Annals of Statistics*, 48(6):3228–3250, 2020.

Michael I. Jordan and Martin J. Wainwright. Semidefinite relaxations for approximate inference on graphs with cycles. *Advances in Neural Information Processing Systems*, 16, 2003.

Jean-Bernard Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

Keiji Matsumoto. A new quantum version of f-divergence. In *Nagoya Winter Workshop: Reality and Measurement in Algebraic Quantum Theory*, pages 229–273. Springer, 2015.

Charles A. Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(12), 2006.

Hà Quang Minh. Quantum Jensen-Shannon divergences between infinite-dimensional positive definite operators. In *International Conference on Geometric Science of Information*, pages 154–162. Springer, 2021.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trend in Machine Learning*, 10 (1-2):1–141, 2017.

Pablo A. Parrilo. Semidefinite programming relaxations for semialgebraic problems. *Mathematical Programming*, 96(2):293–320, 2003.

Dénes Petz. Sufficient subalgebras and the relative entropy of states of a von Neumann algebra. *Communications in Mathematical Physics*, 105(1):123–131, 1986.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. *Advances in Neural Information Processing Systems*, 28, 2015.

Alessandro Rudi, Ulysse Marteau-Ferey, and Francis Bach. Finding global minima via kernel approximations. Technical Report 2012.11978, arXiv, 2020.

Mary Beth Ruskai. Another short and elementary proof of strong subadditivity of quantum entropy. *Reports on Mathematical Physics*, 60(1):1–12, 2007.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2(Nov):67–93, 2001.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

John von Neumann. *Mathematische Grundlagen der Quantenmechanik*. Springer Berlin, 1932.

Martin J. Wainwright and Michael I. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., 2008.

Mark M. Wilde. *Quantum Information Theory*. Cambridge University Press, 2013.