Information Theory with Kernel Methods

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France

July 2022

Studying probability distributions through moments

- **Moments of feature map $\varphi : \mathcal{X} \to \mathcal{H}$ Hilbert space**
  - 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**
  - 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

Studying probability distributions through moments

- **Moments of feature map $\varphi : \mathcal{X} \to \mathcal{H}$ Hilbert space**
  - 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, etc.

Studying probability distributions through moments

- **Moments of feature map $\varphi : \mathcal{X} \to \mathcal{H}$ Hilbert space**
  - 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, etc.
- **Any link with information-theoretic quantities?**

From mean element to covariance operator

- **Covariance operator** $\Sigma_p = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$
  – From $\mathcal{H}$ to $\mathcal{H}$, defined as $\langle f, \Sigma_p g \rangle = \int_{\mathcal{X}} \langle f, \varphi(x) \rangle \langle g, \varphi(x) \rangle dp(x)$
  – Self-adjoint, positive-semidefinite

From mean element to covariance operator

- **Covariance operator** $\Sigma_p = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$
  - From $\mathcal{H}$ to $\mathcal{H}$, defined as $\langle f, \Sigma_p g \rangle = \int_{\mathcal{X}} \langle f, \varphi(x) \rangle \langle g, \varphi(x) \rangle dp(x)$
  - Self-adjoint, 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** $\Sigma_p = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$
  - From $\mathcal{H}$ to $\mathcal{H}$, defined as $\langle f, \Sigma_p g \rangle = \int_{\mathcal{X}} \langle f, \varphi(x)\rangle \langle g, \varphi(x)\rangle dp(x)$
  - Self-adjoint, positive-semidefinite

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

- **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

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 $\forall x \in \mathcal{X}$, $k(x,x) \leqslant 1$
  - Defines a reproducing kernel Hilbert space (RKHS) of functions

$$\begin{aligned} \varphi(x) &= k(\cdot, x) \\ f(x) &= \langle f, \varphi(x) \rangle \text{ with norm } \|f\|^2 \\ k(x,y) &= \langle k(\cdot,x), k(\cdot,y) \rangle = \langle \varphi(x), \varphi(y) \rangle \end{aligned}$$

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 $\forall x \in \mathcal{X}$, $k(x, x) \leqslant 1$
  - Defines a reproducing kernel Hilbert space (RKHS) of functions

$$\begin{aligned} \varphi(x) &= k(\cdot, x) \\ f(x) &= \langle f, \varphi(x) \rangle \text{ with norm } \|f\|^2 \\ k(x, y) &= \langle k(\cdot, x), k(\cdot, y) \rangle = \langle \varphi(x), \varphi(y) \rangle \end{aligned}$$

  - Universal kernel (Steinwart, 2001): RKHS 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 $\Sigma_p = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$

- **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

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

- **Torus** $\mathcal{X} = [0, 1]^d$
  - $k(x, y) = q(x - y)$, q 1-periodic, with positive Fourier series $\hat{q}$
  - Corresponds to $\varphi(x)_\omega = \hat{q}(\omega)^{1/2} e^{i\omega^\top x}$, $\omega \in \mathbb{Z}^d$
  - Example: $\hat{q}(\omega) \propto \exp(-\sigma \|\omega\|_1)$

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

- **Torus $\mathcal{X} = [0,1]^d$**
  - $k(x,y) = q(x-y)$, q 1-periodic, with positive Fourier series $\hat{q}$
  - Corresponds to $\varphi(x)_\omega = \hat{q}(\omega)^{1/2} e^{i\omega^\top x}$, $\omega \in \mathbb{Z}^d$
  - Example: $\hat{q}(\omega) \propto \exp(-\sigma\|\omega\|_1)$

- **Finite sets**
  - Orthonormal embeddings $\langle \varphi(x), \varphi(y) \rangle = 1_{x=y}$
  - $\mathcal{X} = \{-1,1\}^d$, with $\varphi(x)$ composed of monomials

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): $\mathrm{tr}\left[A \log A\right] = \sum_{\lambda \in \Lambda(A)} \lambda \log \lambda$
  - $\Lambda(A)$ set of eigenvalues of A
- **Relative entropy**: $D(A\|B) = \mathrm{tr}[A(\log A - \log B) - A + B]$
  - Kullback-Leibler divergence

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

- **Relative entropy**: $D(A\|B) = \mathrm{tr}[A(\log A - \log B) - A + B]$
  - Kullback-Leibler divergence

- **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
  - $D\Big(\sum_{i=1}^n C_i A C_i^* \Big\| \sum_{i=1}^n C_i B C_i^*\Big) \leqslant D(A\|B)$ if $\sum_{i=1}^n C_i^* C_i = I$
  - Applications to matrix concentration inequalities (Tropp, 2015)

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

- **Properties**
  - Finite if $\big\|\frac{dp}{dq}\big\|_\infty$ 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 entropies 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** $\forall x \in \mathcal{X}$, $\|\varphi(x)\|^2 \leqslant 1$

$$
\begin{aligned}
D(\Sigma_p \| \Sigma_q) &= D\Big(\int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x) \Big\| \int_{\mathcal{X}} \frac{dq}{dp}(x)\varphi(x)\varphi(x)^* dp(x) \Big) \\
&\leqslant \int_{\mathcal{X}} D\Big(\varphi(x)\varphi(x)^* \Big\| \frac{dq}{dp}(x)\varphi(x)\varphi(x)^* \Big) dp(x) \\
&= \int_{\mathcal{X}} \|\varphi(x)\|^2 D\Big(1 \Big\| \frac{dq}{dp}(x) \Big) dp(x) \\
&\leqslant \int_{\mathcal{X}} \log\Big(\frac{dp}{dq}(x) \Big) dp(x) = D(p\|q)
\end{aligned}
$$

Lower bound on Shannon relative entropy

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

$$
\begin{aligned}
D(\Sigma_p \| \Sigma_q) &= D\Big(\int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x) \Big\| \int_{\mathcal{X}} \frac{dq}{dp}(x)\varphi(x)\varphi(x)^* dp(x) \Big) \\
&\leqslant \int_{\mathcal{X}} D\Big(\varphi(x)\varphi(x)^* \Big\| \frac{dq}{dp}(x)\varphi(x)\varphi(x)^* \Big) dp(x) \\
&= \int_{\mathcal{X}} \|\varphi(x)\|^2 D\Big(1 \Big\| \frac{dq}{dp}(x) \Big) dp(x) \\
&\leqslant \int_{\mathcal{X}} \log\Big(\frac{dp}{dq}(x) \Big) dp(x) = D(p\|q)
\end{aligned}
$$

- **How tight?**
  - Define Σ the covariance operator for the uniform distribution τ

Lower-bound on kernel relative entropies

- **Quantum measurement**
  - 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 $\displaystyle\int_{\mathcal{X}} D(y)d\tau(y) = I$

Lower-bound on kernel relative entropies

- **Quantum measurement**
  - 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

$$\tilde{p}(y) = \int_{\mathcal{X}} \langle \varphi(x), \Sigma^{-1/2}\varphi(y)\rangle^2 dp(x) = \int_{\mathcal{X}} h(x,y)dp(x)$$

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

Lower-bound on kernel relative entropies

- **Quantum measurement**
  - 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

$$\tilde{p}(y) = \int_{\mathcal{X}} \langle \varphi(x), \Sigma^{-1/2}\varphi(y)\rangle^2 dp(x) = \int_{\mathcal{X}} h(x,y)dp(x)$$

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 continuous distributions

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

$$0 \leqslant \ D(p\|q) - D(\tilde{p}\|\tilde{q}) \ \leqslant E(p,q) \times \ \sup_{x \in \mathcal{X}} \int_{\mathcal{X}} h(x,y) d(x,y)^2 dy$$

  – leading to the same bound for $D(p\|q) - D(\Sigma_p\|\Sigma_q)$
  – Explicit constant $E(p,q)$, see Bach (2022a)

Small-width asymptotics for continuous distributions

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

$$0 \leqslant \ D(p\|q) - D(\tilde{p}\|\tilde{q}) \ \leqslant E(p,q) \times \ \sup_{x \in \mathcal{X}} \int_{\mathcal{X}} h(x,y) d(x,y)^2 dy$$

  – leading to the same bound for $D(p\|q) - D(\Sigma_p\|\Sigma_q)$
  – Explicit constant $E(p,q)$, see Bach (2022a)

- **Consequences on the torus**

  – With $\hat{q}(\omega) \propto \exp(-\sigma\|\omega\|_1)$, we have $D(p\|q) - D(\Sigma_p\|\Sigma_q) = O(\sigma^2)$

Estimation from finite sample - I

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

Estimation from finite sample - I

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

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

- **Proposition**: $\mathrm{tr}\left[\hat{\Sigma}_p \log \hat{\Sigma}_p\right] = \mathrm{tr}\left[\frac{1}{n} K \log\left(\frac{1}{n} K\right)\right]$
  - 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)

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 $\frac{dp}{d\tau}(x) \geqslant \alpha$

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

  - No need to regularize

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 $\frac{dp}{d\tau}(x) \geqslant \alpha$

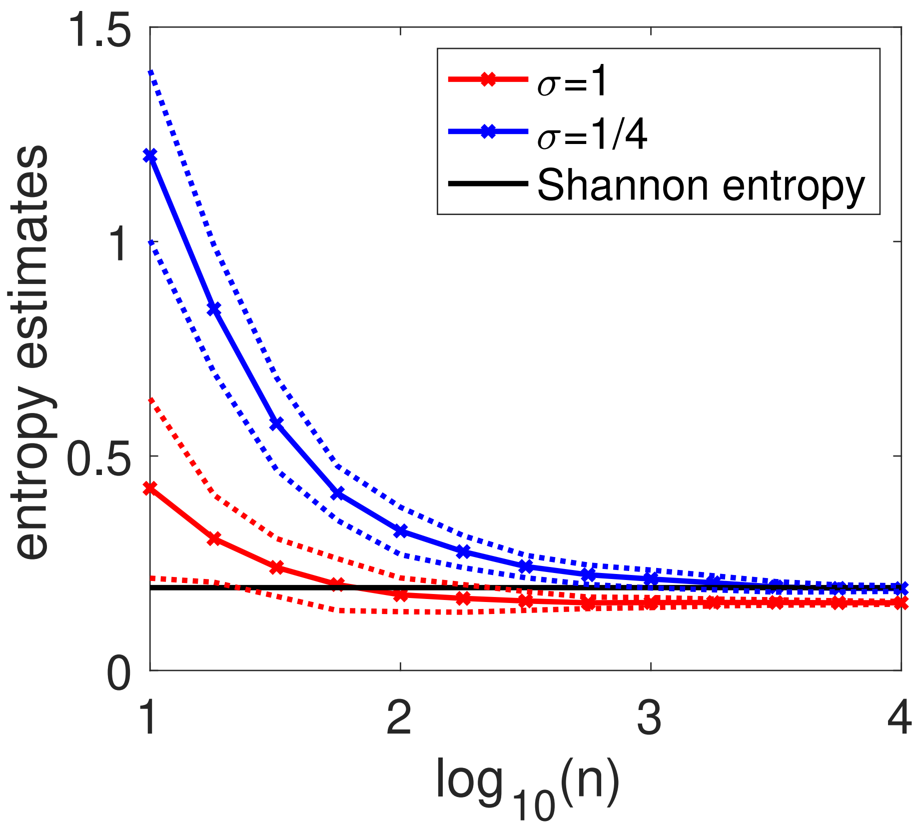
$$\mathbb{E}\Big[\Big|\operatorname{tr}\big[\hat{\Sigma}_p \log \hat{\Sigma}_p\big] - \operatorname{tr}\big[\Sigma_p \log \Sigma_p\big]\Big|\Big] \leqslant \frac{1 + c(8\log n)^2}{n\alpha} + \frac{17}{\sqrt{n}}\big(2\sqrt{c} + \log n\big)$$

  - 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)

Estimation from finite sample - III

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



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

Multivariate probabilistic modelling

- **Product set** $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$
  - Feature space $\mathcal{H}_1 \otimes \mathcal{H}_2$, feature map $\varphi_1 \otimes \varphi_2$
  - Covariance operators $\Sigma_{p_{X_1 X_2}}$ on $\mathcal{H}_1 \otimes \mathcal{H}_2$
  - Covariance operators $\Sigma_{p_{X_1}}$ on $\mathcal{H}_1$, and $\Sigma_{p_{X_2}}$ on $\mathcal{H}_2$

Multivariate probabilistic modelling

- **Product set** $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$
  - Feature space $\mathcal{H}_1 \otimes \mathcal{H}_2$, feature map $\varphi_1 \otimes \varphi_2$
  - Covariance operators $\Sigma_{p_{X_1 X_2}}$ on $\mathcal{H}_1 \otimes \mathcal{H}_2$
  - Covariance operators $\Sigma_{p_{X_1}}$ on $\mathcal{H}_1$, and $\Sigma_{p_{X_2}}$ on $\mathcal{H}_2$
- **Kernel mutual information**
  - Definition: $D(\Sigma_{p_{X_1 X_2}} \| \Sigma_{p_{X_1}} \otimes \Sigma_{p_{X_2}})$
  - Non-negative, equal to zero if and only if X_1 and X_2 are independent

Multivariate probabilistic modelling

- **Product set** $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$
  - Feature space $\mathcal{H}_1 \otimes \mathcal{H}_2$, feature map $\varphi_1 \otimes \varphi_2$
  - Covariance operators $\Sigma_{p_{X_1 X_2}}$ on $\mathcal{H}_1 \otimes \mathcal{H}_2$
  - Covariance operators $\Sigma_{p_{X_1}}$ on $\mathcal{H}_1$, and $\Sigma_{p_{X_2}}$ on $\mathcal{H}_2$

- **Kernel mutual information**
  - Definition: $D(\Sigma_{p_{X_1 X_2}} \| \Sigma_{p_{X_1}} \otimes \Sigma_{p_{X_2}})$
  - Non-negative, equal to zero if and only if X_1 and X_2 are independent

- **Conditional independence**
  - Not as straightforward
  - Data processing inequality $D(\Sigma_{p_{X_1 X_2}} \| \Sigma_{q_{X_1 X_2}}) \geqslant D(\Sigma_{p_{X_1}} \| \Sigma_{q_{X_1}})$

Log-partition functions and variational inference

- **Log-partition function**: given $f : \mathcal{X} \rightarrow \mathbb{R}$ 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)

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)

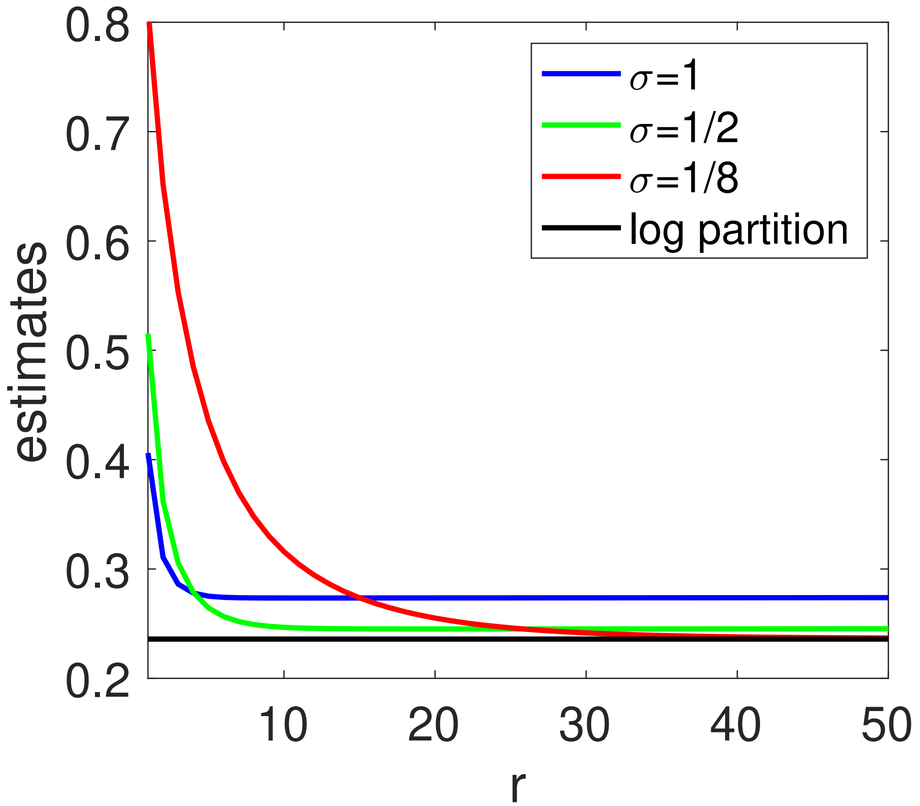
$$b(f) = \sup_{p \text{ measure}} \int_{\mathcal{X}} f(x) dp(x) - D(\Sigma_p\|\Sigma_q)$$

  – 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**: $b_\varepsilon(f) = \sup_{p \text{ measure}} \int_{\mathcal{X}} f(x) dp(x) - \varepsilon D(\Sigma_p \| \Sigma_q)$
- **Convex duality**

$$b_\varepsilon(f) = \inf_M \ \varepsilon \log \operatorname{tr} \exp\Big(\frac{1}{\varepsilon} M + \log \Sigma_q\Big)$$

such that $\forall x \in \mathcal{X}, \ f(x) = \langle \varphi(x), M\varphi(x) \rangle$

Relationship with optimization

- **Adding a temperature**: $b_\varepsilon(f) = \sup_{p \text{ measure}} \int_{\mathcal{X}} f(x) dp(x) - \varepsilon D(\Sigma_p \| \Sigma_q)$
- **Convex duality**

$$b_\varepsilon(f) = \inf_M \ \varepsilon \log \operatorname{tr} \exp\Big(\frac{1}{\varepsilon} M + \log \Sigma_q\Big)$$

$$\text{such that } \forall x \in \mathcal{X}, \ f(x) = \langle \varphi(x), M\varphi(x) \rangle$$

- **Zero temperature limit**: When ε tends to zero, $b_\varepsilon(f)$ converges to

$$\inf_M \ \lambda_{\max}(M) \text{ such that } \ \forall x \in \mathcal{X}, \ f(x) = \langle \varphi(x), M\varphi(x) \rangle$$

$$\Leftrightarrow \inf_{c \in \mathbb{R}, \ A \succcurlyeq 0} \ c \quad \text{such that } \ \forall x \in \mathcal{X}, \ f(x) = c - \langle \varphi(x), A\varphi(x) \rangle$$

  - Optimization formulation of Rudi, Marteau-Ferey, and Bach (2020)
  - Based on "kernel sums-of-squares"

Optimizing bounds

- **Property**: $D(\Sigma_p \| \Sigma_q)$ is concave in the kernel

Optimizing bounds

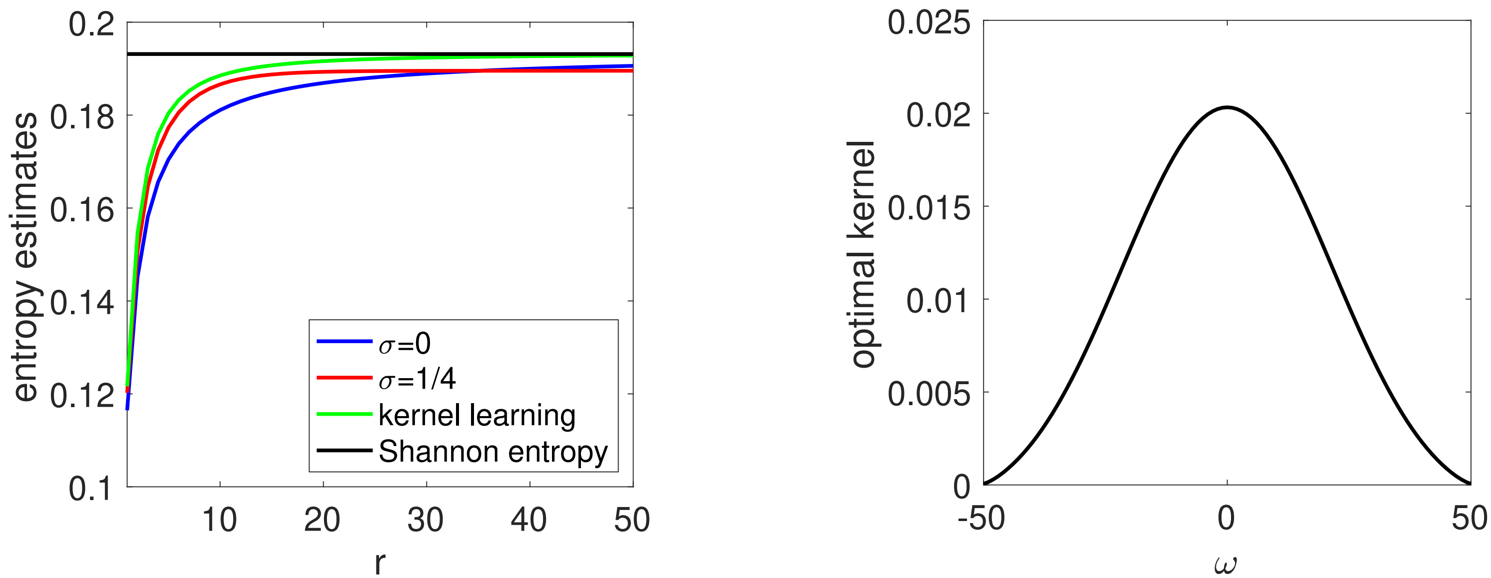
- **Property**: $D(\Sigma_p \| \Sigma_q)$ is concave in the kernel
- **Maximizing lower-bound on entropy**
  - Constraint: $\Lambda \succcurlyeq 0$ such that $\forall x \in \mathcal{X}, \langle \varphi(x), \Lambda \varphi(x) \rangle \leqslant 1$
  - Maximize $D(\Lambda^{1/2} \Sigma_p \Lambda^{1/2} \| \Lambda^{1/2} \Sigma_q \Lambda^{1/2})$

Optimizing bounds

- **Property**: $D(\Sigma_p \| \Sigma_q)$ is concave in the kernel

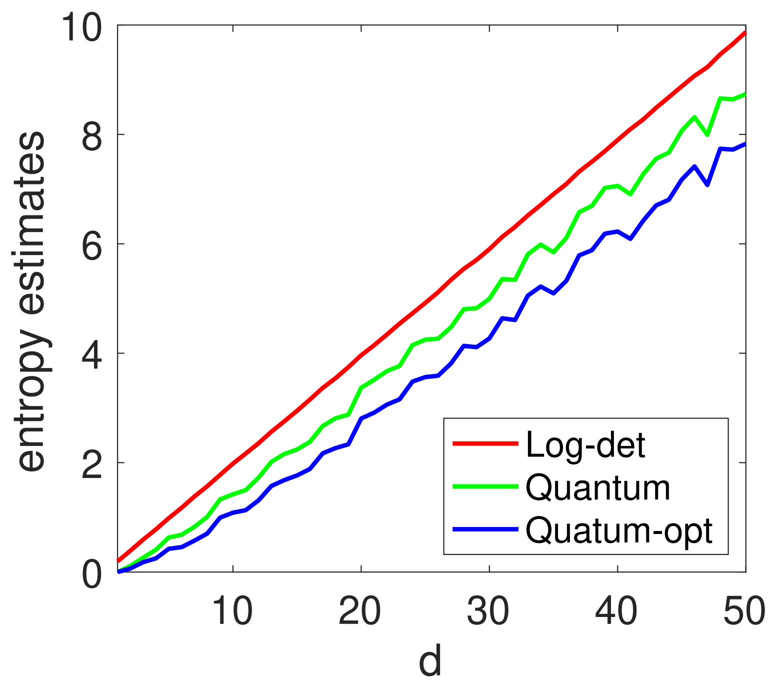
- **Maximizing lower-bound on entropy**
  - Constraint: $\Lambda \succcurlyeq 0$ such that $\forall x \in \mathcal{X}, \langle \varphi(x), \Lambda \varphi(x) \rangle \leqslant 1$
  - Maximize $D(\Lambda^{1/2}\Sigma_p\Lambda^{1/2} \| \Lambda^{1/2}\Sigma_q\Lambda^{1/2})$

- **Illustration for $\mathcal{X} = [0, 1]$**



Optimizing bounds

- **Illustration for** $\mathcal{X} = \{-1, 1\}^d$
  - $\mathcal{X} = \{-1, 1\}^d$, and $\varphi(x) = \mathrm{Diag}(\eta)^{1/2} \begin{pmatrix} x \\ 1 \end{pmatrix} \in \mathbb{R}^{d+1}$
  - Maximize over η in the simplex in $\mathbb{R}^{d+1}$
  - Comparison with log-determinant bound of Jordan and Wainwright (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**: $D(p\|q) = \int_{\mathcal{X}} f\left(\frac{dp}{dq}(x)\right) dq(x)$
  - Need f operator convex (KL, squared Hellinger, Pearson, χ^2)
  - All properties are preserved

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

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

- **Optimal lower-bound**

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

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

Conclusion

- **Information theory with kernel methods**
  - Quantum entropies applied to covariance operators
  - Precise relationships with Shannon entropies
  - Applications to variational inference

Conclusion

- **Information theory with kernel methods**
  - Quantum entropies applied to covariance operators
  - Precise relationships with Shannon entropies
  - Applications to variational inference

- **Extensions**
  - Large-scale algorithms (Bach, 2022b)
  - Structured objects beyond finite sets and $\mathbb{R}^d$

Conclusion

- **Information theory with kernel methods**
  - Quantum entropies applied to covariance operators
  - Precise relationships with Shannon entropies
  - Applications to variational inference
- **Extensions**
  - Large-scale algorithms (Bach, 2022b)
  - Structured objects beyond finite sets and $\mathbb{R}^d$
- **References**
  - `https://arxiv.org/abs/2202.08545`
  - `https://arxiv.org/abs/2206.13285`
  - `https://francisbach.com/information-theory-with-kernel-methods/`

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.

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.

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

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.

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.

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.

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

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.