A spectral framework for closed-form relative density estimation

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France

MLNYC, May 2026

A spectral framework for closed-form relative density estimation
Least-squares is all I need

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France

MLNYC, May 2026

A spectral framework for closed-form relative density estimation
2-minute summary

- **Normalizing constants in machine learning and data science**
  - Probabilistic model: $\frac{dp}{dq}(x) = e^{v(x)-a(v)}$ with $a(v) = \log \int_{\mathcal{X}} e^{v(x)} dq(x)$
  - Learning v is "simple" when $\mathcal{X}$ is finite
  - Difficult otherwise (numerically, and/or statistically)

- **Ubiquitous** (softmax regression, reinforcement learning, transformers, etc.)

A spectral framework for closed-form relative density estimation
2-minute summary

- **Normalizing constants in machine learning and data science**
  - Probabilistic model: $\frac{dp}{dq}(x) = e^{v(x)-a(v)}$ with $a(v) = \log \int_{\mathcal{X}} e^{v(x)} dq(x)$
  - Learning v is "simple" when $\mathcal{X}$ is finite
  - Difficult otherwise (numerically, and/or statistically)
- **Ubiquitous** (softmax regression, reinforcement learning, transformers, etc.)
- **Can the benefits / simplicity / magic of least-squares be used?**
  - Estimation in "closed form" for linear models
  - Many algorithmic and theoretical tools
- **See** `https://arxiv.org/abs/2605.10668` **for all details**

A Kullback-Leibler (KL) divergence view of machine learning

- **KL divergence**: given two distributions p and q on $\mathcal{X}$

$$D(p\|q) = \int_{\mathcal{X}} \log\left(\frac{dp}{dq}(x)\right) dp(x)$$

A Kullback-Leibler (KL) divergence view of machine learning

- **KL divergence**: given two distributions p and q on $\mathcal{X}$

$$D(p\|q) = \int_{\mathcal{X}} \log\left(\frac{dp}{dq}(x)\right) dp(x) = \int_{\mathcal{X}} f\left(\frac{dp}{dq}(x)\right) dq(x) \text{ with } f(t) = t\log t - t + 1$$

  - Example of f-divergence (see, e.g., Polyanskiy and Wu, 2025)
  - Pearson divergence, $f(t) = \frac{1}{2}(t-1)^2$ with

$$D(p\|q) = \frac{1}{2}\int_{\mathcal{X}} \left(\frac{dp}{dq}(x) - 1\right)^2 dq(x) \text{ “}=\text{” } \frac{1}{2}\int_{\mathcal{X}} \frac{(dp(x) - dq(x))^2}{dq(x)}$$

  - NB: different from Bregman divergences

A Kullback-Leibler (KL) divergence view of machine learning

- **Variational formulation**, using $f(t) = \sup_{v \in \mathbb{R}} \{vt - f^*(v)\}$, with f^* Fenchel conjugate of f

$$D(p\|q) = \int_{\mathcal{X}} f\left(\frac{dp}{dq}(x)\right) dq(x) = \sup_{v:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} v(x)dp(x) - \int_{\mathcal{X}} f^*(v(x))dq(x)$$

  - Optimal v defined by $v(x) = f'(dp/dq(x)) = \log(dp/dq(x))$ for KL divergence
  - Broniatowski and Keziou (2006); Nguyen et al. (2010)
  - Equivalent to Donsker-Varadhan variational representation (for KL)

A Kullback-Leibler (KL) divergence view of machine learning

- **Variational formulation**, using $f(t) = \sup_{v \in \mathbb{R}} \{vt - f^*(v)\}$, with f^* Fenchel conjugate of f

$$D(p\|q) = \int_{\mathcal{X}} f\left(\frac{dp}{dq}(x)\right) dq(x) = \sup_{v:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} v(x)dp(x) - \int_{\mathcal{X}} f^*(v(x))dq(x)$$

  - Optimal v defined by $v(x) = f'(dp/dq(x)) = \log(dp/dq(x))$ for KL divergence
  - Broniatowski and Keziou (2006); Nguyen et al. (2010)
  - Equivalent to Donsker-Varadhan variational representation (for KL)

- **Relative density estimation** from samples for unnormalized models

$$\sup_{v:\mathcal{X}\to\mathbb{R}} \frac{1}{n_p}\sum_{i=1}^{n_p} v(x_i) - \frac{1}{n_q}\sum_{j=1}^{n_q} f^*(v(y_j))$$

  - Convex but unstable for KL divergence, where $f^*(v) = e^v - 1$
  - Closed form and stable for linear models?

A Kullback-Leibler (KL) divergence view of machine learning

- **Mutual information** for p distribution on $x = (x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2$
  - If $q(x_1, x_2) = p(x_1)p(x_2) =$ products of marginals, $D(p\|q) = I(p)$

A Kullback-Leibler (KL) divergence view of machine learning

- **Mutual information** for p distribution on $x = (x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2$
  - If $q(x_1, x_2) = p(x_1)p(x_2) =$ products of marginals, $D(p\|q) = I(p)$

- **Variational formulation**

$$D(p\|q) = \sup_{v:\mathcal{X}_1\times\mathcal{X}_2\to\mathbb{R}} \int_{\mathcal{X}_1\times\mathcal{X}_2} v(x_1, x_2)dp(x_1, x_2) - \int_{\mathcal{X}_1\times\mathcal{X}_2} f^*(v(x_1, x_2))dp(x_1)dp(x_2)$$

  - Optimal v defined by $v(x_1, x_2) = f'\left(\frac{dp(x_1,x_2)}{dp(x_1)dp(x_2)}\right) = f'\left(\frac{dp(x_2|x_1)}{dp(x_2)}\right)$
  - For KL divergence $v(x_1, x_2) = \log \frac{dp(x_2|x_1)}{dp(x_2)}$

- **For finite** $\mathcal{X}_2$ (normalized model), equivalent to $v(x_1, x_2) = \log p(x_2|x_1) - \log p(x_2)$
  - Equivalent to logistic / softmax regression: closed form for linear models?

A detour through Pearson divergence

- For $f(t) = \frac{1}{2}(t-1)^2$, we have $f^*(u) = \frac{u^2}{2} + u$

$$D(p\|q) = \sup_{v:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} v(x)dp(x) - \int_{\mathcal{X}} f^*(v(x))dq(x) = \sup_{v:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} v(x)(dp(x)-dq(x)) - \frac{1}{2}\int_{\mathcal{X}} v(x)^2 dq(x)$$

A detour through Pearson divergence

- For $f(t) = \frac{1}{2}(t-1)^2$, we have $f^*(u) = \frac{u^2}{2} + u$

$$D(p\|q) = \sup_{v:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} v(x)dp(x) - \int_{\mathcal{X}} f^*(v(x))dq(x) = \sup_{v:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} v(x)(dp(x)-dq(x)) - \frac{1}{2}\int_{\mathcal{X}} v(x)^2 dq(x)$$

  - Linear model $v(x) = \theta^\top \varphi(x)$, equivalent to $\sup_{\theta\in\mathbb{R}^m} \theta^\top(\mu_p - \mu_q) - \frac{1}{2}\theta^\top \Sigma_q \theta$
  - Solution in closed-form from moments: $\theta = \Sigma_q^{-1}(\mu_p - \mu_q)$
  - Optimal value: $\frac{1}{2}(\mu_p - \mu_q)^\top \Sigma_q^{-1}(\mu_p - \mu_q)$

A detour through Pearson divergence

- For $f(t) = \frac{1}{2}(t-1)^2$, we have $f^*(u) = \frac{u^2}{2} + u$

$$D(p\|q) = \sup_{v:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} v(x)dp(x) - \int_{\mathcal{X}} f^*(v(x))dq(x) = \sup_{v:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} v(x)(dp(x)-dq(x)) - \frac{1}{2}\int_{\mathcal{X}} v(x)^2 dq(x)$$

  - Linear model $v(x) = \theta^\top \varphi(x)$, equivalent to $\sup_{\theta\in\mathbb{R}^m} \theta^\top(\mu_p - \mu_q) - \frac{1}{2}\theta^\top \Sigma_q \theta$
  - Solution in closed-form from moments: $\theta = \Sigma_q^{-1}(\mu_p - \mu_q)$
  - Optimal value: $\frac{1}{2}(\mu_p - \mu_q)^\top \Sigma_q^{-1}(\mu_p - \mu_q)$

- **Consequences**

(1) closed-form formula based on first- and second-order moments
(2) classical simple algorithms (regularization or (S)GD) with sharp analysis
(3) usual extensions (kernelization, neural networks)

A detour through Pearson divergence

- For $f(t) = \frac{1}{2}(t-1)^2$, we have $f^*(u) = \frac{u^2}{2} + u$

$$D(p\|q) = \sup_{v:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} v(x)dp(x) - \int_{\mathcal{X}} f^*(v(x))dq(x) = \sup_{v:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} v(x)(dp(x)-dq(x)) - \frac{1}{2}\int_{\mathcal{X}} v(x)^2 dq(x)$$

  - Linear model $v(x) = \theta^\top \varphi(x)$, equivalent to $\sup_{\theta\in\mathbb{R}^m} \theta^\top(\mu_p - \mu_q) - \frac{1}{2}\theta^\top \Sigma_q \theta$
  - Solution in closed-form from moments: $\theta = \Sigma_q^{-1}(\mu_p - \mu_q)$
  - Optimal value: $\frac{1}{2}(\mu_p - \mu_q)^\top \Sigma_q^{-1}(\mu_p - \mu_q)$

- **Consequences**

(1) closed-form formula based on first- and second-order moments
(2) classical simple algorithms (regularization or (S)GD) with sharp analysis
(3) usual extensions (kernelization, neural networks)

- **Extension to other f-divergences, including the KL divergence?**

Variational formulation with two functions

- **Alternative expression**

$$
\begin{aligned}
D(p\|q) &= \sup_{v:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} v(x)dp(x) - \int_{\mathcal{X}} f^*(v(x))dq(x) \\
&= \sup_{v,w:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} v(x)dp(x) + \int_{\mathcal{X}} w(x)dq(x) \quad \text{such that} \quad \forall x \in \mathcal{X},\ w(x) \leqslant -f^*(v(x))
\end{aligned}
$$

Variational formulation with two functions

- **Alternative expression**

$$D(p\|q) = \sup_{v:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} v(x)dp(x) - \int_{\mathcal{X}} f^*(v(x))dq(x)$$

$$= \sup_{v,w:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} v(x)dp(x) + \int_{\mathcal{X}} w(x)dq(x) \text{ such that } \forall x \in \mathcal{X},\ w(x) \leqslant -f^*(v(x))$$

  - Optimal $v(x) = f'(dp/dq(x))$
  - Optimal $w(x) = -f^*(v(x)) = g'(dq/dp(x))$ with $g(t) = tf(1/t)$
  - Preserves symmetry
  - Still exact, but potentially more flexible

Weighted chi-squared divergence ($\rho \in [0, 1]$)

- For $f(t) = \dfrac{1}{2}\dfrac{(t-1)^2}{\rho t + 1 - \rho} = \sup\limits_{u \in \mathbb{R}} (t-1)u - \dfrac{u^2}{2}(\rho t + 1 - \rho)$

Weighted chi-squared divergence ($\rho \in [0,1]$)

- For $f(t) = \dfrac{1}{2}\dfrac{(t-1)^2}{\rho t + 1 - \rho} = \sup_{u \in \mathbb{R}} (t-1)u - \dfrac{u^2}{2}(\rho t + 1 - \rho)$, we have

$$\begin{aligned} D(p\|q) &= \int_{\mathcal{X}} f\Big(\frac{dp}{dq}(x)\Big) dq(x) \\ &= \sup_{u:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} \Big[u(x)\Big(\frac{dp}{dq}(x) - 1\Big) - \frac{u(x)^2}{2}\Big(\rho\frac{dp}{dq}(x) + 1 - \rho\Big)\Big] dq(x) \end{aligned}$$

Weighted chi-squared divergence ($\rho \in [0,1]$)

- For $f(t) = \dfrac{1}{2}\dfrac{(t-1)^2}{\rho t + 1 - \rho} = \sup_{u \in \mathbb{R}} (t-1)u - \dfrac{u^2}{2}(\rho t + 1 - \rho)$, we have

$$\begin{aligned} D(p\|q) &= \int_{\mathcal{X}} f\Big(\frac{dp}{dq}(x)\Big) dq(x) \\ &= \sup_{u:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} \Big[u(x)\Big(\frac{dp}{dq}(x) - 1\Big) - \frac{u(x)^2}{2}\Big(\rho\frac{dp}{dq}(x) + 1 - \rho\Big)\Big] dq(x) \\ &= \sup_{u:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} \Big[u(x) - \frac{\rho}{2}u(x)^2\Big] dp(x) + \int_{\mathcal{X}} \Big[-u(x) - \frac{1-\rho}{2}u(x)^2\Big] dq(x), \end{aligned}$$

  - Potentials $v(x) = u(x) - \frac{\rho}{2}u(x)^2$ and $w(x) = -u(x) - \frac{1-\rho}{2}u(x)^2$

Weighted chi-squared divergence ($\rho \in [0,1]$)

- For $f(t) = \dfrac{1}{2}\dfrac{(t-1)^2}{\rho t + 1 - \rho} = \sup_{u \in \mathbb{R}} (t-1)u - \dfrac{u^2}{2}(\rho t + 1 - \rho)$, we have

$$D(p\|q) = \sup_{u:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} \left[u(x) - \frac{\rho}{2}u(x)^2\right] dp(x) + \int_{\mathcal{X}} \left[-u(x) - \frac{1-\rho}{2}u(x)^2\right] dq(x),$$

  - Potentials $v(x) = u(x) - \frac{\rho}{2}u(x)^2$ and $w(x) = -u(x) - \frac{1-\rho}{2}u(x)^2$

Weighted chi-squared divergence ($\rho \in [0, 1]$)

- For $f(t) = \dfrac{1}{2}\dfrac{(t-1)^2}{\rho t + 1 - \rho} = \sup_{u \in \mathbb{R}} (t-1)u - \dfrac{u^2}{2}(\rho t + 1 - \rho)$, we have

$$D(p\|q) = \sup_{u : \mathcal{X} \to \mathbb{R}} \int_{\mathcal{X}} \left[u(x) - \frac{\rho}{2} u(x)^2 \right] dp(x) + \int_{\mathcal{X}} \left[-u(x) - \frac{1-\rho}{2} u(x)^2 \right] dq(x),$$

  - Potentials $v(x) = u(x) - \frac{\rho}{2}u(x)^2$ and $w(x) = -u(x) - \frac{1-\rho}{2}u(x)^2$

- **Linear model** $v(x) = \theta^\top \varphi(x)$, equivalent to $\sup_{\theta \in \mathbb{R}^m} \theta^\top (\mu_p - \mu_q) - \dfrac{1}{2}\theta^\top (\rho \Sigma_p + (1-\rho)\Sigma_q)\theta$

  - Solution in closed-form from moments: $\theta = (\rho \Sigma_p + (1-\rho)\Sigma_q)^{-1}(\mu_p - \mu_q)$
  - Optimal value: $\frac{1}{2}(\mu_p - \mu_q)^\top (\rho \Sigma_p + (1-\rho)\Sigma_q)^{-1}(\mu_p - \mu_q)$

- **Same benefits from least-squares estimation**

Integral representations of f-divergences

- **Which functions can be written as** $\int_0^1 \frac{1}{2}\frac{(t-1)^2}{\rho t + 1 - \rho} d\nu(\rho)$ **for a probability measure** ν**?**

Integral representations of f-divergences

- **Which functions can be written as $\int_0^1 \frac{1}{2}\frac{(t-1)^2}{\rho t + 1 - \rho} d\nu(\rho)$ for a probability measure ν?**
  - All normalized $(f(1) = f'(1) = 0,\ f''(1) = 1)$ operator-convex functions

$$\forall \text{ symmetric } A, B,\ \forall \lambda \in [0,1],\ f(\lambda A + (1-\lambda)B) \preccurlyeq \lambda f(A) + (1-\lambda) f(B)$$

  - Lesniewski and Ruskai (1999); Bach (2025)

Integral representations of f-divergences

- **Which functions can be written as $\int_0^1 \frac{1}{2}\frac{(t-1)^2}{\rho t+1-\rho} d\nu(\rho)$ for a probability measure ν?**
  - All normalized $(f(1)=f'(1)=0,\ f''(1)=1)$ operator-convex functions

Divergence	$f(t)$	$f^*(u)$	$d\nu(\rho)$
α-divergence, $\alpha\in[-1,2]$	$\frac{t^\alpha-\alpha t+(\alpha-1)}{\alpha(\alpha-1)}$	$\frac{-1+(1+(\alpha-1)u)^{\alpha/(\alpha-1)}}{\alpha}$	$\frac{2}{\alpha}\frac{\sin(\alpha-1)\pi}{(\alpha-1)\pi}(1-\rho)^\alpha\rho^{1-\alpha}d\rho$
Kullback-Leibler, $\alpha=1$	$t\log t-t+1$	e^u-1	$2(1-\rho)d\rho$
Reverse KL, $\alpha=0$	$-\log t+t-1$	$-\log(1-u)$	$2\rho d\rho$
squared Hellinger, $\alpha=\frac{1}{2}$	$2(\sqrt{t}-1)^2$	$\frac{u}{1-u/2}$	$\frac{8}{\pi}\sqrt{\rho(1-\rho)}d\rho$
Pearson χ^2, $\alpha=2$	$\frac{1}{2}(t-1)^2$	$\frac{u^2}{2}+u$	$\delta_0(\rho)$
Reverse Pearson, $\alpha=-1$	$\frac{1}{2t}(t-1)^2$	$1-\sqrt{1-2u}$	$\delta_1(\rho)$
Le Cam	$\frac{(t-1)^2}{t+1}$	$4-u-4\sqrt{1-u}$	$\delta_{1/2}(\rho)$
Jensen-Shannon	$2t\log\frac{2t}{t+1}+2\log\frac{2}{t+1}$	$-2\log(2-e^{u/2})$	$(2-2\lvert 1-2\rho\rvert)d\rho$

Key new result

$$D(p\|q) = \int_0^1 \frac{1}{2} \frac{(\frac{dp}{dq}(x) - 1)^2}{\rho \frac{dp}{dq}(x) + 1 - \rho} d\nu(\rho)$$

Key new result

$$
\begin{aligned}
D(p\|q) &= \int_0^1 \frac{1}{2} \frac{(\frac{dp}{dq}(x) - 1)^2}{\rho \frac{dp}{dq}(x) + 1 - \rho} d\nu(\rho) \\
&= \int_0^1 \sup_{u_\rho : \mathcal{X} \to \mathbb{R}} \left\{ \int_{\mathcal{X}} \left[u_\rho(x) - \frac{\rho}{2} u_\rho(x)^2 \right] dp(x) + \int_{\mathcal{X}} \left[-u_\rho(x) - \frac{1-\rho}{2} u_\rho(x)^2 \right] dq(x) \right\} d\nu(\rho)
\end{aligned}
$$

Key new result

$$
\begin{aligned}
D(p\|q) &= \int_0^1 \frac{1}{2} \frac{(\frac{dp}{dq}(x)-1)^2}{\rho \frac{dp}{dq}(x) + 1 - \rho} d\nu(\rho) \\
&= \int_0^1 \sup_{u_\rho : \mathcal{X} \to \mathbb{R}} \left\{ \int_{\mathcal{X}} \left[u_\rho(x) - \frac{\rho}{2} u_\rho(x)^2 \right] dp(x) + \int_{\mathcal{X}} \left[-u_\rho(x) - \frac{1-\rho}{2} u_\rho(x)^2 \right] dq(x) \right\} d\nu(\rho) \\
&= \sup_{v,w : \mathcal{X} \to \mathbb{R}} \int_{\mathcal{X}} v(x) dp(x) + \int_{\mathcal{X}} w(x) dq(x) \quad \text{such that} \quad \forall x \in \mathcal{X},\ w(x) \leqslant -f^*(v(x))
\end{aligned}
$$

– With $v(x) = \int_0^1 \left\{ u_\rho(x) - \frac{\rho}{2} u_\rho(x)^2 \right\} d\nu(\rho)$ and $w(x) = \int_0^1 \left\{ -u_\rho(x) - \frac{1-\rho}{2} u_\rho(x)^2 \right\} d\nu(\rho)$

Key new result

$$
\begin{aligned}
D(p\|q) &= \int_0^1 \frac{1}{2} \frac{(\frac{dp}{dq}(x) - 1)^2}{\rho \frac{dp}{dq}(x) + 1 - \rho} d\nu(\rho) \\
&= \int_0^1 \sup_{u_\rho : \mathcal{X} \to \mathbb{R}} \left\{ \int_{\mathcal{X}} \left[u_\rho(x) - \frac{\rho}{2} u_\rho(x)^2 \right] dp(x) + \int_{\mathcal{X}} \left[-u_\rho(x) - \frac{1-\rho}{2} u_\rho(x)^2 \right] dq(x) \right\} d\nu(\rho) \\
&= \sup_{v, w : \mathcal{X} \to \mathbb{R}} \int_{\mathcal{X}} v(x) dp(x) + \int_{\mathcal{X}} w(x) dq(x) \quad \text{such that} \quad \forall x \in \mathcal{X},\ w(x) \leqslant -f^*(v(x))
\end{aligned}
$$

- With $v(x) = \int_0^1 \left\{ u_\rho(x) - \frac{\rho}{2} u_\rho(x)^2 \right\} d\nu(\rho)$ and $w(x) = \int_0^1 \left\{ -u_\rho(x) - \frac{1-\rho}{2} u_\rho(x)^2 \right\} d\nu(\rho)$
- For each $\rho \in [0, 1]$, u_ρ can be estimated in parallel by least-squares
- New divergence when restricting all u_ρ to linear models in φ

$$
F(p\|q, \varphi) = \frac{1}{2} \int_0^1 (\mu_p - \mu_q)^\top (\rho \Sigma_p + (1-\rho) \Sigma_q)^{-1} (\mu_p - \mu_q) d\nu(\rho)
$$

Properties of the new divergence

$$F(p\|q,\varphi) = \frac{1}{2}\int_0^1 (\mu_p - \mu_q)^\top (\rho\Sigma_p + (1-\rho)\Sigma_q)^{-1}(\mu_p - \mu_q) d\nu(\rho)$$

- **Based on moments** $\mu_p = \mathbb{E}_p[\varphi]$, $\mu_q = \mathbb{E}_q[\varphi]$, $\Sigma_p = \mathbb{E}_p[\varphi\varphi^\top]$, $\Sigma_q = \mathbb{E}_q[\varphi\varphi^\top]$

Properties of the new divergence

$$F(p\|q,\varphi) = \frac{1}{2}\int_0^1 (\mu_p - \mu_q)^\top (\rho\Sigma_p + (1-\rho)\Sigma_q)^{-1}(\mu_p - \mu_q)d\nu(\rho)$$

- **Based on moments** $\mu_p = \mathbb{E}_p[\varphi]$, $\mu_q = \mathbb{E}_q[\varphi]$, $\Sigma_p = \mathbb{E}_p[\varphi\varphi^\top]$, $\Sigma_q = \mathbb{E}_q[\varphi\varphi^\top]$
- **Convexity, monotonicity, linear invariance**

Properties of the new divergence

$$F(p\|q,\varphi) = \frac{1}{2}\int_0^1 (\mu_p - \mu_q)^\top (\rho\Sigma_p + (1-\rho)\Sigma_q)^{-1}(\mu_p - \mu_q) d\nu(\rho)$$

- **Based on moments** $\mu_p = \mathbb{E}_p[\varphi]$, $\mu_q = \mathbb{E}_q[\varphi]$, $\Sigma_p = \mathbb{E}_p[\varphi\varphi^\top]$, $\Sigma_q = \mathbb{E}_q[\varphi\varphi^\top]$
- **Convexity, monotonicity, linear invariance**
- **Tightness**: $F(p\|q,\varphi) = D(p\|q) \Leftrightarrow \forall\rho,\ \exists\theta_\rho,\ \forall x \in \mathcal{X}, \dfrac{\frac{dp}{dq}(x) - 1}{\rho\frac{dp}{dq}(x) + 1 - \rho} = \theta_\rho^\top \varphi(x)$

  (all least-squares problems for $\rho \in [0,1]$ are tight)

Properties of the new divergence

$$F(p\|q,\varphi) = \frac{1}{2}\int_0^1 (\mu_p - \mu_q)^\top (\rho\Sigma_p + (1-\rho)\Sigma_q)^{-1}(\mu_p - \mu_q) d\nu(\rho)$$

- **Based on moments** $\mu_p = \mathbb{E}_p[\varphi]$, $\mu_q = \mathbb{E}_q[\varphi]$, $\Sigma_p = \mathbb{E}_p[\varphi\varphi^\top]$, $\Sigma_q = \mathbb{E}_q[\varphi\varphi^\top]$
- **Convexity, monotonicity, linear invariance**
- **Tightness**: $F(p\|q,\varphi) = D(p\|q) \Leftrightarrow \forall\rho,\ \exists\theta_\rho,\ \forall x \in \mathcal{X}, \dfrac{\frac{dp}{dq}(x) - 1}{\rho\frac{dp}{dq}(x) + 1 - \rho} = \theta_\rho^\top \varphi(x)$

  (all least-squares problems for $\rho \in [0,1]$ are tight)
- **Importation of existing least-squares results by integration**

Properties of the new divergence

$$F(p\|q,\varphi) = \frac{1}{2}\int_0^1 (\mu_p - \mu_q)^\top (\rho\Sigma_p + (1-\rho)\Sigma_q)^{-1}(\mu_p - \mu_q) d\nu(\rho)$$

- **Based on moments** $\mu_p = \mathbb{E}_p[\varphi]$, $\mu_q = \mathbb{E}_q[\varphi]$, $\Sigma_p = \mathbb{E}_p[\varphi\varphi^\top]$, $\Sigma_q = \mathbb{E}_q[\varphi\varphi^\top]$
- **Convexity, monotonicity, linear invariance**
- **Tightness**: $F(p\|q,\varphi) = D(p\|q) \Leftrightarrow \forall\rho,\ \exists\theta_\rho,\ \forall x \in \mathcal{X}, \dfrac{\frac{dp}{dq}(x) - 1}{\rho\frac{dp}{dq}(x) + 1 - \rho} = \theta_\rho^\top \varphi(x)$

  (all least-squares problems for $\rho \in [0,1]$ are tight)
- **Importation of existing least-squares results by integration**
- **Algorithmic feasibility?**
  - Avoiding integration with respect to ρ
  - Beyond linear models

Algorithmic feasibility: avoiding integral in ρ

- **Main computational result** for m-dimensional feature maps
  - Generalized eigenvalue decomposition of the pair (Σ_p, Σ_q)
  - Basis $(v_i)_{i \in \{1,\dots,m\}}$ such that $\forall i, j,\ v_i^\top \Sigma_q v_j = 1_{i=j}$ and $\Sigma_p v_i = \lambda_i \Sigma_q v_i$

Algorithmic feasibility: avoiding integral in ρ

- **Main computational result** for m-dimensional feature maps
  - Generalized eigenvalue decomposition of the pair (Σ_p, Σ_q)
  - Basis $(v_i)_{i\in\{1,\dots,m\}}$ such that $\forall i,j,\ v_i^\top \Sigma_q v_j = 1_{i=j}$ and $\Sigma_p v_i = \lambda_i \Sigma_q v_i$

$$F(p\|q,\varphi) = \sum_{i=1}^{m} \frac{f(\lambda_i)}{(\lambda_i - 1)^2} \big((\mu_p - \mu_q)^\top v_i\big)^2$$

Algorithmic feasibility: avoiding integral in ρ

- **Main computational result** for m-dimensional feature maps
  - Generalized eigenvalue decomposition of the pair (Σ_p, Σ_q)
  - Basis $(v_i)_{i\in\{1,\dots,m\}}$ such that $\forall i, j,\ v_i^\top \Sigma_q v_j = 1_{i=j}$ and $\Sigma_p v_i = \lambda_i \Sigma_q v_i$

$$F(p\|q,\varphi) = \sum_{i=1}^{m} \frac{f(\lambda_i)}{(\lambda_i - 1)^2} \big((\mu_p - \mu_q)^\top v_i\big)^2$$

- **Closed form:** $v(x) = \varphi(x)^\top M \varphi(x) + 2c^\top \varphi(x)$ **and** $w(x) = \varphi(x)^\top N \varphi(x) - 2c^\top \varphi(x)$
  - "Quadratic + linear" functions
  - Corresponds to derivatives of $F(p\|q,\varphi)$ with respect to $\Sigma_p, \Sigma_q, \frac{1}{2}(\mu_p - \mu_q)$

Algorithmic feasibility: avoiding integral in ρ

- **Main computational result** for m-dimensional feature maps
  - Generalized eigenvalue decomposition of the pair (Σ_p, Σ_q)
  - Basis $(v_i)_{i\in\{1,\dots,m\}}$ such that $\forall i, j,\ v_i^\top \Sigma_q v_j = 1_{i=j}$ and $\Sigma_p v_i = \lambda_i \Sigma_q v_i$

$$F(p\|q,\varphi) = \sum_{i=1}^{m} \frac{f(\lambda_i)}{(\lambda_i - 1)^2} \big((\mu_p - \mu_q)^\top v_i\big)^2$$

- **Closed form:** $v(x) = \varphi(x)^\top M \varphi(x) + 2c^\top \varphi(x)$ **and** $w(x) = \varphi(x)^\top N \varphi(x) - 2c^\top \varphi(x)$
  - "Quadratic + linear" functions
  - Corresponds to derivatives of $F(p\|q,\varphi)$ with respect to $\Sigma_p, \Sigma_q, \frac{1}{2}(\mu_p - \mu_q)$

- **Computational complexity**: $O(m^3)$ given moments
  - Can be made lower with feature learning

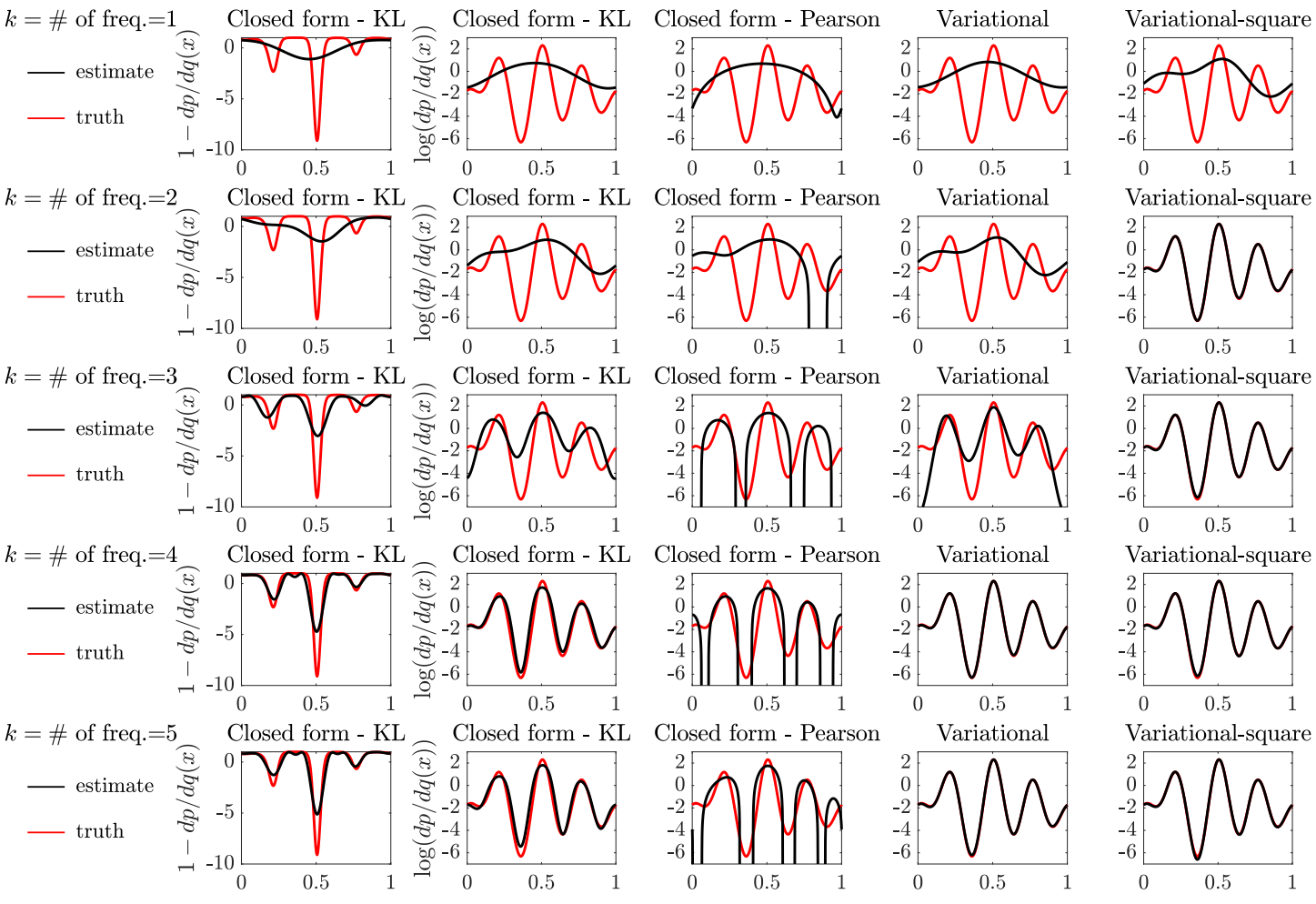
Algorithmic feasibility: avoiding integral in ρ

- **Closed form:** $v(x) = \varphi(x)^\top M \varphi(x) + 2c^\top \varphi(x)$ **and** $w(x) = \varphi(x)^\top N \varphi(x) - 2c^\top \varphi(x)$

$$
\begin{aligned}
c &= \sum_{i=1}^{m} \frac{f(\lambda_i)}{(\lambda_i - 1)^2} v_i v_i^\top (\mu_p - \mu_q) \\
M &= \sum_{i,j=1}^{m} (\mu_p - \mu_q)^\top v_i v_j^\top (\mu_p - \mu_q) \frac{f(\lambda_i)/(\lambda_i - 1)^2 - f(\lambda_j)/(\lambda_j - 1)^2}{\lambda_i - \lambda_j} v_i v_j^\top \\
N &= -\sum_{i,j=1}^{m} (\mu_p - \mu_q)^\top v_i v_j^\top (\mu_p - \mu_q) \frac{f(\lambda_i)\lambda_i/(\lambda_i - 1)^2 - f(\lambda_j)\lambda_j/(\lambda_j - 1)^2}{\lambda_i - \lambda_j} v_i v_j^\top
\end{aligned}
$$

  - Corresponds to derivatives of $F(p\|q, \varphi)$ with respect to $\Sigma_p, \Sigma_q, \frac{1}{2}(\mu_p - \mu_q)$
  - NB: for Pearson divergence, $M = 0$

Illustration



Estimation of $\log dp/dq(x)$ for $x \in [0, 1]$, with cosines

- Closed form - KL
- Closed form - Pearson
- Variational (regular and square features)

Estimation from data

- **Regularization**: $F_\lambda(p\|q,\varphi) = \frac{1}{2}\int_0^1 (\mu_p-\mu_q)^\top(\rho\Sigma_p+(1-\rho)\Sigma_q+\lambda I)^{-1}(\mu_p-\mu_q)d\nu(\rho)$
  - Corresponds to adding penalty $-\frac{\lambda}{2}\int_0^1 \|\theta_\rho\|^2 d\nu(\rho)$ to variational formulation

Estimation from data

- **Regularization**: $F_\lambda(p\|q,\varphi) = \frac{1}{2}\int_0^1 (\mu_p-\mu_q)^\top(\rho\Sigma_p+(1-\rho)\Sigma_q+\lambda I)^{-1}(\mu_p-\mu_q)d\nu(\rho)$
  - Corresponds to adding penalty $-\frac{\lambda}{2}\int_0^1 \|\theta_\rho\|^2 d\nu(\rho)$ to variational formulation
- **Data**: n_p i.i.d. observations $x_1,\ldots,x_{n_p}$ from p, n_q i.i.d. observations $y_1,\ldots,y_{n_q}$ from q
  - Feature matrices $\Phi_p \in \mathbb{R}^{n_p\times m}$ and $\Phi_q \in \mathbb{R}^{n_q\times m}$
  - Empirical moments $\hat{\mu}_p = \frac{1}{n_p}\Phi_p^\top 1_{n_p}$, $\hat{\mu}_q = \frac{1}{n_q}\Phi_q^\top 1_{n_q}$, $\hat{\Sigma}_p = \frac{1}{n_p}\Phi_p^\top\Phi_p$, $\hat{\Sigma}_q = \frac{1}{n_q}\Phi_q^\top\Phi_q$
- **Natural estimator**: $F_\lambda(\hat{p}\|\hat{q},\varphi)$
  - Running cost with spectral formulation: $O(m^2n+m^3)$

Estimation from data

- **Regularization**: $F_\lambda(p\|q,\varphi) = \frac{1}{2}\int_0^1 (\mu_p-\mu_q)^\top(\rho\Sigma_p+(1-\rho)\Sigma_q+\lambda I)^{-1}(\mu_p-\mu_q)d\nu(\rho)$
  - Corresponds to adding penalty $-\frac{\lambda}{2}\int_0^1 \|\theta_\rho\|^2 d\nu(\rho)$ to variational formulation
- **Data**: n_p i.i.d. observations $x_1,\ldots,x_{n_p}$ from p, n_q i.i.d. observations $y_1,\ldots,y_{n_q}$ from q
  - Feature matrices $\Phi_p \in \mathbb{R}^{n_p\times m}$ and $\Phi_q \in \mathbb{R}^{n_q\times m}$
  - Empirical moments $\hat{\mu}_p = \frac{1}{n_p}\Phi_p^\top 1_{n_p}$, $\hat{\mu}_q = \frac{1}{n_q}\Phi_q^\top 1_{n_q}$, $\hat{\Sigma}_p = \frac{1}{n_p}\Phi_p^\top\Phi_p$, $\hat{\Sigma}_q = \frac{1}{n_q}\Phi_q^\top\Phi_q$
- **Natural estimator**: $F_\lambda(\hat{p}\|\hat{q},\varphi)$
  - Running cost with spectral formulation: $O(m^2n+m^3)$
- **Kernelization**: function of only dot-products $\varphi(z)^\top\varphi(z')$ between observations z, z'
  - Running cost: $O(n^3)$ or $O(m^2n)$ using random features

Mutual information

- **Mutual information** for p distribution on $x = (x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2$
  - If $q(x_1, x_2) = p(x_1)p(x_2) =$ products of marginals, $D(p\|q) = I(p)$
- **Potential** $v(x_1, x_2)$ estimate of $f'\left(\frac{dp(x_1,x_2)}{dp(x_1)dp(x_2)}\right) = f'\left(\frac{dp(x_2|x_1)}{dp(x_2)}\right)$

Mutual information

- **Mutual information** for p distribution on $x = (x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2$
  - If $q(x_1, x_2) = p(x_1)p(x_2) =$ products of marginals, $D(p\|q) = I(p)$
- **Potential** $v(x_1, x_2)$ estimate of $f'\big(\frac{dp(x_1,x_2)}{dp(x_1)dp(x_2)}\big) = f'\big(\frac{dp(x_2|x_1)}{dp(x_2)}\big)$
- **Factorized moments** for feature map $\varphi_2(x_2) \otimes \varphi_1(x_1)$

$$\mu_p = \mathbb{E}[\varphi_2(x_2) \otimes \varphi_1(x_1)], \qquad \mu_q = \mathbb{E}[\varphi_2(x_2)] \otimes \mathbb{E}[\varphi_1(x_1)]$$
$$\Sigma_p = \mathbb{E}[\varphi_2(x_2)\varphi_2(x_2)^\top \otimes \varphi_1(x_1)\varphi_1(x_1)^\top], \quad \Sigma_q = \mathbb{E}[\varphi_2(x_2)\varphi_2(x_2)^\top] \otimes \mathbb{E}[\varphi_1(x_1)\varphi_1(x_1)^\top]$$

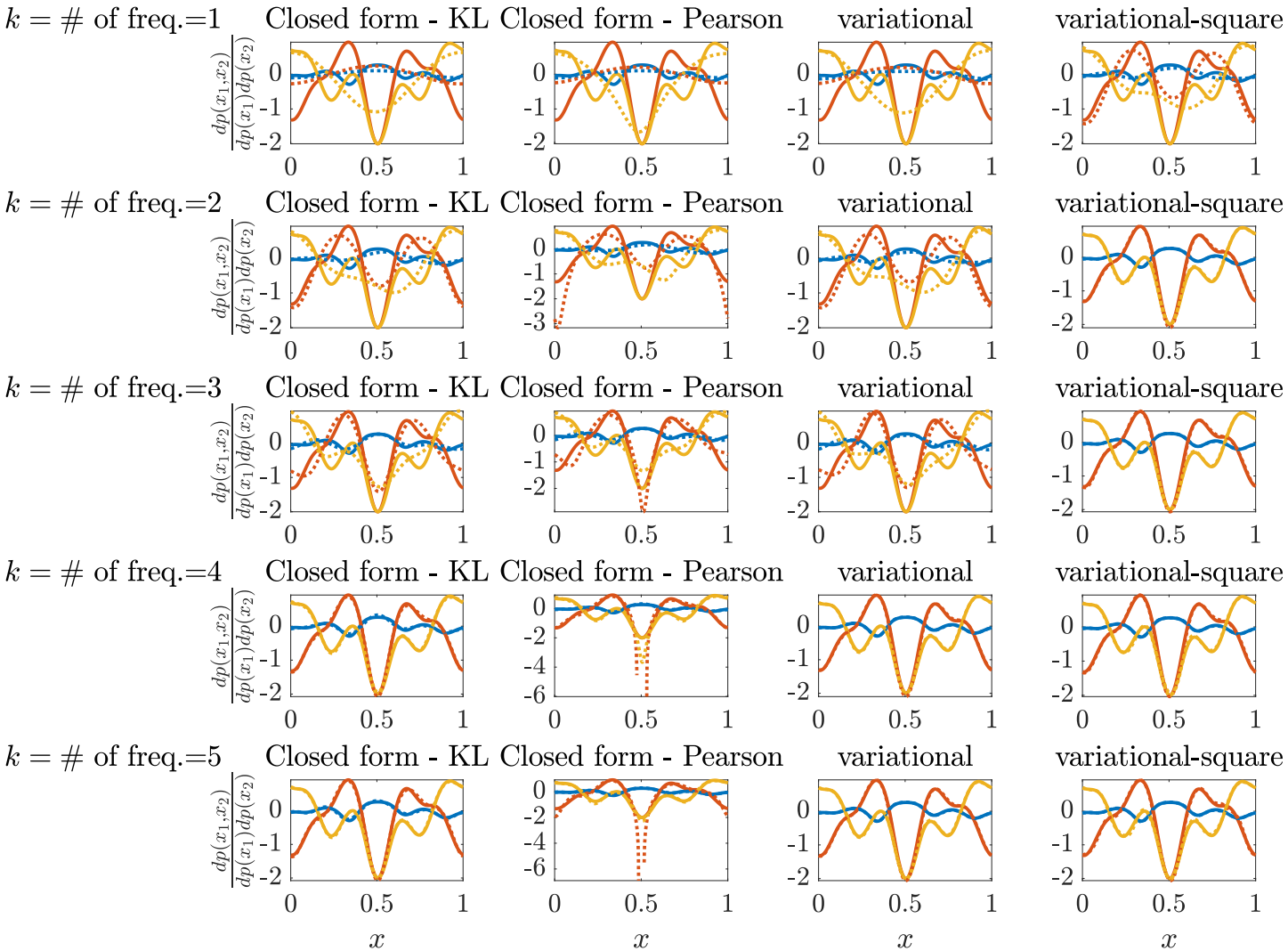
Mutual information

- **Mutual information** for p distribution on $x = (x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2$
  - If $q(x_1, x_2) = p(x_1)p(x_2) =$ products of marginals, $D(p\|q) = I(p)$
- **Potential** $v(x_1, x_2)$ estimate of $f'\left(\frac{dp(x_1,x_2)}{dp(x_1)dp(x_2)}\right) = f'\left(\frac{dp(x_2|x_1)}{dp(x_2)}\right)$
- **Factorized moments** for feature map $\varphi_2(x_2) \otimes \varphi_1(x_1)$

$$\mu_p = \mathbb{E}[\varphi_2(x_2) \otimes \varphi_1(x_1)], \qquad \mu_q = \mathbb{E}[\varphi_2(x_2)] \otimes \mathbb{E}[\varphi_1(x_1)]$$
$$\Sigma_p = \mathbb{E}[\varphi_2(x_2)\varphi_2(x_2)^\top \otimes \varphi_1(x_1)\varphi_1(x_1)^\top], \quad \Sigma_q = \mathbb{E}[\varphi_2(x_2)\varphi_2(x_2)^\top] \otimes \mathbb{E}[\varphi_1(x_1)\varphi_1(x_1)^\top]$$

- **Finite** $\mathcal{X}_2$ (with m_2 elements): Closed-form estimate for logistic / softmax regression
  - Complexity: $O(m_1^2 n + m_2 m_1^3)$ for eigenvalue decompositions
  - Softmax regression using Newton's method: $O(m_2^2 m_1^2 n + m_2^3 m_1^3)$

Illustration

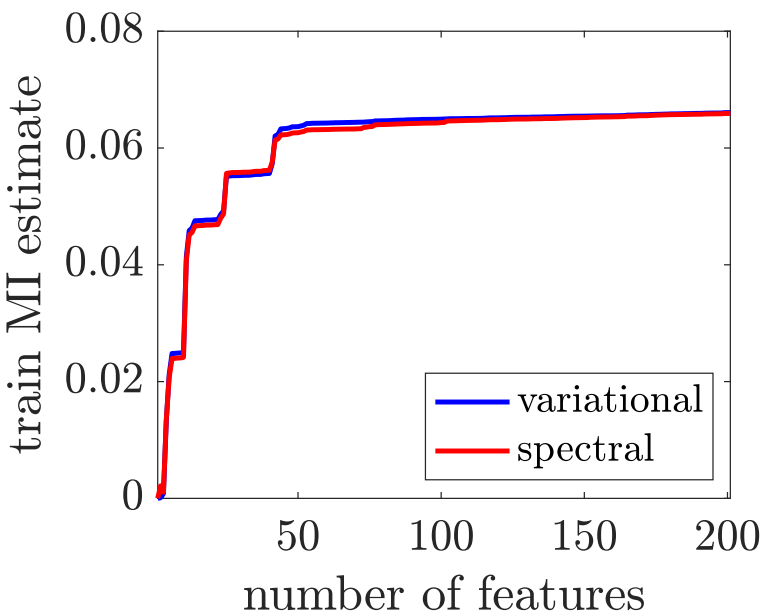


Softmax regression for $x \in [0, 1]$, with cosines

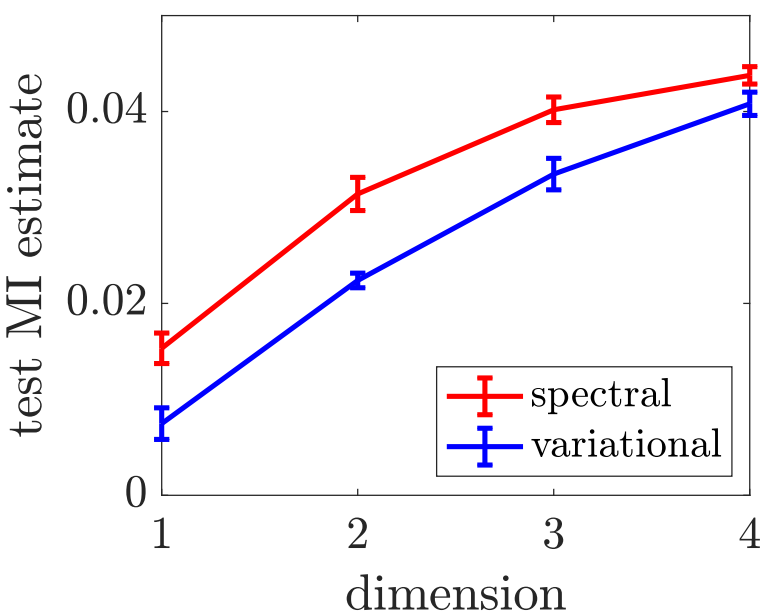
- Closed form - KL
- Closed form - Pearson
- Variational (regular and square features)

Empirical comparison with softmax regression ($\Leftrightarrow$ variational)

- **Synthetic experiments with random neural network features**
  - Training (left): sometimes worse, sometimes better
  - Testing (right): (slightly) worse



Training criterion



Testing criterion

Theoretical results (linear models)

- **Importing results from kernel least-squares regression**
  - Caponnetto and De Vito (2007); Harchaoui et al. (2008); Lin et al. (2020)
- **Simplest setup** $\mathcal{X}$ unit Euclidean ball on $\mathbb{R}^d$
  - p and q with $t \leqslant \kappa + \frac{d+1}{2}$ square-integrable strictly positive densities
  - Kernel obtained from random ReLU features $(w^\top x + b)_+^\kappa$

Theoretical results (linear models)

- **Importing results from kernel least-squares regression**
  - Caponnetto and De Vito (2007); Harchaoui et al. (2008); Lin et al. (2020)
- **Simplest setup** $\mathcal{X}$ unit Euclidean ball on $\mathbb{R}^d$
  - p and q with $t \leqslant \kappa + \frac{d+1}{2}$ square-integrable strictly positive densities
  - Kernel obtained from random ReLU features $(w^\top x + b)_+^\kappa$
- **Proposition:** If $\lambda \propto 1/n^{\frac{\kappa+(d+1)/2}{t+d/2}}$

$$\mathbb{E}\big[|F_\lambda(\hat{p}, \hat{q}, \varphi) - D(p\|q)|\big] = O\Big(n^{-\frac{t}{t+d/2}} + D(p\|q) n^{-\frac{t/2}{t+d/2}} \sqrt{\log n} + n^{-1/2}\sqrt{D(p\|q)}\Big)$$

  - Additive and multiplicative terms
  - Not minimax optimal
  - Adaptivity to smoothness

Debiasing

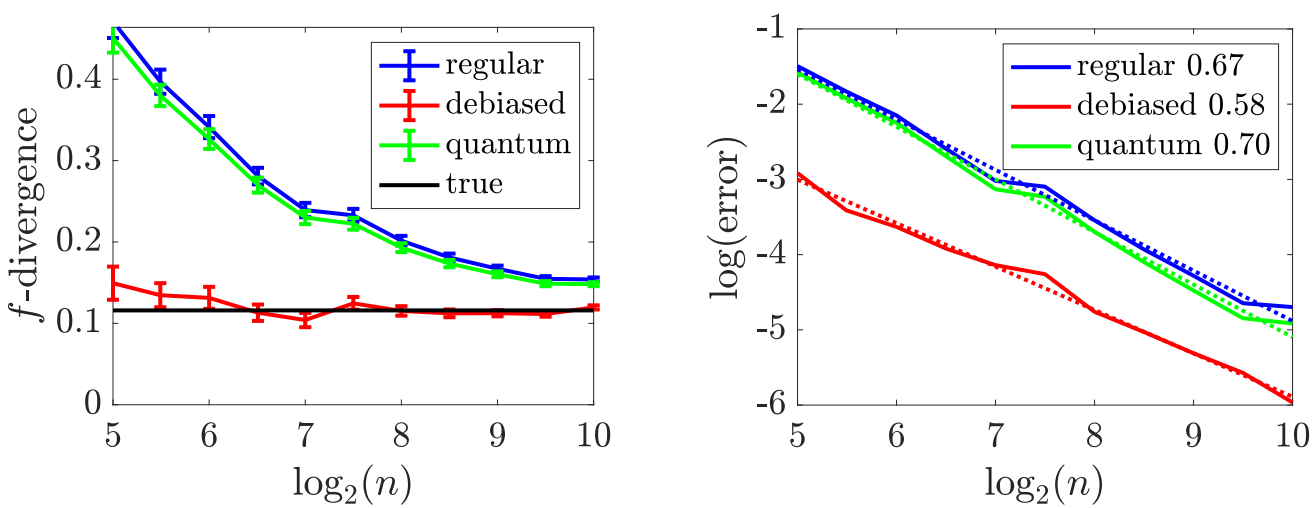
- **Bias due to V-statistic**: $\frac{1}{2}\int_0^1 (\hat{\mu}_p - \hat{\mu}_q)^\top (\rho\hat{\Sigma}_p + (1-\rho)\hat{\Sigma}_q + \lambda I)^{-1}(\hat{\mu}_p - \hat{\mu}_q) d\nu(\rho)$
  - Quadratic forms of means are not unbiased (Laurent, 1996)

Debiasing

- **Bias due to V-statistic**: $\frac{1}{2}\int_0^1 (\hat{\mu}_p - \hat{\mu}_q)^\top (\rho\hat{\Sigma}_p + (1-\rho)\hat{\Sigma}_q + \lambda I)^{-1}(\hat{\mu}_p - \hat{\mu}_q) d\nu(\rho)$
  - Quadratic forms of means are not unbiased (Laurent, 1996)
  - Estimation of bias: $\mathrm{tr}[\hat{C}(\rho\hat{\Sigma}_p + (1-\rho)\hat{\Sigma}_q + \lambda I)^{-1}]$
  - With $\hat{C} = (\hat{\Sigma}_p - \hat{\mu}_p\hat{\mu}_p^\top)/(n_p - 1) + (\hat{\Sigma}_q - \hat{\mu}_q\hat{\mu}_q^\top)/(n_q - 1)$
  - Can be subtracted for improved behavior (and now minimax optimal additive term)

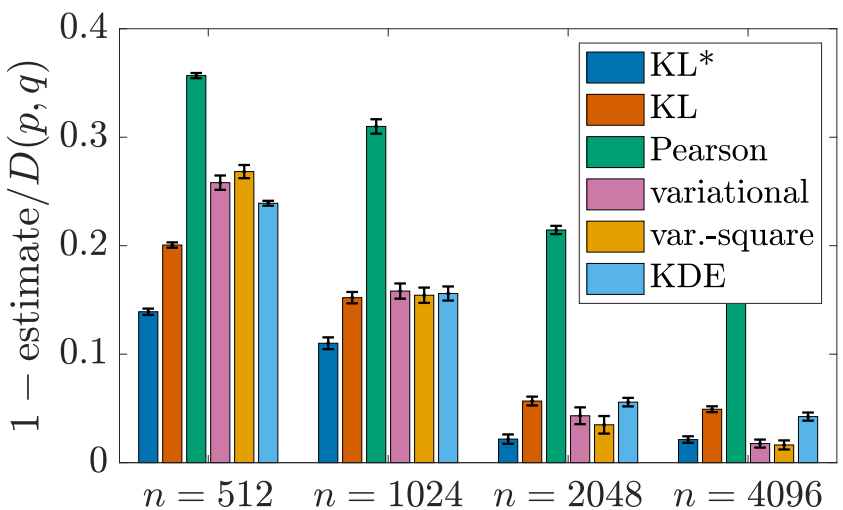
Debiasing

- **Bias due to V-statistic**: $\frac{1}{2}\int_0^1 (\hat{\mu}_p - \hat{\mu}_q)^\top (\rho\hat{\Sigma}_p + (1-\rho)\hat{\Sigma}_q + \lambda I)^{-1}(\hat{\mu}_p - \hat{\mu}_q) d\nu(\rho)$
  - Quadratic forms of means are not unbiased (Laurent, 1996)
  - Estimation of bias: $\mathrm{tr}[\hat{C}(\rho\hat{\Sigma}_p + (1-\rho)\hat{\Sigma}_q + \lambda I)^{-1}]$
  - With $\hat{C} = (\hat{\Sigma}_p - \hat{\mu}_p\hat{\mu}_p^\top)/(n_p - 1) + (\hat{\Sigma}_q - \hat{\mu}_q\hat{\mu}_q^\top)/(n_q - 1)$
  - Can be subtracted for improved behavior (and now minimax optimal additive term)
- **Simulations in 1D** with fixed schedule for λ

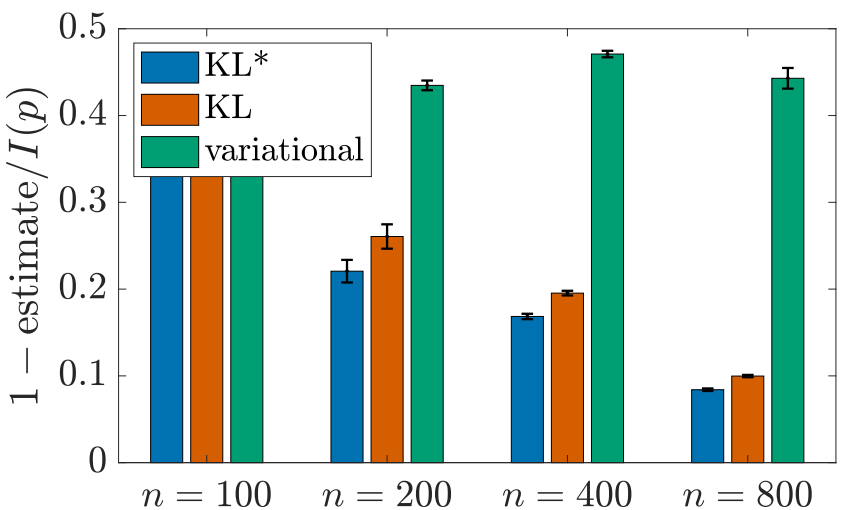


Experiments

- **Comparison to existing estimators on two-dimensional problems** (random ReLU features)
  - KL estimation with/without $w(x) = -f^*(v(x))$
  - Pearson estimate of $\frac{dp}{dq}(x) - 1$
  - Variational (with or without square features)
  - Kernel density estimation



Regular estimation



Mutual information

Summary for linear models

- **Closed-form estimation from m-dimensional features**
  - Running-time $O(m^3 + m^2 n)$
  - Convergence rates for estimation for high-dimensional feature maps
- **Relative density estimation**
  - Normalized models (softmax regression)
  - Unnormalized models

Summary for linear models

- **Closed-form estimation from m-dimensional features**
  - Running-time $O(m^3 + m^2 n)$
  - Convergence rates for estimation for high-dimensional feature maps
- **Relative density estimation**
  - Normalized models (softmax regression)
  - Unnormalized models
- **Extension to feature learning**
  - More efficient linear models (avoid $O(m^3)$)
  - Extension to neural network

Feature learning - view 1

- **Shared features** for all $u_\rho : \mathcal{X} \to \mathbb{R}$, $\rho \in [0, 1]$

$$D(p\|q) = \int_0^1 \sup_{u_\rho : \mathcal{X} \to \mathbb{R}} \left\{ \int_{\mathcal{X}} \left[u_\rho(x) - \frac{\rho}{2} u_\rho(x)^2 \right] dp(x) + \int_{\mathcal{X}} \left[-u_\rho(x) - \frac{1-\rho}{2} u_\rho(x)^2 \right] dq(x) \right\} d\nu(\rho)$$

  - Consider $u_\rho(x) = \theta_\rho^\top \psi_\Gamma(x)$ for some parameterized feature map ψ_Γ
  - Still infinitely many ρ's
  - Interpretation through data-processing inequality

Feature learning - view 1

- **Shared features** for all $u_\rho : \mathcal{X} \to \mathbb{R}$, $\rho \in [0,1]$

$$D(p\|q) = \int_0^1 \sup_{u_\rho : \mathcal{X} \to \mathbb{R}} \left\{ \int_{\mathcal{X}} \left[u_\rho(x) - \frac{\rho}{2} u_\rho(x)^2 \right] dp(x) + \int_{\mathcal{X}} \left[-u_\rho(x) - \frac{1-\rho}{2} u_\rho(x)^2 \right] dq(x) \right\} d\nu(\rho)$$

  - Consider $u_\rho(x) = \theta_\rho^\top \psi_\Gamma(x)$ for some parameterized feature map ψ_Γ
  - Still infinitely many ρ's
  - Interpretation through data-processing inequality

- **Examples**
  - Linear feature learning: $\psi_\Gamma(x) = \Gamma^\top \varphi(x)$ for $\varphi : \mathcal{X} \to \mathbb{R}^m$ and $\Gamma \in \mathbb{R}^{m \times r}$
  - Neural network: $\psi_\Gamma(x) = ((w_i^\top x + b_i)_+)_{i \in \{1,\dots,m\}} \in \mathbb{R}^m$
  - Efficient neural network: $\psi_\Gamma(x) = \Gamma^\top ((w_i^\top x + b_i)_+)_{i \in \{1,\dots,m\}} \in \mathbb{R}^r$ and $\Gamma \in \mathbb{R}^{m \times r}$

Feature learning - view 2

- **Spectral variational algorithm**: $\max_{\Gamma} F(\hat{p}\|\hat{q}, \psi_\Gamma)$
  - Minorization-maximization: lower bound the convex function F by

$$\mathrm{tr}\left[M\int_{\mathcal{X}}\psi_\Gamma(x)\psi_\Gamma(x)^\top\hat{p}(x)\right]+\mathrm{tr}\left[N\int_{\mathcal{X}}\psi_\Gamma(x)\psi_\Gamma(x)^\top\hat{q}(x)\right]+2c^\top\int_{\mathcal{X}}\psi_\Gamma(x)(d\hat{p}(x)-d\hat{q}(x))$$

  - Alternating between computing (M, N, c) and maximizing with respect to Γ
  - Need for regularization

Feature learning - view 2

- **Spectral variational algorithm**: $\max_{\Gamma} F(\hat{p}\|\hat{q}, \psi_\Gamma)$
  - Minorization-maximization: lower bound the convex function F by

$$\mathrm{tr}\left[M\int_{\mathcal{X}}\psi_\Gamma(x)\psi_\Gamma(x)^\top\hat{p}(x)\right]+\mathrm{tr}\left[N\int_{\mathcal{X}}\psi_\Gamma(x)\psi_\Gamma(x)^\top\hat{q}(x)\right]+2c^\top\int_{\mathcal{X}}\psi_\Gamma(x)(d\hat{p}(x)-d\hat{q}(x))$$

  - Alternating between computing (M, N, c) and maximizing with respect to Γ
  - Need for regularization

- **Stochastic approximation** extension "à la" online EM (Cappé and Moulines, 2009)

- **Running-time complexity**
  - Similar to classical supervised learning with ψ_Γ is low-dimensional

Theoretical results (neural networks)

- **Idealized estimator**
  - m-dimensional feature map $\hat{\varphi}$ with components $(w_j^\top x + b_j)_+$, $j = 1, \ldots, m$
  - Exact minimization of $F(\hat{p} \| \hat{q}, \hat{\varphi})$ with weight decay penalty

Theoretical results (neural networks)

- **Idealized estimator**
  - m-dimensional feature map $\hat{\varphi}$ with components $(w_j^\top x + b_j)_+$, $j = 1, \dots, m$
  - Exact minimization of $F(\hat{p} \| \hat{q}, \hat{\varphi})$ with weight decay penalty

- **Proposition**: Assume that dp/dq strictly positive, has t square-integrable derivatives, and depends only on a projection onto a subspace of dimension d_{eff}.

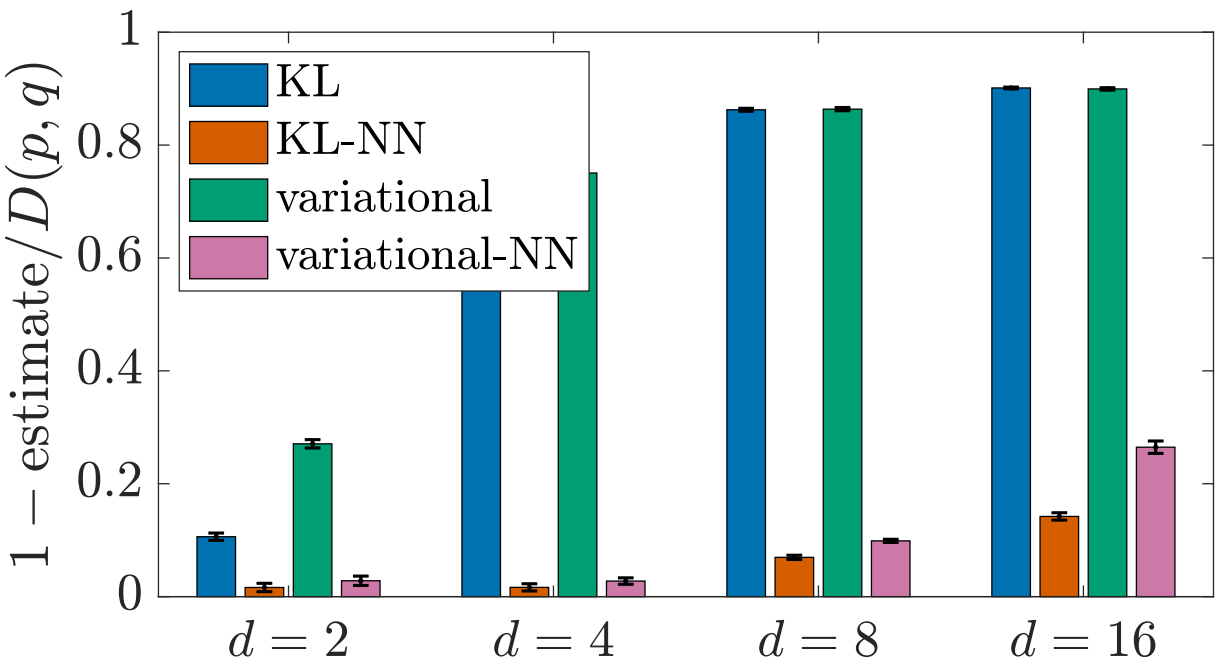
  If $t > \dfrac{d_{\text{eff}} + 3}{2}$, for $\lambda \propto 1/\sqrt{n}$ and $n = O(m^2)$

$$\left| F_\lambda(\hat{p} \| \hat{q}, \hat{\varphi}) - D(p \| q) \right| = O(n^{-1/2})$$

  - Adaptivity to linear latent variables (Bach, 2017)

Neural network experiments

- **Synthetic example**
  - $n = 10000$ training observations, and $m = 50$ hidden neurons
  - $dp/dq(x)$ depends only on a single variable



Related work

- **Density ratio estimation with kernel methods**
  - Link with Fisher discriminant analysis (Harchaoui, Bach, and Moulines, 2008)
  - (weighted) chi-square divergences (Kanamori et al., 2009; Sugiyama et al., 2012; Yamada et al., 2013; Ribero et al., 2026)
  - KL divergence extension (Sugiyama et al., 2007)

- **Quantum divergences** (Bach, 2023, 2025)
  - Similar variational formulation *with normalization constraint*: $\forall x,\ \|\varphi(x)\|^2 = 1$
  - Similar results when applicable, feature learning not as flexible

- **Variational formulation for mutual information**
  - Suzuki and Sugiyama (2010); Belghazi et al. (2018); Poole et al. (2019)

- **Link between f-divergences and binary classification**
  - Reid and Williamson (2011); Nowozin et al. (2016)

Summary - conclusion

$$t \log t - t + 1 = \int_0^1 \frac{(t-1)^2}{\rho t + 1 - \rho}(1-\rho)d\rho$$

- **A spectral framework for closed-form relative density estimation**
  - KL divergence as integrals of weighted chi-squared divergences
  - Decoupled least-squares estimation with spectral formula for fixed features
  - Efficient feature learning
  - Potential replacement for output layer for large/complex output spaces

Summary - conclusion

$$t \log t - t + 1 = \int_0^1 \frac{(t-1)^2}{\rho t + 1 - \rho}(1-\rho) d\rho$$

- **A spectral framework for closed-form relative density estimation**
  - KL divergence as integrals of weighted chi-squared divergences
  - Decoupled least-squares estimation with spectral formula for fixed features
  - Efficient feature learning
  - Potential replacement for output layer for large/complex output spaces

- **Future work**
  - Provable improved robustness compared to plain variational inference
  - Parallelization (Schubert and Gertz, 2018) or sketching (Gribonval et al., 2021).
  - Applications in variational inference (Wainwright and Jordan, 2008), sampling (Ribero et al., 2026) or independent component analysis (Bach and Jordan, 2002).
  - Normalization issues beyond probabilistic models

References

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19): 1–53, 2017.

Francis Bach. Information theory with kernel methods. *IEEE Transactions on Information Theory*, 69(2):752–775, 2023.

Francis Bach. Sum-of-squares relaxations for information theory and variational inference. *Foundations of Computational Mathematics*, 25(3):865–903, 2025.

Francis Bach and Michael I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3(Jul):1–48, 2002.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning*, 2018.

Michel Broniatowski and Amor Keziou. Minimization of φ-divergences on sets of signed measures. *Studia Scientiarum Mathematicarum Hungarica*, 43(4):403–442, 2006.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Olivier Cappé and Eric Moulines. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(3):593–613, 2009.

Rémi Gribonval, Antoine Chatalic, Nicolas Keriven, Vincent Schellekens, Laurent Jacques, and Philip Schniter. Sketching data sets for large-scale learning: Keeping only what you need. *IEEE Signal Processing Magazine*, 38(5):12–36, 2021.

Zaid Harchaoui, Francis Bach, and Eric Moulines. Testing for homogeneity with kernel Fisher discriminant analysis. Technical Report 0804.1026, arXiv, 2008.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.

Béatrice Laurent. Efficient estimation of integral functionals of a density. *The Annals of Statistics*, 24(2):659–681, 1996.

Andrew Lesniewski and Mary Beth Ruskai. Monotone Riemannian metrics and relative entropy on noncommutative probability spaces. *Journal of Mathematical Physics*, 40(11):5702–5724, 1999.

Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, and Volkan Cevher. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces. *Applied and Computational Harmonic Analysis*, 48(3):868–890, 2020.

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, 2016.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2025.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, 2019.

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12(22):731–817, 2011.

Mónica Ribero, Antonin Schrab, and Arthur Gretton. Regularized f-divergence kernel tests. Technical Report 2601.19755, arXiv,

2026.

Erich Schubert and Michael Gertz. Numerically stable parallel computation of (co-)variance. In *International Conference on Scientific and Statistical Database Management*, 2018.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems*, 2007.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

Taiji Suzuki and Masashi Sugiyama. Sufficient dimension reduction via squared-loss mutual information estimation. In *International Conference on Artificial Intelligence and Statistics*, 2010.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural Computation*, 25(5):1324–1370, 2013.