Some novel kernel-based divergences between probability distributions

Anna Korba (ENSAE, IP Paris)

Statistical and Numerical Methods for Non-commutative Optimal Transport, 2025

Context

Comparing probability distributions is a fundamental task, e.g. in

Context

Comparing probability distributions is a fundamental task, e.g. in

- testing (goodness of fit tests)

Context

Comparing probability distributions is a fundamental task, e.g. in

- testing (goodness of fit tests)
- sampling as optimization: minimize functional $\mathcal{F} : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$

$$\min_{p \in \mathcal{P}(R^d)} \mathcal{F}(p),$$

where $\mathcal{P}(\mathbb{R}^d)$ denotes the space of probability distributions over $\mathbb{R}^d$.

It is an infinite-dimensional optimization problem.

Context

Comparing probability distributions is a fundamental task, e.g. in

- testing (goodness of fit tests)
- sampling as optimization: minimize functional $\mathcal{F} : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$

$$\min_{p \in \mathcal{P}(R^d)} \mathcal{F}(p),$$

where $\mathcal{P}(\mathbb{R}^d)$ denotes the space of probability distributions over $\mathbb{R}^d$.

It is an infinite-dimensional optimization problem.

Applications:

1. Bayesian inference (learn complex posteriors for parametric models, $q = \tilde{q}/Z$ with Z unknown)
2. Generative modeling (learn data distributions)

Learn to **sample** from a probability distribution q: $z_1, \ldots z_m \sim q$.

Sampling as optimization

The sampling problem can be rewritten as minimizing $\mathcal{F}(p) = \mathcal{D}(p|q)$

- where $q \in \mathcal{P}(\mathbb{R}^d)$ is a target distribution
- and $\mathcal{D}$ a loss objective that cancels only for $p = q$.

The choice of $\mathcal{D}/\mathcal{F}$ depends on the information on the target q.

Sampling as optimization

The sampling problem can be rewritten as minimizing $\mathcal{F}(p) = \mathcal{D}(p|q)$

- where $q \in \mathcal{P}(\mathbb{R}^d)$ is a target distribution
- and $\mathcal{D}$ a loss objective that cancels only for $p = q$.

The choice of $\mathcal{D}/\mathcal{F}$ depends on the information on the target q.

Examples:

- $\mathcal{F}(p) = \mathrm{KL}(p|q)$, where $\mathrm{KL}(p|q) = \int \log\left(\frac{dp}{dq}(x)\right) dp(x)$ if p absolutely continuous w.r.t. q (with density dp/dq), $+\infty$ else.

  Convenient when the unnormalized density of q is known since the minimization objective **does not depend on the normalization constant!**

  Indeed writing $q(x) = e^{-V(x)}/Z$ we have:

  $$\mathrm{KL}(p|q) = \int_{\mathbb{R}^d} \log\left(\frac{p}{e^{-V}}(x)\right) dp(x) + \log(Z).$$

  But, it is not convenient when p or q are discrete, because the KL is $+\infty$ unless $supp(p) \subset supp(q)$.

The sampling problem can be rewritten as minimizing $\mathcal{F}(p) = \mathcal{D}(p|q)$

- where $q \in \mathcal{P}(\mathcal{Z})$ is a target distribution
- and $\mathcal{D}$ a loss objective that cancels only for $p = q$.

The choice of $\mathcal{D}/\mathcal{F}$ depends on the information on the target q.

Examples:

- **When we have samples of q (or a discrete measure), it is convenient to choose $\mathcal{D}$ as an integral probability metric (IPM)**

  For instance, $\mathcal{D}$ could be the MMD (Maximum Mean Discrepancy)[1]:

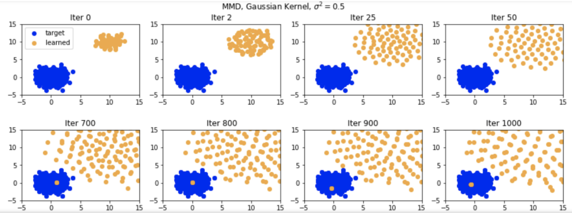
$$
\begin{aligned}
\mathrm{MMD}^2(p,q) &= \sup_{f \in \mathrm{H}_k, \|f\|_{\mathcal{H}_k} \leq 1} \left| \int f dp - \int f dq \right| \\
&= \|m_p - m_q\|^2_{\mathrm{H}_k}, \quad \text{where } m_p = \int k(x,\cdot) dp(x) \\
&= \mathbb{E}_{x,y \sim p}[k(x,y)] + \mathbb{E}_{x,y \sim q}[k(x,y)] - 2\,\mathbb{E}_{\substack{x \sim p \\ y \sim q}}[k(x,y)]
\end{aligned}
$$

[1] $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ a p.s.d. kernel (e.g. $k(x,y) = e^{-\|x-y\|^2}$) with RKHS H_k, $\langle f, k(x,.)\rangle_{\mathrm{H}_k} = f(x)$ for $f \in \mathrm{H}_k$.

Are all functionals good optimization objectives?

Example: Take $k(x,y) = e^{-\frac{\|x-y\|^2}{\sigma^2}}$, $p = \frac{1}{n}\sum_{i=1}^{n} \delta_{x^i}$, $q = \frac{1}{m}\sum_{j=1}^{m} \delta_{y^j}$.

$$\mathrm{MMD}^2(p,q) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(x^i,x^j) + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} k(y^i,y^j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(x^i,y^j).$$



Optimizing MMD with gradient descent can miserably fail (Arbel et al., 2019).

Remark: works much better choosing $k(x,y) = -\|x-y\|$, where the MMD is known as Energy distance see (Hertrich et al., 2024).

Statistical and Geometrical Properties of Regularized Kernel Kullback-Leibler Divergence

Joint work with Clémentine Chazal (ENSAE) and Francis Bach (INRIA).

Published at Neurips 2024.

Kernel Kullback-Leibler (KKL) divergence (Bach, 2022)

Let $q \in \mathcal{P}(\mathbb{R}^d)$. The covariance operator w.r.t. q is defined as
$\Sigma_q = \int k(\cdot, x) \otimes k(\cdot, x) dq(x)$, where $(a \otimes b)c = \langle b, c \rangle_{\mathrm{H}_k} a$ for $a, b, c \in \mathrm{H}_k$.

[1] if k^2 is characteristic. See paper for sufficient conditions

Kernel Kullback-Leibler (KKL) divergence (Bach, 2022)

Let $q \in \mathcal{P}(\mathbb{R}^d)$. The covariance operator w.r.t. q is defined as $\Sigma_q = \int k(\cdot, x) \otimes k(\cdot, x) dq(x)$, where $(a \otimes b)c = \langle b, c \rangle_{\mathrm{H}_k} a$ for $a, b, c \in \mathrm{H}_k$.

For $p, q \in \mathcal{P}(\mathbb{R}^d)$, the KKL is defined as:

$$\mathrm{KKL}(p|q) := \mathrm{Tr}(\Sigma_p \log \Sigma_p) - \mathrm{Tr}(\Sigma_p \log \Sigma_q) = \sum_{\substack{(\lambda,\gamma) \\ \in \Lambda_p \times \Lambda_q}} \lambda \log\left(\frac{\lambda}{\gamma}\right) \langle f_\lambda, g_\gamma \rangle^2_{\mathrm{H}_k},$$

where Λ_p and Λ_q are the set of eigenvalues of the covariance operators Σ_p and Σ_q, with associated eigenvectors $(f_\lambda)_{\lambda \in \Lambda_p}$ and $(g_\gamma)_{\gamma \in \Lambda_q}$.

[1] if k^2 is characteristic. See paper for sufficient conditions

Kernel Kullback-Leibler (KKL) divergence (Bach, 2022)

Let $q \in \mathcal{P}(\mathbb{R}^d)$. The covariance operator w.r.t. q is defined as $\Sigma_q = \int k(\cdot, x) \otimes k(\cdot, x) dq(x)$, where $(a \otimes b)c = \langle b, c \rangle_{\mathrm{H}_k} a$ for $a, b, c \in \mathrm{H}_k$.

For $p, q \in \mathcal{P}(\mathbb{R}^d)$, the KKL is defined as:

$$\mathrm{KKL}(p|q) := \mathrm{Tr}(\Sigma_p \log \Sigma_p) - \mathrm{Tr}(\Sigma_p \log \Sigma_q) = \sum_{\substack{(\lambda, \gamma) \\ \in \Lambda_p \times \Lambda_q}} \lambda \log\left(\frac{\lambda}{\gamma}\right) \langle f_\lambda, g_\gamma \rangle^2_{\mathrm{H}_k},$$

where Λ_p and Λ_q are the set of eigenvalues of the covariance operators Σ_p and Σ_q, with associated eigenvectors $(f_\lambda)_{\lambda \in \Lambda_p}$ and $(g_\gamma)_{\gamma \in \Lambda_q}$.

- can be seen as **second-order embeddings** of probability distributions, in contrast with first-order kernel mean embeddings (as used in MMD)

[1] if k^2 is characteristic. See paper for sufficient conditions

Kernel Kullback-Leibler (KKL) divergence (Bach, 2022)

Let $q \in \mathcal{P}(\mathbb{R}^d)$. The covariance operator w.r.t. q is defined as $\Sigma_q = \int k(\cdot, x) \otimes k(\cdot, x) dq(x)$, where $(a \otimes b)c = \langle b, c \rangle_{\mathrm{H}_k} a$ for $a, b, c \in \mathrm{H}_k$.

For $p, q \in \mathcal{P}(\mathbb{R}^d)$, the KKL is defined as:

$$\mathrm{KKL}(p|q) := \mathrm{Tr}(\Sigma_p \log \Sigma_p) - \mathrm{Tr}(\Sigma_p \log \Sigma_q) = \sum_{\substack{(\lambda,\gamma) \\ \in \Lambda_p \times \Lambda_q}} \lambda \log \left(\frac{\lambda}{\gamma} \right) \langle f_\lambda, g_\gamma \rangle^2_{\mathrm{H}_k},$$

where Λ_p and Λ_q are the set of eigenvalues of the covariance operators Σ_p and Σ_q, with associated eigenvectors $(f_\lambda)_{\lambda \in \Lambda_p}$ and $(g_\gamma)_{\gamma \in \Lambda_q}$.

- can be seen as **second-order embeddings** of probability distributions, in contrast with first-order kernel mean embeddings (as used in MMD)
- $\mathrm{KL}(\tilde{k} \star p | \tilde{k} \star q) \leq \mathrm{KKL}(p|q) \leq \mathrm{KL}(p|q)$ for some smoothing kernel $\tilde{k}$

[1] if k^2 is characteristic. See paper for sufficient conditions

Kernel Kullback-Leibler (KKL) divergence (Bach, 2022)

Let $q \in \mathcal{P}(\mathbb{R}^d)$. The covariance operator w.r.t. q is defined as $\Sigma_q = \int k(\cdot, x) \otimes k(\cdot, x) dq(x)$, where $(a \otimes b)c = \langle b, c\rangle_{\mathrm{H}_k} a$ for $a, b, c \in \mathrm{H}_k$.

For $p, q \in \mathcal{P}(\mathbb{R}^d)$, the KKL is defined as:

$$\mathrm{KKL}(p|q) := \mathrm{Tr}(\Sigma_p \log \Sigma_p) - \mathrm{Tr}(\Sigma_p \log \Sigma_q) = \sum_{\substack{(\lambda,\gamma) \\ \in \Lambda_p \times \Lambda_q}} \lambda \log \left(\frac{\lambda}{\gamma}\right) \langle f_\lambda, g_\gamma \rangle_{\mathrm{H}_k}^2,$$

where Λ_p and Λ_q are the set of eigenvalues of the covariance operators Σ_p and Σ_q, with associated eigenvectors $(f_\lambda)_{\lambda \in \Lambda_p}$ and $(g_\gamma)_{\gamma \in \Lambda_q}$.

- can be seen as **second-order embeddings** of probability distributions, in contrast with first-order kernel mean embeddings (as used in MMD)
- $\mathrm{KL}(\tilde{k} \star p | \tilde{k} \star q) \leq \mathrm{KKL}(p|q) \leq \mathrm{KL}(p|q)$ for some smoothing kernel $\tilde{k}$
- $\mathrm{KKL}(p|q) = 0$ if and only if $p = q$[1] Bach (2022, Proposition 4)

[1] if k^2 is characteristic. See paper for sufficient conditions

Questions:

- what is the behavior of KKL for empirical measures? does it admit a tractable closed-form expression ?
- is it a suitable optimization objective?

Regularized KKL (Chazal et al., 2024)

$$\mathrm{KKL}(p|q) := \mathrm{Tr}(\Sigma_p \log \Sigma_p) - \mathrm{Tr}(\Sigma_p \log \Sigma_q) = \sum_{(\lambda,\gamma)\in\Lambda_p\times\Lambda_q} \lambda \log\left(\frac{\lambda}{\gamma}\right) \langle f_\lambda, g_\gamma\rangle^2_{\mathrm{H}_k}.$$

- $\mathrm{KKL}(p|q) < \infty$ requires $\mathrm{Ker}(\Sigma_q) \subset \mathrm{Ker}(\Sigma_p)$
- True if $\mathrm{Supp}(p) \subset \mathrm{Supp}(q)$: if $f \in \mathrm{Ker}(\Sigma_q)$, then

$$\langle f, \Sigma_q f\rangle_{\mathrm{H}_k} = \int \langle f, k(x,\cdot) \otimes k(x,\cdot) f\rangle_{\mathrm{H}_k} dq(x) = \int_{\mathbb{R}^d} f(x)^2 dq(x) = 0$$

  and so f is zero on the support of q, then also on the support of p
- Hence the KKL is not convenient if p, q are discrete with different supports

Regularized KKL (Chazal et al., 2024)

$$\mathrm{KKL}(p|q) := \mathrm{Tr}(\Sigma_p \log \Sigma_p) - \mathrm{Tr}(\Sigma_p \log \Sigma_q) = \sum_{(\lambda,\gamma)\in\Lambda_p\times\Lambda_q} \lambda \log\left(\frac{\lambda}{\gamma}\right) \langle f_\lambda, g_\gamma\rangle^2_{\mathrm{H}_k}.$$

- $\mathrm{KKL}(p|q) < \infty$ requires $\mathrm{Ker}(\Sigma_q) \subset \mathrm{Ker}(\Sigma_p)$
- True if $\mathrm{Supp}(p) \subset \mathrm{Supp}(q)$: if $f \in \mathrm{Ker}(\Sigma_q)$, then

$$\langle f, \Sigma_q f\rangle_{\mathrm{H}_k} = \int \langle f, k(x,\cdot)\otimes k(x,\cdot) f\rangle_{\mathrm{H}_k} dq(x) = \int_{\mathbb{R}^d} f(x)^2 dq(x) = 0$$

  and so f is zero on the support of q, then also on the support of p
- Hence the KKL is not convenient if p, q are discrete with different supports

A simple fix that we propose is to consider a regularized version of KKL which is, for $\alpha \in]0,1[$,

$$\begin{aligned}\mathrm{KKL}_\alpha(p|q) &:= \mathrm{KKL}(p|(1-\alpha)q + \alpha p)\\ &= \mathrm{Tr}(\Sigma_p \log \Sigma_p) - \mathrm{Tr}(\Sigma_p \log((1-\alpha)\Sigma_q + \alpha\Sigma_p)).\end{aligned}$$

and which recovers KKL as $\alpha \to 0$ (it goes to 0 when $\alpha \to 1$).
Note it still cancels for $p = q$.

Skewness of the regularized KKL

Proposition

Let $p \ll q$. The function $\alpha \mapsto \mathrm{KKL}_\alpha(p|q)$ is decreasing on $[0,1]$.

Proposition

Let $p, q \in \mathcal{P}(\mathbb{R}^d)$. Assume that $p \ll q$ and that $\frac{dp}{dq} \leqslant \frac{1}{\mu}$ for some $\mu > 0$. Then,

$$|\mathrm{KKL}_\alpha(p|q) - \mathrm{KKL}(p|q)| \leqslant \left(\alpha\left(1+\frac{1}{\mu}\right) + \frac{\alpha^2}{1-\alpha}\left(1+\frac{1}{\mu^2}\right)\right) |\operatorname{Tr}(\Sigma_p \log \Sigma_q)|.$$

Skewness of the regularized KKL

Proposition

Let $p \ll q$. The function $\alpha \mapsto \mathrm{KKL}_\alpha(p|q)$ is decreasing on $[0, 1]$.

Proposition

Let $p, q \in \mathcal{P}(\mathbb{R}^d)$. Assume that $p \ll q$ and that $\frac{dp}{dq} \leqslant \frac{1}{\mu}$ for some $\mu > 0$. Then,

$$|\mathrm{KKL}_\alpha(p|q) - \mathrm{KKL}(p|q)| \leqslant \left(\alpha\left(1 + \frac{1}{\mu}\right) + \frac{\alpha^2}{1-\alpha}\left(1 + \frac{1}{\mu^2}\right)\right) |\operatorname{Tr}\left(\Sigma_p \log \Sigma_q\right)|.$$

These two propositions above show that the regularized KKL shares a similar behavior than the regularized (standard) KL:

Skewness of the regularized KKL

Proposition

Let $p \ll q$. The function $\alpha \mapsto \mathrm{KKL}_\alpha(p|q)$ is decreasing on $[0, 1]$.

Proposition

Let $p, q \in \mathcal{P}(\mathbb{R}^d)$. Assume that $p \ll q$ and that $\frac{dp}{dq} \leqslant \frac{1}{\mu}$ for some $\mu > 0$. Then,

$$|\mathrm{KKL}_\alpha(p|q) - \mathrm{KKL}(p|q)| \leqslant \left(\alpha\left(1 + \frac{1}{\mu}\right) + \frac{\alpha^2}{1-\alpha}\left(1 + \frac{1}{\mu^2}\right)\right) |\operatorname{Tr}(\Sigma_p \log \Sigma_q)|.$$

These two propositions above show that the regularized KKL shares a similar behavior than the regularized (standard) KL:

- which is also monotone decreasing in α

Skewness of the regularized KKL

Proposition

Let $p \ll q$. The function $\alpha \mapsto \mathrm{KKL}_\alpha(p|q)$ is decreasing on $[0,1]$.

Proposition

Let $p, q \in \mathcal{P}(\mathbb{R}^d)$. Assume that $p \ll q$ and that $\frac{dp}{dq} \leqslant \frac{1}{\mu}$ for some $\mu > 0$. Then,

$$|\mathrm{KKL}_\alpha(p|q) - \mathrm{KKL}(p|q)| \leqslant \left(\alpha\left(1+\frac{1}{\mu}\right)+\frac{\alpha^2}{1-\alpha}\left(1+\frac{1}{\mu^2}\right)\right)\left|\operatorname{Tr}\left(\Sigma_p \log \Sigma_q\right)\right|.$$

These two propositions above show that the regularized KKL shares a similar behavior than the regularized (standard) KL:

- which is also monotone decreasing in α
- same bound, replacing $\left|\operatorname{Tr}\left(\Sigma_p \log \Sigma_q\right)\right|$ by $\int \log q dp$

Skewness of the regularized KKL

Proposition

Let $p \ll q$. The function $\alpha \mapsto \mathrm{KKL}_\alpha(p|q)$ is decreasing on $[0,1]$.

Proposition

Let $p, q \in \mathcal{P}(\mathbb{R}^d)$. Assume that $p \ll q$ and that $\frac{dp}{dq} \leqslant \frac{1}{\mu}$ for some $\mu > 0$. Then,

$$|\mathrm{KKL}_\alpha(p|q) - \mathrm{KKL}(p|q)| \leqslant \left(\alpha\left(1+\frac{1}{\mu}\right) + \frac{\alpha^2}{1-\alpha}\left(1+\frac{1}{\mu^2}\right)\right) |\operatorname{Tr}(\Sigma_p \log \Sigma_q)|.$$

These two propositions above show that the regularized KKL shares a similar behavior than the regularized (standard) KL:

- which is also monotone decreasing in α
- same bound, replacing $|\operatorname{Tr}(\Sigma_p \log \Sigma_q)|$ by $\int \log q dp$
- yet the tools used to derive these are completely different by nature than for the KL case, e.g. identities like

$$\operatorname{Tr}(\Sigma_p(\log \Sigma_p - \log \Sigma_q)) = \int_0^{+\infty} \operatorname{Tr}\Big(\Sigma_p(\Sigma_p + \beta I)^{-1}\Big) - \operatorname{Tr}\Big(\Sigma_q(\Sigma_q + \beta I)^{-1}\Big) d\beta$$

  and operator monotony.

Concentration of the regularized KKL

Proposition

Let $p, q \in \mathcal{P}(\mathbb{R}^d)$. Assume that $p \ll q$ with $\frac{dp}{dq} \leqslant \frac{1}{\mu}$ for some $0 < \mu \leqslant 1$ and let $\alpha \leqslant \frac{1}{2}$, and that $c = \int_0^{+\infty} \sup_{x \in \mathbb{R}^d} \langle k(x, \cdot), (\Sigma_p + \beta I)^{-1} k(x, \cdot) \rangle^2_{\mathrm{H}_k} d\beta$ is finite. Let $\hat{p}$, $\hat{q}$ supported on n, m i.i.d. samples from p and q respectively. We have:

$$\mathbb{E}|\mathrm{KKL}_\alpha(\hat{p}|\hat{q}) - \mathrm{KKL}_\alpha(p|q)| \leqslant \frac{35}{\sqrt{m \wedge n}} \frac{1}{\alpha\mu}(2\sqrt{c} + \log n) + \frac{1}{m \wedge n}\left(1 + \frac{1}{\mu} + \frac{c(24 \log n)^2}{\alpha\mu^2}(1 + \frac{n}{m \wedge n})\right).$$

Concentration of the regularized KKL

Proposition

Let $p, q \in \mathcal{P}(\mathbb{R}^d)$. Assume that $p \ll q$ with $\frac{dp}{dq} \leqslant \frac{1}{\mu}$ for some $0 < \mu \leqslant 1$ and let $\alpha \leqslant \frac{1}{2}$, and that $c = \int_0^{+\infty} \sup_{x \in \mathbb{R}^d} \langle k(x,\cdot), (\Sigma_p + \beta I)^{-1} k(x,\cdot) \rangle^2_{\mathrm{H}_k} d\beta$ is finite. Let $\hat{p}$, $\hat{q}$ supported on n, m i.i.d. samples from p and q respectively. We have:

$$\mathbb{E}|\mathrm{KKL}_\alpha(\hat{p}|\hat{q}) - \mathrm{KKL}_\alpha(p|q)| \leqslant \frac{35}{\sqrt{m \wedge n}} \frac{1}{\alpha\mu}(2\sqrt{c} + \log n) + \frac{1}{m \wedge n}\left(1 + \frac{1}{\mu} + \frac{c(24 \log n)^2}{\alpha\mu^2}(1 + \frac{n}{m \wedge n})\right).$$

Remarks:

- It is possible to derive a similar bound which does not require the condition $p \ll q$; yet it scales in $O(\frac{1}{\alpha^2})$ instead of $O(\frac{1}{\alpha})$ above.
- if $n = m$, the bound above scales as $\mathcal{O}\left(\frac{(\log n)^2}{n} + \frac{\log n}{\sqrt{n}}\right)$
- proof involves technical intermediate results: concentration of sums of random self-adjoint operators and estimation of degrees of freedom.

Regularized KKL closed-form for discrete measures

Proposition

Let $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ *and* $\hat{q} = \frac{1}{m}\sum_{j=1}^{m} \delta_{y_j}$ *two discrete distributions.*

Define $K_{\hat{p}} = (k(x_i, x_j))_{i,j=1}^{n} \in \mathbb{R}^{n\times n}$, $K_{\hat{q}} = (k(y_i, y_j))_{i,j=1}^{m} \in \mathbb{R}^{m\times m}$, $K_{\hat{p},\hat{q}} = (k(x_i, y_j))_{i,j=1}^{n,m} \in \mathbb{R}^{n\times m}$.

Then, for any $\alpha \in]0,1[$, *we have:*

$$\mathrm{KKL}_\alpha(\hat{p}||\hat{q}) = \mathrm{Tr}\left(\frac{1}{n}K_{\hat{p}} \log \frac{1}{n}K_{\hat{p}}\right) - \mathrm{Tr}\left(I_\alpha K \log(K)\right),$$

$$\textit{where } I_\alpha = \begin{pmatrix} \frac{1}{\alpha}I & 0 \\ 0 & 0 \end{pmatrix} \textit{ and } K = \begin{pmatrix} \frac{\alpha}{n}K_{\hat{p}} & \sqrt{\frac{\alpha(1-\alpha)}{nm}}K_{\hat{p},\hat{q}} \\ \sqrt{\frac{\alpha(1-\alpha)}{nm}}K_{\hat{q},\hat{p}} & \frac{1-\alpha}{m}K_{\hat{q}} \end{pmatrix}.$$

Computational cost (due to the singular value decomposition): $\mathcal{O}((n+m)^3)$.

Illustrations of skewness and concentration of the KKL

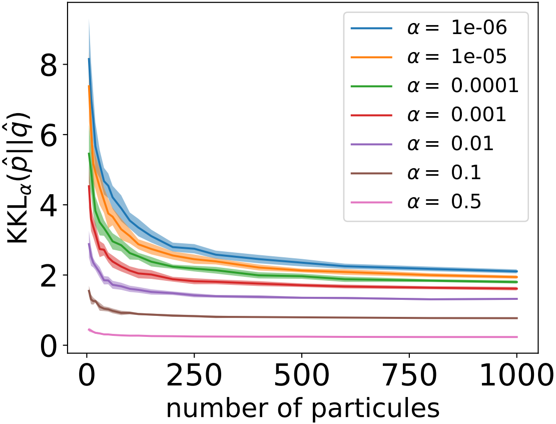
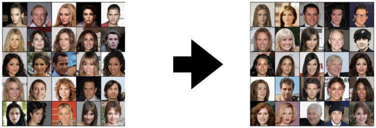


Figure: Concentration of empirical KKL_α for $d = 10$, $\sigma = 10$, with Gaussian kernel $k(x, y) = \exp\left(-\|x - y\|^2/\sigma^2\right)$. p, q different anisotropic Gaussians. Computed over 50 runs.

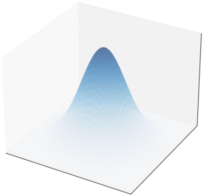
Dynamical measure transport and gradient flows

Motivation of generative modeling:



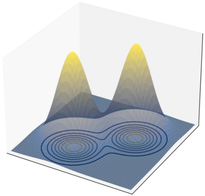
Idea: transport an initial, tractable measure p_0 onto q by minimizing $\mathcal{F} = \mathcal{D}(\cdot|q)$

Initial Measure p_0



$\Longrightarrow$

Target distribution q



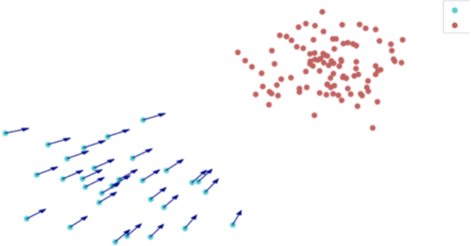
What about $\mathcal{D} = \mathrm{KKL}_\alpha$?

KKL minimization in practice

Introduce a particle system $x_0^1, \dots, x_0^n \sim p_0$, a step-size γ, and at each step[1]:

$$x_{l+1}^i = x_l^i - \gamma \nabla_{W_2} \mathcal{F}(\hat{p}_l)(x_l^i) \quad \text{for } i = 1, \dots, n, \text{ where } \hat{p}_l = \frac{1}{n} \sum_{i=1}^n \delta_{x_l^i}. \tag{1}$$

In particular, as $\mathcal{F}(p) = \mathrm{KKL}_\alpha(p|q)$ is well-defined for discrete measures p, Algorithm (1) **simply corresponds to gradient descent** of $F : \mathbb{R}^{N \times d} \to \mathbb{R}$, $F(x^1, \dots, x^n) := \mathcal{F}(p^n)$ where $p^n = \frac{1}{n} \sum_{i=1}^n \delta_{x^i}$.



[1] $\nabla_{W_2} \mathcal{F}(p) := \nabla \mathcal{F}'(p) : \mathbb{R}^d \to \mathbb{R}^d$ denotes the Wasserstein gradient of $\mathcal{F}$, and $\mathcal{F}'(p)$ denotes the first variation of $\mathcal{F}$ at p defined by:
$\lim_{\epsilon \to 0} \frac{1}{\epsilon}(\mathcal{F}(p + \epsilon(\nu - p)) - \mathcal{F}(p)) = \int_{\mathbb{R}^d} \mathcal{F}'(p)(x)(d\nu - dp)(x)$, $\mathcal{F}'(p) : \mathbb{R}^d \to \mathbb{R}$..

KKL minimization in practice

Introduce a particle system $x_0^1, \ldots, x_0^n \sim p_0$, a step-size γ, and at each step[1]:

$$x_{l+1}^i = x_l^i - \gamma \nabla_{W_2} \mathcal{F}(\hat{p}_l)(x_l^i) \quad \text{for } i = 1, \ldots, n, \text{ where } \hat{p}_l = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_l^i}. \qquad (1)$$

In particular, as $\mathcal{F}(p) = \mathrm{KKL}_\alpha(p|q)$ is well-defined for discrete measures p, Algorithm (1) **simply corresponds to gradient descent** of $F : \mathbb{R}^{N \times d} \to \mathbb{R}$, $F(x^1, \ldots, x^n) := \mathcal{F}(p^n)$ where $p^n = \frac{1}{n} \sum_{i=1}^n \delta_{x^i}$.

Proposition

Consider $\hat{p}, \hat{q}$ and the matrices $K_{\hat{p}},\ K$. Let $g(x) = \frac{\log x}{x}$. Then, the first variation of $\mathcal{F} = \mathrm{KKL}_\alpha(\cdot|\hat{q})$ at $\hat{p}$ is, for any $x \in \mathbb{R}^d$:

$$\mathcal{F}'(\hat{p})(x) = 1 + S(x)^T g(K_{\hat{p}}) S(x) - T(x)^T g(K) T(x) - T(x)^T A T(x),$$

where $S(x) = (\frac{1}{\sqrt{n}} k(x, x_1), .., \frac{1}{\sqrt{n}} k(x, x_n))$, $T(x) = (\sqrt{\frac{\alpha}{n}} k(x, x_1), .., \sqrt{\frac{1-\alpha}{m}} k(x, y_1), \ldots)$, and A is a matrix constructed from the eigenvectors and eigenvalues of both K and $\alpha \Sigma_{\hat{p}} + (1-\alpha)\Sigma_{\hat{q}}$.

---

[1] $\nabla_{W_2} \mathcal{F}(p) := \nabla \mathcal{F}'(p) : \mathbb{R}^d \to \mathbb{R}^d$ denotes the Wasserstein gradient of $\mathcal{F}$, and $\mathcal{F}'(p)$ denotes the first variation of $\mathcal{F}$ at p defined by:
$\lim_{\epsilon \to 0} \frac{1}{\epsilon}(\mathcal{F}(p + \epsilon(\nu - p)) - \mathcal{F}(p)) = \int_{\mathbb{R}^d} \mathcal{F}'(p)(x)(d\nu - dp)(x),\ \mathcal{F}'(p) : \mathbb{R}^d \to \mathbb{R}.$.

Related divergences (competitors)

Recall that f-divergences write $D(p|q) = \int f\left(\frac{p}{q}\right) dq$, f convex, $f(1) = 0$. They admit a variational form [Nguyen et al. (2010)]:

$$D(p|q) = \sup_{h:\mathbb{R}^d \to \mathbb{R}} \int hdp - \int f^\star(h)dq$$

where $f^\star(y) = \sup_{x\in\mathbb{R}^d} \langle x, y\rangle - f(x)$ is the convex conjugate (or Legendre transform) of f and h measurable.

Examples:

- $\mathrm{KL}(p|q)$: $f(x) = x\log(x) - x + 1$, $f^\star(y) = e^y - 1$
- $\chi^2(p|q)$: $f(x) = (x-1)^2$, $f^\star(y) = y + \frac{1}{4}y^2$

Related divergences (competitors)

Recall that f-divergences write $D(p|q) = \int f\left(\frac{p}{q}\right) dq$, f convex, $f(1) = 0$.
They admit a variational form [Nguyen et al. (2010)]:

$$D(p|q) = \sup_{h:\mathbb{R}^d \to \mathbb{R}} \int h dp - \int f^\star(h) dq$$

where $f^\star(y) = \sup_{x \in \mathbb{R}^d} \langle x, y \rangle - f(x)$ is the convex conjugate (or Legendre transform) of f and h measurable.

Examples:

- $\mathrm{KL}(p|q)$: $f(x) = x\log(x) - x + 1$, $f^\star(y) = e^y - 1$
- $\chi^2(p|q)$: $f(x) = (x-1)^2$, $f^\star(y) = y + \frac{1}{4}y^2$

Idea: restrict the search space to a RKHS !

- for the KL $\Longrightarrow$ Glaser, P., Arbel, M., & Gretton, A. *KALE flow: A relaxed KL gradient flow for probabilities with disjoint support.* (Neurips 2021).
- for the χ^2 $\Longrightarrow$ Chen, Z., Mustafi, A., Glaser, P., Korba, A., Gretton, A., & Sriperumbudur, B. K. *(De)-regularized Maximum Mean Discrepancy Gradient Flow.* (2024, arXiv preprint arXiv:2409.14980).

Kale (Glaser et al., 2021)

$$\mathrm{KALE}(p|q) = (1+\lambda)\max_{h\in\mathrm{H}_k}\int h dp - \int e^h dq - \frac{\lambda}{2}\|h\|^2_{\mathrm{H}_k}.$$

- interpolates between a KL ($\lambda \to 0$) and and MMD ($\lambda \to \infty$)
- For discrete distributions p and q supported on n atoms, the KALE divergence does not admit a closed-form
- But it can be written as a strongly convex n-dimensional problem and solved with, e.g., Newton
- In constrast, KKL has a closed-form, and can be optimized with L-BFGS(Liu and Nocedal, 1989)

Experiments

$\alpha = 0.01,\ \sigma^2 = 0.1,\ n = 100.$

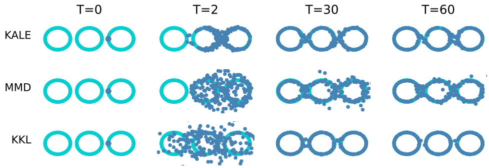


Figure: MMD, KALE and KKL flow for 3 rings target.

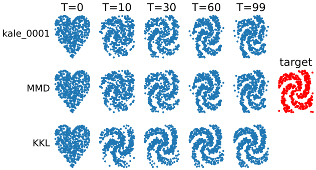
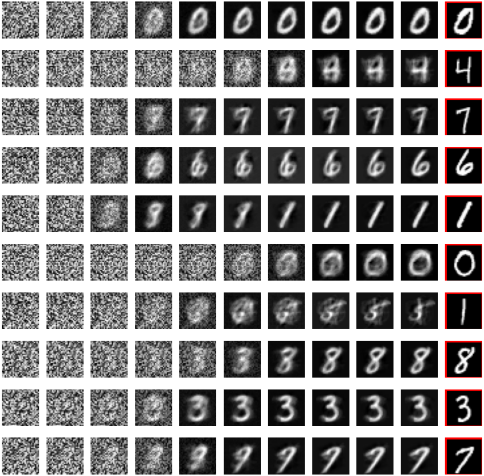


Figure: Shape transfer

Higher dimensional experiments on synthetic Gaussian mixtures in the paper.



De-Regularized MMD: Interpolate between MMD and χ^2-divergence[1]

$$\mathrm{DMMD}(p||q) = (1+\lambda)\left\{ \max_{h \in \mathrm{H}_k} \int h dp - \int (h + \frac{1}{4}h^2) dq - \frac{1}{4}\lambda \|h\|^2_{\mathrm{H}_k} \right\} \qquad (2)$$

[1] Joint work with Zonghao Chen, Aratrika Mustafi, Pierre Glaser, Arthur Gretton, Bharath K. Sriperumbudur. *https://arxiv.org/abs/2409.14980*

De-Regularized MMD: Interpolate between MMD and χ^2-divergence[1]

$$\mathrm{DMMD}(p||q) = (1+\lambda)\left\{ \max_{h \in \mathrm{H}_k} \int h dp - \int (h + \frac{1}{4}h^2) dq - \frac{1}{4}\lambda \|h\|^2_{\mathrm{H}_k} \right\} \tag{2}$$

- It is a divergence for any λ, recovers χ^2 for $\lambda = 0$ and MMD for $\lambda = +\infty$.

[1] Joint work with Zonghao Chen, Aratrika Mustafi, Pierre Glaser, Arthur Gretton, Bharath K. Sriperumbudur. *https://arxiv.org/abs/2409.14980*

De-Regularized MMD: Interpolate between MMD and χ^2-divergence[1]

$$\mathrm{DMMD}(p||q) = (1+\lambda)\left\{ \max_{h \in \mathrm{H}_k} \int h dp - \int (h + \frac{1}{4}h^2) dq - \frac{1}{4}\lambda \|h\|^2_{\mathrm{H}_k} \right\} \tag{2}$$

- It is a divergence for any λ, recovers χ^2 for $\lambda = 0$ and MMD for $\lambda = +\infty$.
- DMMD **and its gradient can be written in closed-form**

$$\mathrm{DMMD}(p|q) = (1+\lambda) \left\| (\Sigma_q + \lambda \, \mathrm{Id})^{-\frac{1}{2}} (m_p - m_q) \right\|^2_{\mathrm{H}_k}, \quad \nabla \, \mathrm{DMMD}(p|q) = \nabla h_{p,q}$$

where $\Sigma_q = \int k(\cdot, x) \otimes k(\cdot, x) dq(x)$, and $h_{p,q}$ solves (2).

[1] Joint work with Zonghao Chen, Aratrika Mustafi, Pierre Glaser, Arthur Gretton, Bharath K. Sriperumbudur. *https://arxiv.org/abs/2409.14980*

De-Regularized MMD: Interpolate between MMD and χ^2-divergence[1]

$$\mathrm{DMMD}(p||q) = (1+\lambda)\left\{ \max_{h \in \mathrm{H}_k} \int h dp - \int (h + \frac{1}{4}h^2) dq - \frac{1}{4}\lambda \|h\|^2_{\mathrm{H}_k} \right\} \qquad (2)$$

- It is a divergence for any λ, recovers χ^2 for $\lambda = 0$ and MMD for $\lambda = +\infty$.
- DMMD **and its gradient can be written in closed-form**

$$\mathrm{DMMD}(p|q) = (1+\lambda) \left\| (\Sigma_q + \lambda \, \mathrm{Id})^{-\frac{1}{2}} (m_p - m_q) \right\|^2_{\mathrm{H}_k}, \quad \nabla \, \mathrm{DMMD}(p|q) = \nabla h_{p,q}$$

where $\Sigma_q = \int k(\cdot, x) \otimes k(\cdot, x) dq(x)$, and $h_{p,q}$ solves (2).

- In particular for p, q discrete (supported on n, m samples respectively), it writes with kernel Gram matrices over samples of p, p^* in complexity $\mathcal{O}(m^3 + nm)$.

[1] Joint work with Zonghao Chen, Aratrika Mustafi, Pierre Glaser, Arthur Gretton, Bharath K. Sriperumbudur. *https://arxiv.org/abs/2409.14980*

De-Regularized MMD: Interpolate between MMD and χ^2-divergence[1]

$$\mathrm{DMMD}(p||q) = (1+\lambda)\left\{ \max_{h \in \mathrm{H}_k} \int h dp - \int (h + \frac{1}{4}h^2) dq - \frac{1}{4}\lambda \|h\|^2_{\mathrm{H}_k} \right\} \qquad (2)$$

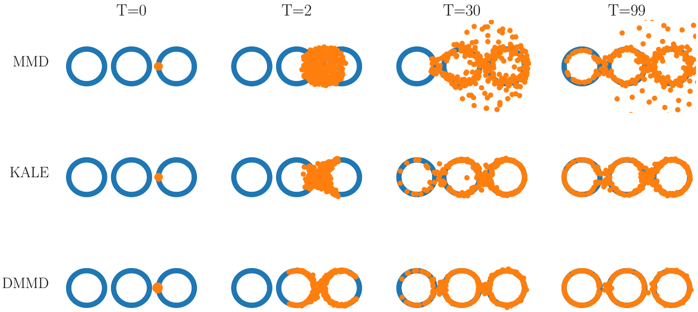
- It is a divergence for any λ, recovers χ^2 for $\lambda = 0$ and MMD for $\lambda = +\infty$.
- DMMD **and its gradient can be written in closed-form**

  $$\mathrm{DMMD}(p|q) = (1+\lambda) \left\| (\Sigma_q + \lambda \,\mathrm{Id})^{-\frac{1}{2}} (m_p - m_q) \right\|^2_{\mathrm{H}_k}, \quad \nabla \mathrm{DMMD}(p|q) = \nabla h_{p,q}$$

  where $\Sigma_q = \int k(\cdot, x) \otimes k(\cdot, x) dq(x)$, and $h_{p,q}$ solves (2).
- In particular for p, q discrete (supported on n, m samples respectively), it writes with kernel Gram matrices over samples of p, p^* in complexity $\mathcal{O}(m^3 + nm)$.
- It is an MMD with a regularized kernel: $\tilde{k}(x, x') = \sum_{i \geq 1} \frac{\varrho_i}{\varrho_i + \lambda} e_i(x) e_i(x')$ which is a regularized version of the original kernel $k(x, x') = \sum_{i \geq 1} \varrho_i e_i(x) e_i(x')$

  $\implies$ we inherit the statistical rates $\mathcal{O}(n^{-1/2})$ (Gretton et al., 2012)

[1] Joint work with Zonghao Chen, Aratrika Mustafi, Pierre Glaser, Arthur Gretton, Bharath K. Sriperumbudur. *https://arxiv.org/abs/2409.14980*


T=0
T=2
T=30
T=99
MMD
KALE
DMMD

Kernel Trace Distance: Quantum Statistical Metric between Measures through RKHS Density Operators

Joint work with Arturo Castellanos, Pavlo mozharovskyi and Hicham Janati (Télécom ParisTech).



Kernel Trace distance

Let $q \in \mathcal{P}(\mathbb{R}^d)$. Recall that the covariance operator w.r.t. q is defined as $\Sigma_q = \int k(\cdot, x) \otimes k(\cdot, x) dq(x)$, where $(a \otimes b)c = \langle b, c \rangle_{\mathrm{H}_k} a$ for $a, b, c \in \mathrm{H}_k$.

We define the ***kernel trace distance*** between two probability measures p, q on $\mathcal{X}$ is defined as:

$$d_{KT}(p, q) = ||\Sigma_p - \Sigma_q||_1,$$

where $\|T\|_1 = (\mathrm{Tr}(|T|))$ denotes the Schatten-1 norm, where $|T| = \sqrt{T^* T}$.

Similarly to the KKL, if k^2 is universal, d_{KT} is a well defined distance.

Comparison with other divergences

- Recall that $\|T\|_1 \geq \|T\|_2$, and that $\|\Sigma_p - \Sigma_q\|_2 = \mathrm{MMD}_{k^2}(p, q)$, so $\mathrm{MMD}_{k^2}(p, q) \leq d_{KT}(p, q)$

Comparison with other divergences

- Recall that $\|T\|_1 \geq \|T\|_2$, and that $\|\Sigma_p - \Sigma_q\|_2 = \mathrm{MMD}_{k^2}(p, q)$, so $\mathrm{MMD}_{k^2}(p, q) \leq d_{KT}(p, q)$
- we proved that $d_{KT}(p, q)$ can be written as an Integral Probability Metric over $\mathcal{F}_1 = \{f : x \mapsto \varphi(x)^* U \varphi(x) | U \in \mathcal{L}(\mathcal{H}), ||U||_\infty = 1\}$, which are functions with values in [-1,1] so $d_{KT}(p, q) \leq TV(p, q)$

Comparison with other divergences

- Recall that $\|T\|_1 \geq \|T\|_2$, and that $\|\Sigma_p - \Sigma_q\|_2 = \mathrm{MMD}_{k^2}(p, q)$, so $\mathrm{MMD}_{k^2}(p, q) \leq d_{KT}(p, q)$
- we proved that $d_{KT}(p, q)$ can be written as an Integral Probability Metric over $\mathcal{F}_1 = \{f : x \mapsto \varphi(x)^* U \varphi(x) | U \in \mathcal{L}(\mathcal{H}), ||U||_\infty = 1\}$, which are functions with values in [-1,1] so $d_{KT}(p, q) \leq TV(p, q)$
- Fuchs–van de Graaf inequality yields $d_{KBW}(p, q)^2 \leq d_{KT}(p, q) \leq 2 d_{KBW}(p, q)$ where $d_{KBW}(p, q)$ is the Bures distance between Σ_p and Σ_q:

$$d_{KBW}(p, q) = \sqrt{\mathrm{Tr}\,\Sigma_p + \mathrm{Tr}\,\Sigma_q - 2F(\Sigma_p, \Sigma_q)}$$

  where $F(A, B) = \mathrm{Tr}\left(A^{1/2} B A^{1/2}\right)^{1/2}$ is called the fidelity.

Comparison with other divergences

- Recall that $\|T\|_1 \geq \|T\|_2$, and that $\|\Sigma_p - \Sigma_q\|_2 = \mathrm{MMD}_{k^2}(p, q)$, so $\mathrm{MMD}_{k^2}(p, q) \leq d_{KT}(p, q)$
- we proved that $d_{KT}(p, q)$ can be written as an Integral Probability Metric over $\mathcal{F}_1 = \{f : x \mapsto \varphi(x)^* U \varphi(x) | U \in \mathcal{L}(\mathcal{H}), ||U||_\infty = 1\}$, which are functions with values in [-1,1] so $d_{KT}(p, q) \leq TV(p, q)$
- Fuchs–van de Graaf inequality yields $d_{KBW}(p, q)^2 \leq d_{KT}(p, q) \leq 2 d_{KBW}(p, q)$ where $d_{KBW}(p, q)$ is the Bures distance between Σ_p and Σ_q:

$$d_{KBW}(p, q) = \sqrt{\operatorname{Tr} \Sigma_p + \operatorname{Tr} \Sigma_q - 2F(\Sigma_p, \Sigma_q)}$$

  where $F(A, B) = \operatorname{Tr}\left(A^{1/2} B A^{1/2}\right)^{1/2}$ is called the fidelity.

- Pinsker's inequality yields $d_{KT}(p, q) \leq \sqrt{2\mathrm{KKL}(p|q)}$, where $\mathrm{KKL}(p|q) = \mathrm{KL}(\Sigma_p | \Sigma_q) = \operatorname{Tr}(\Sigma_p(\log \Sigma_p - \log \Sigma_q))$

Computation of d_{KT} in practice

Recall that $d_{KT}(p,q) = \|\Sigma_p - \Sigma_q\|_1 = \sum_{i=1}^{\infty} \lambda_i$, where (λ_i) are the singular values of $\Sigma_p - \Sigma_q = \Sigma_{p-q}$.

Let $x_1, \ldots, x_n \sim p$ and $y_1, \ldots, y_m \sim q$. Denote $X = (x_1, \ldots, x_n)$ and $\hat{p}_n$ the samples and empirical distributions, similarly Y and $\hat{q}_m$.

$\Sigma_{p_n - q_m}$ has the same eigenvalues as:

$$K = \left[\begin{array}{c|c} \frac{1}{n} K_{XX} & \frac{i}{\sqrt{mn}} K_{XY} \\ \hline \frac{i}{\sqrt{mn}} K_{YX} & -\frac{1}{m} K_{YY} \end{array}\right]$$

$\implies$ Get the eigenvalues by Singular Value decomposition, and compute their 1-norm (complexity $\mathcal{O}(n+m)^2$).

Concentration of d_{KT}

We note $A \lesssim_p b$ when for any $\delta > 0$, $\exists c_\delta < \infty$ s.t. $p(A \leq c_\delta b) \geq \delta$.

Theorem

- *If the eigenvalues of Σ_p follow a polynomial decay rate of order $\alpha > 1$:*

$$\underline{A} i^{-\alpha} \leq \lambda_i \leq \bar{A} i^{-\alpha} \text{ for some } \alpha > 1 \text{ and } 0 < \underline{A} < \bar{A} < \infty \tag{P}$$

$$\text{then: } d_{KT}(p, p_n) \lesssim_{p^{\otimes n}} n^{-\frac{1}{2}+\frac{1}{2\alpha}}.$$

- *If the eigenvalues of Σ_p follow an exponential decay rate:*

$$\underline{B} e^{-\tau i} \leq \lambda_i \leq \bar{B} e^{-\tau i} \text{ for some } \tau > 0 \text{ and } \underline{B}, \bar{B} \in (0, \infty), \tag{E}$$

$$\text{then: } d_{KT}(p, p_n) \lesssim_{p^{\otimes n}} \frac{(\log n)^{\frac{3}{2}}}{\sqrt{n}}.$$

Concentration of d_{KT}

We note $A \lesssim_p b$ when for any $\delta > 0$, $\exists c_\delta < \infty$ s.t. $p(A \leq c_\delta b) \geq \delta$.

Theorem

- *If the eigenvalues of Σ_p follow a polynomial decay rate of order $\alpha > 1$:*

$$\underline{A} i^{-\alpha} \leq \lambda_i \leq \bar{A} i^{-\alpha} \text{ for some } \alpha > 1 \text{ and } 0 < \underline{A} < \bar{A} < \infty \qquad \text{(P)}$$

$$\textit{then: } d_{KT}(p, p_n) \lesssim_{p^{\otimes n}} n^{-\frac{1}{2}+\frac{1}{2\alpha}}.$$

- *If the eigenvalues of Σ_p follow an exponential decay rate:*

$$\underline{B} e^{-\tau i} \leq \lambda_i \leq \bar{B} e^{-\tau i} \text{ for some } \tau > 0 \text{ and } \underline{B}, \bar{B} \in (0, \infty), \qquad \text{(E)}$$

$$\textit{then: } d_{KT}(p, p_n) \lesssim_{p^{\otimes n}} \frac{(\log n)^{\frac{3}{2}}}{\sqrt{n}}.$$

Assuming some decay rate on the eigenvalues, we can focus on the convergence of the operators on a subspace of the top eigenvectors, using results from the Kernel PCA literature (Blanchard et al., 2007; Rudi et al., 2013).

Concentration of d_{KT}

We note $A \lesssim_p b$ when for any $\delta > 0$, $\exists c_\delta < \infty$ s.t. $p(A \leq c_\delta b) \geq \delta$.

Theorem

- *If the eigenvalues of Σ_p follow a polynomial decay rate of order $\alpha > 1$:*

$$\underline{A} i^{-\alpha} \leq \lambda_i \leq \bar{A} i^{-\alpha} \text{ for some } \alpha > 1 \text{ and } 0 < \underline{A} < \bar{A} < \infty \tag{P}$$

$$\text{then: } d_{KT}(p, p_n) \lesssim_{p^{\otimes n}} n^{-\frac{1}{2} + \frac{1}{2\alpha}}.$$

- *If the eigenvalues of Σ_p follow an exponential decay rate:*

$$\underline{B} e^{-\tau i} \leq \lambda_i \leq \bar{B} e^{-\tau i} \text{ for some } \tau > 0 \text{ and } \underline{B}, \bar{B} \in (0, \infty), \tag{E}$$

$$\text{then: } d_{KT}(p, p_n) \lesssim_{p^{\otimes n}} \frac{(\log n)^{\frac{3}{2}}}{\sqrt{n}}.$$

Assuming some decay rate on the eigenvalues, we can focus on the convergence of the operators on a subspace of the top eigenvectors, using results from the Kernel PCA literature (Blanchard et al., 2007; Rudi et al., 2013).

Corollary

If Assumption (P) verified: $d_{KBW}(p, p_n) \lesssim_{p^{\otimes n}} n^{-\frac{1}{4} + \frac{1}{4\alpha}}$,
If Assumption (E) is verified: $d_{KBW}(p, p_n) \lesssim_{p^{\otimes n}} (\log n)^{\frac{3}{4}} n^{-\frac{1}{4}}$.

Experiments - Particle flows

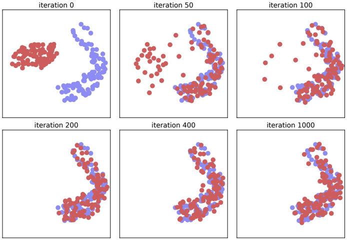


Figure: Particle flow with d_{KT}

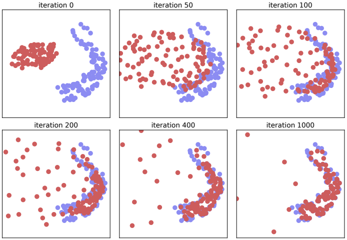


Figure: Particle flow with MMD

Experiments - ABC computation

Fact: Denote $P_\varepsilon = (1-\varepsilon)P + \varepsilon C$ where C is some contamination distribution. If $k(x,x) = 1$, $|d_{KT}(P_\varepsilon, Q) - d_{KT}(P, Q)| \leq 2\varepsilon$.

$\Longrightarrow d_{KT}$ is pretty robust to contamination, in contrast to the W_2 !

Experiments - ABC computation

Fact: Denote $P_\varepsilon = (1-\varepsilon)P + \varepsilon C$ where C is some contamination distribution. If $k(x,x) = 1$, $|d_{KT}(P_\varepsilon, Q) - d_{KT}(P, Q)| \leq 2\varepsilon$.

$\Longrightarrow d_{KT}$ is pretty robust to contamination, in contrast to the W_2 !

ABC: performing Bayesian inference in a likelihood-free fashion

$$\text{ABC posterior: } \pi(\theta|X^n) \propto \int \pi(\theta)\mathbb{1}_{\{d(X^n,Y^m)<\epsilon\}} p_\theta(Y^m)\mathrm{d}Y^m, \text{ where}$$

- $\pi(\theta)$ is a prior over the parameter space Θ
- $\epsilon > 0$ is a tolerance threshold
- Y^m are synthetic data generated according to $p_\theta(Y^m) = \prod_{j=1}^m p_\theta(Y_j)$.

Experiments - ABC computation

Fact: Denote $P_\varepsilon = (1-\varepsilon)P + \varepsilon C$ where C is some contamination distribution. If $k(x,x) = 1$, $|d_{KT}(P_\varepsilon, Q) - d_{KT}(P,Q)| \leq 2\varepsilon$.

$\implies d_{KT}$ is pretty robust to contamination, in contrast to the W_2 !

ABC: performing Bayesian inference in a likelihood-free fashion

$$\text{ABC posterior: } \pi(\theta|X^n) \propto \int \pi(\theta)\mathbb{1}_{\{d(X^n,Y^m)<\epsilon\}} p_\theta(Y^m)\mathrm{d}Y^m, \text{ where}$$

- $\pi(\theta)$ is a prior over the parameter space Θ
- $\epsilon > 0$ is a tolerance threshold
- Y^m are synthetic data generated according to $p_\theta(Y^m) = \prod_{j=1}^m p_\theta(Y_j)$.

computation:

- draw $\theta_i \sim \pi$ for $i = 1, ..., T$
- simulate synthetic data $Y^m \sim p_{\theta_i}$
- accept θ_i if the synthetic data is close to the real data

The result is a list L_θ of all accepted θ_i.

Experiments- ABC computation

True posterior (linear regression setting):

- prior $\pi = \mathcal{N}(0, \sigma_0^2)$ on θ
- real data consist of $n = 100$ samples $X^n = x_1, \ldots, x_n$ following $\mu^* = \mathcal{N}(1, 1)$ where 10% of the samples are replaced by contaminations from $\mathcal{N}(20, 1)$
- we can compute the (expected) true posterior mean as $\mathbb{E}[\sum_{i=1}^n x_i] \frac{n}{n+(\sigma_0^2)^{-1}}$, where $\mathbb{E}[\sum_{i=1}^n x_i] = 0.9 \times 1 + 0.1 \times 20 = 2.9$

Experiments- ABC computation

True posterior (linear regression setting):

- prior $\pi = \mathcal{N}(0, \sigma_0^2)$ on θ
- real data consist of $n = 100$ samples $X^n = x_1, \ldots, x_n$ following $\mu^* = \mathcal{N}(1, 1)$ where 10% of the samples are replaced by contaminations from $\mathcal{N}(20, 1)$
- we can compute the (expected) true posterior mean as $\mathbb{E}[\sum_{i=1}^n x_i] \frac{n}{n+(\sigma_0^2)^{-1}}$, where $\mathbb{E}[\sum_{i=1}^n x_i] = 0.9 \times 1 + 0.1 \times 20 = 2.9$

Method

- we fit the model $p_\theta = \mathcal{N}(\theta, 1)$ by picking the best θ possible.
- we carry out $T = 10000$ iterations, generating each times $m = n$ synthetic data.

Experiments- ABC computation

True posterior (linear regression setting):

- prior $\pi = \mathcal{N}(0, \sigma_0^2)$ on θ
- real data consist of $n = 100$ samples $X^n = x_1, \ldots, x_n$ following $\mu^* = \mathcal{N}(1, 1)$ where 10% of the samples are replaced by contaminations from $\mathcal{N}(20, 1)$
- we can compute the (expected) true posterior mean as $\mathbb{E}[\sum_{i=1}^n x_i] \frac{n}{n+(\sigma_0^2)^{-1}}$, where $\mathbb{E}[\sum_{i=1}^n x_i] = 0.9 \times 1 + 0.1 \times 20 = 2.9$

Method

- we fit the model $p_\theta = \mathcal{N}(\theta, 1)$ by picking the best θ possible.
- we carry out $T = 10000$ iterations, generating each times $m = n$ synthetic data.

Evaluation:

- we measure the average Mean Square Error between the target parameter $\theta^* = 1$ and the accepted $\theta_i \in L_\theta$: $\widehat{MSE} = \frac{1}{|L_\theta|} \sum_{\theta_i \in L_\theta} ||\theta_i - \theta^*||^2$

ABC - Results

Table: Average MSE of ABC Results. Gaussian kernel is used with $\sigma = 1$. $\mathrm{MMD_E}$ denotes MMD with the energy kernel $k(x, y) = -\|x - y\|$.

ε	distance	#accept. (*std*)	MSE (*std*)
0.05	MMD	1092 (*45*)	0.19 (*0.02*)
	$\mathrm{MMD_E}$	0	N/A
	d_{KT}	0	N/A
0.25	MMD	2964 (*92*)	1.29 (*0.06*)
	$\mathrm{MMD_E}$	0	N/A
	d_{KT}	58 (*25*)	**0.03** (*0.01*)
0.5	MMD	6168 (*406*)	7.47 (*1.83*)
	$\mathrm{MMD_E}$	846 (*35*)	0.17 (*0.05*)
	d_{KT}	828 (*34*)	0.12 (*0.01*)
1	MMD	10000 (*0*)	26.0 (*0.18*)
	$\mathrm{MMD_E}$	2926 (*52*)	1.33 (*0.6*)
	d_{KT}	2067 (*93*)	0.63 (*0.04*)

Conclusion:

- MMD is too lenient to accept most sampled θ_i leading to a high average MSE unless ε is carefully chosen
- d_{KT} discriminates between the correct and the wrong θ_i for a wide range of ϵ (even larger than the contamination threshold $\epsilon = 0.1$).

Conclusion

We introduced novel divergences between probability distributions

- that are closed-form for discrete measures
- enjoy nice statistical rates
- are more expensive than the MMD, but perform better on a wide variety of tasks

Conclusion

We introduced novel divergences between probability distributions

- that are closed-form for discrete measures
- enjoy nice statistical rates
- are more expensive than the MMD, but perform better on a wide variety of tasks

What is missing:

- Statistical lower bounds

Conclusion

We introduced novel divergences between probability distributions

- that are closed-form for discrete measures
- enjoy nice statistical rates
- are more expensive than the MMD, but perform better on a wide variety of tasks

What is missing:

- Statistical lower bounds
- Quantization rates

Conclusion

We introduced novel divergences between probability distributions

- that are closed-form for discrete measures
- enjoy nice statistical rates
- are more expensive than the MMD, but perform better on a wide variety of tasks

What is missing:

- Statistical lower bounds
- Quantization rates
- Propagation of chaos/descent lemma: "standard" proof requires Lipschitzness of the vector field

Conclusion

We introduced novel divergences between probability distributions

- that are closed-form for discrete measures
- enjoy nice statistical rates
- are more expensive than the MMD, but perform better on a wide variety of tasks

What is missing:

- Statistical lower bounds
- Quantization rates
- Propagation of chaos/descent lemma: "standard" proof requires Lipschitzness of the vector field
- Characterize the convexity, i.e. get a lower bound on the (Wasserstein) Hessian of the loss, or use a functional inequality?

Conclusion

We introduced novel divergences between probability distributions

- that are closed-form for discrete measures
- enjoy nice statistical rates
- are more expensive than the MMD, but perform better on a wide variety of tasks

What is missing:

- Statistical lower bounds
- Quantization rates
- Propagation of chaos/descent lemma: ”standard” proof requires Lipschitzness of the vector field
- Characterize the convexity, i.e. get a lower bound on the (Wasserstein) Hessian of the loss, or use a functional inequality?

(In constrast, for the MMD these things are known (Arbel et al., 2019), and partly for the kernel regularized variational approximations such as KALE or De-regularized MMD (Neumayer et al., 2024; Chen et al., 2024))

References I

Ambrosio, L., Gigli, N., and Savaré, G. (2005). *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media.

Arbel, M., Korba, A., Salim, A., and Gretton, A. (2019). Maximum mean discrepancy gradient flow. *Advances in Neural Information Processing Systems*, 32.

Bach, F. (2022). Information theory with kernel methods. *IEEE Transactions on Information Theory*, 69(2):752–775.

Blanchard, G., Bousquet, O., and Zwald, L. (2007). Statistical properties of kernel principal component analysis. *Machine Learning*, 66:259–294.

Chazal, C., Korba, A., and Bach, F. (2024). Statistical and geometrical properties of regularized kernel kullback-leibler divergence. *Advances in neural information processing systems*.

Chen, Z., Mustafi, A., Glaser, P., Korba, A., Gretton, A., and Sriperumbudur, B. K. (2024). (De)-regularized maximum mean discrepancy gradient flow. *arXiv preprint arXiv:2409.14980*.

References II

Glaser, P., Arbel, M., and Gretton, A. (2021). KALE flow: A relaxed KL gradient flow for probabilities with disjoint support. *Advances in Neural Information Processing Systems*, 34:8018–8031.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Hertrich, J., Wald, C., Altekrüger, F., and Hagemann, P. (2024). Generative sliced MMD flows with Riesz kernels. *International Conference on Learning Representations*.

Liu, D. C. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1-3):503–528.

Neumayer, S., Stein, V., and Steidl, G. (2024). Wasserstein gradient flows for Moreau envelopes of f-divergences in reproducing kernel hilbert spaces. *arXiv preprint arXiv:2402.04613*.

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

References III

Rudi, A., Canas, G. D., and Rosasco, L. (2013). On the Sample Complexity of Subspace Learning. *Advances in Neural Information Processing Systems*, 26.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer Science & Business Media.