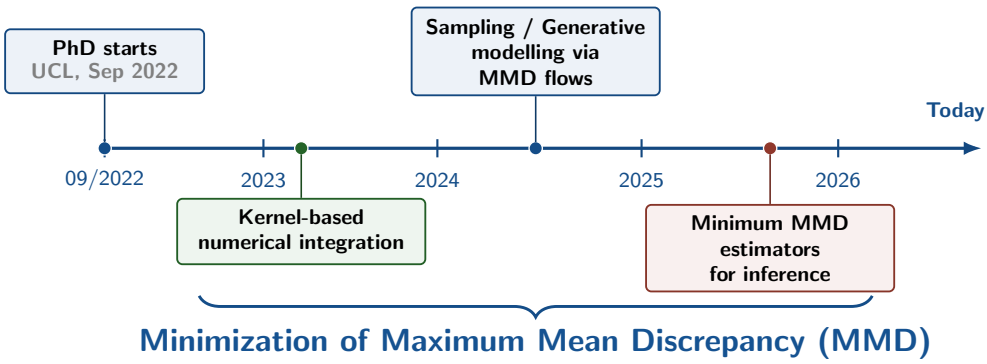
Minimization of Maximum Mean Discrepancy (MMD)

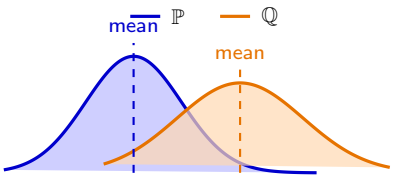
Zonghao (Hudson) Chen

Research timeline



Background: what is MMD?

- Given two distributions $\mathbb{P}$ and $\mathbb{Q}$, how to compare them?
- Compute the difference of their means: $|\mathbb{E}_{\mathbb{P}}[X] - \mathbb{E}_{\mathbb{Q}}[X]|$.
- Compute the difference of their second moments: $|\mathbb{E}_{\mathbb{P}}[X^2] - \mathbb{E}_{\mathbb{Q}}[X^2]|$.
- The first two moments are not enough!



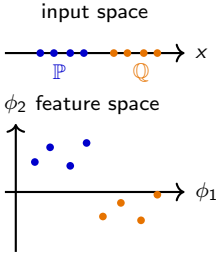
Goal: Compare $\mathbb{P}$ and $\mathbb{Q}$ with a general f, i.e., $|\mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(X)]|$?

Background: Maximum Mean Discrepancy (MMD) [GBR+12]

- $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) with kernel k.

$$\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) := \sup_{f \in \mathcal{H}} |\mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(X)]|.$$

- The RKHS contains functions that are linear combinations of the feature maps $\phi : \mathcal{X} \to \mathcal{H}$, $f = \sum_{i=1}^{n} \alpha_i \phi(x_i)$.



Theorem

MMD admits a closed-form expression:

$$\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{X,X' \sim \mathbb{P}}[k(X, X')] + \mathbb{E}_{Y,Y' \sim \mathbb{Q}}[k(Y, Y')] - 2\mathbb{E}_{X \sim \mathbb{P},\, Y \sim \mathbb{Q}}[k(X, Y)].$$

- MMD is a probability divergence ($\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = 0 \iff \mathbb{P} = \mathbb{Q}$) if k is characteristic (Gaussian) [SFL11].

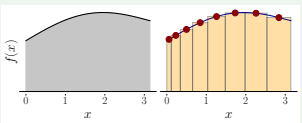
We now have MMD

A kernel-based way to compare probability distributions $\mathbb{P}$ and $\mathbb{Q}$.

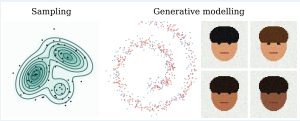
What can we use it for?

Numerical Integration

$$\int f(x)\mathrm{d}\mathbb{P}(x)$$

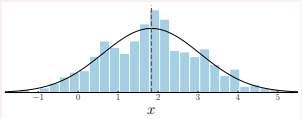


Sampling / generative modelling



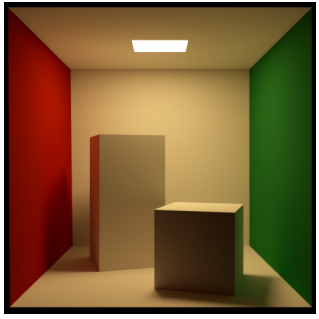
Statistical inference

$$\theta_{\mathrm{MMD}}$$

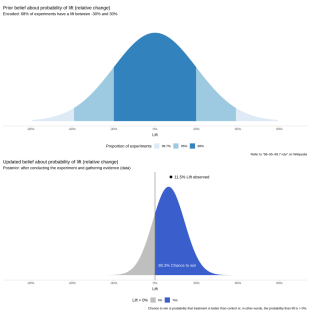


Numerical Integration

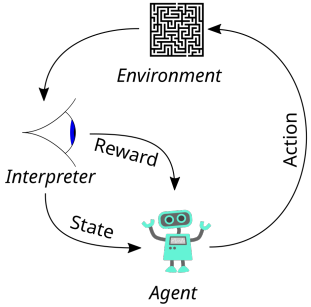
Where do integrals appear in computer science?



Computer graphics
Rendering



Bayesian inference
Posterior expectations



Reinforcement learning
Expected returns

Part I: Numerical Integration

- Given a function f and a distribution $\mathbb{P}$, how to compute $\int f(x)\mathrm{d}\mathbb{P}(x)$?
- Given a set of points $\{x_i\}_{i=1}^n$ and function evaluations $\{f(x_i)\}_{i=1}^n$.
- Quadrature rule:

$$\hat{I} = \sum_{i=1}^{n} w_i f(x_i).$$

- When $w_1 = \cdots = w_n = 1/n$, this is Monte Carlo. $\hat{I}_{MC} = \frac{1}{n}\sum_{i=1}^n f(x_i)$.

$$\mathbb{E}\left[|\hat{I}_{MC} - I|\right] = \mathcal{O}(n^{-1/2}).$$

- Or equivalently, to obtain an ϵ-accurate estimate, we need $n = \mathcal{O}(\epsilon^{-2})$ samples.

Question: Can we do better than $\mathcal{O}(n^{-1/2})$?

Answer: Find better weights $w_1, \ldots, w_n$.

Part I: Numerical Integration

- Suppose $f \in \mathcal{H}$ is in the RKHS associated with k.
- Denote $\hat{\mathbb{P}} = \sum_{i=1}^n w_i \delta_{x_i}$ as the weighted empirical distribution.

$$\left| I - \hat{I} \right|^2 = \left| \int f(x) \mathrm{d}\mathbb{P}(x) - \sum_{i=1}^n w_i f(x_i) \right|^2 \leq \|f\|_{\mathcal{H}}^2 \cdot \mathrm{MMD}^2(\hat{\mathbb{P}}, \mathbb{P}).$$

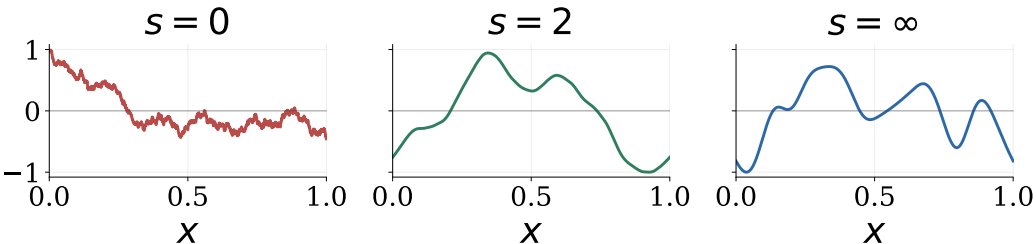
- The right-hand side is the worst-case error of the quadrature rule.

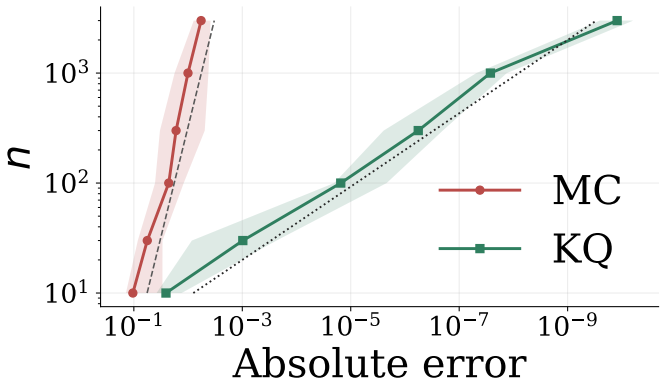
$$w_{1:n} := \arg \min_{w_{1:n} \in \mathbb{R}^n} \mathrm{MMD} \left(\sum_{i=1}^n w_i \delta_{x_i}, \mathbb{P} \right),$$
$$w_{1:n} = \mathbb{E}_{X \sim \mathbb{P}}[k(X, x_{1:n})] K(x_{1:n}, x_{1:n})^{-1}.$$

- The kernel quadrature (KQ) rule: $\hat{I}_{\mathrm{KQ}} = \sum_{i=1}^n w_i f(x_i)$.

Rate: KQ achieves $\mathcal{O}(n^{-s/d})$ rate.

- s is the smoothness of f, d is the dimension of input x.
- Better than $\mathcal{O}(n^{-1/2})$ when $s > d/2$.





- $s = 4$ and $d = 1$.

Conditional Bayesian Quadrature (UAI 2025)

Z. Chen, M. Naslidnyk, A. Gretton, F.-X. Briol

- Extension to conditional expectations: $I(\theta) = \mathbb{E}_{X \sim p(\cdot|\theta)}[g(X, \theta)]$.
- Weighted rule: $\hat{I}_{\mathrm{CBQ}}(\theta) := \sum_{t=1}^{T} \sum_{n=1}^{N} w_{n,t} g(x_n^{(t)}, \theta_t)$.

Nested Expectations with Kernel Quadrature (ICML 2025)

Z. Chen, M. Naslidnyk, F.-X. Briol

- Extension to nested expectations: $I := \mathbb{E}_{\theta \sim \mathbb{Q}} \left[f \left(\mathbb{E}_{X \sim \mathbb{P}_\theta} [g(X, \theta)] \right) \right]$.
- Weighted rule: $\hat{I}_{\mathrm{NKQ}} = \sum_{t=1}^{T} w_t^{\Theta} f \left(\sum_{n=1}^{N} w_{n,t}^{\mathcal{X}} g(x_n^{(t)}, \theta_t) \right)$.

BayesSum: Bayesian Quadrature in Discrete Spaces (ProbNum 2026)

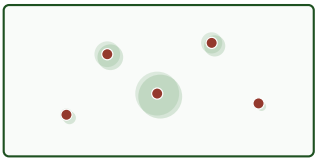
S. S. Kang, F.-X. Briol, T. Karvonen, Z. Chen

- Extension to sums: $I = \mathbb{E}_{X \sim \mathbb{P}}[f(X)] = \sum_{x \in \mathcal{X}} f(x) p(x)$.

Sampling / Generative Modelling

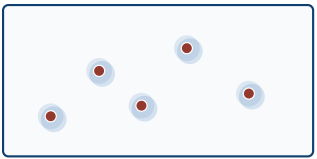
Quadrature

fixed locations, choose weights



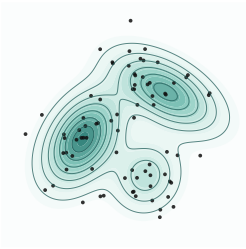
Sampling / generation

equal weights, choose locations

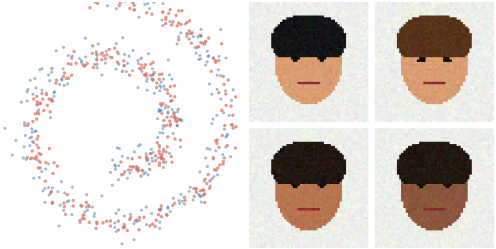


Part II: Sampling / Generative modelling

Sampling



Generative modelling



- Given a target distribution $\mathbb{Q}$, how to draw samples $\{x_i\}_{i=1}^n$ from it?
  - When $\mathbb{Q}$ is known up to normalization, this is sampling.
  - When $\mathbb{Q}$ is known with i.i.d samples, this is generative modelling.

Problem: How to learn a target probability distribution $\mathbb{Q}$ on $\mathbb{R}^d$.

- This problem can be written as an optimization problem on $\mathcal{P}(\mathbb{R}^d)$, the space of probability distributions.

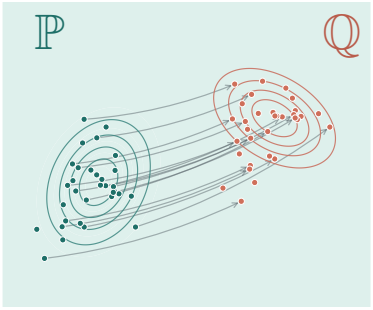
$$\arg \min_{\mathbb{P} \in \mathcal{P}(\mathbb{R}^d)} D(\mathbb{P}, \mathbb{Q}).$$

- Here D is a similarity metric or distance, e.g. Kullback–Leibler divergence / MMD.
  - $D(\mathbb{P}_1, \mathbb{P}_2) = 0$ if and only if $\mathbb{P}_1 = \mathbb{P}_2$.
- How to find the minimum? Gradient descent!

Question: Wait... How to define gradients on $\mathcal{P}(\mathbb{R}^d)$?

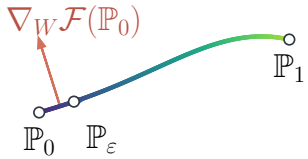
Answer: Wassserstein gradient ! [Ott01]

Optimal transport



Wasserstein gradient

$\mathcal{F} : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$



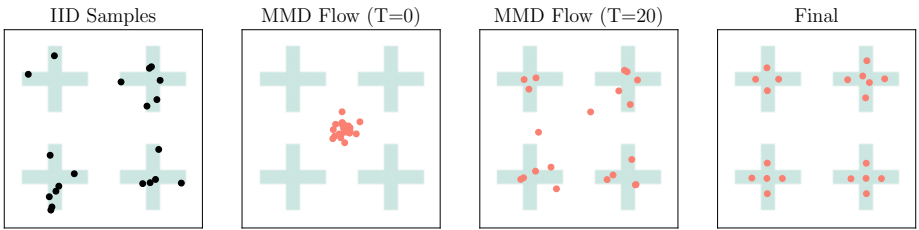
$$\nabla_W \mathcal{F}(\mathbb{P}_0) = \lim_{\varepsilon \to 0} \frac{\mathcal{F}(\mathbb{P}_\varepsilon) - \mathcal{F}(\mathbb{P}_0)}{\varepsilon}$$

Part II: Sampling / Generative modelling

- Unsurprisingly, we use MMD as the divergence.

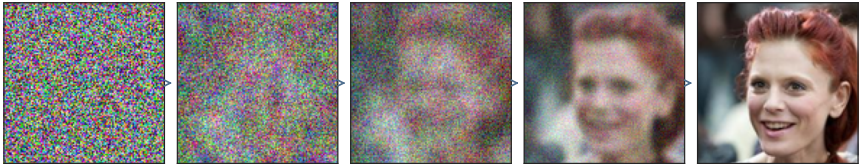
$$\arg\min_{\mathbb{P}\in\mathcal{P}(\mathbb{R}^d)} \mathrm{MMD}^2(\mathbb{P},\mathbb{Q}).$$

- A (Wasserstein) gradient descent of $\mathrm{MMD}^2(\mathbb{P},\mathbb{Q})$ gives a sequence of distributions: $\mathbb{P}_0, \mathbb{P}_1, \ldots, \mathbb{P}_t$ [AKSG19].
  - In the limit, $\mathbb{P}_t$ converges to $\mathbb{Q}$ and $\mathrm{MMD}(\mathbb{P}_t,\mathbb{Q}) \to 0$ (or not?).



- MMD flow generates more uniform samples.

Part II: Sampling / Generative modelling



- MMD flow does not outperform diffusion-based generative models [GBG25].
- A novel framework of generative modelling & sampling.

(De)-regularized MMD Gradient Flow (JMLR 2025)

Z. Chen, A. Mustafi, P. Glaser, A. Korba, A. Gretton, B. K. Sriperumbudur

- DrMMD interpolates between MMD and χ^2 divergence.
- Convergence in KL divergence: $\lim_{t\to\infty} \mathrm{KL}(\mathbb{P}_t, \mathbb{Q}) = 0$.

Sobolev Regularized MMD Gradient Flow (Arxiv)

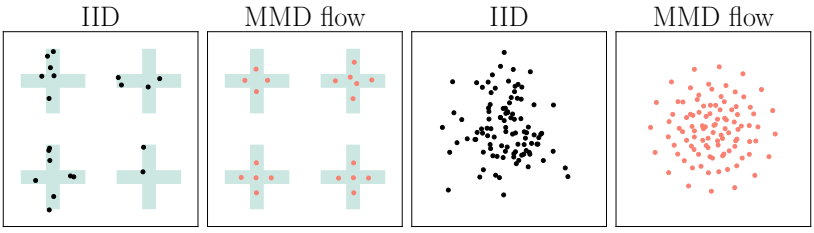
C. Tian, B. K. Sriperumbudur, A. Gretton, Z. Chen

- Gradient regularization on the critic [MSR19].

$$\mathrm{SrMMD}(\mu, \pi) = \sup_{f\in\mathcal{H}:\|\nabla f\|^2_{L^d_2(\mu)}+\lambda\|f\|^2_{\mathcal{H}}\leq 1} \int f \, \mathrm{d}(\mu - \pi).$$

- Convergence in MMD: $\lim_{t\to\infty} \mathrm{MMD}(\mathbb{P}_t, \mathbb{Q}) = 0$.
- No conditions on π.

Empirical observation:
(Wasserstein) gradient flows give samples more uniform than i.i.d samples [XKS22].



- However, theoretical rate can hardly beat Monte Carlo: $n^{-1/2}$.

Stationary MMD Points (ICML 2026)

Z. Chen, T. Karvonen, H. Kanagawa, F.-X. Briol, C. J. Oates

- Samples from MMD flow are better for numerical integration than IID samples.
- IID samples: $\mathcal{O}(n^{-1/2})$, stationary MMD points: $o(n^{-1/2})$.

Thinned Mean Field Langevin Dynamics (ICML 2026)

Z. Chen, H. Kanagawa, F.-X. Briol, C. J. Oates, L. Mackey

- Computational cost of MMD flow is $\mathcal{O}(n^2)$ due to pair-wise interaction.
- Reduce cost to $\mathcal{O}(n^{1.5})$ without sacrificing the sample quality.
- Kernel thinning!

Statistical Inference

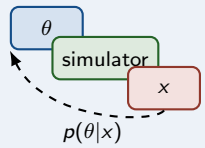
MMD flow

$$\arg \min_{\mathbb{P} \in \mathcal{P}(\mathbb{R}^d)} \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q})$$

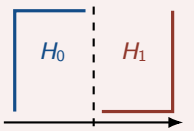
Minimum MMD inference

$$\theta_{\mathrm{MMD}} = \arg \min_{\theta \in \Theta} \mathrm{MMD}^2(\mathbb{P}_\theta, \mathbb{Q})$$

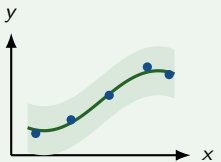
Where does statistical inference appear in computer science?



Simulation-based inference
Infer parameters from simulator outputs.



Hypothesis testing
Is the data compatible with a class of models?



Uncertainty
How confident should a prediction be?

Part III: Statistical inference

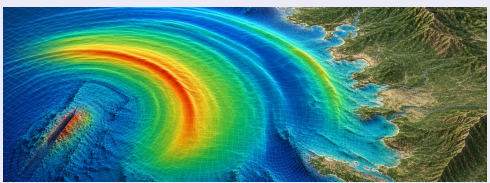
- Given a parametric model $\{\mathbb{P}_\theta\}_{\theta\in\Theta\subseteq\mathbb{R}^p}$.
- Infer the parameter θ that best explains an unknown data distribution $\mathbb{Q}$.
- Maximum likelihood estimation is the most popular:

$$\theta_{\mathrm{MLE}} = \arg\max_{\theta\in\Theta} \mathbb{E}_{X\sim\mathbb{Q}}[\log p_\theta(X)].$$

What is not captured by maximum likelihood?

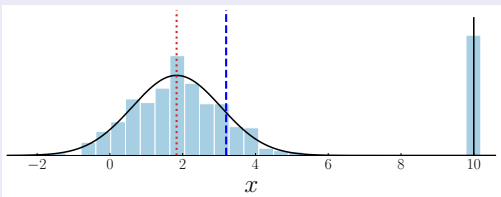
Likelihood-free simulators

When $\mathbb{P}_\theta$ is a scientific simulator: we can sample $x \sim p_\theta(x)$, but cannot evaluate the density $p_\theta(x)$.



Robustness

A few contaminated observations can strongly move the fitted parameter.



Part III: Statistical inference

- Minimum MMD estimator [BBDG19, CAA22]

$$\begin{aligned}\theta_{\mathrm{MMD}} &= \arg\min_{\theta\in\Theta} \mathrm{MMD}^2(\mathbb{P}_\theta, \mathbb{Q}) \\ &= \arg\min_{\theta\in\Theta} \Big\{ \mathbb{E}_{x,x'\sim\mathbb{P}_\theta}[k(x,x')] - 2\mathbb{E}_{x\sim\mathbb{P}_\theta, y\sim\mathbb{Q}}[k(x,y)] + \mathrm{const} \Big\}.\end{aligned}$$

Challenge 1: Likelihood-free inference?

✓ Only requires samples from $\mathbb{P}_\theta$ and $\mathbb{Q}$.

Challenge 2: Robustness to outliers?

✓ Bounded effects under bounded kernels.

Question: How to reliably find θ_{MMD}?

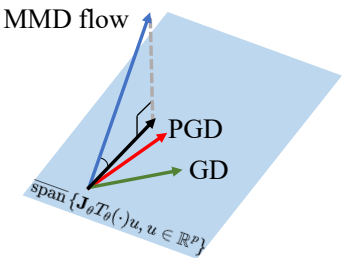
- MMD flow:

$$\mathbb{P}_{t+1} = (\mathrm{Id} - \gamma \nabla f_{\mathbb{P}_t, \mathbb{Q}})_{\#} \mathbb{P}_t.$$

- MMD flow converges globally.
- Consider $\mathbb{P}_\theta = (T_\theta)_{\#}\rho$:

$$\mathbb{P}_{\theta_{t+1}} = (T_{\theta_t - \gamma \Delta \theta_t})_{\#}\rho.$$

- Choose $\mathbb{P}_{\theta_{t+1}}$ closest to the MMD flow iterate $\mathbb{P}_{t+1}$.
- $\lim_{t\to\infty} \mathrm{MMD}(\mathbb{P}_{\theta_t}, \mathbb{Q}) = \min_\theta \mathrm{MMD}(\mathbb{P}_\theta, \mathbb{Q})$.



What next: causal inference.

Minimum MMD for counterfactual inference

Question: What would have happened under a different intervention?

- Treatment A.
- Counterfactual outcome: $Y(a)$ under treatment $A = a$.
- Missing data/inference problem.

Conformal Counterfactual Inference under Hidden Confounding (KDD 2024)

Z. Chen, R. Guo, J.-F. Ton, Y. Liu

Neural Networks for NPIV: Optimization and Generalization (Arxiv)

Z. Chen, A. Nitanda, A. Gretton, T. Suzuki

NPIV with Observed Covariates (Arxiv)

Z. Shen*, Z. Chen*, D. Meunier, I. Steinwart, A. Gretton, Z. Li

Summary

Summary

- MMD is a powerful tool for comparing probability distributions:

$$\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) := \sup_{f \in \mathcal{H}} |\mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(X)]|.$$

- Minimization of MMD in ....

$\int$

Numerical Integration

$$\min_{w_{1:n} \in \mathbb{R}^n} \mathrm{MMD}\left(\sum_{i=1}^{n} w_i \delta_{x_i}, \mathbb{P}\right)$$

$\sim$

Sampling / generative modelling

$$\min_{\mathbb{P} \in \mathcal{P}(\mathbb{R}^d)} \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q})$$

θ

Statistical inference

$$\min_{\theta \in \Theta} \mathrm{MMD}^2(\mathbb{P}_\theta, \mathbb{Q})$$

[AKSG19] Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[BBDG19] Francois-Xavier Briol, Alessandro Barp, Andrew B Duncan, and Mark Girolami. Statistical inference for generative models with maximum mean discrepancy. *arXiv preprint arXiv:1906.05944*, 2019.

[CAA22] Badr-Eddine Ch'erief-Abdellatif and Pierre Alquier. Finite sample properties of parametric mmd estimation: robustness to misspecification and dependence. *Bernoulli*, 28(1):181–213, 2022.

[GBG25] Alexandre Galashov, Valentin De Bortoli, and Arthur Gretton. Deep MMD gradient flow without adversarial training. In *The Thirteenth International Conference on Learning Representations*, 2025.

[GBR+12] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

[MSR19] Youssef Mroueh, Tom Sercu, and Anant Raj. Sobolev descent. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2976–2985. PMLR, 2019.

[Ott01] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. 2001.

[SFL11] Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(7):2389–2410, 2011.

[XKS22] Lantian Xu, Anna Korba, and Dejan Slepcev. Accurate quantization of measures via interacting particle-based optimization. In *International Conference on Machine Learning*, pages 24576–24595. PMLR, 2022.