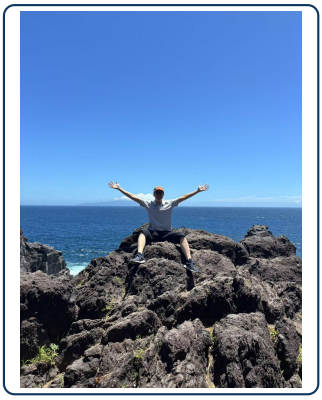
Stationary MMD Points

ICML 2026

Zonghao (Hudson) Chen

About Me



Zonghao (Hudson) Chen

4th year PhD student, University College London

Foundational AI Centre	François-Xavier Briol
Gatsby Unit	Arthur Gretton

Tsinghua Uni., B.Eng. in Electronic Engineering, 2022

Research Interests

- Kernel and nonparametric methods
- Statistical learning theory

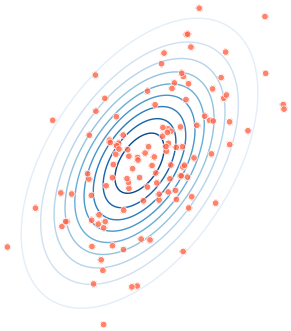
Motivation

Problem: Approximation of a target distribution μ with a finite set of points $\{x_i\}_{i=1}^n$.

- Denote the empirical distribution of $\{x_i\}_{i=1}^n$ as

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}.$$

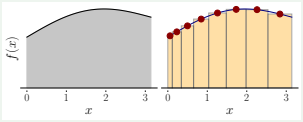
- We want μ_n to be close to μ.
- Uniform weights.



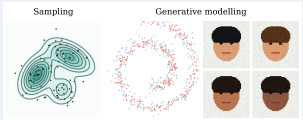
Motivation

Problem: Approximation of a target distribution μ with a finite set of points $\{x_i\}_{i=1}^n$.

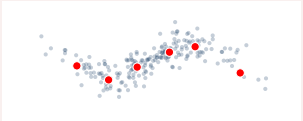
Numerical Integration



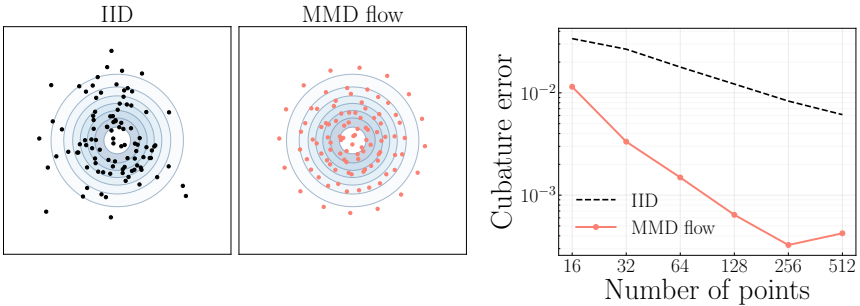
Sampling



Coreset Selection



An empirical observation



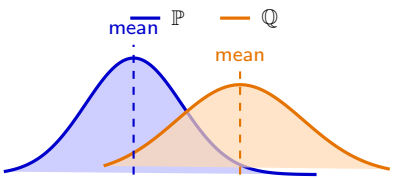
- Samples from MMD flow are more uniform than IID samples [XKS22].
- Lower numerical integration error for fixed n.

$$\left| \int f(x)\, \mathrm{d}\mu(x) - \int f(x)\, \mathrm{d}\mu_n(x) \right|.$$

Goal: Why?

Background: What is MMD?

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



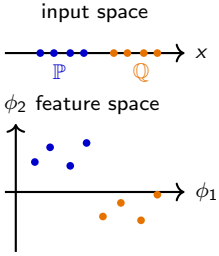
Goal: Compare $\mathbb{P}$ and $\mathbb{Q}$ with a general f, i.e., $|\mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(X)]|$?

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:

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

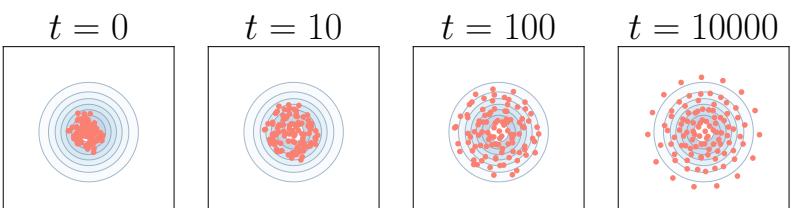
- MMD is a probability divergence ($\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = 0 \iff \mathbb{P} = \mathbb{Q}$) if k is c_0-universal (Gaussian) [SFL11].

Background: MMD gradient flow [AKSG19]

- Start with an initial set of n particles $\{x_i^{(0)}\}_{i=1}^n$ at time 0.
- Evolve the particles according to the MMD gradient flow:

$$x_i^{(t+1)} = x_i^{(t)} - \gamma \cdot \nabla_{x_i^{(t)}} \mathrm{MMD}^2 \left(\mu, \frac{1}{n} \sum_{j=1}^{n} \delta_{x_j^{(t)}} \right), \qquad i = 1, \ldots, n.$$

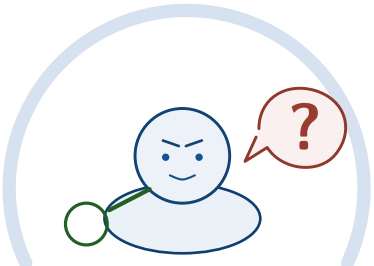
  - $\gamma > 0$ is the step size.



- A Wasserstein gradient flow.

Motivation

Observation: MMD flow samples are more uniform than IID samples [XKS22].



Existing work

- Existing finite-sample rates in MMD gradient flows [AKSG19, CMG+25].
- Let $\mu_n^{(t)} = \frac{1}{n}\sum_{j=1}^{n} \delta_{x_j^{(t)}}$ be the empirical distribution of n particles at iterate t.

$$\mathrm{MMD}\left(\mu, \mu_n^{(t)}\right) \leq C(t) n^{-\frac{1}{2}}, \quad W_2\left(\mu, \mu_n^{(t)}\right) \leq C(t) n^{-\frac{1}{2}}.$$

Limitations

- Same rate as IID sampling: $n^{-\frac{1}{2}}$.
- The constant $C(t)$ grows exponentially in t.

Existing work

- Minimum MMD points

$$\{x_i^*\}_{i=1}^n = \arg\min_{\{x_i\}_{i=1}^n} \mathrm{MMD}^2\left(\mu, \frac{1}{n}\sum_{j=1}^n \delta_{x_j}\right).$$

- Let $\mu_n^* = \frac{1}{n}\sum_{j=1}^n \delta_{x_j^*}$ be the empirical distribution [XKS22, DP10]:

$$\mathrm{MMD}\left(\mu, \mu_n^*\right) \leq Cn^{-1}.$$

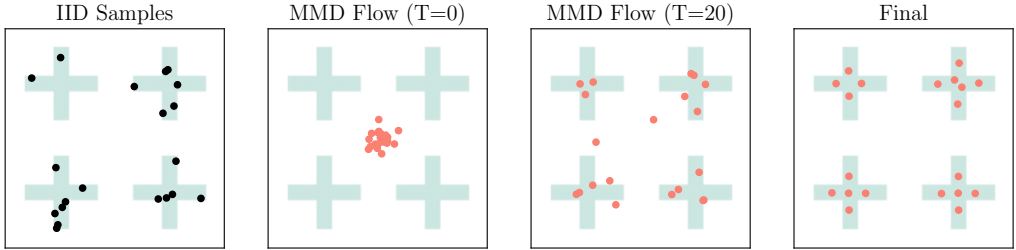
- Quasi Monte Carlo point set achieve this rate [DP10].

Limitations

- Minimum MMD points are hard to compute, due to the non-convexity of MMD.
- Gradient descent only finds a stationary point instead.

Motivation

× ~~**Minimum MMD points**~~ but ✓ **Stationary MMD points**



- A four-cross distribution.
- An ideal minimum MMD points would have 5 points in each cross.
- Stationary MMD points have $(4, 6, 5, 5)$ points in each cross.

Stationary MMD points

Definition (Stationary MMD points)

A sequence of point sets $\{x_i\}_{i=1}^n$ *with* $\mu_n := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, *that satisfies:*

1. **Convergence:** $\mathrm{MMD}(\mu, \mu_n) \to 0$ *as* $n \to \infty$.
2. **Stationarity:** $\nabla_{x_1,\ldots,x_n} \mathrm{MMD}^2(\mu, \mu_n) = 0$.

- Minimum MMD points are a special case of stationary MMD points.
- Stationary MMD points are easier to compute, e.g., via MMD gradient flow.

Question: Why does stationary MMD points exhibit superior numerical integraion performance than IID points?

Stationary MMD points

- Simulate MMD flow for long enough, $t \to \infty$.
- For any $i \in \{1, \dots, n\}$:

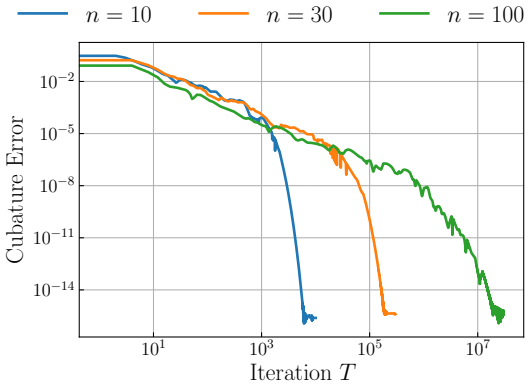
$$\nabla_{x_i} \mathrm{MMD}^2 \left(\mu, \ \frac{1}{n} \sum_{j=1}^{n} \delta_{x_j} \right) = 0 \Longrightarrow \frac{1}{n} \sum_{j=1}^{n} \nabla_1 k\left(x_i, x_j\right) - \int \nabla_1 k\left(x_i, y\right) \mathrm{d}\mu(y) = 0.$$

- Exact integration on $x \mapsto \nabla_1 k\left(x_i, x\right)$.
- Exact integration on $\{1\}$.

Exact integration subspace:

$$\mathcal{F}_n := \mathrm{span}\Big(\{1\} \cup \big\{ \nabla_1 k(x_i, \cdot) : i \in \{1, \dots, n\} \big\} \Big) \subset \mathcal{H}.$$

Stationary MMD points



- Exact integration on $\mathcal{F}_n$.

Stationary MMD points

Assumption

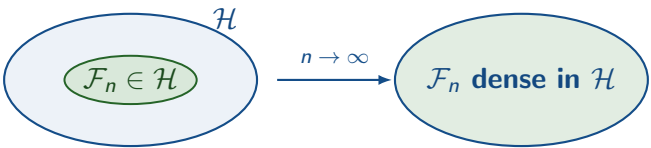
The kernel k is smooth and c_0-universal.

Proposition (Approximation of $\mathcal{F}_n$ towards $\mathcal{H}$)

Given $f \in \mathcal{H}$. Consider the semi-norm

$$|f|_n := \inf \{ \|r_n\|_{\mathcal{H}} : f = f_n + r_n,\ f_n \in \mathcal{F}_n,\ r_n \in \mathcal{H} \}$$

Then, $\lim_{n \to \infty} |f|_n = 0$.



Stationary MMD points

Theorem (Super-convergence)

For a stationary MMD point set $\{x_i\}_{i=1}^n$ *and* $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$,

$$\left|\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \int f \,\mathrm{d}\mu\right| \leq |f|_n \cdot \mathrm{MMD}\,(\mu, \mu_n)$$

Proof sketch

- Decompose $f = f_n + r_n$ with $f_n \in \mathcal{F}_n$.

$$\left|\frac{1}{n}\sum_{i=1}^{n} f\,(x_i) - \int f \,\mathrm{d}\mu\right| = \left|\frac{1}{n}\sum_{i=1}^{n} r_n\,(x_i) - \int r_n \,\mathrm{d}\mu\right| \leq \|r_n\|_{\mathcal{H}} \cdot \mathrm{MMD}\,(\mu, \mu_n)\,,$$

Stationary MMD points

Theorem (Super-convergence)

Let $f \in \mathcal{H}$ be fixed. For a stationary MMD point set $\{x_i\}_{i=1}^n$ and $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$,

$$\left| \frac{1}{n}\sum_{i=1}^{n} f(x_i) - \int f \, \mathrm{d}\mu \right| \leq |f|_n \cdot \mathrm{MMD}(\mu, \mu_n).$$

$|f|_n \to 0$ **Stationarity**

$\mathrm{MMD}(\mu, \mu_n) \to 0$ **MMD flow**

- Proved already (✓): $|f|_n \to 0$ due to stationarity.
- Next step (➜): $\mathrm{MMD}(\mu, \mu_n) \to 0$.
- Key bit (★): Avoid $C(t)$ exponential in time.

If we can prove that, $\mathrm{MMD}(\mu, \mu_n) = \mathcal{O}(n^{-\frac{1}{2}})$, then we obtain $o(n^{-\frac{1}{2}})$ rate.

Stationary MMD points

- At iterate $t \in \mathbb{N}$
- MMD gradient flow: for $i = 1, \ldots, n$,

$$x_i^{(t+1)} = x_i^{(t)} - \gamma \left\{ \frac{1}{n} \sum_{j=1}^{n} \nabla_1 k \left(x_i^{(t)},\, x_j^{(t)} \right) - \mathbb{E}_{Y \sim \mu} \left[\nabla_1 k \left(x_i^{(t)},\, Y \right) \right] \right\}$$

- Noisy MMD gradient flow: for $i = 1, \ldots, n$,
  - $u_i^{(t)} \sim \mathcal{N}(0, \mathrm{I}_d)$ and $\beta_t > 0$.

$$x_i^{(t+1)} = x_i^{(t)} - \gamma \left\{ \frac{1}{n} \sum_{j=1}^{n} \nabla_1 k \left(x_i^{(t)} + \beta_t u_i^{(t)},\, x_j^{(t)} \right) - \mathbb{E}_{Y \sim \mu} \left[\nabla_1 k \left(x_i^{(t)} + \beta_t u_i^{(t)},\, Y \right) \right] \right\}$$

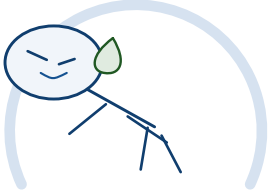
Stationary MMD points

Assumption (Noise level)

Define $\Phi(z, w) = \mathbb{E}_{Y\sim\mu}[\nabla_1 k(z, Y)] - \nabla_1 k(z, w)$. *For each* $t \in \mathbb{N}$, $\beta_t \in (0, 1)$ *satisfies,*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{u_i^{(t)}}\left[\left\|\frac{1}{n}\sum_{j=1}^{n}\Phi\left(x_i^{(t)} + \beta_t u_i^{(t)}, x_j^{(t)}\right)\right\|^2\right] \geq \beta_t^2\,\mathrm{MMD}^2\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i^{(t)}}, \mu\right).$$

- In expectation, a gradient dominance condition.
- Seems unavoidable for convergence of MMD flow [AKSG19, CMG+25].



Stationary MMD points

Theorem

Suppose that there exists $\beta_t \propto t^{-\alpha}$ for $0 \leq \alpha < 1/2$ such that the condition holds. C' depends only on kernel k and dimension d. Then, for any $T \leq n^{\frac{1}{\alpha}}$, with high probability,

$$\mathrm{MMD}\left(\mu_n^{(T)}, \mu\right) = \mathcal{O}\left(\exp\left(-C'\gamma T^{1-2\alpha}\right) + \frac{(\log T)^2 T^{\alpha}}{\sqrt{n}}\right).$$

- Reduce the exponential dependence on T to a polynomial dependence.
- First end-to-end finite-particle convergence bound for MMD flow.
- Take $T \propto (\log n)^{\frac{1}{1-2\alpha}}$, then $\mathrm{MMD}\left(\mu_n^{(T)}, \mu\right) = \mathcal{O}(n^{-\frac{1}{2}})$ up to logarithmic factors.

Stationary MMD points

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \int f \, \mathrm{d}\mu \right| \leq |f|_n \cdot \mathrm{MMD}(\mu, \mu_n).$$

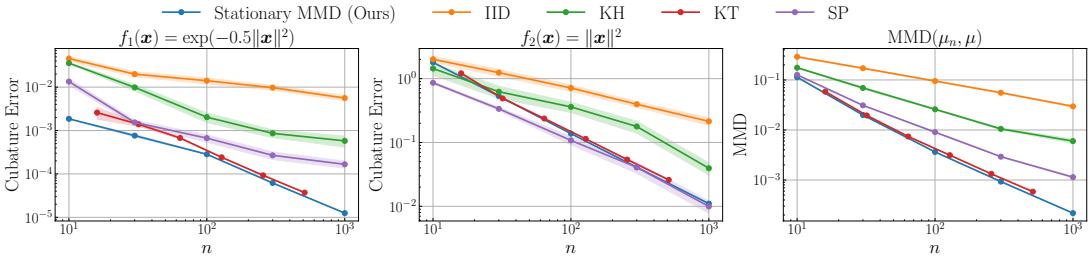
- $\lim_{n\to\infty} |f|_n = 0$ due to stationarity.
- $\mathrm{MMD}(\mu, \mu_n) = \mathcal{O}(n^{-\frac{1}{2}}(\log n)^b)$ due to MMD flow convergence.

$o(n^{-\frac{1}{2}}(\log n)^b)$ rate for stationary MMD points.

Remark. If started with some point sets with $\mathrm{MMD}(\mu, \mu_n) = \mathcal{O}(n^{-r})$, then we can run MMD flow for $T = \mathcal{O}(1)$ iterations to obtain a new point set with $o(n^{-r})$ rate.

Experiments

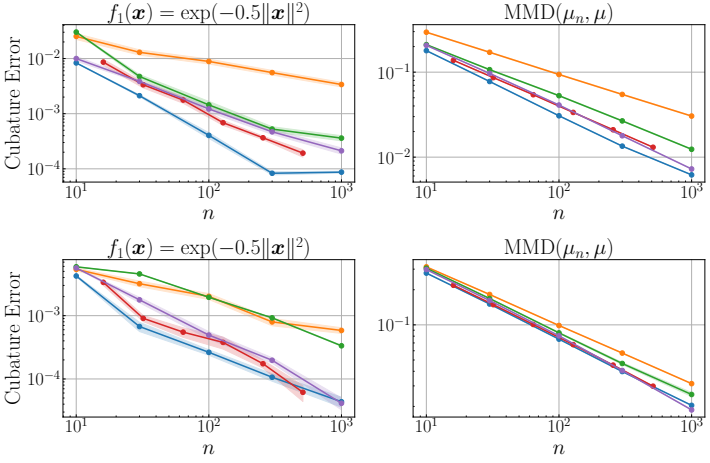
- Mixture of Gaussians.
- KH is kernel herding [CWS12]; KT is kernel thinning [DM24]; SP is support points [MJ18].



- $f_1 \in \mathcal{H}$ while $f_2 \notin \mathcal{H}$.
- Stationary MMD points along with KT are the best.
- SP works well on f_2 since it uses $k(x, y) = \|x - y\|$.

Experiments

- Coreset selection from a large dataset
- House (Top) has $22,784$ and Elevator (Bottom) has $16,599$ data points.



Summary

Contributions

- We propose stationary MMD points.
- A novel framework for low-discrepancy points.
- Better rate from stationarity.

Limitations

- We only prove a $o(n^{-\frac{1}{2}}(\log n)^b)$ rate, unclear if better than IID $\mathcal{O}(n^{-\frac{1}{2}})$ or not.
- Noise injection in MMD flow is not ideal.

References I

[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.

[CMG+25] Zonghao Chen, Atalanti Mustafi, Pierre Glaser, Anna Korba, Arthur Gretton, and Bharath K. Sriperumbudur. (De)-regularized MMD Gradient Flow. *Journal of Machine Learning Research*, 2025.

[CWS12] Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel herding. *arXiv preprint arXiv:1203.3472*, 2012.

[DM24] Raaz Dwivedi and Lester Mackey. Kernel thinning. *Journal of Machine Learning Research*, 25(152):1–77, 2024.

[DP10] Josef Dick and Friedrich Pillichshammer. *Digital nets and sequences: discrepancy theory and quasi–Monte Carlo integration*. Cambridge University Press, 2010.

References II

[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.

[MJ18] Simon Mak and V Roshan Joseph. Support points. 2018.

[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.