De-regularized Maximum Mean Discrepancy Gradient Flow

Zonghao Chen[1] Aratrika Mustafi[2] Pierre Glaser[1] Anna Korba[3] Arthur Gretton[1] Bharath K. Sriperumbudur[2]

[1] University College London [2] Pennsylvania State University [3] ENSAE, CREST, IP Paris

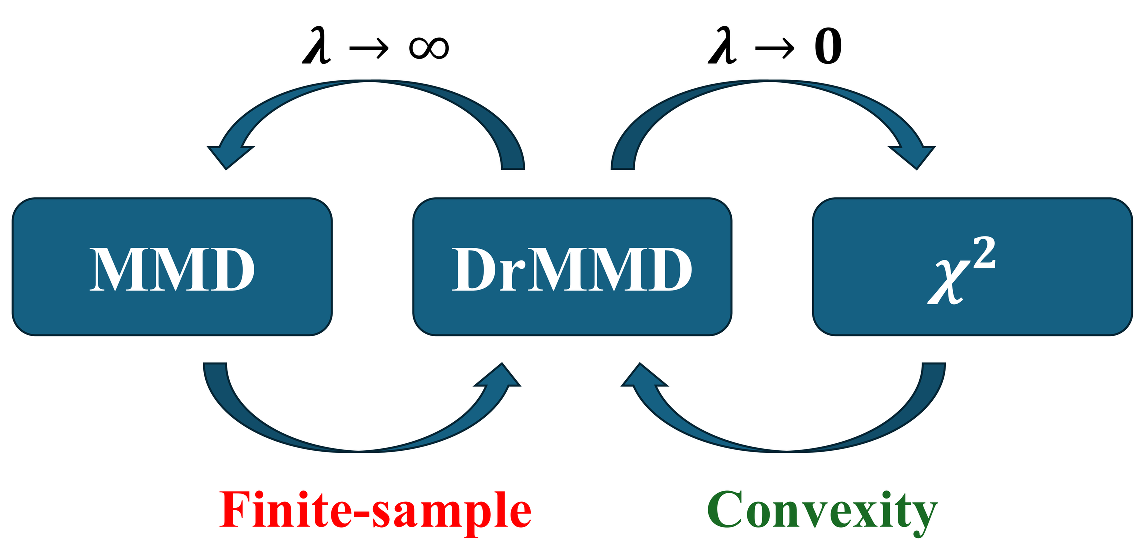
@Hudson19990518

TL;DR

We propose a new Wasserstein gradient flow of a de-regularization of maximum mean discrepancy (DrMMD).

- Generative modeling where π is known with M i.i.d samples $\{y_i\}_{i=1}^M$.

1. DrMMD interpolates between MMD^2 and χ^2-divergence.
2. DrMMD flow admits tractable finite sample implementations.
3. DrMMD flow enjoys global convergence in KL divergence.
4. DrMMD flow theoretically justify using adaptive kernels in MMD based generative models.



$$\mathrm{MMD}^2 = \left\| \mathcal{T}_\pi^{\frac{1}{2}} \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1 \right) \right\|_{L^2(\pi)}^2, \quad \chi^2 = \left\| \frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1 \right\|_{L^2(\pi)}^2$$

$$\mathrm{DrMMD} = (1+\lambda) \left\| \left((\mathcal{T}_\pi + \lambda)^{-1} \mathcal{T}_\pi \right)^{\frac{1}{2}} \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1 \right) \right\|_{L^2(\pi)}^2$$

- $\mathcal{T}_\pi : L^2(\pi) \to L^2(\pi), f \mapsto \int k(x,\cdot) f(x) \mathrm{d}\pi(x)$ is the kernel integral operator.
- $(\mathcal{T}_\pi + \lambda)^{-1} \mathcal{T}_\pi$ is Tikhonov regularization.

Accepted (minor revision) to JMLR

DrMMD Flow $(\mu_t)_{t \geq 0}$

Continuity Equation

$$\partial_t \mu_t + \nabla \cdot (\mu_t v_t) = 0, \quad v_t = (1+\lambda) \nabla h_{\mu_t, \pi}$$
$$h_{\mu_t,\pi} = (\Sigma_\pi + \lambda)^{-1} (m_\mu - m_\pi).$$

Here, $\Sigma_\pi : \mathcal{H} \mapsto \mathcal{H}$ with $\Sigma_\pi = \mathbb{E}_{X \sim \pi}[k(X,\cdot) \otimes k(X,\cdot)]$ is the kernel covariance operator. $m_\pi = \mathbb{E}_{X \sim \pi}[k(X,\cdot)] \in \mathcal{H}$ is the kernel mean embedding.

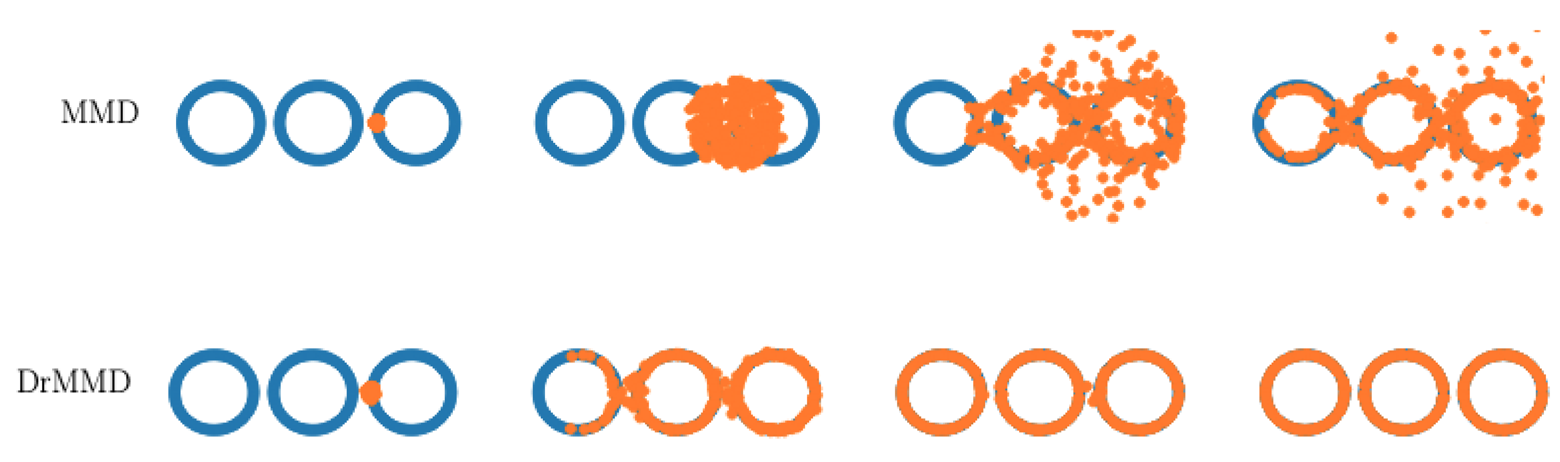
Tractable finite-sample implementation

- Both the covariance operator Σ_π and the embedding m_π admit consistent finite-sample estimators.
- Given empirical distributions $\hat{\mu} = \frac{1}{N} \sum_{i=1}^N x_i$ and $\hat{\pi} = \frac{1}{M} \sum_{i=1}^M y_i$. Given the Gram matrices $K_{xx} \in \mathbb{R}^{N \times N}$, $K_{yy} \in \mathbb{R}^{M \times M}$, $K_{xy} \in \mathbb{R}^{N \times M}$.

$$\begin{aligned} h_{\hat{\mu},\hat{\pi}}(\cdot) = & \frac{1}{N\lambda} k(\cdot, x_{1:N}) 1_N - \frac{1}{M\lambda} k(\cdot, y_{1:M}) 1_M \\ & - \frac{1}{M\lambda} k(\cdot, y_{1:M}) (M\lambda + K_{yy})^{-1} K_{yx} 1_N \\ & + \frac{1}{M\lambda} k(\cdot, y_{1:M}) (M\lambda + K_{yy})^{-1} K_{yy} 1_M. \end{aligned}$$

- Unlike diffusion models or flow matching, the velocity field $\nabla h_{\hat{\mu},\hat{\pi}}$ of DrMMD flow is available in closed form and does not need to be learned.

Empirical Evaluations



Global Convergence

Ass. 1. *$k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is bounded, continuous and c_0-universal. The kernel has bounded first- and second-order derivatives.*

Ass. 2. *$\pi \propto \exp(-V)$ is Poincaré with C_P.*

Continuous-time convergence

Suppose $\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi} - 1 \in \mathrm{Ran}(\mathcal{T}_\pi^r)$ with $r > 0$ with a pre-image q_t, $\|\nabla (\log \pi)^\top \nabla (\frac{d\mu_t}{d\pi})\|_{L^2(\pi)} \leq \mathcal{J}$ and $\|\Delta(\frac{d\mu_t}{d\pi})\|_{L^2(\pi)} \leq \mathcal{I}$. Then,

$$\partial_t \mathrm{KL}(\mu_t \| \pi) \leq -C_P^{-1} \mathrm{KL}(\mu_t \| \pi) + \lambda^r (\mathcal{J} + \mathcal{I}).$$

- Recovers χ^2 flow convergence whe $\lambda = 0$.
- $r > 0$ tells the regularity of DrMMD flow.

Discrete-time convergence

Ass. 3. *$\pi \propto \exp(-V)$ with $\mathbf{H}V \leq \beta$.*

$$\mathrm{KL}(\mu_{n+1} \| \pi) - \mathrm{KL}(\mu_n \| \pi) \leq -C_P^{-1} \chi^2(\mu_n \| \pi) \gamma + \underbrace{\gamma \lambda^r Q(\mathcal{J} + \mathcal{I})}_{\text{Approximation error}} + \underbrace{\gamma^2 \lambda^{-1} \beta \chi^2(\mu_n \| \pi)}_{\text{Discretization error}}.$$

- $\gamma > 0$ is the step size.
- Trade-off between Approximation error and time-discretization error.
- Adaptive regularization $\lambda_n \propto \chi^2(\mu_n \| \pi)^{\frac{1}{r+1}}$
- To reach error $\mathrm{KL}(\mu_n \| \pi) \leq \delta$, it takes $n = \mathcal{O}((\frac{1}{\delta})^{\frac{r+1}{r}} \log \frac{1}{\delta})$ iterations.
- In contrast, Langevin Monte Carlo takes $n = \mathcal{O}((\frac{1}{\delta}) \log \frac{1}{\delta})$ iterations.