Theoretical Insights for GANs using Gradient Flows

Pierre Glaser, Gatsby Computaional Neuroscience Unit, University College London
Joint work with Michael Arbel and Arthur Gretton

Introduction: Divergences in ML and Statistics

Many ML tasks are divergence optimization problems in disguise!

Learning using Maximum Likelihood

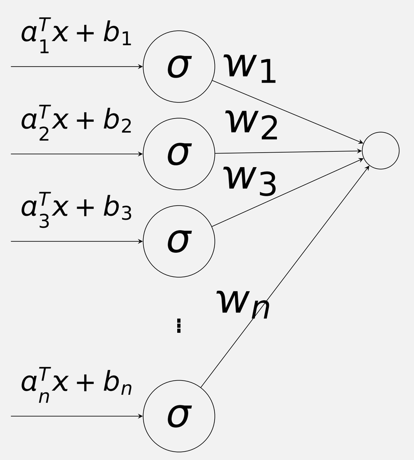
$$\max_\theta \frac{1}{N}\sum_{i=1}^{N} \log p_\theta(X^{(i)}) \underset{N\to\infty}{\to} -\min_\theta \mathrm{KL}(p_\theta \,||\, p) + C$$

Sampling using Langevin Dynamics [1]

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \mathrm{d}W_t, \quad X_0 \sim \mathbb{P}_0$$

Law(X_t) follows the "Gradient descent" trajectory of $\mathrm{KL}(\cdot \,||\, e^{-V(\cdot)}/Z)$ starting from $\mathbb{P}_0$

Nonlinear regression using 2-layer neural networks [2]



- Input-output pair: $(X, y) \sim \rho \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$.
- Predictor class: $f_{\{a_i, b_i, w_i\}}(x) = \frac{1}{N}\sum_{i=1}^{N} w_i \sigma(a_i^\top x + b_i)$, with integral form $f_\mu(x) = \int w\sigma(a^\top x + b)\mathrm{d}\mu(a, b, w)$
- Least-squares objective: $R(f) = \mathbb{E}_\rho \|f(X) - Y\|^2$

Assume the well specified case: $\exists \mu^\star : y = f_{\mu^\star}(X), \forall X$. Then:

$$R(f_\mu) = \mathrm{MMD}(\mu \,||\, \mu^\star), \quad \text{for some RKHS } \mathcal{H}$$

Implicit generative Models (IGMs)

Definition

An IGM passes a *simple* random variable through a complex map:

Initial draw $U \sim \mathcal{N}(0, \sigma^2 I) \longrightarrow$ Final Sample: $X_\theta = f_\theta(U)$

Implicitly defined probability measure: $\mathrm{d}\mathbb{P}_\theta(x) = (f_\theta)_\# \mathrm{d}\mathcal{N}(0, I)(x)$

GAN: Sampling using IGMs

IGMs can be trained to sample from a *unknown* target distribution $\mathbb{P}$ with *known* samples $\{X^{(i)}\}_{i=1}^N$ through the alternate training of:

- A model D_ϕ separating IGM samples $\{X_\theta^{(i)}\}_{i=1}^N$ from $\{X^{(i)}\}_{i=1}^N$ using MLE
- The actual IGM, that should *minimize* D_ϕ's final likelihood

$$\min_\theta J(\theta) := \min_\theta \max_\phi \mathbb{E}_{\{X^{(i)}, T\}_{i=1}^N, \{X_\theta^{(i)}, F\}_{i=1}^N} \log l_\phi(X, y)$$

Theorem [3]: GAN training minimizes the Jensen-Shannon divergence

Assume: $\mathrm{d}\mathbb{P} = p(x)\mathrm{d}x$, $\mathrm{d}\mathbb{P}_\theta(x) = p_\theta(x)\mathrm{d}x$, sufficiently expressive D_ϕ. Then the *population version* of GAN training objective reduces to:

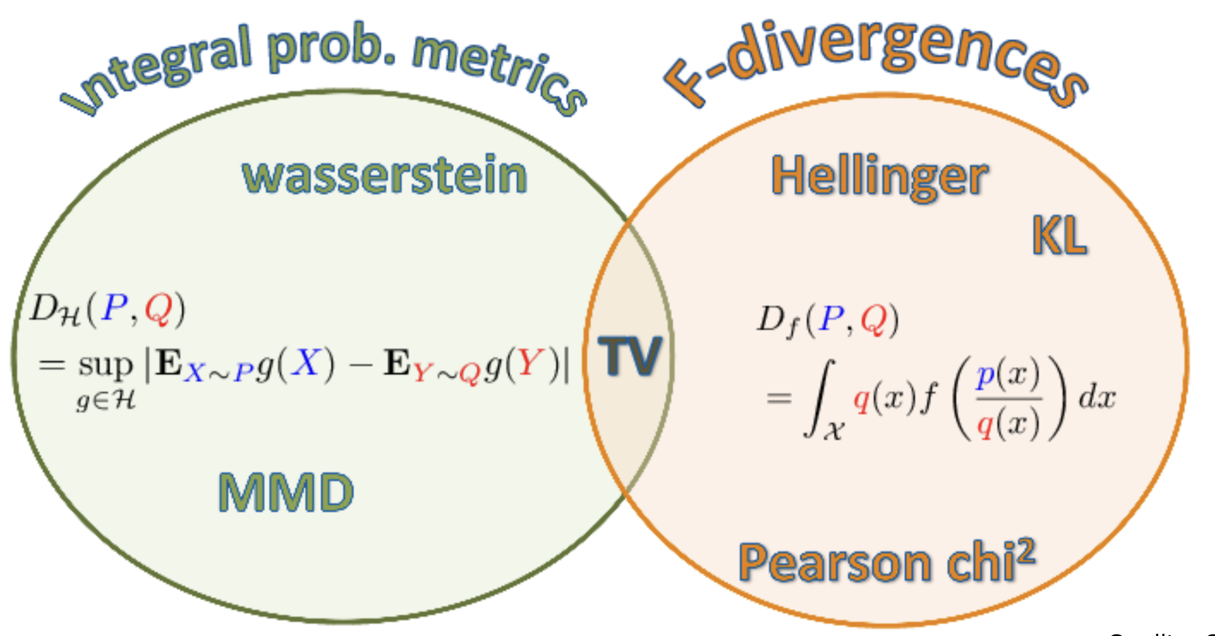
$$J(\theta) = \mathrm{JS}(\mathbb{P} \,||\, \mathbb{P}_\theta)$$

Suggests designing generative models using $J(\theta) = D(\mathbb{P}_\theta \,||\, \mathbb{P})$

→Led to Wasserstein GAN, MMD GAN f-gan... Training often retains a generator/discriminator structure.

Integral Probability Metrics (IPMs) vs f-divergences

IPMs and f-divergences are two classes accounting for many divergences:



Credits: Gretton et. al

Induced Topology	**Weak**: $D_{\mathcal{H}}(\mathbb{P} \,||\, \mathbb{Q}) < +\infty$ for any $\mathbb{P}$, $\mathbb{Q}$, for most $\mathcal{H}$.	**Strong**: If $\mathbb{P} \not\ll \mathbb{Q}$, then $D_f(\mathbb{P} \,||\, \mathbb{Q}) = +\infty$
Variational Formulation	$\sup_{g\in\mathcal{G}} \int g\mathrm{d}\mathbb{P} - \int g\mathrm{d}\mathbb{Q}$	$\sup_{g\in\mathcal{C}_b} \int g\mathrm{d}\mathbb{P} - \int f^*(g)\mathrm{d}\mathbb{Q}$

GAN training as Gradient Flows

- **Gradient Flow** (GF) = continuous-time limit of gradient descent:

$$\frac{d\theta}{dt} = -\nabla_\theta D(\mathbb{P}_\theta \,||\, \mathbb{P}), \text{ given } \theta_0$$

- **Wasserstein**: Geometry in which the GAN training path $t \longrightarrow \mathbb{P}_{\theta_t}$ is continuous. Associated (nonparametric) proximal minimizing dynamics:

$$\frac{\partial \mathbb{P}}{\partial t} + \mathrm{div}(\mathbb{P}\nabla\frac{\delta D}{\delta \mathbb{P}}) = 0, \text{ given } \mathbb{P}_0$$

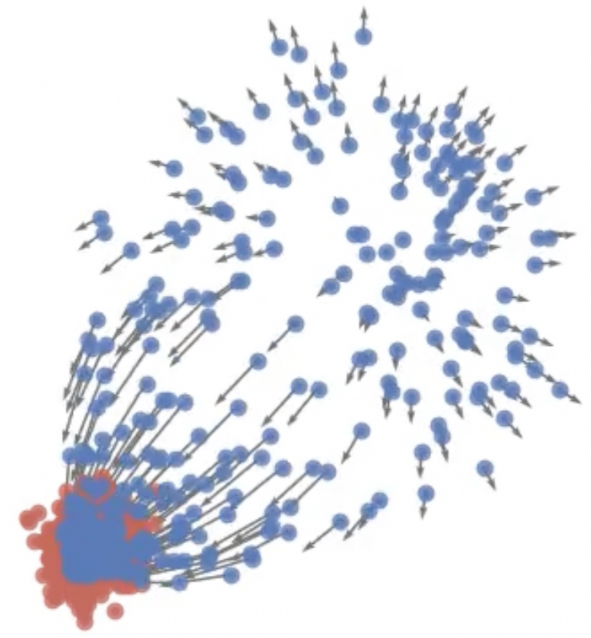


Figure 1: Numerical MMD Flow simulations using samples from $\mathbb{P}_0$ and $\mathbb{Q}$

Research Goals

GAN training using D ≃ Wasserstein Gradient Flow of D

→ Study Wasserstein Gradient Flows as idealized GAN training dynamics to design, improve and guide GAN training algorithms.

References

[1] Richard Jordan, David Kinderlehrer, and Felix Otto. The Variational Formulation of the Fokker–Planck Equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, January 1998.

[2] Lenaic Chizat and Francis Bach. On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport. *arXiv:1805.09545 [cs, math, stat]*, October 2018. arXiv: 1805.09545.

[3] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *arXiv:1406.2661 [cs, stat]*, June 2014. arXiv: 1406.2661.

[4] Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum Mean Discrepancy Gradient Flow. *arXiv:1906.04370 [cs, stat]*, December 2019. arXiv: 1906.04370.

[5] Pierre Glaser, Michael Arbel, and Arthur Gretton. KALE Flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support. *arXiv:2106.08929 [cs, stat]*, June 2021. arXiv: 2106.08929.

Lessons from MMD Flow

- MMD: IPM using $\mathcal{B}(0_{\mathcal{H}}, 1)$ of a RKHS $\mathcal{H}$ as its critic class.
- RKHS functions are very smooth: RKHS-norm convergence implies point-wise convergence, making the MMD sometimes "too weak" to train generative models
- Global Convergence of the MMD Gradient Flow can be ensured using "Noise Injection" [4]

$$Z_{t+1}^{(i)} = Z_t^{(i)} - \gamma\nabla f_{\mathbb{P}_t, \mathbb{Q}}(X_t^{(i)})$$
$$\downarrow$$
$$Z_{t+1}^{(i)} = Z_t^{(i)} - \gamma\nabla f_{\mathbb{P}_t, \mathbb{Q}}(X_t^{(i)} + \beta_t U_t^{(i)}),\ U_i \sim \mathcal{N}(0, 1)$$

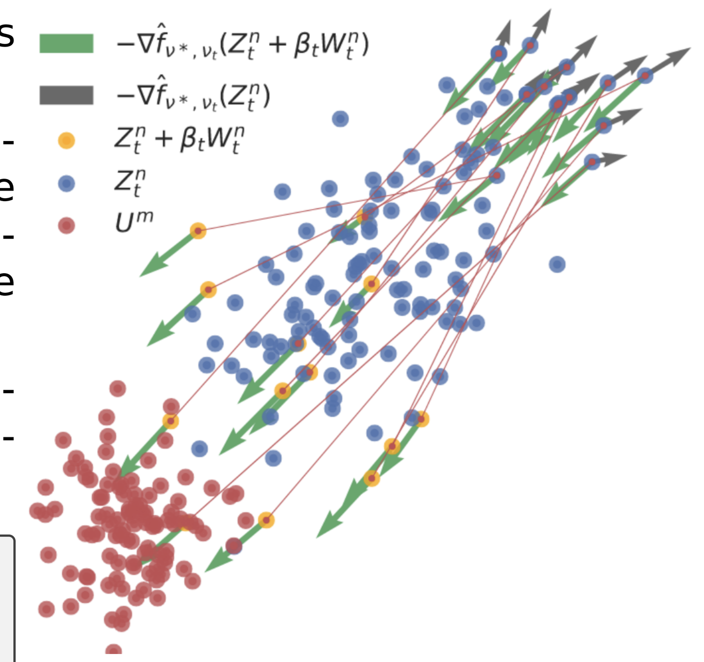


Figure 2: Noise injection in MMD Flow

Theorem (Informal)

Noise-injection $\iff$ using a time-dependent kernel $k_t = k \star \mathcal{N}(0, \beta_t^2)$

Insights for MMD Gans

- Adaptive kernel width can ensure global MMD flow convergence!
- The best noise schedules maximize the signal sent by the target
- $\implies$ Can serve as a regularization criterion for MMD GANs!

RKHS smoothing of f-divergences (KALE)

- f-divergences are highly sensitive to support $\mathbb{P} \not\ll \mathbb{Q} \implies D_f(\mathbb{P} \,||\, \mathbb{Q}) = +\infty$
- Idea get a ***smoothed* support sensitivity** by **kernelizing** f-divergences:

$$D_{f,\mathcal{H}}(\mathbb{P} \,||\, \mathbb{Q}) = \sup_{f\in\mathcal{H}} \int \mathrm{d}\mathbb{P} - \int f^*(h)\mathrm{d}\mathbb{Q} - \frac{\lambda}{2}\|h\|^2$$

Example: $\mathrm{KALE}(\mathbb{P} \,||\, \mathbb{Q}) = 1 + \sup_{h\in\mathcal{H}} \int h\mathrm{d}\mathbb{P} - \int e^h \mathrm{d}\mathbb{Q} + \frac{\lambda}{2}\|h\|_{\mathcal{H}}^2$. Can be used in Generative Models!

Theorem: KALE interpolates between KL and MMD [5]

Assume $\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q} \in \mathcal{H}$. Then:

$$\mathrm{KALE}(\mathbb{P} \,||\, \mathbb{Q}) \underset{\lambda\to 0}{\to} \mathrm{KL}(\mathbb{P} \,||\, \mathbb{Q}) \quad \text{and} \quad \mathrm{KALE}(\mathbb{P} \,||\, \mathbb{Q}) \underset{\lambda\to\infty}{\to} \mathrm{MMD}(\mathbb{P} \,||\, \mathbb{Q})$$

→ λ makes KALE *interpolate* between the KL and the MMD

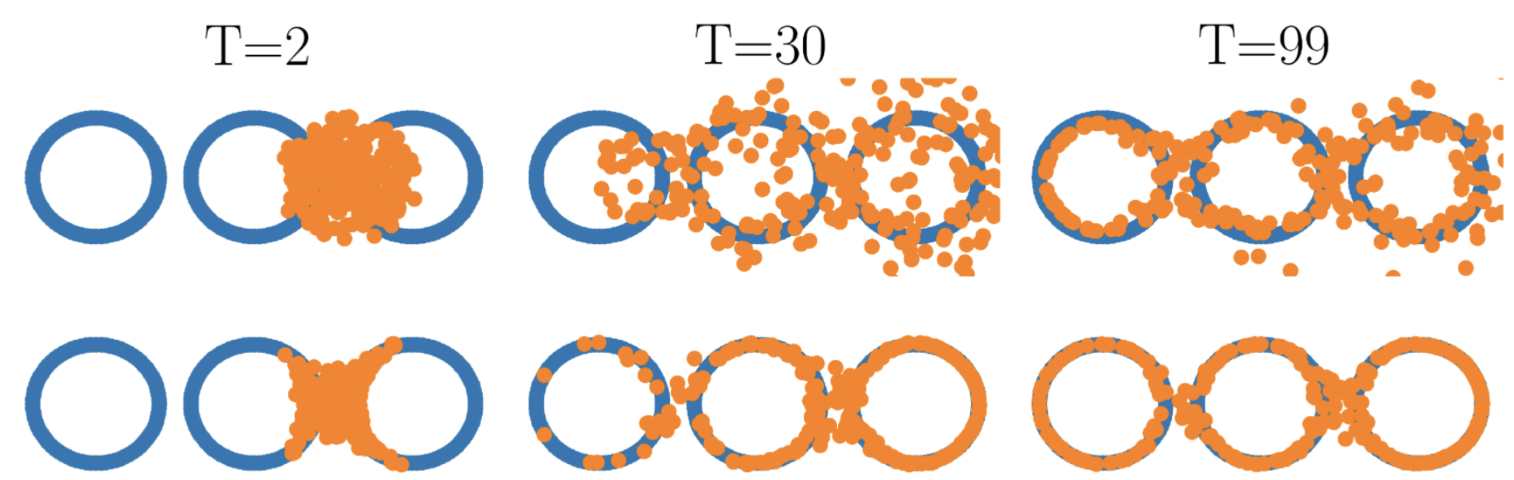


Figure 3: MMD Flow (top), KALE flow (bottom): KALE exhibits a higher sensitivity to disjoint supports, leading to better behaved flows