HAL open science

Unifying GANs and Score-Based Diffusion as Generative Particle Models

Jean-Yves Franceschi, Mike Gartrell, Ludovic Dos Santos, Thibaut Issenhuth, Emmanuel de Bézenac, Mickaël Chen, Alain Rakotomamonjy

To cite this version:

Jean-Yves Franceschi, Mike Gartrell, Ludovic Dos Santos, Thibaut Issenhuth, Emmanuel de Bézenac, et al.. Unifying GANs and Score-Based Diffusion as Generative Particle Models. Alice Oh; Tristan Naumann; Amir Globerson; Kate Saenko; Moritz Hardt; Sergey Levine. 37th Conference on Neural Information Processing Systems, Dec 2023, New Orleans, United States. Curran Associates, Inc., Advances in Neural Information Processing Systems, 36, 2023. hal-04322365

HAL Id: hal-04322365

https://hal.science/hal-04322365v1

Submitted on 4 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

Unifying GANs and Score-Based Diffusion as Generative Particle Models

J.-Y. Franceschi,[1] M. Gartrell,[1] L. Dos Santos,[1,*] T. Issenhuth,[1,2,*] E. de Bézenac,[3,*] M. Chen,[4,*] A. Rakotomamonjy[1,*]

[1]Criteo AI Lab [2]LIGM, Ecole des Ponts, Univ Gustave Eiffel, CNRS [3]SAM, D-MATH, ETH Zürich [4]Valeo.ai

Our Contributions

- We unify gradient flows, score-based diffusion models, and GANs in a single framework.
- We represent generated data as moving particles. A model is defined by:
  - a gradient vector field that the particles follow either at inference or training time;
  - the possibility of incorporating a generator that smooths this movement.
- This suggests the existence of hybrid models:
  - a generator trained with diffusion guidance (Score GANs);
  - a GAN trained without a generator (Discriminator Flows).
- We experimentally verify our findings.

GANs vs Diffusion

Traditional opposition in the literature.

GANs → Generator trained by discriminating true vs fake data.
- Generator (manifold learning).
- Close to SOTA performance.
- Harder to optimize.
- Fast inference.

Diffusion → Learns to progressively reverse a data degradation process.
- No generator (on the data space).
- SOTA performance.
- Easier to optimize.
- Slow inference.

From PMs to Int-PMs

- We assign to each generated particle $x = g_\theta(z)$ the same loss as in PMs: $\mathcal{L}_{\text{gen}}(\theta) = -\mathbb{E}_{z\sim p_z}\Big[h_{\rho_t}\big(g_\theta(z)\big)\Big]$.
- We do not take into account the dependency of ρ_t w.r.t. θ_t, to mimic PMs: $\rho = \mathrm{StopGradient}(g_\theta \sharp p_z)$.
- Continuous-time gradient descent:
$$\frac{\mathrm{d}\theta_t}{\mathrm{d}t} = -\nabla_{\theta_t}\mathcal{L}_{\text{gen}}(\theta_t) = \eta\nabla_{\theta_t}\mathbb{E}_{z\sim p_z}\Big[h_{\rho_t}\big(g_{\theta_t}(z)\big)\Big] = \eta\mathbb{E}_{z\sim p_z}\Big[\nabla_{\theta_t}g_{\theta_t}(z)\nabla h_{\rho_t}\big(g_{\theta_t}(z)\big)\Big].$$
- Evolution of particles:
$$\frac{\mathrm{d}g_{\theta_t}(z)}{\mathrm{d}t} = \nabla_{\theta_t}g_{\theta_t}(z)^\top\frac{\mathrm{d}\theta_t}{\mathrm{d}t} = \eta\mathbb{E}_{z'\sim p_z}\Big[\nabla_{\theta_t}g_{\theta_t}(z)^\top\nabla_{\theta_t}g_{\theta_t}(z')\nabla h_{\rho_t}\big(g_{\theta_t}(z')\big)\Big].$$

Particle-Based Framework

Generated particles $x_t \sim \rho_t$ follow a gradient vector field ∇h_{ρ_t}, i.e. optimize an objective h_{ρ_t}.

Wasserstein Gradient

$$-\nabla_W\mathcal{F}(\rho_t) = -\nabla\frac{\partial\mathcal{F}(\rho_t)}{\partial\rho_t}$$

Log Ratio Gradient

$$\alpha_t\nabla\log\Big[p_{\text{data}} \star k^{\sigma(t)}_{\text{RBF}}\Big] - \beta_t\nabla\log\rho_t$$

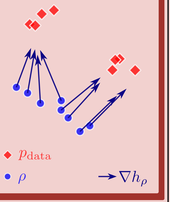
Discriminator Gradient

$$-\nabla\big(c\circ f_{\rho_t}\big)$$

where f_{ρ_t} discriminates ρ_t from p_{data}

Particle Models (No Generator)

- At generation / inference time t:
$$x_0 \sim \pi = \rho_0,\ \mathrm{d}x_t = \nabla h_{\rho_t}(x_t)\,\mathrm{d}t.$$
- Independently moving particles.
- Each x_t individually follows a **gradient ascent** path on $h_{\rho_t}(x_t)$.
- h_ρ is usually a predefined functional approximated with neural networks.

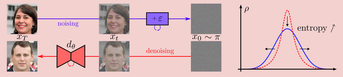


Wasserstein Gradient Flows

- Gradient descent for functionals over distributions $\mathcal{F}$ (Santambrogio, 2017).

Objective $\mathcal{F}(\rho)$		h_ρ
Forward KL	$\mathbb{E}_\rho \log \rho/p_{\text{data}}$	$-\log \rho/p_{\text{data}}$
f-divergence	$\mathbb{E}_{p_{\text{data}}} f\big(\rho/p_{\text{data}}\big)$	$-f'\big(\rho/p_{\text{data}}\big)$
Squared MMD w.r.t. kernel k	$\mathbb{E}_{\substack{x,x'\sim\rho\\ y,y'\sim p_{\text{data}}}}\begin{bmatrix}k(x,x')\\+k(y,y')\\-2k(x,y)\end{bmatrix}$	$\mathbb{E}_{y\sim p_{\text{data}}}\big[k(y,\cdot)\big] - \mathbb{E}_{x\sim\rho}\big[k(x,\cdot)\big]$
Entropy	$\mathbb{E}_\rho \log\rho$	$-\log\rho$

Score-Based Diffusion



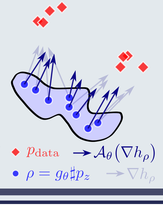
- Using Jordan et al. (1998) in Song et al. (2019) and Karras et al. (2022):
$$\mathrm{d}x_t = \alpha_t\nabla\log\Big[p_{\text{data}} \star k^{\sigma(t)}_{\text{RBF}}\Big](x_t)\,\mathrm{d}t + \sqrt{2\beta_t}\,\mathrm{d}W_t \quad\Leftrightarrow\quad -\beta_t\nabla\log\rho_t\,\mathrm{d}t.$$

Discriminator Flows

- Particles directly follow the discriminator gradient.
- The discriminator is simultaneously trained and used to generate data.

Interacting Particle Models (Generator)

- Training with **the same loss**:
$$\mathcal{L}_{\text{gen}}(\theta) = -\mathbb{E}_{z\sim p_z}\Big[h_{\rho_t}\big(g_\theta(z)\big)\Big].$$
- At training time t:
$$\mathrm{d}g_{\theta_t}(z) = \eta\big[\mathcal{A}_{\theta_t}(z)\big]\big(\nabla h_{\rho_t}\big)\,\mathrm{d}t.$$
- Generalization of PMs where particles interact with each other.



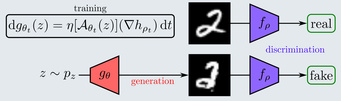
Stein Gradient Flows

- Stein gradient flows (Liu, 2017) are kernelized Wasserstein gradient flows:
$$\mathrm{d}x_t = \mathbb{E}_{x'_t\sim\rho_t}\Big[k\big(x_t,x'_t\big)\nabla h_{\rho_t}\big(x'_t\big)\Big]\,\mathrm{d}t.$$
- Int-PMs under mild hypotheses (generalization of Durr et al. (2022)).
- Hint towards the same h_ρ being used in a PM and an Int-PM.

Score GANs

- Estimate ∇h_ρ with two score matching networks and use it in parameter update equation of generator training.
- Data score (pretrained like diffusion): $s^{p_{\text{data}}}_\psi(\cdot,\sigma) \equiv \nabla\log\big[p_{\text{data}} \star k^\sigma_{\text{RBF}}\big]$.
- Gen. score (continuously updated like a discriminator): $s^\rho_\phi \equiv \nabla\log\rho$.

GANs



- Gradient descent-ascent on the min-max objective yields the generator loss:
$$\mathcal{L}_{\text{GAN}}(g_\theta) = \mathbb{E}_{z\sim p_z}\Big[\big(c\circ f_\rho\big)\big(g_\theta(z)\big)\Big].$$

Smoothing Operator

- $\mathcal{A}_{\theta_t}(z)$ is a linear operator on vector fields (kernel integral operator):
$$\big[\mathcal{A}_{\theta_t}(z)\big](V) \triangleq \mathbb{E}_{z'\sim p_z}\Big[k_{g_{\theta_t}}(z,z')V\big(g_{\theta_t}(z')\big)\Big],\ k_{g_{\theta_t}}(z,z') \triangleq \nabla_{\theta_t}g_{\theta_t}(z')^\top\nabla_{\theta_t}g_{\theta_t}(z).$$
- $k_{g_{\theta_t}}$ is the generator's matrix Neural Tangent Kernel (NTK, Jacot et al., 2018).
- Special case: $k_{g_{\theta_t}}(z,z') = \delta_{z=z'}I_d$ (generator with infinite capacity).
  - No interaction between particles: $\big[\mathcal{A}_{\theta_t}(z)\big](V) = V\big(g_\theta(z)\big)$.
  - $\mathrm{d}g_{\theta_t}(z) = \nabla h_{\rho_t}\big(g_{\theta_t}(z)\big)\,\mathrm{d}t$: we retrieve PMs.
- General case: $\mathcal{A}_{\theta_t}$ represents the parameterization of ρ as a manifold.
  - $\mathcal{A}_{\theta_t}$ smooths the original vector field ∇h_{ρ_t} by convolving it with k.
  - Particles interact with each other through generator parameterization.

Other Models & Flows

- Int-PMs and Stein (generalization of Durr et al. (2022)): $k\Big(g_{\theta_t}(z), g_{\theta_t}(z')\Big) = k_{g_{\theta_t}}(z,z')$ in the NTK regime.
- Langevin diffusion (Song et al., 2019) is a KL flow.
- Under some hypotheses, GANs are Stein flows (Franceschi et al., 2022; Yi et al., 2023): KL flow for f-divergence GANs, squared MMD for IPM GANs.
- As a consequence, under similar hypotheses, Discriminator Flows with the same losses are Wasserstein flows.
- Many methods use neural networks to approximate the flow (Alvarez-Melis et al., 2022; Heng et al., 2023).

Score GANs in Practice

- Two practical issues:
  - sliced score matching to train s^ρ_ϕ;
  - scheduling σs w.r.t. training time t.
- We randomly sample σ and also noise the particles:
$$\nabla h_\rho = \nabla\log[p_{\text{data}} \star k^\sigma_{\text{RBF}}] - \nabla\log[\rho_t \star k^\sigma_{\text{RBF}}],$$
$$\equiv \widetilde{\nabla h}_\rho(\cdot,\sigma) = s^{p_{\text{data}}}_\psi(\cdot,\sigma) - s^\rho_\phi(\cdot,\sigma).$$
- Generator update:
  - few-step training of s^ρ_ϕ with denoising score matching;
  - gradient descent step:
$$\theta \leftarrow \theta + \eta\mathop{\mathbb{E}}_{\sigma\sim p_\sigma,\varepsilon\sim\mathcal{N}(0,\sigma I_D),z\sim p_z}\Big[\nabla_\theta g_\theta(z)\widetilde{\nabla h}_\rho\big(g_\theta(z)+\varepsilon,\sigma\big)\Big].$$

Discr. Flows in Practice

- Discriminator loss:
$$\mathcal{L}_{\text{d}}(f;\rho,p_{\text{data}}) = \mathbb{E}_\rho[a\circ f] - \mathbb{E}_{p_{\text{data}}}[b\circ f] + \mathcal{R}(f;\rho,p_{\text{data}}).$$
- Naive training: successive f_{ρ_t} trainings and ρ_t updates.
- For efficiency purposes, we simultaneously learn all time-parameterized discriminators: $f_{\rho_t} = f_\phi(\cdot,t)$.
- Training step:
  - sample $t \sim \mathcal{U}\big([0,1]\big)$, $x_0 \sim \pi$;
  - compute $x_t = -\eta\int_0^t\nabla\big(c\circ f_\phi(\cdot,s)\big)(x_s)\,\mathrm{d}s$;
  - train $f_\phi(\cdot,t)$ to discriminate between x_t and p_{data}.
- Generalization of some gradient flows.

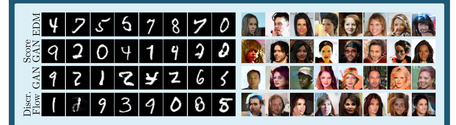
Links

Article

Code and samples

- Code for all models and baselines.
- Animated samples.

Experimental Results



Dataset	PMs (no generator)		Int-PMs (generator)	
	EDM	Discr. Flow	GAN	Score GAN
MNIST	3	4	3	15
CelebA	10	41	19	35

- Hybrid models are viable, support theory.
- EDM: diffusion (Karras et al., 2022).

Properties

- **Discr. flows** learn a path to the data distribution, unlike **diffusion**.
- **PMs vs Int-PMs**: Int-PMs are prone to mode collapse but are faster than PMs at inference and have better latent space properties.

Perspectives

- Our work paves the way for new hybrid models.
- Model improvements: Score GANs for score distillation, Discr. Flows for generation efficiency.
- Framework improvements: convergence, second-order and discrete-time optimization, more accurate GAN modeling.

Thirty-seventh Conference on Neural Information Processing Systems — NeurIPS 2023, New Orleans, LA, USA