Score-of-Mixture Training

One-Step Generative Model Training Made Simple via Score Estimation of Mixture Distributions

Tejas Jayashankar*, Jongha (Jon) Ryu*, Gregory W. Wornell

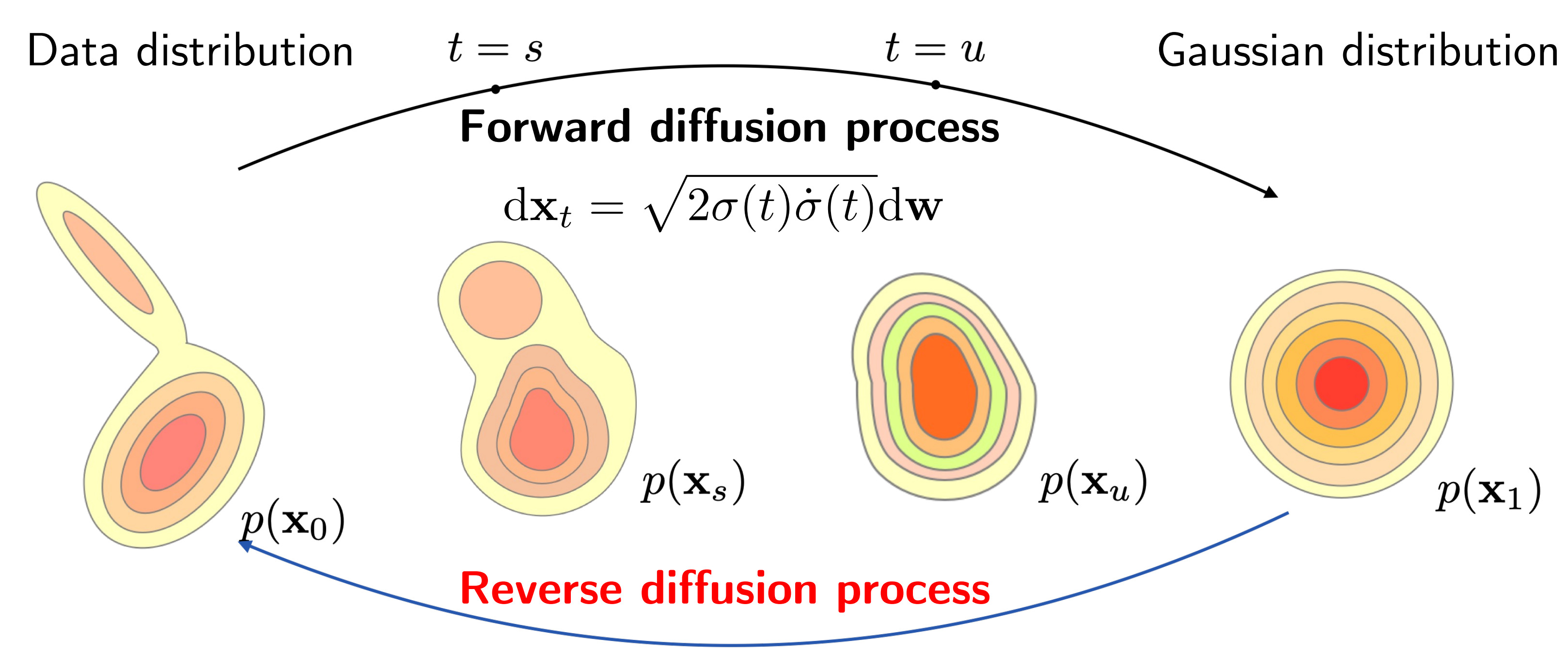
MIT EECS | {tejasj,jongha,gww}@mit.edu

SCAN ME!

ICML International Conference On Machine Learning

TL;DR: Stable training from scratch, SOTA samples via one-step sampling (cf. GAN, diffusion distillation, consistency training)

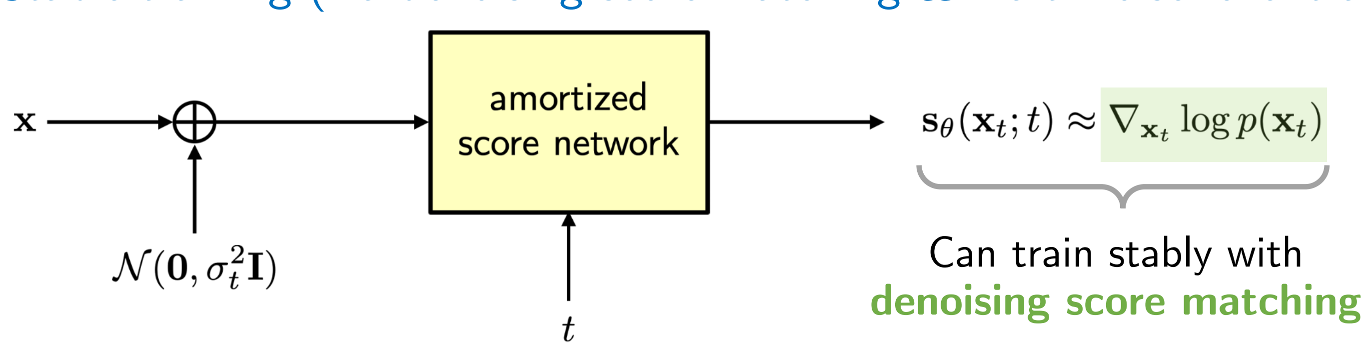
Preliminaries: Diffusion Models



$$\mathrm{d}\mathbf{x}_t = -2\sigma(t)\dot{\sigma}(t)\underbrace{\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t)}_{\text{"noisy scores"}}\mathrm{d}t + \sqrt{2\sigma(t)\dot{\sigma}(t)}\mathrm{d}\mathbf{w}$$

Idea: Learn noisy scores & emulate reverse process

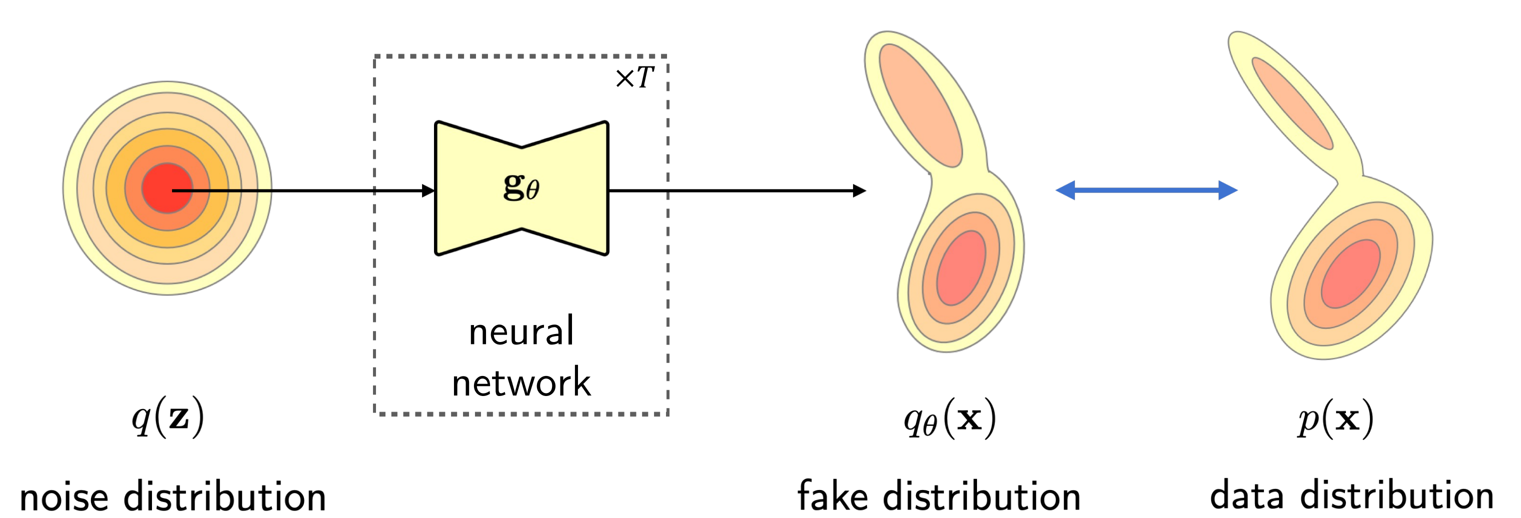
(+) Stable training (via denoising score matching & multi-noise-level training)



(+) High-quality samples (noisy scores can be estimated accurately)

(−) Emulating reverse process is slow and expensive

Goal: One-Step Generative Modeling



We want $q_\theta \approx p$ with $T = 1$

(cf. T ≈ (a few hundreds) for diffusion models)

Solutions for One-Step Generative Modeling

	Learning principle	Generation complexity	Training dynamics	Requires pretrained model?
Diffusion model	Denoising score matching	$T > 1$	Stable	No
Diffusion distillation	Reverse KLD minimization	$T \geq 1$	Stable	Yes
GAN	Minimize JSD w/ discriminator	$T = 1$	Unstable	No
Consistency training **Consistency distillation**	Emulate PF ODE paths	$T \geq 1$	Unstable Stable	No Yes
SMT (from scratch) **SMD (distillation)**	Minimize α-JSD w/ score of mixture	$T = 1$	Stable	No Yes

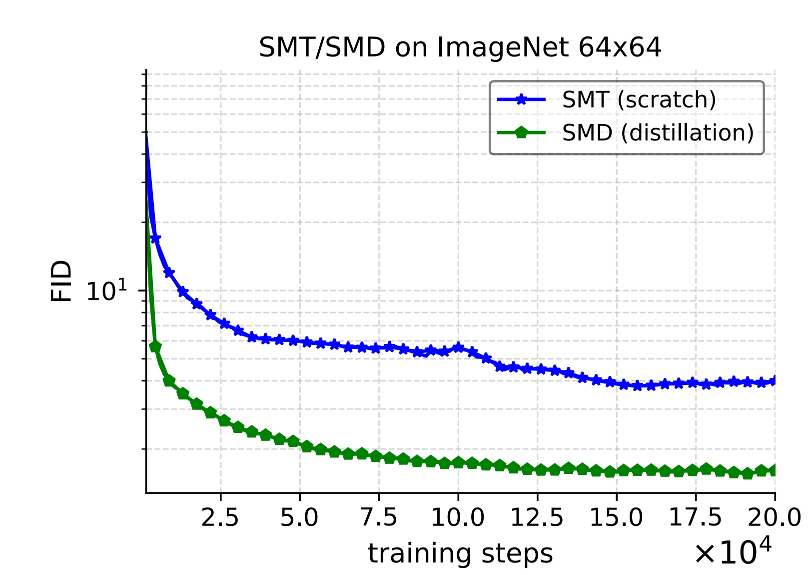
We can achieve the BEST of ALL WORLDS!

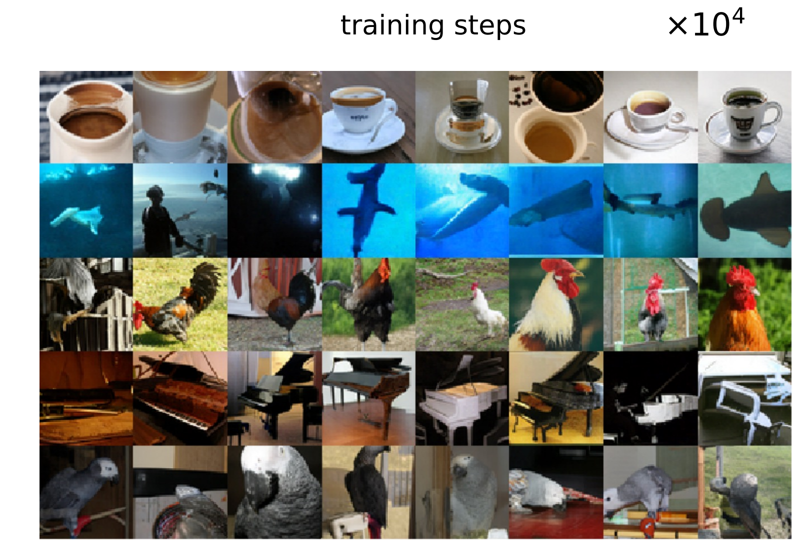
Idea 1. Minimizing New Statistical Divergences

Idea 2. Gradient Update with Score of Mixture Estimation

Experiments

Method	ImageNet 64x64 # params	ImageNet 64x64 NFE	ImageNet 64x64 FID↓	CIFAR-10 32x32 # params	CIFAR-10 32x32 NFE	CIFAR-10 32x32 FID↓
Training from scratch: Diffusion models						
DDPM (Ho et al., 2020)	-	-	-	56M	1000	3.17
ADM (Dhariwal & Nichol, 2021)	296M	250	2.07	-	-	-
EDM (Karras et al., 2022b)	296M	512	**1.36**	56M	35	**1.97**
Training from scratch: One-step models						
CT (Song et al., 2023)	296M	1	13.0	56M	1	8.70
iCT (Song & Dhariwal, 2024a)	296M	1	4.02	56M	1	2.83
iCT-deep (Song & Dhariwal, 2024a)	592M	1	3.25	112M	1	**2.51**
ECT (Geng et al., 2024)	280M	1	5.51	56M	1	3.60
SMT (ours)	296M	1	**3.23**	56M	1	3.13
Diffusion distillation						
PD (Salimans & Ho, 2022)	296M	1	10.7	60M	1	9.12
TRACT (Berthelot et al., 2023)	296M	1	7.43	56M	1	3.78
CD (LPIPS) (Song et al., 2023)	296M	1	6.20	56M	1	4.53
Diff-Instruct (Luo et al., 2024a)	296M	1	5.57	56M	1	4.53
MultiStep-CD (Heek et al., 2024)	1200M	1	3.20	-	-	-
DMD w/o reg (Yin et al., 2024b)	296M	1	5.60	56M	1	5.58
DMD2 w/ GAN (Yin et al., 2024a)	296M	1	1.51	56M	1	2.43
MMD (Salimans et al., 2024)	400M	1	3.00	-	-	-
SiD (Zhou et al., 2024)	296M	1	1.52	56M	1	**1.92**
SiM (Luo et al., 2024b)	-	-	-	56M	1	2.02
SMD (ours)	296M	1	**1.48**	56M	1	2.22
w/ expensive regularizer or finetuning						
CTM (Kim et al., 2024)	296M	1	1.92	56M	1	**1.98**
DMD w/ reg (Yin et al., 2024b)	296M	1	2.62	56M	1	2.66
DMD2 (finetuned) (Yin et al., 2024a)	296M	1	**1.23**	-	-	-





Samples from SMT on ImageNet64x64. (Unique class per row.)

SOTA for ImageNet 64x64, competitive result for CIFAR-10

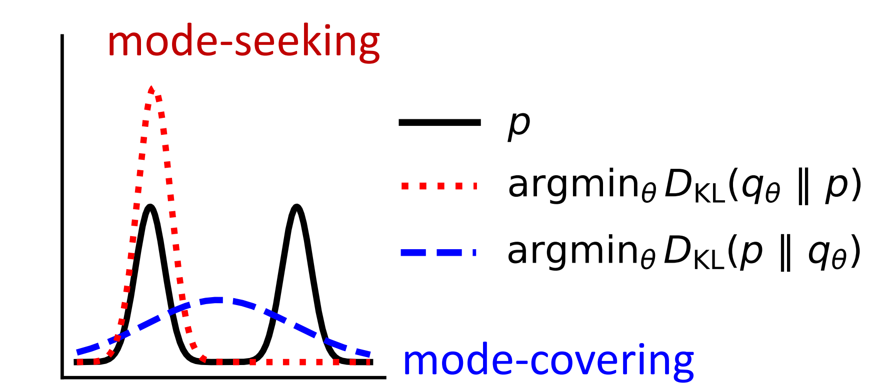
(Bonus: we can also support *distillation*!)

Idea 1 (for Generator Training Objective): Minimizing New Statistical Divergences

New statistical divergence: α-skew Jensen-Shannon divergence ($\alpha \in [0,1]$)

$$D_{\mathrm{JS}}^{(\alpha)}(p,q_\theta) \triangleq \frac{1}{1-\alpha}D_{\mathrm{KL}}\Big(p \,\Big\|\, \alpha p + (1-\alpha)q_\theta\Big) + \frac{1}{\alpha}D_{\mathrm{KL}}\Big(q_\theta \,\Big\|\, \alpha p + (1-\alpha)q_\theta\Big)$$

- Well-defined even for non-overlapping supports
- Interpolating
  - $\alpha = 0$: $D_{\mathrm{KL}}(q_\theta \,||\, p)$ (reverse KL)
  - $\alpha = \frac{1}{2}$: $D_{\mathrm{JS}}(p, q_\theta)$ (Jensen-Shannon)
  - $\alpha = 1$: $D_{\mathrm{KL}}(p \,||\, q_\theta)$ (forward KL)
- Different α enforce different support matching property



Our generator objective function:

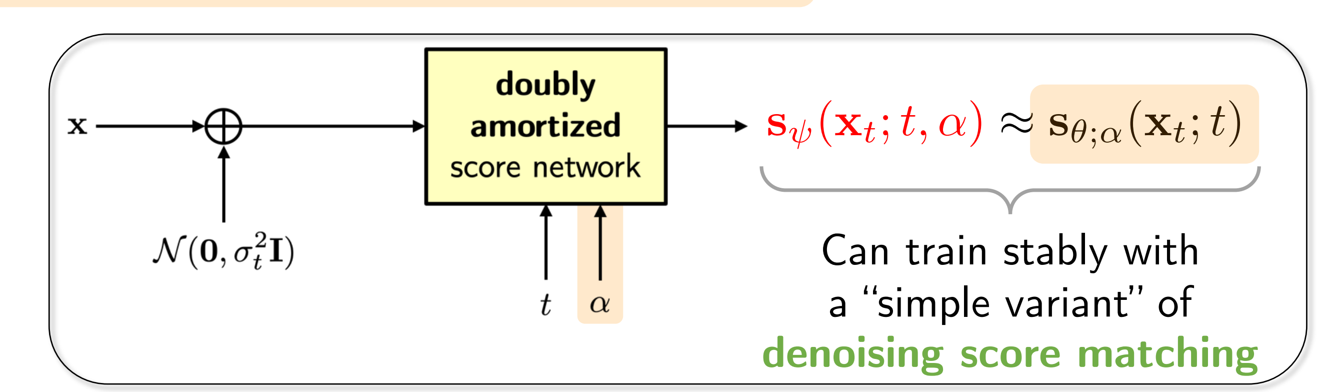
$$\boxed{\mathcal{L}_{\mathrm{gen}}(\theta) \triangleq \mathbb{E}_{p(\alpha)p(t)}\Big[D_{\mathrm{JS}}^{(\alpha)}\Big(p * \mathcal{N}(\mathbf{0},\sigma_t^2\mathbf{I}_D), q_\theta * \mathcal{N}(\mathbf{0},\sigma_t^2\mathbf{I}_D)\Big)\Big]}$$

Idea 2 (for Generator Gradient Estimation): Gradient Estimate with Score of Mixture Distribution

$$\nabla_\theta D_{\mathrm{JS}}^{(\alpha)}(p,q_\theta) = \frac{1}{\alpha}\mathbb{E}_{q(\mathbf{z})}\Big[\nabla_\theta \underbrace{\mathbf{g}_\theta(\mathbf{z})}_{\text{generator}}\Big(\mathbf{s}_{\theta;0}(\mathbf{x}) - \mathbf{s}_{\theta;\alpha}(\mathbf{x})\Big)\Big|_{\mathbf{x}=\mathbf{g}_\theta(\mathbf{z})}\Big]$$

$$\mathbf{s}_{\theta;\alpha}(\mathbf{x}) \triangleq \nabla_{\mathbf{x}}\log(\alpha p(\mathbf{x}) + (1-\alpha)q_\theta(\mathbf{x}))$$

score of α-mixture distribution



Our Proposal: Score-of-Mixture Training (SMT)

- Generator ($\mathbf{g}_\theta$) update with

freeze

$$\nabla_\theta\mathcal{L}_{\mathrm{gen}}(\theta) \approx \mathbb{E}_{p(\alpha)p(t)q(\mathbf{z})q(\boldsymbol{\epsilon})}\Big[\frac{1}{\alpha}\nabla_\theta\mathbf{g}_\theta(\mathbf{z})\Big(\mathbf{s}_\psi(\mathbf{x}_t;t,\alpha) - \mathbf{s}_\psi(\mathbf{x}_t;t,\alpha)\Big)\Big|_{\mathbf{x}_t=\mathbf{g}_\theta(\mathbf{z})+\sigma_t\boldsymbol{\epsilon}}\Big]$$

- Amortized score model ($\mathbf{s}_\psi(\mathbf{x}_t;t,\alpha)$) update with

freeze

$$\nabla_\psi\mathbb{E}_{p(\alpha)p(t)}\Big[\mathcal{L}_{\mathrm{score}}(\psi,\theta;t,\alpha)\Big]$$