Statistical Distances and Optimization in Generative Models

Donghwan Kim

KAIST

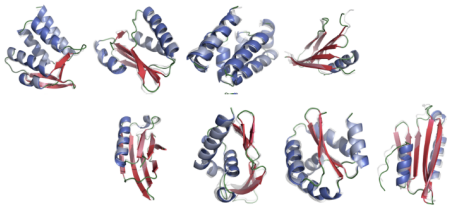
KIAS Center for AI and Natural Sciences 2026 Winter Workshop

Jan 6, 2026

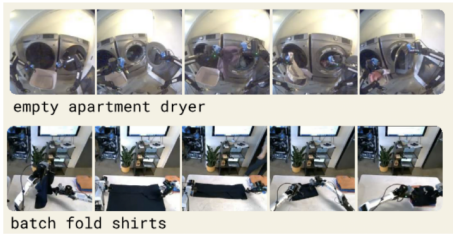
Generative Models

Text-2-Video
MovieGen, Meta

Text-2-Image
Stable Diffusion 3



Protein Generation
Huguet el a. 24



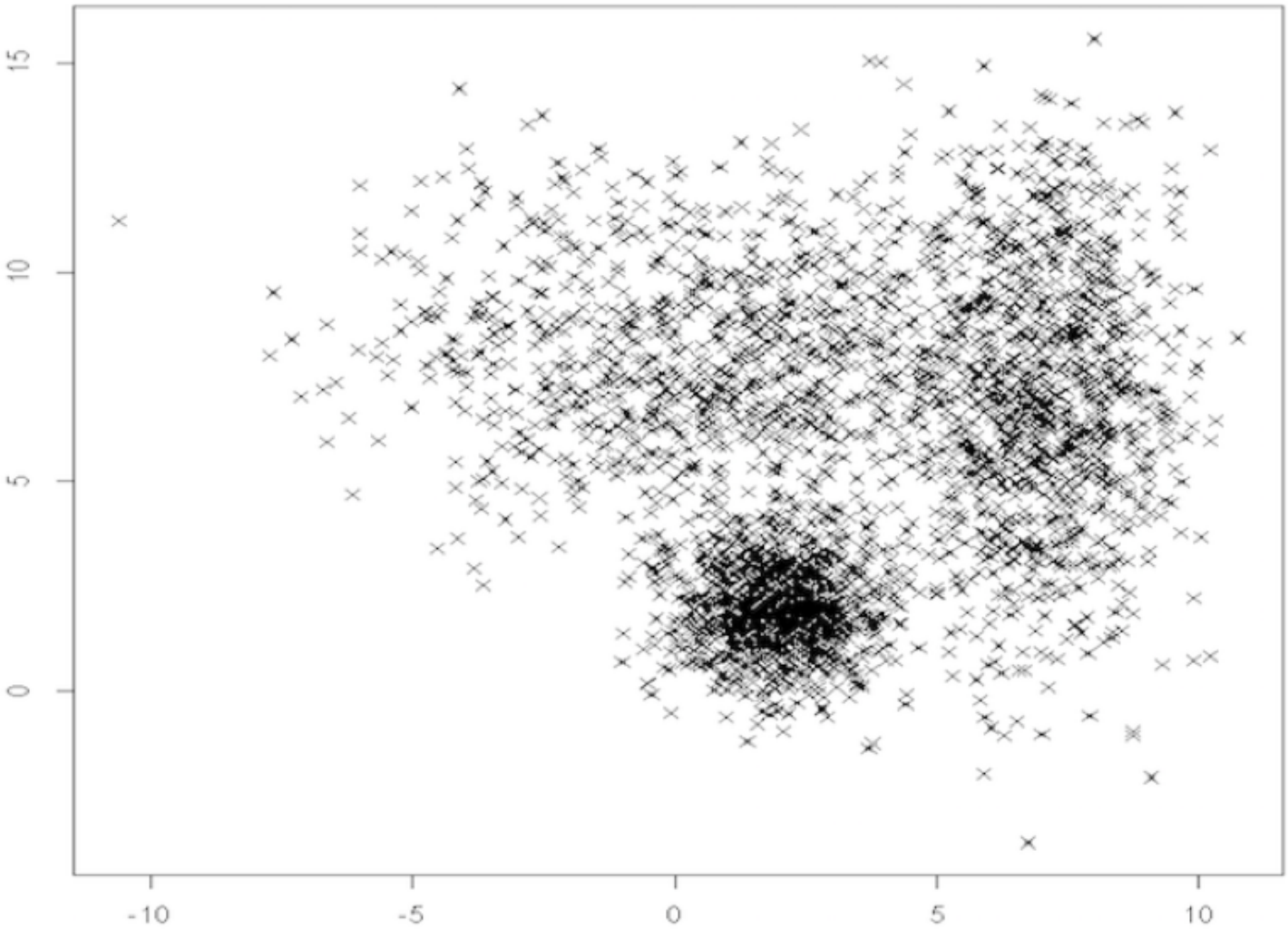
Robot Action Model
Black et a. 24

- Let ν denote the *target* distribution and $\{x_i\}_{i=1}^n$ be its n i.i.d. samples.
- Let μ_θ denote the *generated* distribution parameterized by θ.

- **Goal:** train θ, using $\{x_i\}_{i=1}^n$, so that $\nu \approx \mu_\theta$.

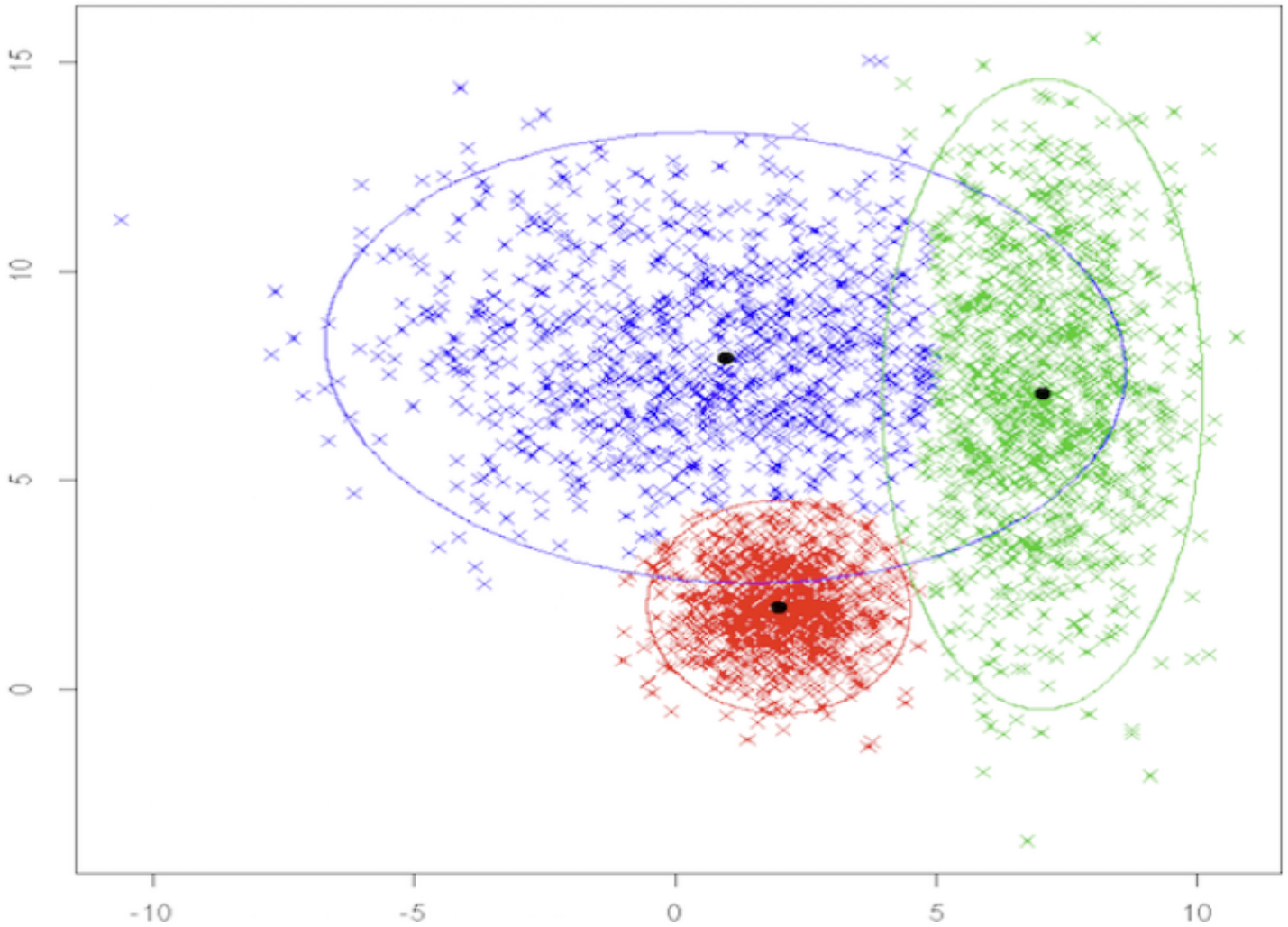
Statistical Distance and Optimization!

Density Fitting



- Density fitting with a Gaussian mixture model.

Density Fitting



- Density fitting with a Gaussian mixture model.

Density Fitting: Maximum Likelihood Estimation

- Maximum Likelihood Estimation (MLE)

$$\min_{\theta} -\frac{1}{n}\sum_{i=1}^{n} \log \mu_\theta(x_i)$$

$\{x_i\}_{i=1}^n$: n i.i.d. samples of ν

μ_θ: generated distribution

- Gaussian mixture model:

$$\mu_\theta = \sum_{i=1}^{K} w_i \mathcal{N}(\mu_i, \Sigma_i)$$

$$\theta = \{(w_i, \mu_i, \Sigma_i)\}_{i=1}^K$$

Density Fitting: KL Divergence Minimization

- Kullback-Leibler (KL) Divergence between ν and μ_θ

$$\begin{aligned}\mathcal{D}_{\mathrm{KL}}(\nu\|\mu_\theta) &:= \int \nu(x)\log\frac{\nu(x)}{\mu_\theta(x)}dx \\ &= -\int \nu(x)\log\mu_\theta(x)dx + \mathrm{Const.}\end{aligned}$$

- In practice, ν is not known, and only the empirical distribution is given

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

Density Fitting: KL Divergence Minimization

- Kullback-Leibler (KL) Divergence between ν and μ_θ

$$\begin{aligned}\mathcal{D}_{\mathrm{KL}}(\nu\|\mu_\theta) &:= \int \nu(x)\log\frac{\nu(x)}{\mu_\theta(x)}dx \\ &= -\int \nu(x)\log\mu_\theta(x)dx + \mathrm{Const.}\end{aligned}$$

- In practice, ν is not known, and only the empirical distribution is given

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

Density Fitting: Empirical KL Divergence Minimization

- *Empirical* KL Divergence

$$\begin{aligned}\mathcal{D}_{\mathrm{KL}}(\hat{\nu}\|\mu_\theta) &= -\int \hat{\nu}(x)\log\mu_\theta(x)dx + \text{Const.} \\ &= -\frac{1}{n}\sum_{i=1}^{n}\log\mu_\theta(x_i) + \text{Const.}\end{aligned}$$

- Density Fitting: MLE = Empirical KL Divergence Minimization

Are we done?

Density Fitting: Empirical KL Divergence Minimization

- *Empirical* KL Divergence

$$\begin{aligned}\mathcal{D}_{\mathrm{KL}}(\hat{\nu}\|\mu_\theta) &= -\int \hat{\nu}(x)\log\mu_\theta(x)dx + \text{Const.} \\ &= -\frac{1}{n}\sum_{i=1}^{n}\log\mu_\theta(x_i) + \text{Const.}\end{aligned}$$

- Density Fitting: MLE = Empirical KL Divergence Minimization

Are we done?

Density Fitting: Empirical KL Divergence Minimization

- *Empirical* KL Divergence

$$\begin{aligned}\mathcal{D}_{\mathrm{KL}}(\hat{\nu}\|\mu_\theta) &= -\int \hat{\nu}(x)\log\mu_\theta(x)dx + \text{Const.} \\ &= -\frac{1}{n}\sum_{i=1}^{n}\log\mu_\theta(x_i) + \text{Const.}\end{aligned}$$

- Density Fitting: MLE = Empirical KL Divergence Minimization

Are we done?

Generative Model in High Dimension Space

- Recall Gaussian mixture model: $\theta = \{(w_i, \mu_i, \Sigma_i)\}_{i=1}^K$

$$\mu_\theta = \sum_{i=1}^K w_i \mathcal{N}(\mu_i, \Sigma_i)$$

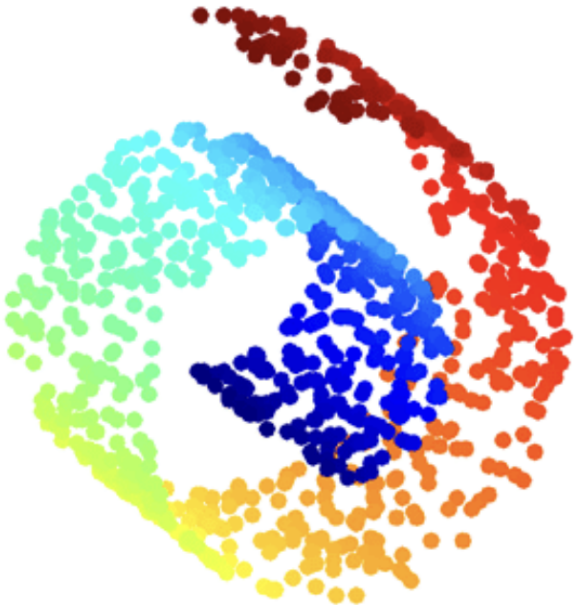
- Target distribution ν is usually supported on a low-dimensional manifold!

Generative Model in High Dimension Space

- Recall Gaussian mixture model: $\theta = \{(w_i, \mu_i, \Sigma_i)\}_{i=1}^K$

$$\mu_\theta = \sum_{i=1}^K w_i \mathcal{N}(\mu_i, \Sigma_i)$$

- Target distribution ν is usually supported on a low-dimensional manifold!



Different μ_θ?

Generative Model in High Dimension Space (cont'd)

- Let ξ denote the reference distibution on a *low-dimensional* latent space.

  (*e.g.*, Gaussian distribution)

- Let μ_θ be the pushforward of ξ by G_θ

$$\mu_\theta = (G_\theta)_\sharp \xi.$$

- μ_θ does not admit a density with respect to Lebesgue measure, which means $\mu_\theta(x_i) = 0$ almost everywhere.

  MLE (and KL Div. Minimization) cannot be used in this context!

Generative Model in High Dimension Space (cont'd)

- Let ξ denote the reference distibution on a *low-dimensional* latent space.

  (*e.g.*, Gaussian distribution)

- Let μ_θ be the pushforward of ξ by G_θ

$$\mu_\theta = (G_\theta)_\sharp \xi.$$

- μ_θ does not admit a density with respect to Lebesgue measure, which means $\mu_\theta(x_i) = 0$ almost everywhere.

MLE (and KL Div. Minimization) cannot be used in this context!

Jensen-Shannon Divergence

- KL Divergence $\mathcal{D}_{\mathrm{KL}}(\nu\|\mu_\theta)$ takes ∞ almost everywhere.

- How about Jensen-Shannon (JS) Divergence?

$$\mathcal{D}_{\mathrm{JS}}(\nu,\mu_\theta) := \frac{1}{2}\mathcal{D}_{\mathrm{KL}}\left(\nu\middle\|\frac{\nu+\mu_\theta}{2}\right) + \frac{1}{2}\mathcal{D}_{\mathrm{KL}}\left(\mu_\theta\middle\|\frac{\nu+\mu_\theta}{2}\right)$$

$$\mathcal{D}_{\mathrm{JS}}(\nu\|\mu_\theta) = \log 2 \text{ almost everywhere}$$

Jensen-Shannon Divergence

- KL Divergence $\mathcal{D}_{\mathrm{KL}}(\nu\|\mu_\theta)$ takes ∞ almost everywhere.

- How about Jensen-Shannon (JS) Divergence?

$$\mathcal{D}_{\mathrm{JS}}(\nu,\mu_\theta) := \frac{1}{2}\mathcal{D}_{\mathrm{KL}}\left(\nu\middle\|\frac{\nu+\mu_\theta}{2}\right) + \frac{1}{2}\mathcal{D}_{\mathrm{KL}}\left(\mu_\theta\middle\|\frac{\nu+\mu_\theta}{2}\right)$$

$\mathcal{D}_{\mathrm{JS}}(\nu\|\mu_\theta) = \log 2$ almost everywhere

Jensen-Shannon Divergence: Minimax Formulation

- Original GAN[1] minimizes JS Divergence via *minimax* formulation

$$\begin{aligned} & \min_{\theta} 2(\mathcal{D}_{\mathrm{JS}}(\nu, \mu_\theta) - \log 2) \\ = \ & \min_{\theta} \left(\max_{w} \{ \mathcal{L}(\theta, w) := -\mathbb{E}_{x\sim\nu}[l(f_w(x))] - \mathbb{E}_{z\sim\mu_\theta}[l(-f_w(z))] \} \right), \end{aligned}$$

  where $l(t) = \log(1 + \exp(-t))$ is the logistic loss. $\mu_\theta = (G_\theta)_\sharp \xi$

- For a fixed generator θ, inner maximization is a classification problem, where the classifier (discriminator) f_w distinguishes between true x with label 1 and generated z with label -1.

[1] Goodfellow et al., Generative Adversarial Nets, NeurIPS, 2014.

JS Divergence: Minimax Formulation (cont'd)

- Although $\mathcal{D}_{\mathrm{JS}}(\nu\|\mu_\theta) = \log 2$ almost everywhere, the original GAN showed promising results! How?

- Trained by an approximation of gradient descent on the JS divergence

$$\begin{aligned}
\theta_{k+1} &= \theta_k - \eta\nabla_\theta\mathcal{L}(\theta_k, w_k) && \approx\ \theta_k - \eta\nabla_\theta\mathcal{D}_{\mathrm{JS}}(\nu\|\mu_{\theta_k}) \\
w_{k,1} &= w_k + \eta\nabla_w\mathcal{L}(\theta_{k+1}, w_k) \\
w_{k,2} &= w_{k,1} + \eta\nabla_w\mathcal{L}(\theta_{k+1}, w_{k,1}) \\
&\vdots \\
w_{k+1} &= w_{k,M-1} + \eta\nabla_w\mathcal{L}(\theta_{k+1}, w_{k,M-1})
\end{aligned}$$

- $M = 1$ and 5 are found to be magic(?) numbers.

JS Divergence: Minimax Formulation (cont'd)

- Although $\mathcal{D}_{\mathrm{JS}}(\nu\|\mu_\theta) = \log 2$ almost everywhere, the original GAN showed promising results! How?

- Trained by an approximation of gradient descent on the JS divergence

$$\begin{aligned}
\theta_{k+1} &= \theta_k - \eta\nabla_\theta\mathcal{L}(\theta_k, w_k) && \approx\ \theta_k - \eta\nabla_\theta\mathcal{D}_{\mathrm{JS}}(\nu\|\mu_{\theta_k}) \\
w_{k,1} &= w_k + \eta\nabla_w\mathcal{L}(\theta_{k+1}, w_k) \\
w_{k,2} &= w_{k,1} + \eta\nabla_w\mathcal{L}(\theta_{k+1}, w_{k,1}) \\
&\vdots \\
w_{k+1} &= w_{k,M-1} + \eta\nabla_w\mathcal{L}(\theta_{k+1}, w_{k,M-1})
\end{aligned}$$

- $M = 1$ and 5 are found to be magic(?) numbers.

JS Divergence: Minimax Formulation (cont'd)

- Although $\mathcal{D}_{\mathrm{JS}}(\nu\|\mu_\theta) = \log 2$ almost everywhere, the original GAN showed promising results! How?

- Trained by an approximation of gradient descent on the JS divergence

$$\begin{aligned}
\theta_{k+1} &= \theta_k - \eta\nabla_\theta\mathcal{L}(\theta_k, w_k) && \approx\ \theta_k - \eta\nabla_\theta\mathcal{D}_{\mathrm{JS}}(\nu\|\mu_{\theta_k}) \\
w_{k,1} &= w_k + \eta\nabla_w\mathcal{L}(\theta_{k+1}, w_k) \\
w_{k,2} &= w_{k,1} + \eta\nabla_w\mathcal{L}(\theta_{k+1}, w_{k,1}) \\
&\vdots \\
w_{k+1} &= w_{k,M-1} + \eta\nabla_w\mathcal{L}(\theta_{k+1}, w_{k,M-1})
\end{aligned}$$

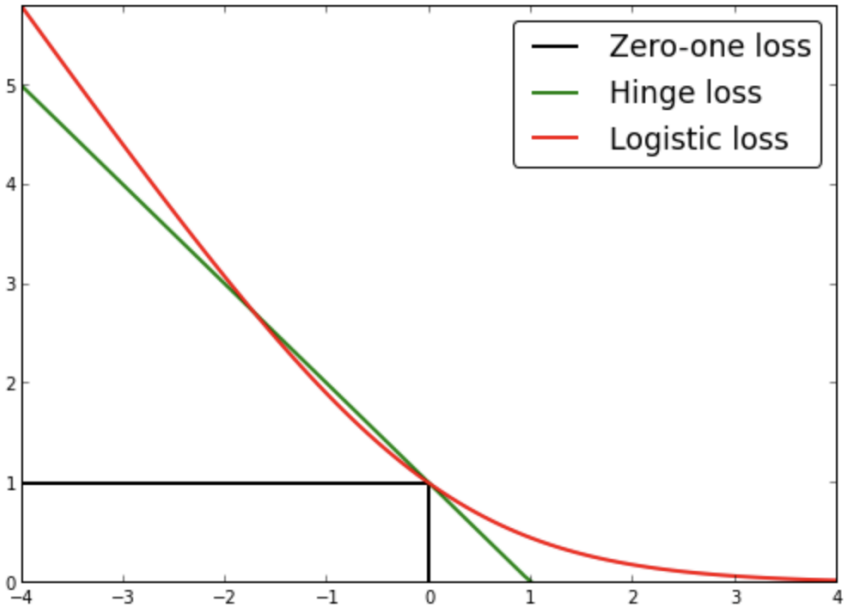
- $M = 1$ and 5 are found to be magic(?) numbers.

JS Divergence: Minimax Formulation (cont'd)

- Due to the logistic loss $l(t) = \log(1 + \exp(-t))$ in

$$\min_{\theta} \max_{w} \{\mathcal{L}(\theta, w) := -\mathbb{E}_{x\sim\nu}[l(f_w(x))] - \mathbb{E}_{z\sim\mu_\theta}[l(-f_w(z))]\},$$

the original GAN suffered from a *vanishing gradient* issue.



Wasserstein Distance

- Wasserstein distance

$$\mathcal{D}_{\mathrm{W}}(\nu, \mu_\theta) := \inf_{\gamma \in \Pi(\nu, \mu_\theta)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|]$$

- Gradient is meaningful almost everywhere

$$\nabla_\theta \mathcal{D}_{\mathrm{W}}(\nu, \mu_\theta),$$

but computing this is very expensive.

Wasserstein Distance

- Wasserstein distance

$$\mathcal{D}_{\mathrm{W}}(\nu, \mu_\theta) := \inf_{\gamma \in \Pi(\nu, \mu_\theta)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|]$$

- Gradient is meaningful almost everywhere

$$\nabla_\theta \mathcal{D}_{\mathrm{W}}(\nu, \mu_\theta),$$

  but computing this is very expensive.

Wasserstein Distance: Minimax formulation

- Wasserstein GAN[2] minimizes Wasserstein distance via minimax formulation

$$\min_{\theta} \mathcal{D}_{\mathrm{W}}(\nu, \mu_\theta) = \min_{\theta} \left(\max_{w:\|f_w\|_L \leq 1} \mathbb{E}_{x\sim\nu}[f_w(x)] - \mathbb{E}_{z\sim\mu_\theta}[f_w(z)] \right).$$

no vanishing gradient issue!

- However, the constraint $\|f_w\|_L \leq 1$ is complicated, so it is widely replaced by the gradient penalty[3]

$$\lambda \mathbb{E}_{\tilde{x}\sim\zeta}[(\|\nabla_x f_w(\tilde{x})\| - 1)^2]$$

useful but not good enough!

[2] Arjovsky, Chintala, Bottou, Wasserstein Generative Adversarial Networks, ICML, 2017

[3] Gulrajani, Ahmed,Arjovsky, Dumoulin, Courville, Improved Training of Wasserstein GANs, NeurIPS, 2017

Wasserstein Distance: Minimax formulation

- Wasserstein GAN[2] minimizes Wasserstein distance via minimax formulation

$$\min_{\theta} \mathcal{D}_{\mathrm{W}}(\nu, \mu_\theta) = \min_{\theta} \left(\max_{w:\|f_w\|_L \leq 1} \mathbb{E}_{x\sim\nu}[f_w(x)] - \mathbb{E}_{z\sim\mu_\theta}[f_w(z)] \right).$$

no vanishing gradient issue!

- However, the constraint $\|f_w\|_L \leq 1$ is complicated, so it is widely replaced by the gradient penalty[3]

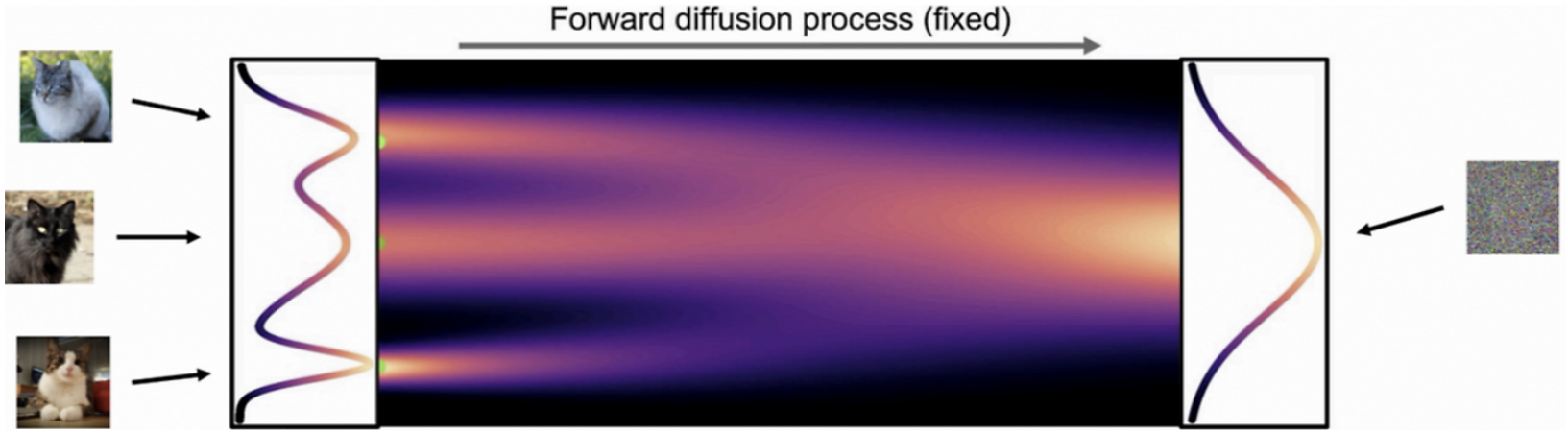
$$\lambda \mathbb{E}_{\tilde{x}\sim\zeta}[(\|\nabla_x f_w(\tilde{x})\| - 1)^2]$$

useful but not good enough!

[2] Arjovsky, Chintala, Bottou, Wasserstein Generative Adversarial Networks, ICML, 2017

[3] Gulrajani, Ahmed,Arjovsky, Dumoulin, Courville, Improved Training of Wasserstein GANs, NeurIPS, 2017

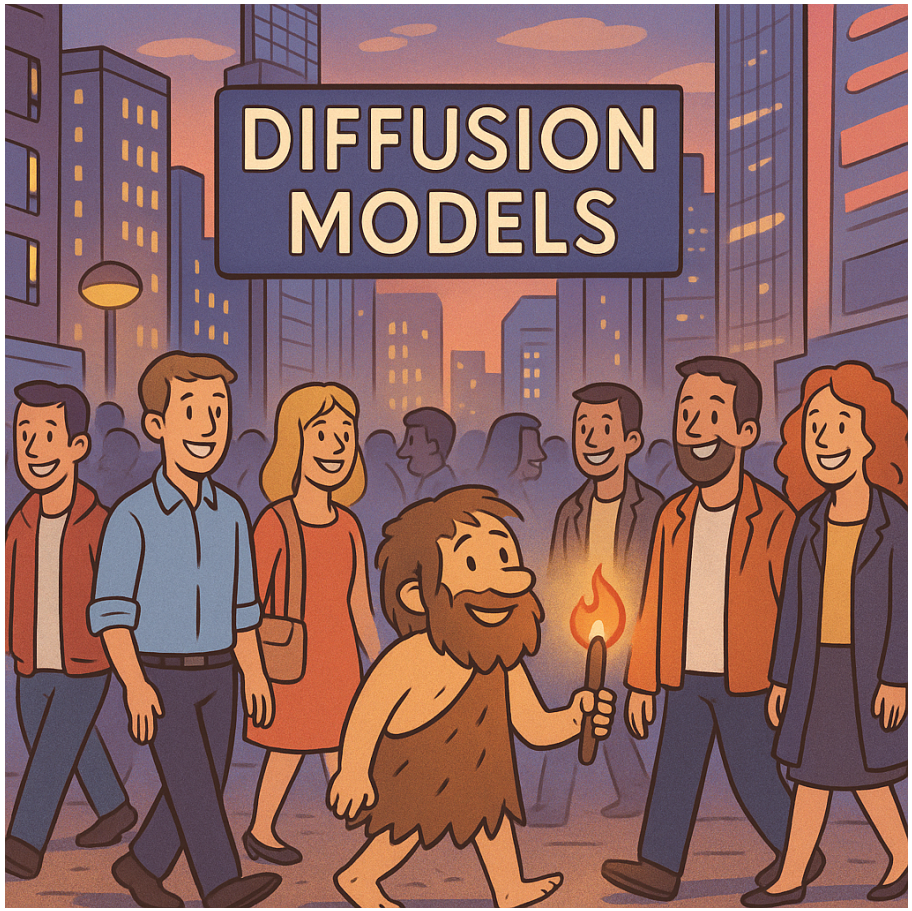
Rise of Diffusion and Flow Matching Models



- μ_θ: output distribution of reverse diffusion process $\mu_\theta = (G_\theta)_\sharp \xi$?
- Generating data requires multiple sampling steps,

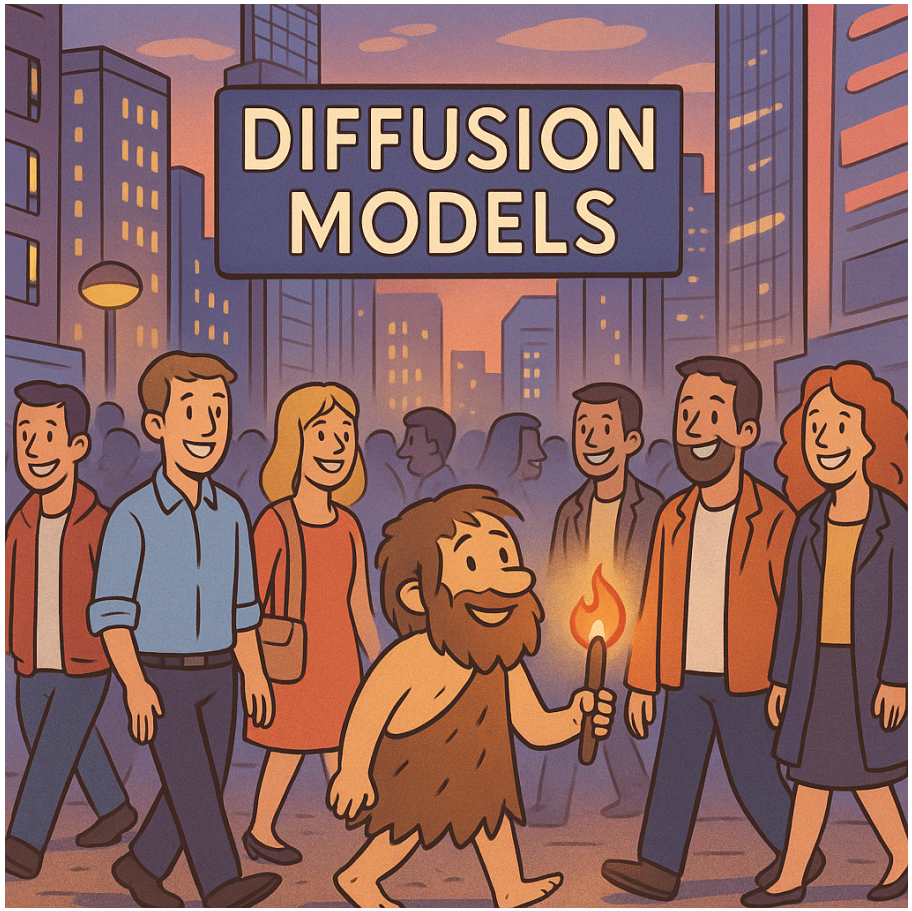
  but training only involves minimization!

Rise of Diffusion and Flow Matching Models (cont'd)



Q. Is it really impossible to *stably* train the *direct* map G_θ?

Rise of Diffusion and Flow Matching Models (cont'd)



Q. Is it really impossible to *stably* train the *direct* map G_θ?

Wasserstein Distance Minimization

- Squared Wasserstein-2 distance

$$\mathcal{D}^2_{\mathrm{W}_2}(\nu, \mu_\theta) := \inf_{\gamma \in \Pi(\nu, \mu_\theta)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|^2]$$

piece-wise quadratic!

- For any $\alpha \in \Pi(\nu, \xi)$, we have[4]

$$\mathcal{D}^2_{\mathrm{W}_2}(\nu, \mu_\theta) \leq \mathbb{E}_{(x,z)\sim\alpha}[\|x - G_\theta(z)\|^2].$$

In the empirical setting, we considered the discrete OT map α.

Given α, this is a regression problem! (benign overfitting and implicit bias)

[4] Chae, Kim, K., Rethinking memorization-generalization trade-off in generative models, ICML HiLD Workshop, 2025.

Wasserstein Distance Minimization

- Squared Wasserstein-2 distance

$$\mathcal{D}^2_{\mathrm{W}_2}(\nu, \mu_\theta) := \inf_{\gamma \in \Pi(\nu, \mu_\theta)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|^2]$$

piece-wise quadratic!

- For any $\alpha \in \Pi(\nu, \xi)$, we have[4]

$$\mathcal{D}^2_{\mathrm{W}_2}(\nu, \mu_\theta) \leq \mathbb{E}_{(x,z)\sim\alpha}[\|x - G_\theta(z)\|^2].$$

In the empirical setting, we considered the discrete OT map α.

Given α, this is a regression problem! (benign overfitting and implicit bias)

[4] Chae, Kim, K., Rethinking memorization-generalization trade-off in generative models, ICML HiLD Workshop, 2025.

How About Other Statistical Distances?

- Kullback-Leibler divergence
- Jensen-Shannon divergence
- Wassersten distance

- Reverse Kullback-Leibler divergence
- Jeffreys divergence
- Total variation distance
- Hellinger distance
- Rényi divergence
- Pearson χ^2 divergence

$$\vdots$$

none are good enough...

Zero-Infinity Distance

- Zero-Infinity distance:

$$\mathcal{D}_{\mathrm{ZI}}(\nu, \mu_\theta) = \begin{cases} 0, & \nu = \mu_\theta, \\ \infty, & \text{otherwise.} \end{cases}$$

Isn't this worse than KL divergence?

Zero-Infinity Distance: Minimax formulation

- Zero-Infinity GAN:[5] $=$ (WGAN w/o Lipschitz constraint)

$$\min_{\theta} \mathcal{D}_{\mathrm{ZI}}(\nu, \mu_\theta) = \min_{\theta} \max_{w} \left(\mathbb{E}_{x\sim\nu}[f_w(x)] - \mathbb{E}_{z\sim\mu_\theta}[f_w(z)]\right)$$

no constraint and no vanishing gradient!

[5] Lee, K., Zero-Infinity GAN: Stable dynamics and implicit bias of extragradient, NeurIPS OPT Workshop, 2025

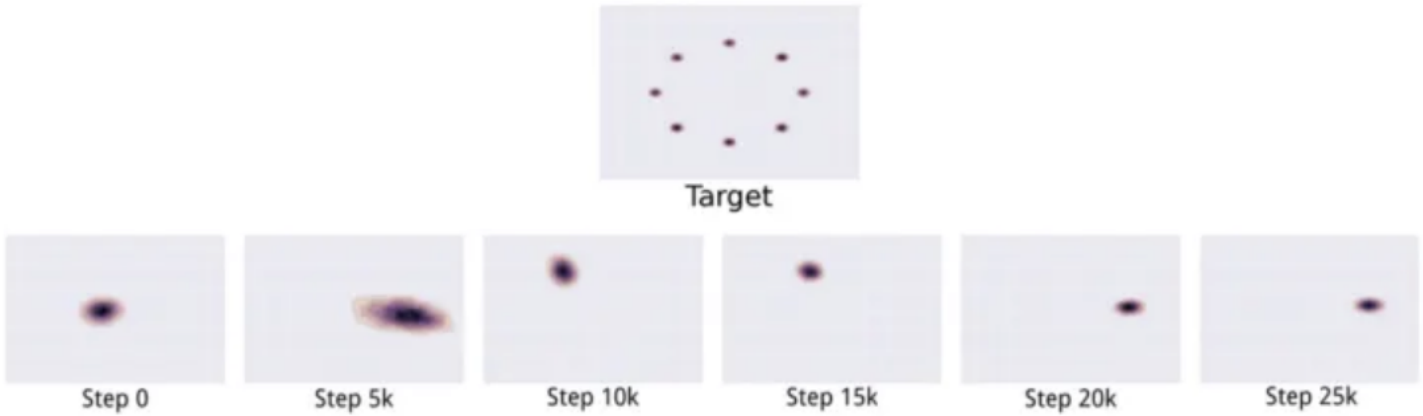
Zero-Infinity Distance: Minimax formulation

- Zero-Infinity GAN:[5] = (WGAN w/o Lipschitz constraint)

$$\min_{\theta} \mathcal{D}_{\mathrm{ZI}}(\nu, \mu_\theta) = \min_{\theta} \max_{w} \left(\mathbb{E}_{x\sim\nu}[f_w(x)] - \mathbb{E}_{z\sim\mu_\theta}[f_w(z)]\right)$$

no constraint and no vanishing gradient!

- How about mode collapse?



[5] Lee, K., Zero-Infinity GAN: Stable dynamics and implicit bias of extragradient, NeurIPS OPT Workshop, 2025

Toy Example: Dirac GAN

- Let $\nu = \delta_0$ and $\mu_\theta = \delta_\theta$.
- Dirac GAN[6] with $f_w(x) = wx$

$$\min_\theta \max_{w \in \mathcal{W}} -l(0) - l(-w\theta)$$

Distance	Loss $l(t)$	Constraint $\mathcal{W}$
Jensen-Shannon	$\log(1+\exp(-t))$	$\mathbb{R}$
Wasserstein	$-t$	$\{w : \lvert w\rvert \leq 1\}$
Zero-Infinity	$-t$	$\mathbb{R}$

$\min_\theta \max_w w\theta$

Approximate gradient descent (minimizing the distance) does not converge!

[6] Mescheder, Geiger and Nowozin. Which training methods for GANs do actually converge?, ICML, 2018.

Toy Example: Dirac GAN

- Let $\nu = \delta_0$ and $\mu_\theta = \delta_\theta$.

- Dirac GAN[6] with $f_w(x) = wx$

$$\min_\theta \max_{w \in \mathcal{W}} -l(0) - l(-w\theta)$$

Distance	Loss $l(t)$	Constraint $\mathcal{W}$
Jensen-Shannon	$\log(1+\exp(-t))$	$\mathbb{R}$
Wasserstein	$-t$	$\{w : \lvert w\rvert \leq 1\}$
Zero-Infinity	$-t$	$\mathbb{R}$

$\min_\theta \max_w w\theta$

Approximate gradient descent (minimizing the distance) does not converge!

[6] Mescheder, Geiger and Nowozin. Which training methods for GANs do actually converge?, ICML, 2018.

Toy Example: Dirac GAN (cont'd)

- Let $p := (\theta, w)$ and $F := (\nabla_\theta \mathcal{L}, -\nabla_w \mathcal{L})$
- Gradient descent ascent (GDA) $p_{k+1} = p_k - \eta F(p_k)$ does not converge.
- Extragradient (EG):

$$p_{k+1} = p_k - \eta F(p_k - \eta F(p_k))$$

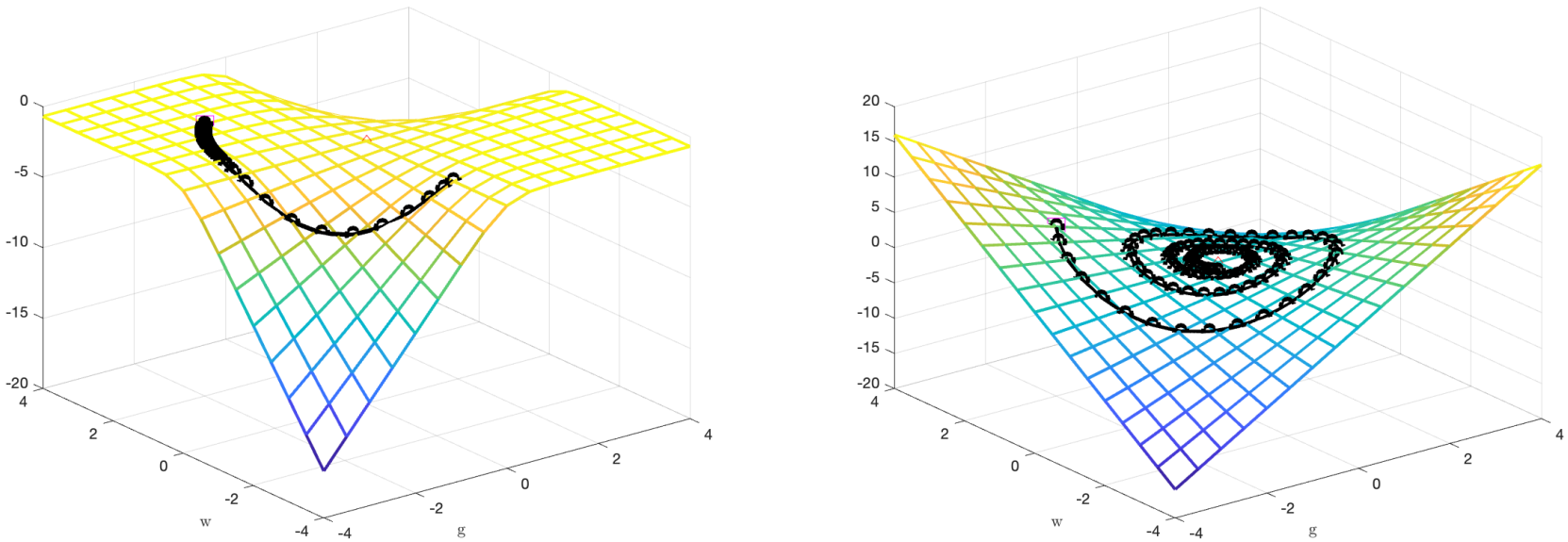
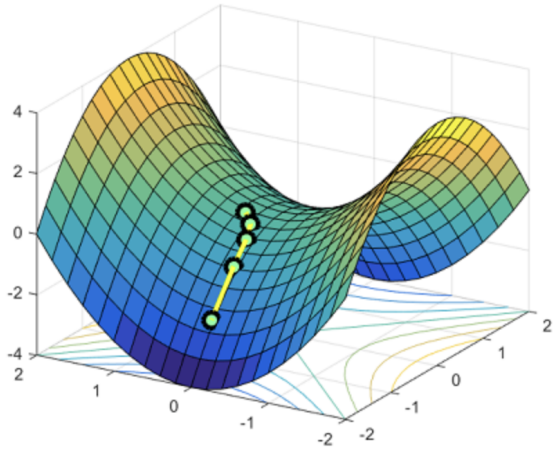


Figure: Trajectories of extragradient: (Left) Jensen-Shannon and (Right) Zero-Infinity

Gradient Methods under Nonconvexity

- Gradient descent escapes *strict saddle* points almost surely.[7]



- **Two-timescale extragradient** escapes **strict non-minimax** points almost surely (for sufficiently large τ).[8]

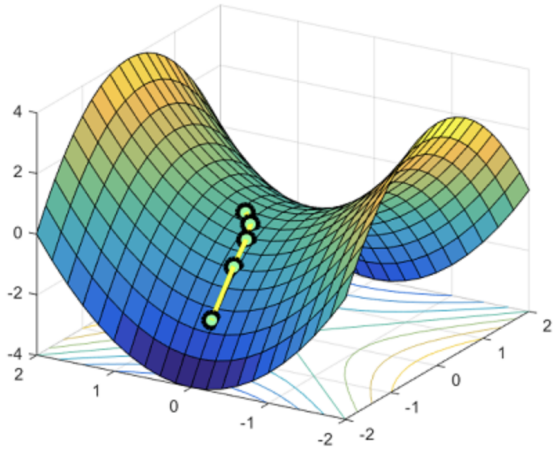
$$p_{k+1} = p_k - \eta \Lambda_\tau F(p_k - \eta \Lambda_\tau F(p_k)), \quad \text{where } \Lambda_\tau = \begin{bmatrix} \frac{1}{\tau} I & 0 \\ 0 & I \end{bmatrix} \text{ for } \tau > 1.$$

[7] Lee, Simchowitz, Jordan, Recht, Gradient descent only converges to minimizers, COLT, 2016.

[8] Chae, Kim, K., Two-timescale extragradient for finding local minimax points, ICLR, 2024.

Gradient Methods under Nonconvexity

- Gradient descent escapes *strict saddle* points almost surely.[7]



- **Two-timescale extragradient** escapes **strict non-minimax** points almost surely (for sufficiently large τ).[8]

$$p_{k+1} = p_k - \eta \Lambda_\tau F(p_k - \eta \Lambda_\tau F(p_k)), \quad \text{where } \Lambda_\tau = \begin{bmatrix} \frac{1}{\tau} I & 0 \\ 0 & I \end{bmatrix} \text{ for } \tau > 1.$$

[7] Lee, Simchowitz, Jordan, Recht, Gradient descent only converges to minimizers, COLT, 2016.

[8] Chae, Kim, K., Two-timescale extragradient for finding local minimax points, ICLR, 2024.

Populations and Empirical Risks in Generative Models

- Population risk:

$$\mathcal{R}(\theta) := \mathcal{D}(\nu, (G_\theta)_\sharp \xi)$$

- Empirical risk: $(\hat{\nu} := \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{x}_i})$

$$\hat{\mathcal{R}}(\theta) := \mathcal{D}(\hat{\nu}, (G_\theta)_\sharp \xi)$$

Overfitting = Pure Memorization

- Another empirical risk:[9] [10] $(\hat{\xi} := \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{z}_i})$

$$\tilde{\mathcal{R}}(\theta) := \mathcal{D}(\hat{\nu}, (G_\theta)_\sharp \hat{\xi})$$

Benign overfitting and implicit bias can happen!

[9] Chae, Kim, K., Rethinking memorization-generalization trade-off in generative models, ICML HiLD Workshop, 2025.

[10] Lee, K., Zero-Infinity GAN: Stable dynamics and implicit bias of extragradient, NeurIPS OPT Workshop, 2025

Populations and Empirical Risks in Generative Models

- Population risk:

$$\mathcal{R}(\theta) := \mathcal{D}(\nu, (G_\theta)_\sharp \xi)$$

- Empirical risk: $(\hat{\nu} := \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{x}_i})$

$$\hat{\mathcal{R}}(\theta) := \mathcal{D}(\hat{\nu}, (G_\theta)_\sharp \xi)$$

Overfitting = Pure Memorization

- Another empirical risk:[9] [10] $(\hat{\xi} := \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{z}_i})$

$$\tilde{\mathcal{R}}(\theta) := \mathcal{D}(\hat{\nu}, (G_\theta)_\sharp \hat{\xi})$$

Benign overfitting and implicit bias can happen!

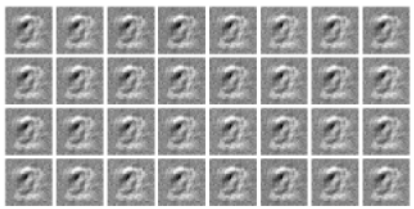
[9] Chae, Kim, K., Rethinking memorization-generalization trade-off in generative models, ICML HiLD Workshop, 2025.

[10] Lee, K., Zero-Infinity GAN: Stable dynamics and implicit bias of extragradient, NeurIPS OPT Workshop, 2025

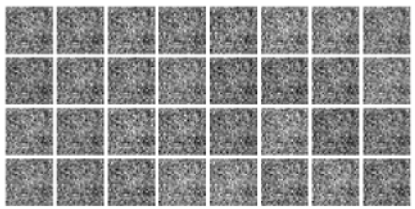
Preliminary Experiments

- 32 MNIST training data
- Compare whether the algorithms reach global solutions (memorizing training data).

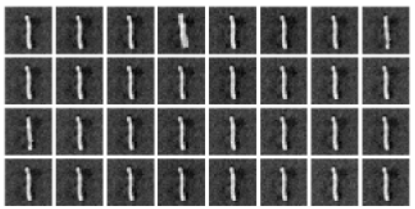
  (Implicit bias exists, but understanding generalization requires further investigation.)



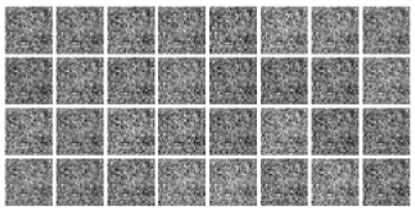
(a) JS-GDA($\tau = 1$)



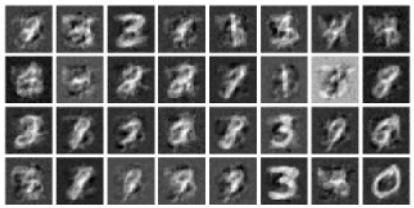
(b) JS-GDA($\tau = 5$)



(c) JS-EG($\tau = 1$)

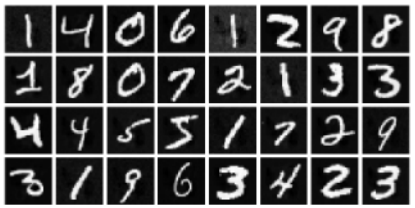


(d) JS-EG($\tau = 5$)

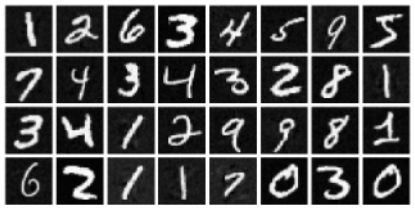


(e) ZI-GDA($\tau = 1$)

(f) ZI-GDA($\tau = 5$)



(g) ZI-EG($\tau = 1$)



(h) ZI-EG($\tau = 5$)

Observed similar results across different datasets and initializations!

Summary

1. Directly minimizing the statistical distance has long been the core of generative models.

2. The Zero-Infinity distance enables *stable* minimax training (without any heuristic regularization), when trained by a proper gradient method.

3. Benign overfitting and implicit bias can happen in generative models.