ICML22 Spotlight Presentation

Soft Truncation: A Universal Training Technique of Score-based Diffusion Model for High Precision Score Estimation

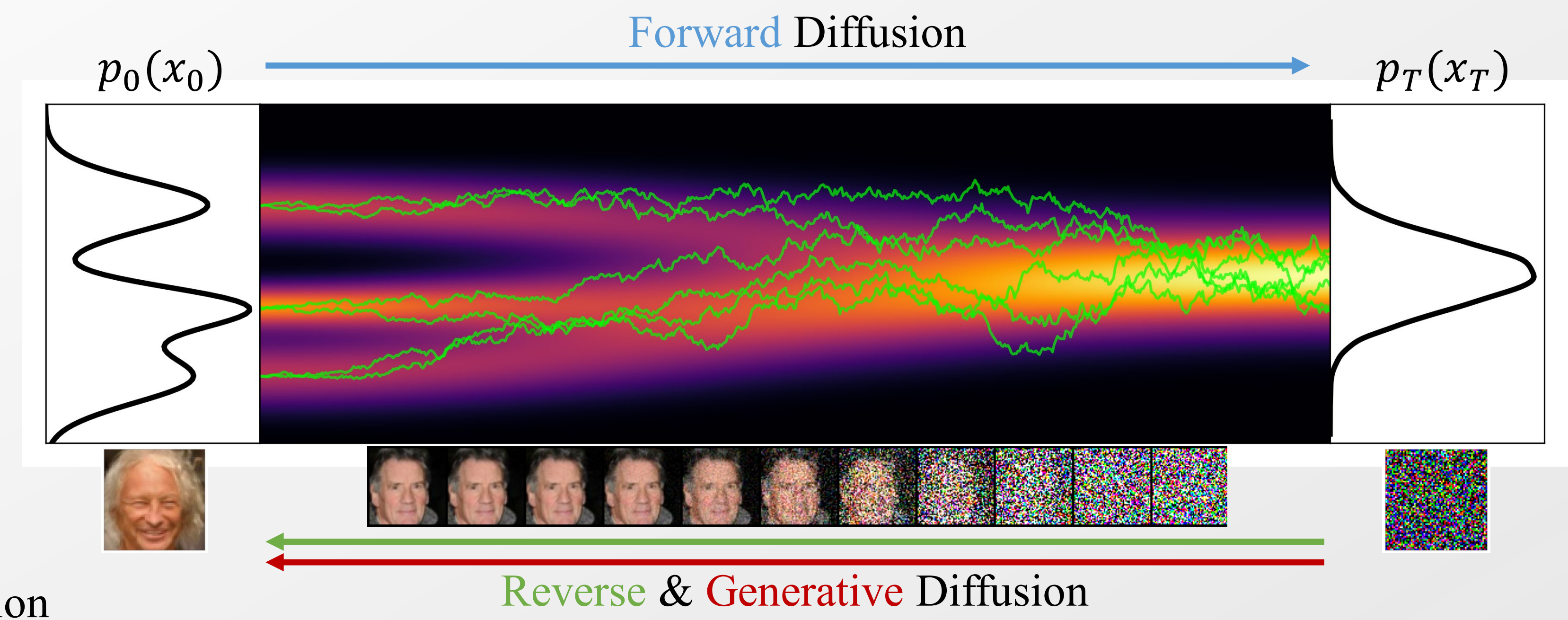
Dongjun Kim[1] Seungjae Shin[1] Kyungwoo Song[2]
Wanmo Kang[1] Il-Chul Moon[1, 3]

1 KAIST

2 University of Seoul 1918 서울시립대학교

3 summary.ai

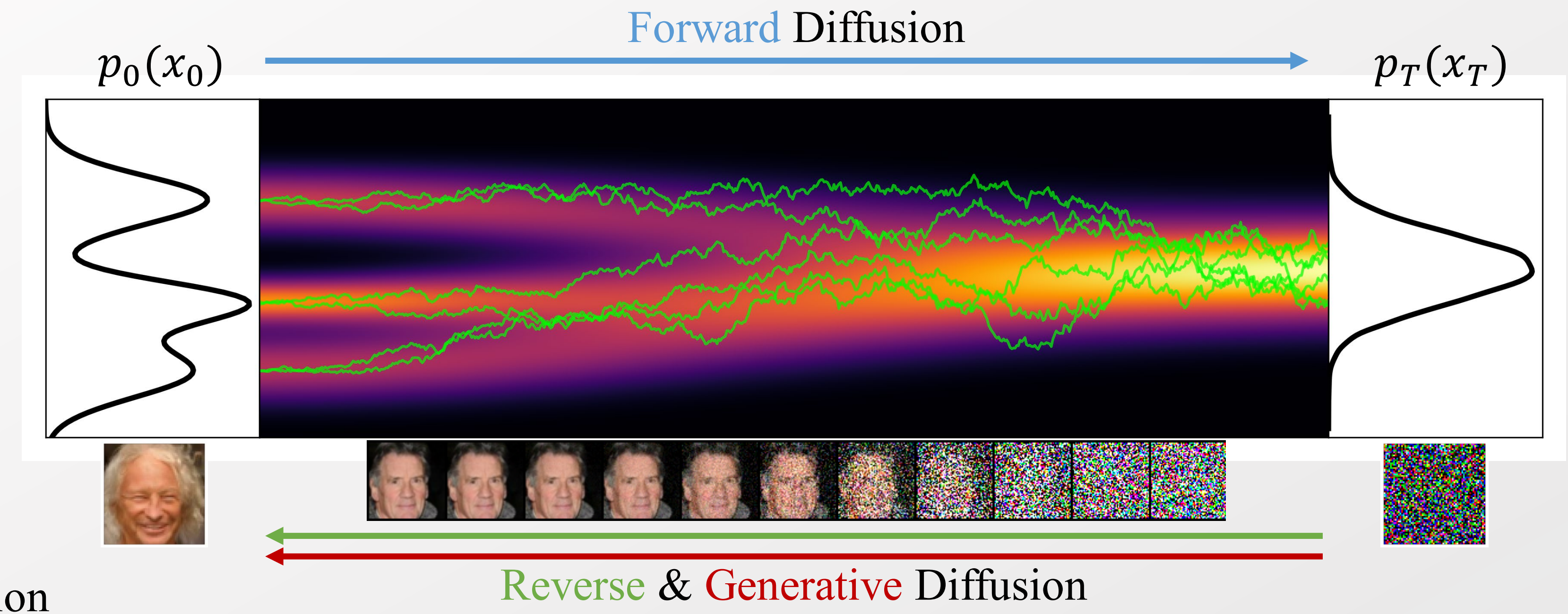
Introduction to Diffusion Model



- Forward Diffusion
  - $\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)\,\mathrm{d}t + g(t)\,\mathrm{d}\mathbf{w}_t$

	$\mathbf{f}(x_t, t)$	$g(x_t, t)$	$p_{0t}(x_t \mid x_0)$
VESDE	0	$\sigma_{min}\left(\frac{\sigma_{max}}{\sigma_{min}}\right)^t \sqrt{\frac{\sigma_{max}}{\sigma_{min}}}$	$\mathcal{N}(x_t; x_0, \sigma_{VE}^2(t)I)$
VPSDE	$-\frac{1}{2}\beta(t)x_t$	$\sqrt{\beta(t)}$	$\mathcal{N}(x_t, \mu_{VP}(t)x_0, \sigma_{VP}^2(t)I)$

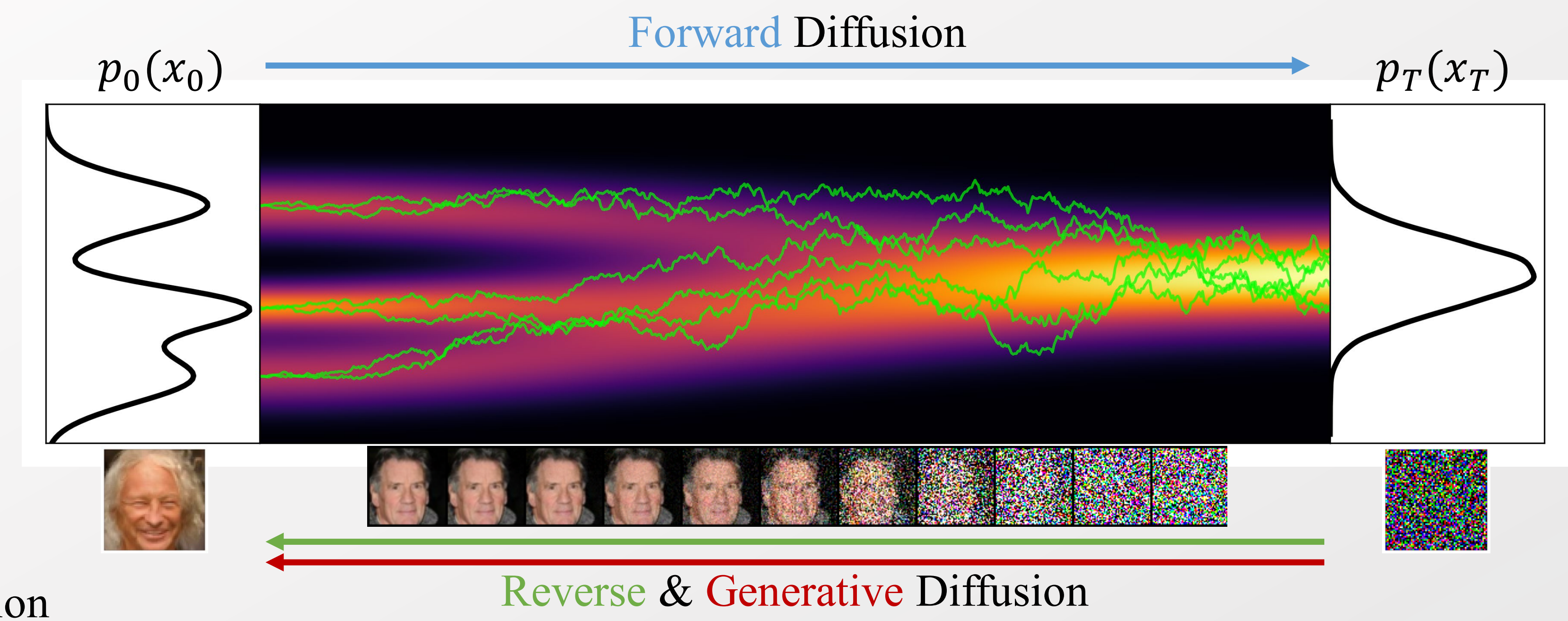
Introduction to Diffusion Model



- Forward Diffusion
  - $\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)\,\mathrm{d}t + g(t)\,\mathrm{d}\mathbf{w}_t$
- Reverse Diffusion
  - $\mathrm{d}\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}_t, t) - g^2(t)\nabla \log p_t(\mathbf{x})\right]\mathrm{d}\bar{t} + g(t)\,\mathrm{d}\bar{\mathbf{w}}_t$

	$\mathbf{f}(x_t, t)$	$\boldsymbol{g}(x_t, t)$	$\boldsymbol{p}_{0t}(x_t \vert x_0)$
VESDE	0	$\sigma_{min}\left(\frac{\sigma_{max}}{\sigma_{min}}\right)^t \sqrt{\frac{\sigma_{max}}{\sigma_{min}}}$	$\mathcal{N}(x_t; x_0, \sigma_{VE}^2(t)I)$
VPSDE	$-\frac{1}{2}\beta(t)x_t$	$\sqrt{\beta(t)}$	$\mathcal{N}(x_t, \mu_{VP}(t)x_0, \sigma_{VP}^2(t)I)$

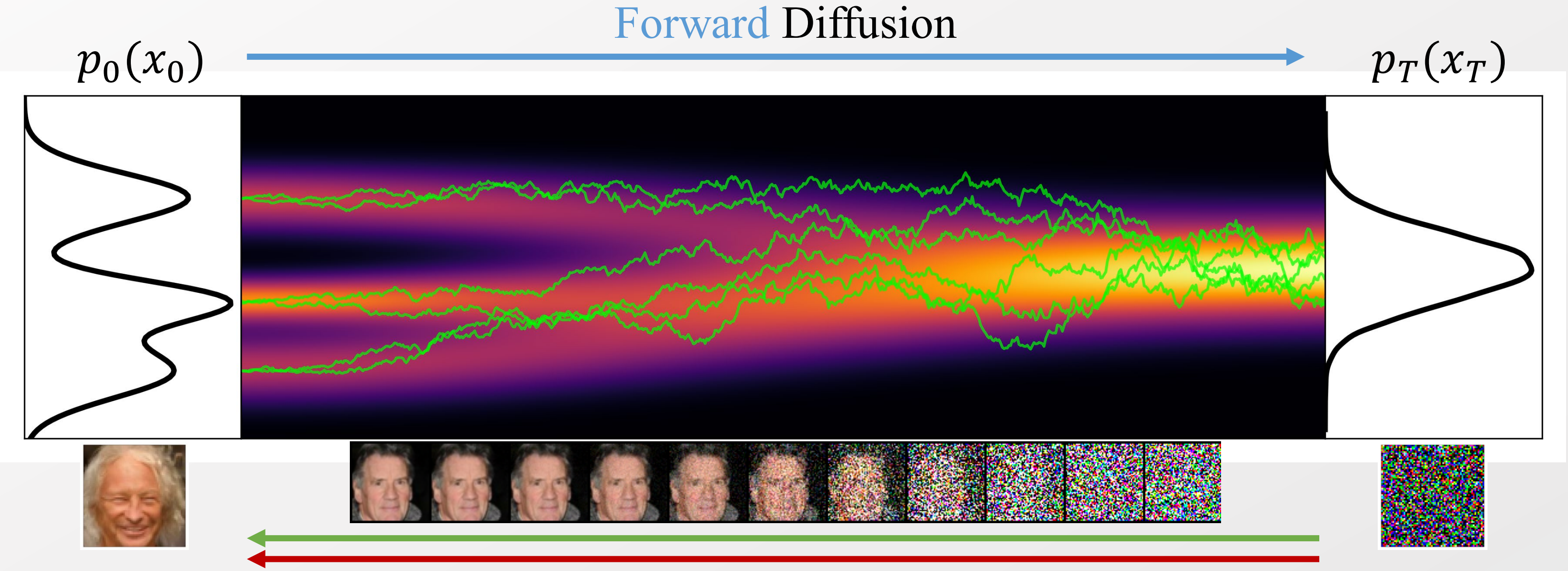
Introduction to Diffusion Model



- Forward Diffusion
  - $\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)\,\mathrm{d}t + g(t)\,\mathrm{d}\mathbf{w}_t$
- Reverse Diffusion
  - $\mathrm{d}\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}_t, t) - g^2(t)\nabla \log p_t(\mathbf{x})\right]\mathrm{d}\bar{t} + g(t)\,\mathrm{d}\bar{\mathbf{w}}_t$
- Generative Diffusion
  - $\mathrm{d}\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}_t, t) - g^2(t)\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\right]\mathrm{d}\bar{t} + g(t)\,\mathrm{d}\bar{\mathbf{w}}_t$

	$\mathbf{f}(x_t, t)$	$g(x_t, t)$	$p_{0t}(x_t \mid x_0)$
VESDE	0	$\sigma_{min}\left(\frac{\sigma_{max}}{\sigma_{min}}\right)^t \sqrt{\frac{\sigma_{max}}{\sigma_{min}}}$	$\mathcal{N}(x_t; x_0, \sigma_{VE}^2(t)I)$
VPSDE	$-\frac{1}{2}\beta(t)x_t$	$\sqrt{\beta(t)}$	$\mathcal{N}(x_t, \mu_{VP}(t)x_0, \sigma_{VP}^2(t)I)$

Introduction to Diffusion Model



- Forward Diffusion
  - $\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)\,\mathrm{d}t + g(t)\,\mathrm{d}\mathbf{w}_t$
- Reverse Diffusion
  - $\mathrm{d}\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}_t, t) - g^2(t)\nabla \log p_t(\mathbf{x})\right]\mathrm{d}\bar{t} + g(t)\,\mathrm{d}\bar{\mathbf{w}}_t$
- Generative Diffusion
  - $\mathrm{d}\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}_t, t) - g^2(t)\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\right]\mathrm{d}\bar{t} + g(t)\,\mathrm{d}\bar{\mathbf{w}}_t$
- Score Estimation Loss
  - $\mathcal{L}(\boldsymbol{\theta}; \lambda, \epsilon) = \frac{1}{2}\int_{\epsilon}^{T} \lambda(t)\mathbb{E}_{p_r(\mathbf{x}_0)}\mathbb{E}_{p_{0t}(\mathbf{x}_t|\mathbf{x}_0)}\left[\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) - \nabla \log p_{0t}(\mathbf{x}_t|\mathbf{x}_0)\|_2^2\right]\mathrm{d}t$

	$\boldsymbol{f(x_t, t)}$	$\boldsymbol{g(x_t, t)}$	$\boldsymbol{p_{0t}(x_t\|x_0)}$
VESDE	0	$\sigma_{min}\left(\frac{\sigma_{max}}{\sigma_{min}}\right)^t\sqrt{\frac{\sigma_{max}}{\sigma_{min}}}$	$\mathcal{N}(x_t; x_0, \sigma_{VE}^2(t)I)$
VPSDE	$-\frac{1}{2}\beta(t)x_t$	$\sqrt{\beta(t)}$	$\mathcal{N}(x_t, \mu_{VP}(t)x_0, \sigma_{VP}^2(t)I)$

Contribution of This Paper

APPLIED ARTIFICIAL INTELLIGENCE LAB

Central Questions

[**Q1**] How is $\mathcal{L}(\theta;\lambda,\epsilon)$ connected to log-likelihood?
[**Q2**] How to optimize $\mathcal{L}(\theta;\lambda,\epsilon)$ well?

Central Questions

[**Q1**] How is $\mathcal{L}(\theta; \lambda, \epsilon)$ connected to log-likelihood?
[**Q2**] How to optimize $\mathcal{L}(\theta; \lambda, \epsilon)$ well?

- Question 1
  - Partial answer to Q1
    - [Corollary 1 (Song21Maximum)] $\mathbb{E}_{\mathbf{x}_0}\big[-\log p_0^{\boldsymbol{\theta}}(\mathbf{x}_0)\big] \leq \mathcal{L}(\boldsymbol{\theta}; \lambda = g^2, \epsilon)$
    - p_t^{θ} is the marginal distribution of the generative process at time t
    - This corollary holds only when $\lambda = g^2$

Variational Bound of General-Weighted Diffusion Loss

Central Questions

[**Q1**] How is $\mathcal{L}(\theta; \lambda, \epsilon)$ connected to log-likelihood?
[**Q2**] How to optimize $\mathcal{L}(\theta; \lambda, \epsilon)$ well?

- Question 1
  - Partial answer to Q1
    - [Corollary 1 (Song21Maximum)] $\mathbb{E}_{\mathbf{x}_0}\left[-\log p_0^{\boldsymbol{\theta}}(\mathbf{x}_0)\right] \leq \mathcal{L}(\boldsymbol{\theta}; \lambda = g^2, \epsilon) \iff D_{KL}(p_r \| p_0^{\boldsymbol{\theta}}) \leq \mathcal{L}(\boldsymbol{\theta}; \lambda = g^2, \epsilon)$
    - p_t^θ is the marginal distribution of the generative process at time t
    - This corollary holds only when $\lambda = g^2$

Variational Bound of General-Weighted Diffusion Loss

Central Questions

[**Q1**] How is $\mathcal{L}(\theta;\lambda,\epsilon)$ connected to log-likelihood?

[**Q2**] How to optimize $\mathcal{L}(\theta;\lambda,\epsilon)$ well?

- Question 1
  - Partial answer to Q1
    - [Corollary 1 (Song21Maximum)] $\mathbb{E}_{\mathbf{x}_0}\big[-\log p_0^{\boldsymbol{\theta}}(\mathbf{x}_0)\big] \le \mathcal{L}(\boldsymbol{\theta};\lambda=g^2,\epsilon) \iff D_{KL}(p_r\|p_0^{\boldsymbol{\theta}}) \le \mathcal{L}(\boldsymbol{\theta};\lambda=g^2,\epsilon)$
    - p_t^θ is the marginal distribution of the generative process at time t
    - This corollary holds only when $\lambda = g^2$
  - Complete answer to Q1

Variational Bound of General-Weighted Diffusion Loss

Central Questions

[**Q1**] How is $\mathcal{L}(\theta; \lambda, \epsilon)$ connected to log-likelihood?
[**Q2**] How to optimize $\mathcal{L}(\theta; \lambda, \epsilon)$ well?

- Question 1
  - Partial answer to Q1
    - [Corollary 1 (Song21Maximum)] $\mathbb{E}_{\mathbf{x}_0}\big[-\log p_0^{\boldsymbol{\theta}}(\mathbf{x}_0)\big] \le \mathcal{L}(\boldsymbol{\theta}; \lambda = g^2, \epsilon) \iff D_{KL}(p_r \| p_0^{\boldsymbol{\theta}}) \le \mathcal{L}(\boldsymbol{\theta}; \lambda = g^2, \epsilon)$
    - p_t^{θ} is the marginal distribution of the generative process at time t
    - This corollary holds only when $\lambda = g^2$
  - Complete answer to Q1
    - [Theorem 1] $\mathbb{E}_{\mathbb{P}_\lambda(\tau)}\Big[\mathbb{E}_{\mathbf{x}_\tau}\big[-\log p_\tau^{\boldsymbol{\theta}}(\mathbf{x}_\tau)\big]\Big] \le \mathcal{L}(\boldsymbol{\theta}; \lambda, \epsilon) \iff \mathbb{E}_{\mathbb{P}_\lambda(\tau)}\big[D_{KL}(p_\tau \| p_\tau^{\boldsymbol{\theta}})\big] \le \mathcal{L}(\boldsymbol{\theta}; \lambda, \epsilon)$

Soft Truncation: Alternative Optimization Method of $\mathcal{L}(\theta;\lambda,\epsilon)$

Central Questions

[**Q1**] How is $\mathcal{L}(\theta;\lambda,\epsilon)$ connected to log-likelihood?
[**Q2**] How to optimize $\mathcal{L}(\theta;\lambda,\epsilon)$ well?

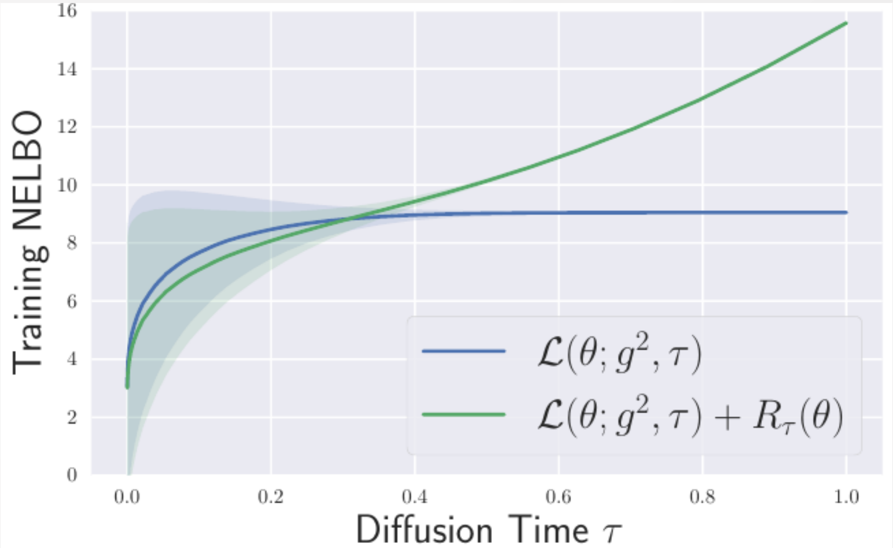
Dataset	Model	NLL	FID
CIFAR-10	DDPM++ (VP, NLL)	**3.03**	6.70
	+ Soft Truncation	**3.03**	**3.45**
	DDPM++ (VP, FID)	3.21	3.90
ImageNet32	DDPM++ (VP, NLL)	3.92	12.68
	+ Soft Truncation	**3.90**	**8.42**
	DDPM++ (VP, FID)	3.95	9.22

Soft Truncation: Alternative Optimization Method of $\mathcal{L}(\theta;\lambda,\epsilon)$

Central Questions

[**Q1**] How is $\mathcal{L}(\theta;\lambda,\epsilon)$ connected to log-likelihood?
[**Q2**] How to optimize $\mathcal{L}(\theta;\lambda,\epsilon)$ well?

- Question 2
  - [Observation 1] Small diffusion time contributes the most of the integration in $\mathcal{L}(\theta;g^2,\epsilon)$



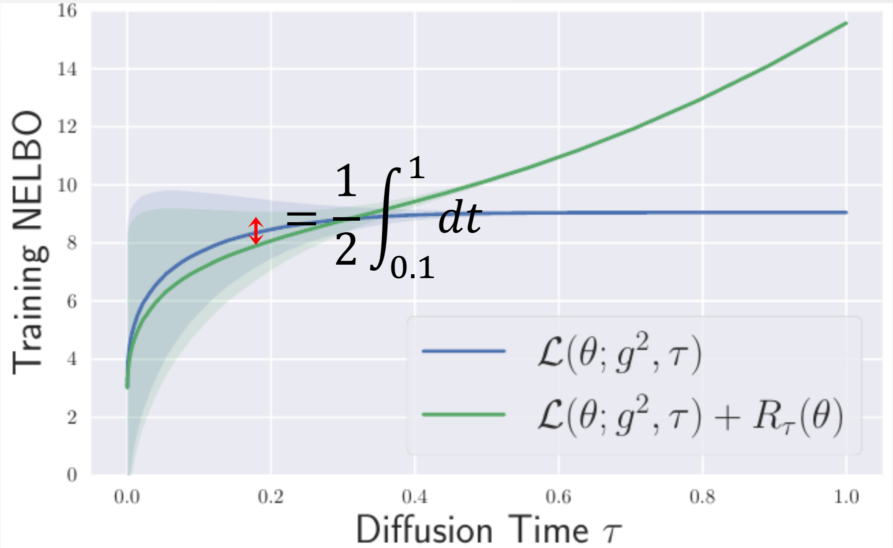
$$\mathcal{L}(\boldsymbol{\theta};g^2,\epsilon) = \frac{1}{2}\int_{\epsilon}^{T} g^2(t)\mathbb{E}_{p_r(\mathbf{x}_0)}\mathbb{E}_{p_{0t}(\mathbf{x}_t|\mathbf{x}_0)}\left[\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t,t)-\nabla\log p_{0t}(\mathbf{x}_t|\mathbf{x}_0)\|_2^2\right]\mathrm{d}t$$

Soft Truncation: Alternative Optimization Method of $\mathcal{L}(\theta; \lambda, \epsilon)$

Central Questions

[**Q1**] How is $\mathcal{L}(\theta; \lambda, \epsilon)$ connected to log-likelihood?
[**Q2**] How to optimize $\mathcal{L}(\theta; \lambda, \epsilon)$ well?

- Question 2
  - [Observation 1] Small diffusion time contributes the most of the integration in $\mathcal{L}(\theta; \lambda, \epsilon)$



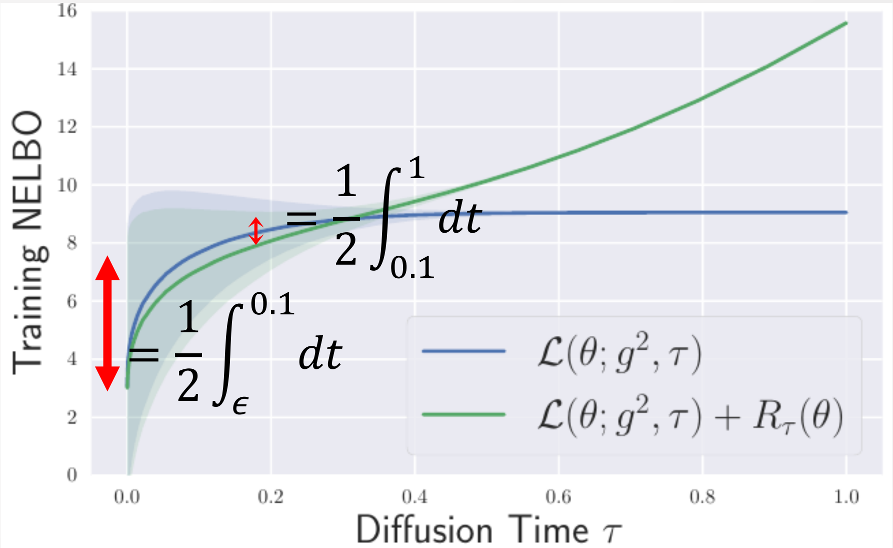
$$\mathcal{L}(\boldsymbol{\theta}; g^2, \epsilon)$$
$$= \frac{1}{2}\int_{\epsilon}^{T} g^2(t)\mathbb{E}_{p_r(\mathbf{x}_0)}\mathbb{E}_{p_{0t}(\mathbf{x}_t|\mathbf{x}_0)}\left[\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) - \nabla \log p_{0t}(\mathbf{x}_t|\mathbf{x}_0)\|_2^2\right] \mathrm{d}t$$

Soft Truncation: Alternative Optimization Method of $\mathcal{L}(\theta;\lambda,\epsilon)$

Central Questions

[**Q1**] How is $\mathcal{L}(\theta;\lambda,\epsilon)$ connected to log-likelihood?
[**Q2**] How to optimize $\mathcal{L}(\theta;\lambda,\epsilon)$ well?

- Question 2
  - [Observation 1] Small diffusion time contributes the most of the integration in $\mathcal{L}(\theta;\lambda,\epsilon)$

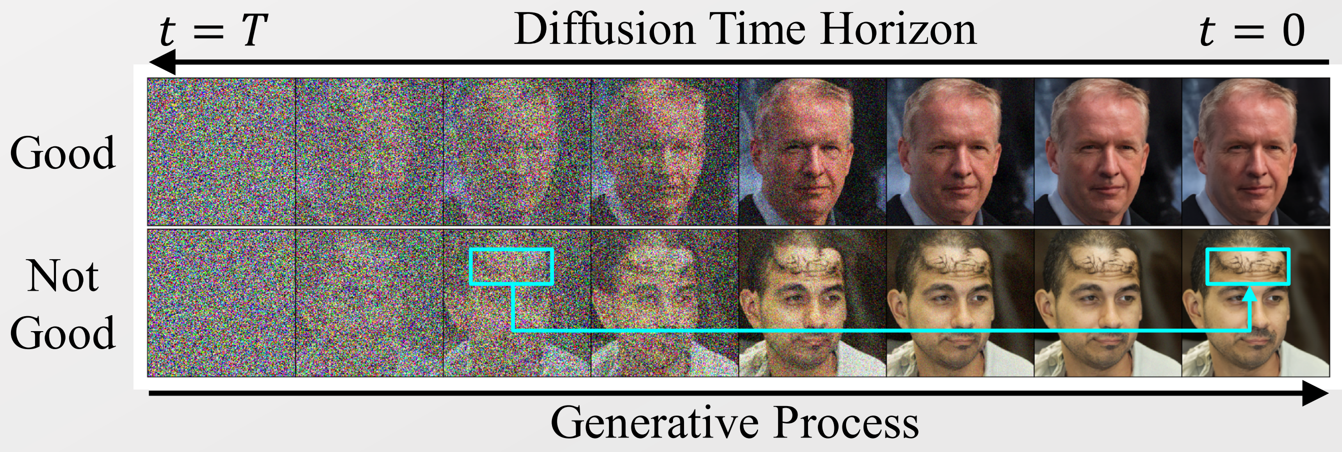


$$\mathcal{L}(\boldsymbol{\theta}; g^2, \epsilon) = \frac{1}{2}\int_{\epsilon}^{T} g^2(t)\mathbb{E}_{p_r(\mathbf{x}_0)}\mathbb{E}_{p_{0t}(\mathbf{x}_t|\mathbf{x}_0)}\left[\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t,t) - \nabla\log p_{0t}(\mathbf{x}_t|\mathbf{x}_0)\|_2^2\right] \mathrm{d}t$$

Central Questions

[**Q1**] How is $\mathcal{L}(\theta;\lambda,\epsilon)$ connected to log-likelihood?
[**Q2**] How to optimize $\mathcal{L}(\theta;\lambda,\epsilon)$ well?

- Question 2
  - [Observation 1] Small diffusion time contributes the most of the integration in $\mathcal{L}(\theta;\lambda,\epsilon)$
  - [Observation 2] Large diffusion time contributes to the global sample fidelity



Soft Truncation: Alternative Optimization Method of $\mathcal{L}(\theta;\lambda,\epsilon)$

Central Questions

[**Q1**] How is $\mathcal{L}(\theta;\lambda,\epsilon)$ connected to log-likelihood?
[**Q2**] How to optimize $\mathcal{L}(\theta;\lambda,\epsilon)$ well?

- Question 2
  - [Observation 1] Small diffusion time contributes the most of the integration in $\mathcal{L}(\theta;\lambda,\epsilon)$
  - [Observation 2] Large diffusion time contributes to the global sample fidelity

  - ⇒ A better optimization method will bring an enhanced score accuracy on **large diffusion time**

Soft Truncation: Alternative Optimization Method of $\mathcal{L}(\theta; \lambda, \epsilon)$

Central Questions

[**Q1**] How is $\mathcal{L}(\theta; \lambda, \epsilon)$ connected to log-likelihood?
[**Q2**] How to optimize $\mathcal{L}(\theta; \lambda, \epsilon)$ well?

- Question 2
  - [Observation 1] Small diffusion time contributes the most of the integration in $\mathcal{L}(\theta; \lambda, \epsilon)$
  - [Observation 2] Large diffusion time contributes to the global sample fidelity
  - ⇒ A better optimization method will bring an enhanced score accuracy on **large diffusion time**
  - (**Soft Truncation**) Optimize $\mathcal{L}(\theta; g^2, \tau) = \int_\tau^T dt$ for $\tau \sim P(\tau)$ in every mini-batch update
    - Softens the static hyper-parameter ϵ with a random variable $\tau \sim P(\tau)$

Central Questions

[**Q1**] How is $\mathcal{L}(\theta;\lambda,\epsilon)$ connected to log-likelihood?
[**Q2**] How to optimize $\mathcal{L}(\theta;\lambda,\epsilon)$ well?

- Question 2
  - [Observation 1] Small diffusion time contributes the most of the integration in $\mathcal{L}(\theta;\lambda,\epsilon)$
  - [Observation 2] Large diffusion time contributes to the global sample fidelity
  - ⇒ A better optimization method will bring an enhanced score accuracy on **large diffusion time**
  - (**Soft Truncation**) Optimize $\mathcal{L}(\theta;g^2,\tau)=\int_\tau^T dt$ for $\tau\sim P(\tau)$ in every mini-batch update
    - Softens the static hyper-parameter ϵ with a random variable $\tau\sim P(\tau)$
  - From $\mathcal{L}(\theta;\lambda,\epsilon)=\mathbb{E}_{P_\lambda(\tau)}[\mathcal{L}(\theta;g^2,\tau)]$, Soft Truncation is an optimization method of general-weighted loss

APPLIED ARTIFICIAL INTELLIGENCE LAB

Central Questions

[**Q1**] How is $\mathcal{L}(\theta;\lambda,\epsilon)$ connected to log-likelihood?
[**Q2**] How to optimize $\mathcal{L}(\theta;\lambda,\epsilon)$ well?

- Question 2
  - Vanilla training can be framed by Maximum Likelihood Estimation by Song21Maximum only when $\lambda = g^2$
    - $\mathbb{E}_{\mathbf{x}_0}\big[-\log p_0^{\boldsymbol{\theta}}(\mathbf{x}_0)\big] \le \mathcal{L}(\boldsymbol{\theta};\lambda=g^2,\epsilon)$

Soft Truncation: Maximum Perturbed Likelihood Estimation

APPLIED ARTIFICIAL INTELLIGENCE LAB

Central Questions

[**Q1**] How is $\mathcal{L}(\theta;\lambda,\epsilon)$ connected to log-likelihood?

[**Q2**] How to optimize $\mathcal{L}(\theta;\lambda,\epsilon)$ well?

- Question 2
  - Vanilla training can be framed by Maximum Likelihood Estimation by Song21Maximum only when $\lambda = g^2$
    - $\mathbb{E}_{\mathbf{x}_0}\big[-\log p_0^{\boldsymbol{\theta}}(\mathbf{x}_0)\big] \le \mathcal{L}(\boldsymbol{\theta};\lambda=g^2,\epsilon)$
  - Soft Truncation can be framed by **Maximum Perturbed Likelihood Estimation**

KAIST

Central Questions

[**Q1**] How is $\mathcal{L}(\theta;\lambda,\epsilon)$ connected to log-likelihood?
[**Q2**] How to optimize $\mathcal{L}(\theta;\lambda,\epsilon)$ well?

- Question 2
  - Vanilla training can be framed by Maximum Likelihood Estimation by Song21Maximum only when $\lambda = g^2$
    - $\mathbb{E}_{\mathbf{x}_0}\left[-\log p_0^{\boldsymbol{\theta}}(\mathbf{x}_0)\right] \le \mathcal{L}(\boldsymbol{\theta};\lambda=g^2,\epsilon)$
  - Soft Truncation can be framed by **Maximum Perturbed Likelihood Estimation**
    - Actual optimization loss at each mini-batch update
      - $D_{KL}(p_\tau \| p_\tau^{\boldsymbol{\theta}}) \le \mathcal{L}(\boldsymbol{\theta};g^2,\tau)$

Soft Truncation: Maximum Perturbed Likelihood Estimation

Central Questions

[**Q1**] How is $\mathcal{L}(\theta; \lambda, \epsilon)$ connected to log-likelihood?
[**Q2**] How to optimize $\mathcal{L}(\theta; \lambda, \epsilon)$ well?

- Question 2
  - Vanilla training can be framed by Maximum Likelihood Estimation by Song21Maximum only when $\lambda = g^2$
    - $\mathbb{E}_{\mathbf{x}_0}\big[-\log p_0^{\boldsymbol{\theta}}(\mathbf{x}_0)\big] \le \mathcal{L}(\boldsymbol{\theta}; \lambda = g^2, \epsilon)$
  - Soft Truncation can be framed by **Maximum Perturbed Likelihood Estimation**
    - Actual optimization loss at each mini-batch update
      - $D_{KL}(p_\tau \| p_\tau^{\boldsymbol{\theta}}) \le \mathcal{L}(\boldsymbol{\theta}; g^2, \tau)$
    - Loss averaged by mini-batches (Theorem 1)
      - $\mathbb{E}_{\mathbb{P}_\lambda(\tau)}\big[D_{KL}(p_\tau \| p_\tau^{\boldsymbol{\theta}})\big] \le \mathcal{L}(\boldsymbol{\theta}; \lambda, \epsilon) = \mathbb{E}_{\mathbb{P}_\lambda(\tau)}\big[\mathcal{L}(\boldsymbol{\theta}; g^2, \tau)\big]$

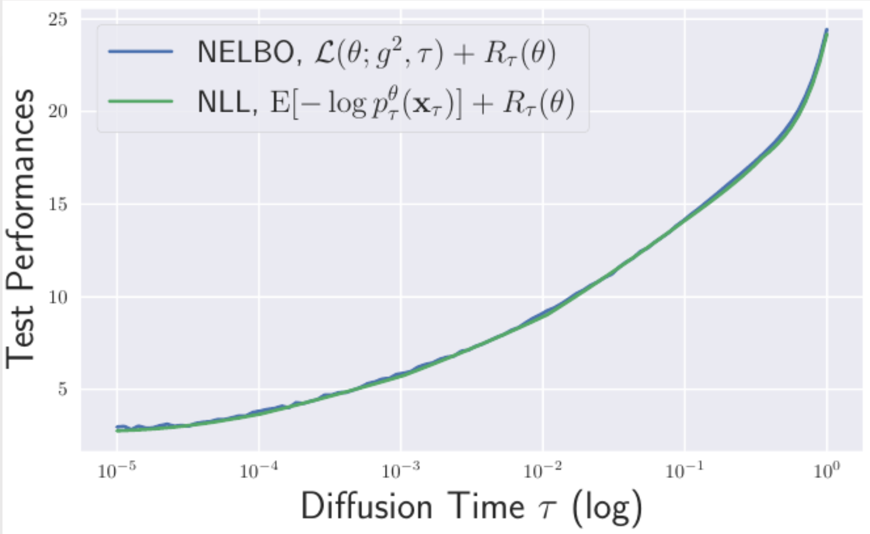
APPLIED ARTIFICIAL INTELLIGENCE LAB

Central Questions

[**Q1**] How is $\mathcal{L}(\theta;\lambda,\epsilon)$ connected to log-likelihood?

[**Q2**] How to optimize $\mathcal{L}(\theta;\lambda,\epsilon)$ well?

- Question 2
  - Vanilla training can be framed by Maximum Likelihood Estimation by Song21Maximum only when $\lambda = g^2$
    - $\mathbb{E}_{\mathbf{x}_0}\big[-\log p_0^{\boldsymbol{\theta}}(\mathbf{x}_0)\big] \le \mathcal{L}(\boldsymbol{\theta};\lambda=g^2,\epsilon)$
  - Soft Truncation can be framed by **Maximum Perturbed Likelihood Estimation**
    - Actual optimization loss at each mini-batch update
      - $D_{KL}(p_\tau \| p_\tau^{\boldsymbol{\theta}}) \le \mathcal{L}(\boldsymbol{\theta};g^2,\tau)$
    - Loss averaged by mini-batches
      - $\mathbb{E}_{\mathbb{P}_\lambda(\tau)}\big[D_{KL}(p_\tau \| p_\tau^{\boldsymbol{\theta}})\big] \le \mathcal{L}(\boldsymbol{\theta};\lambda,\epsilon) = \mathbb{E}_{\mathbb{P}_\lambda(\tau)}\big[\mathcal{L}(\boldsymbol{\theta};g^2,\tau)\big]$
    - Variational bound at each mini-batch is tight



Experimental Result

	Loss	Soft Truncation	NLL	NELBO	FID ODE
CIFAR-10	$\mathcal{L}(\boldsymbol{\theta}; g^2, \epsilon)$	✗	3.03	3.13	6.70
	$\mathcal{L}(\boldsymbol{\theta}; \sigma^2, \epsilon)$	✗	3.21	3.34	3.90
	$\mathcal{L}(\boldsymbol{\theta}; g^2_{\mathbb{P}_1}, \epsilon)$	✗	3.06	3.18	6.11
	$\mathcal{L}_{ST}(\boldsymbol{\theta}; g^2, \mathbb{P}_1)$	✓	**3.01**	**3.08**	3.96
	$\mathcal{L}_{ST}(\boldsymbol{\theta}; g^2, \mathbb{P}_{0.9})$	✓	3.03	3.13	**3.45**
ImageNet32	$\mathcal{L}(\boldsymbol{\theta}; g^2, \epsilon)$	✗	3.92	3.94	12.68
	$\mathcal{L}(\boldsymbol{\theta}; \sigma^2, \epsilon)$	✗	3.95	4.00	9.22
	$\mathcal{L}(\boldsymbol{\theta}; g^2_{\mathbb{P}_1}, \epsilon)$	✗	3.93	3.97	11.89
	$\mathcal{L}_{ST}(\boldsymbol{\theta}; g^2, \mathbb{P}_{0.9})$	✓	**3.90**	**3.91**	**8.42**

- Implications
  - Soft Truncation is a better optimization method against the vanilla optimization

Experimental Result

	Loss	Soft Truncation	NLL	NELBO	FID ODE
CIFAR-10	$\mathcal{L}(\boldsymbol{\theta}; g^2, \epsilon)$	✗	3.03	3.13	6.70
	$\mathcal{L}(\boldsymbol{\theta}; \sigma^2, \epsilon)$	✗	3.21	3.34	3.90
	$\mathcal{L}(\boldsymbol{\theta}; g^2_{\mathbb{P}_1}, \epsilon)$	✗	3.06	3.18	6.11
	$\mathcal{L}_{ST}(\boldsymbol{\theta}; g^2, \mathbb{P}_1)$	✓	**3.01**	**3.08**	3.96
	$\mathcal{L}_{ST}(\boldsymbol{\theta}; g^2, \mathbb{P}_{0.9})$	✓	3.03	3.13	**3.45**
ImageNet32	$\mathcal{L}(\boldsymbol{\theta}; g^2, \epsilon)$	✗	3.92	3.94	12.68
	$\mathcal{L}(\boldsymbol{\theta}; \sigma^2, \epsilon)$	✗	3.95	4.00	9.22
	$\mathcal{L}(\boldsymbol{\theta}; g^2_{\mathbb{P}_1}, \epsilon)$	✗	3.93	3.97	11.89
	$\mathcal{L}_{ST}(\boldsymbol{\theta}; g^2, \mathbb{P}_{0.9})$	✓	**3.90**	**3.91**	**8.42**

- Implications
  - Soft Truncation is a better optimization method against the vanilla optimization
  - Soft Truncation significantly solves the NLL-FID trade-off
    - Soft Truncation achieves comparable FID as much as the case of variance weighting

Experimental Result

	Loss	Soft Truncation	NLL	NELBO	FID ODE
CIFAR-10	$\mathcal{L}(\boldsymbol{\theta}; g^2, \epsilon)$	✗	3.03	3.13	6.70
	$\mathcal{L}(\boldsymbol{\theta}; \sigma^2, \epsilon)$	✗	3.21	3.34	3.90
	$\mathcal{L}(\boldsymbol{\theta}; g^2_{\mathbb{P}_1}, \epsilon)$	✗	3.06	3.18	6.11
	$\mathcal{L}_{ST}(\boldsymbol{\theta}; g^2, \mathbb{P}_1)$	✓	**3.01**	**3.08**	3.96
	$\mathcal{L}_{ST}(\boldsymbol{\theta}; g^2, \mathbb{P}_{0.9})$	✓	3.03	3.13	**3.45**
ImageNet32	$\mathcal{L}(\boldsymbol{\theta}; g^2, \epsilon)$	✗	3.92	3.94	12.68
	$\mathcal{L}(\boldsymbol{\theta}; \sigma^2, \epsilon)$	✗	3.95	4.00	9.22
	$\mathcal{L}(\boldsymbol{\theta}; g^2_{\mathbb{P}_1}, \epsilon)$	✗	3.93	3.97	11.89
	$\mathcal{L}_{ST}(\boldsymbol{\theta}; g^2, \mathbb{P}_{0.9})$	✓	**3.90**	**3.91**	**8.42**

- Implications
  - Soft Truncation is a better optimization method against the vanilla optimization
  - Soft Truncation significantly solves the NLL-FID trade-off
    - Soft Truncation achieves comparable FID as much as the case of variance weighting
    - Soft Truncation keeps NLL at the equivalent level compared to likelihood weighting

Experimental Result

Result on CelebA 64×64

SDE	Model	Loss	NLL	NELBO	FID PC	FID ODE
VE	NCSN++	$\mathcal{L}(\boldsymbol{\theta};\sigma^2,\epsilon)$	3.41	3.42	3.95	-
		$\mathcal{L}_{ST}(\boldsymbol{\theta};\sigma^2,\mathbb{P}_2)$	3.44	3.44	2.68	-
RVE	UNCSN++	$\mathcal{L}(\boldsymbol{\theta};g^2,\epsilon)$	2.01	**2.01**	3.36	-
		$\mathcal{L}_{ST}(\boldsymbol{\theta};g^2,\mathbb{P}_2)$	**1.97**	2.02	**1.92**	-
VP	DDPM++	$\mathcal{L}(\boldsymbol{\theta};\sigma^2,\epsilon)$	2.14	2.21	3.03	2.32
		$\mathcal{L}_{ST}(\boldsymbol{\theta};\sigma^2,\mathbb{P}_1)$	2.17	2.29	2.88	**1.90**
	UDDPM++	$\mathcal{L}(\boldsymbol{\theta};\sigma^2,\epsilon)$	2.11	2.20	3.23	4.72
		$\mathcal{L}_{ST}(\boldsymbol{\theta};\sigma^2,\mathbb{P}_1)$	2.16	2.28	2.22	1.94
	DDPM++	$\mathcal{L}(\boldsymbol{\theta};g^2,\epsilon)$	2.00	2.09	5.31	3.95
		$\mathcal{L}_{ST}(\boldsymbol{\theta};g^2,\mathbb{P}_1)$	2.00	2.11	4.50	2.90
	UDDPM++	$\mathcal{L}(\boldsymbol{\theta};g^2,\epsilon)$	1.98	2.12	4.65	3.98
		$\mathcal{L}_{ST}(\boldsymbol{\theta};g^2,\mathbb{P}_1)$	2.00	2.10	4.45	2.97

Result on CIFAR-10

Loss	NLL	NELBO	FID (ODE)
INDM (VP, NLL)	**2.98**	**2.98**	6.01
INDM (VP, FID)	3.17	3.23	**3.61**
INDM (VP, NLL) + ST	3.01	3.02	3.88

→ Nonlinear SDE

- Implications
  - Soft Truncation is a better optimization method against the vanilla optimization
  - Soft Truncation significantly solves the NLL-FID trade-off
    - Soft Truncation achieves comparable FID as much as the case of variance weighting
    - Soft Truncation keeps NLL at the equivalent level compared to likelihood weighting
  - Soft Truncation is universally applicable to any SDEs and network architectures

KAIST

Experimental Result

Model	CIFAR10 32 × 32			ImageNet32 32 × 32			CelebA 64 × 64		CelebA-HQ 256 × 256	STL-10 48 × 48	
	NLL (↓)	FID (↓)	IS (↑)	NLL	FID	IS	NLL	FID	FID	FID	IS
Likelihood-free Models											
StyleGAN2-ADA+Tuning (Karras et al., 2020)	-	2.92	10.02	-	-	-	-	-	-	-	-
Styleformer (Park & Kim, 2022)	-	2.82	9.94	-	-	-	-	3.66	-	15.17	11.01
Likelihood-based Models											
ARDM-Upscale 4 (Hoogeboom et al., 2021)	**2.64**	-	-	-	-	-	-	-	-	-	-
VDM (Kingma et al., 2021)	2.65	7.41	-	3.72	-	-	-	-	-	-	-
LSGM (FID) (Vahdat et al., 2021)	3.43	**2.10**	-	-	-	-	-	-	-	-	-
NCSN++ cont. (deep, VE) (Song et al., 2021b)	3.45	2.20	9.89	-	-	-	2.39	3.95	7.23	-	-
DDPM++ cont. (deep, sub-VP) (Song et al., 2021b)	2.99	2.41	9.57	-	-	-	-	-	-	-	-
DenseFlow-74-10 (Grcić et al., 2021)	2.98	34.90	-	**3.63**	-	-	1.99	-	-	-	-
ScoreFlow (VP, FID) (Song et al., 2021a)	3.04	3.98	-	3.84	**8.34**	-	-	-	-	-	-
Efficient-VDVAE (Hazami et al., 2022)	2.87	-	-	-	-	-	**1.83**	-	-	-	-
PNDM (Liu et al., 2022)	-	3.26	-	-	-	-	-	2.71	-	-	-
ScoreFlow (deep, sub-VP, NLL) (Song et al., 2021a)	2.81	5.40	-	3.76	10.18	-	-	-	-	-	-
Improved DDPM (L_{simple}) (Nichol & Dhariwal, 2021)	3.37	2.90	-	-	-	-	-	-	-	-	-
UNCSN++ (RVE) + ST	3.04	2.33	**10.11**	-	-	-	1.97	1.92	**7.16**	**7.71**	**13.43**
DDPM++ (VP, FID) + ST	2.91	2.47	9.78	-	-	-	2.10	**1.90**	-	-	-
DDPM++ (VP, NLL) + ST	2.88	3.45	9.19	3.85	8.42	**11.82**	1.96	2.90	-	-	-

Thank you!