Beyond Gaussian

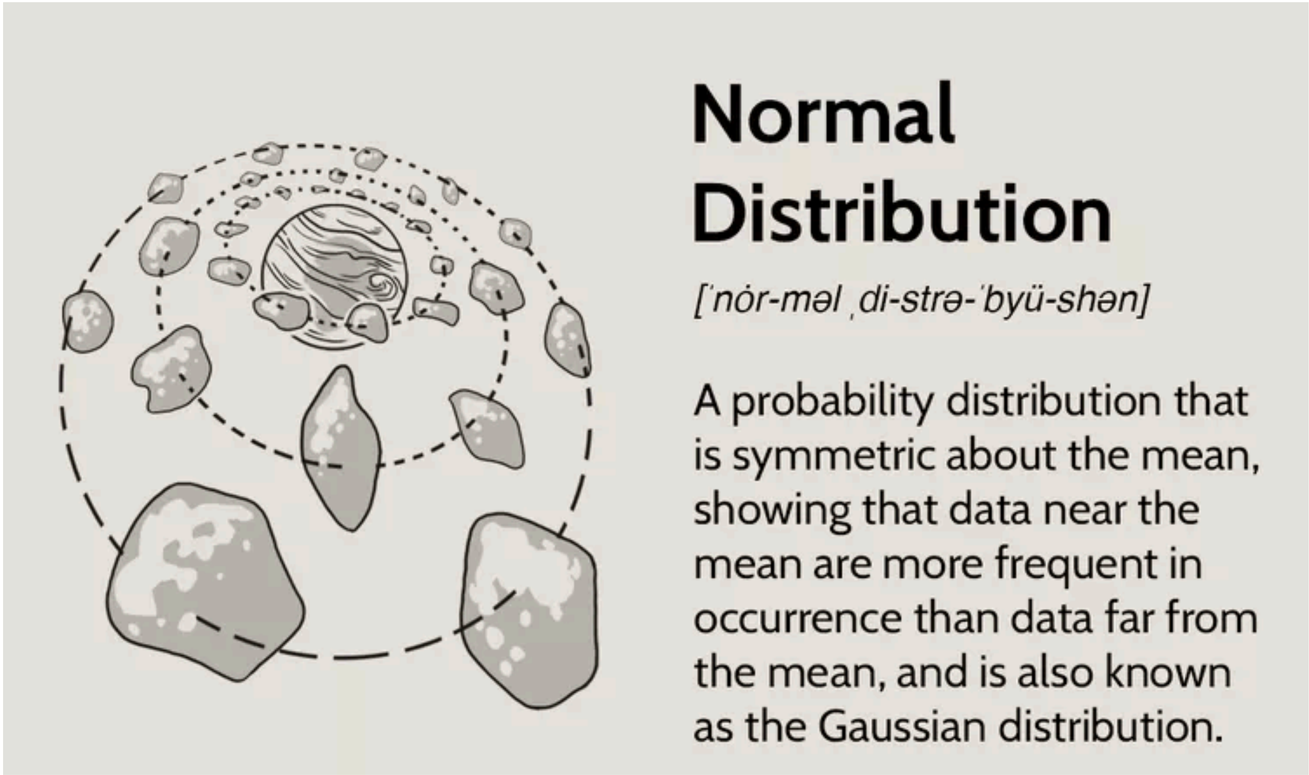
Heavy-tail Distributions in Machine Learning

Artificial Intelligence Graduate School

Sungbin Lim (2022. 11. 18) @ 한국인공지능학회 & NAVER 추계 공동학술대회

Gaussian Distribution in ML

- Many ML scientists prefer Gaussian distribution
  - simple / easy to understand
  - exponential decay
  - central limit theorem
- Applications
  - noise model
  - generative models



Investopedia / Lara Antal

Too Ideal Properties of Gaussian Distribution

- Density function
  - Affine & Fourier transform of Gaussian
  - Exponential decay at tail → cannot model ***heavy-tail*** behavior
- Wiener process (or Brownian motion)
  - Markov property
  - Continuous trajectory → cannot model ***jump*** movement
  - Fokker-Planck Kolmogorov equations
  - Closed under time reversal

too ideal property
→ does not fit real data!

Heavy-tail Distribution

Definition

(**Foss et al.**, 2013) A random variable X is called **heavy-tail** if for all $t > 0$,

$$\underbrace{\mathbb{E}\exp(tX)}_{\text{Moment Generating Function}} = \int_{-\infty}^{\infty} \exp(tx) F(\mathrm{d}x) = \infty$$

CDF of X

(**Resnick**, 2007) A random variable X is heavy-tail if the tail probability is

$$\mathbb{P}(X > x) \sim x^{-\alpha}$$

tail index of X

Heavy-tail Distribution in Statistical Modeling

- Pareto distribution → insurance data
- Log-normal distribution → human behaviors, biomedical, chemistry
- Lévy distribution → geomagnetic reversal, probability theory
- Weibull distribution → failure analysis, extreme value theory
- Log-logistic distribution → survival analysis, economics
- Fréchet distribution → hydrology, decline curve analysis
- Family of stable distributions → finance, spectroscopy
- …

Can we utilize these distributions in ML?

Sub-Gaussian and ERM

- Empirical risk is a good approximation of risk when the data distribution follows sub-Gaussian: for any $h \in \mathscr{H}$,

$$\underbrace{\mathbb{P}(|X| > x) \leq N \exp(-\nu x^2)}_{\text{Sub-Gaussian}} \quad \Rightarrow \quad \left|\widehat{R}(h) - R(h)\right| = O\left(\frac{1}{n^{\alpha}}\right)$$

$$\underbrace{\widehat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i)}_{\text{Empirical Risk}} \qquad \underbrace{R(h) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(h(\mathbf{x}), y)]}_{\text{Risk}}$$

Heavy-Tail and ERM

we need stronger assumptions or different approach instead of ERM

- ~~Empirical risk is a good approximation of risk~~ when the data distribution follows **heavy-tail**: for any $h \in \mathcal{H}$,

$$\mathbb{P}(X > x) \sim x^{-\alpha} \quad \not\Rightarrow \quad \left|\widehat{R}(h) - R(h)\right| = O\left(\frac{1}{n^{\alpha}}\right)$$

Heavy-Tail

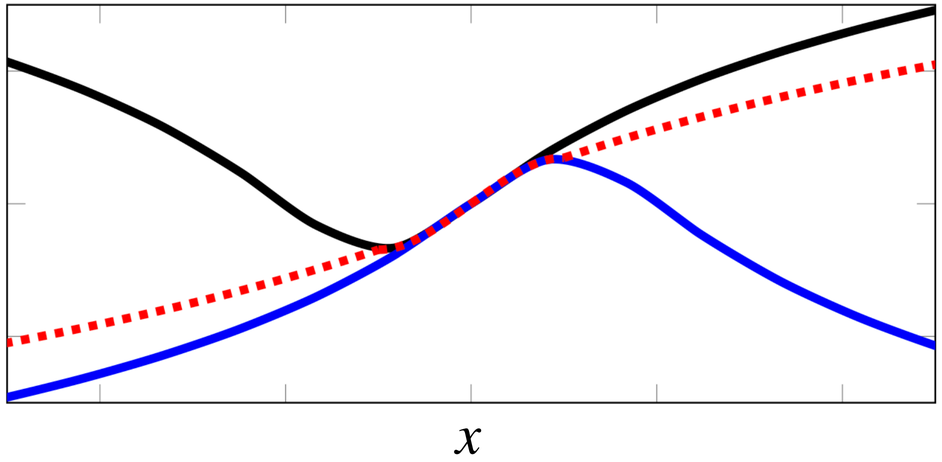
$$\widehat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i) \qquad R(h) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(h(\mathbf{x}), y)]$$

Empirical Risk — Risk

Catoni's estimator

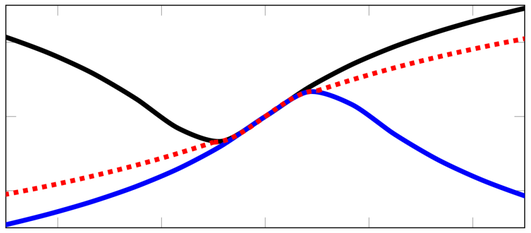
- Influence function: $\psi : \mathbb{R} \to \mathbb{R}$

$$-\log\left(1 - x + \frac{x^2}{2}\right) \leq \psi(x) \leq \log\left(1 + x + \frac{x^2}{2}\right)$$



Challenging the empirical mean and empirical variance, O. Catoni, ***Annales de l'Institut Henri Poincaré*** 2012

Catoni's estimator



- Influence function: $\psi : \mathbb{R} \to \mathbb{R}$

$$-\log\left(1 - x + \frac{x^2}{2}\right) \leq \psi(x) \leq \log\left(1 + x + \frac{x^2}{2}\right)$$

$$\sum_{i=1}^{n} \psi\left[\alpha(x_i - \hat{\theta})\right] = 0$$

positive parameter

what if the variance is *not* finite?

- Catoni (2012) demonstrates that with high probability, the deviation of $\hat{\theta}$ from the mean is upper bounded by $O(\sqrt{\mathrm{Var}(X)/n})$

Challenging the empirical mean and empirical variance, O. Catoni, ***Annales de l'Institut Henri Poincaré*** 2012

Heavy-Tail and Truncated Risk Minimization

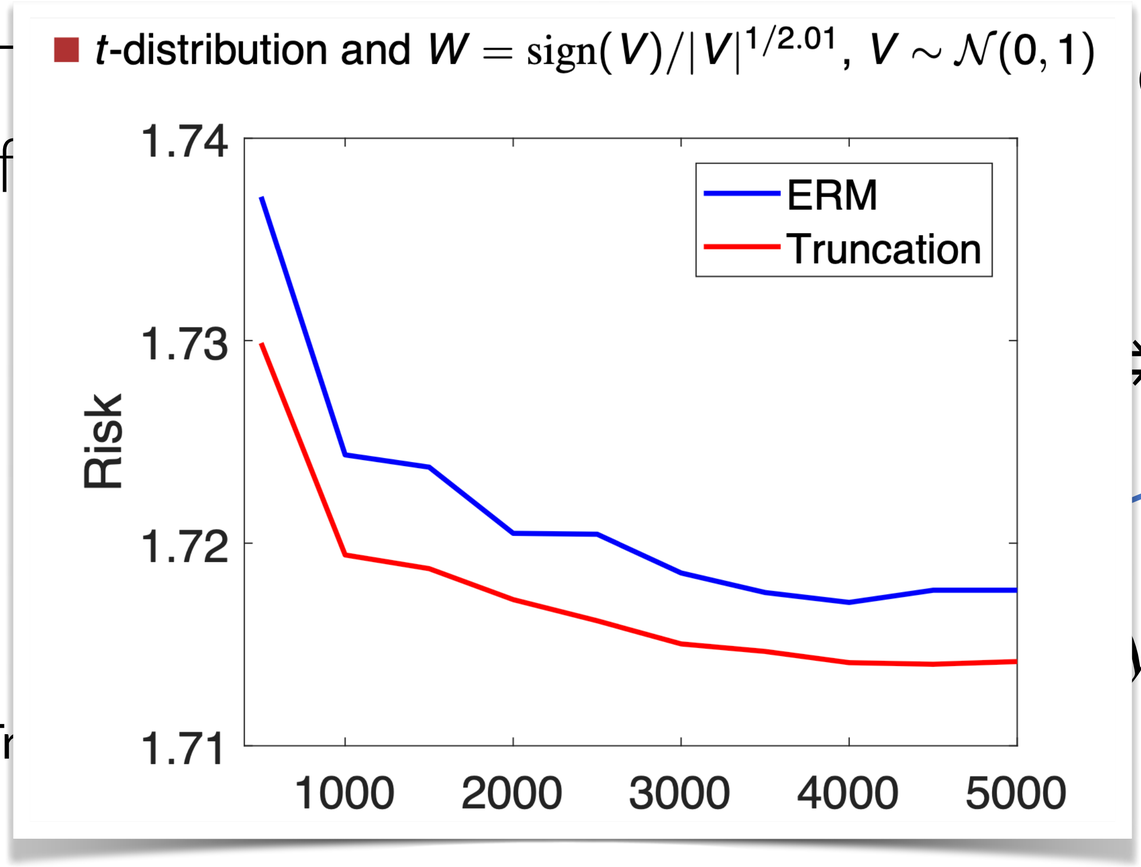
- Truncated risk is a good approximation of risk when the data distribution follows **heavy-tail**: for any $h \in \mathscr{H}$,

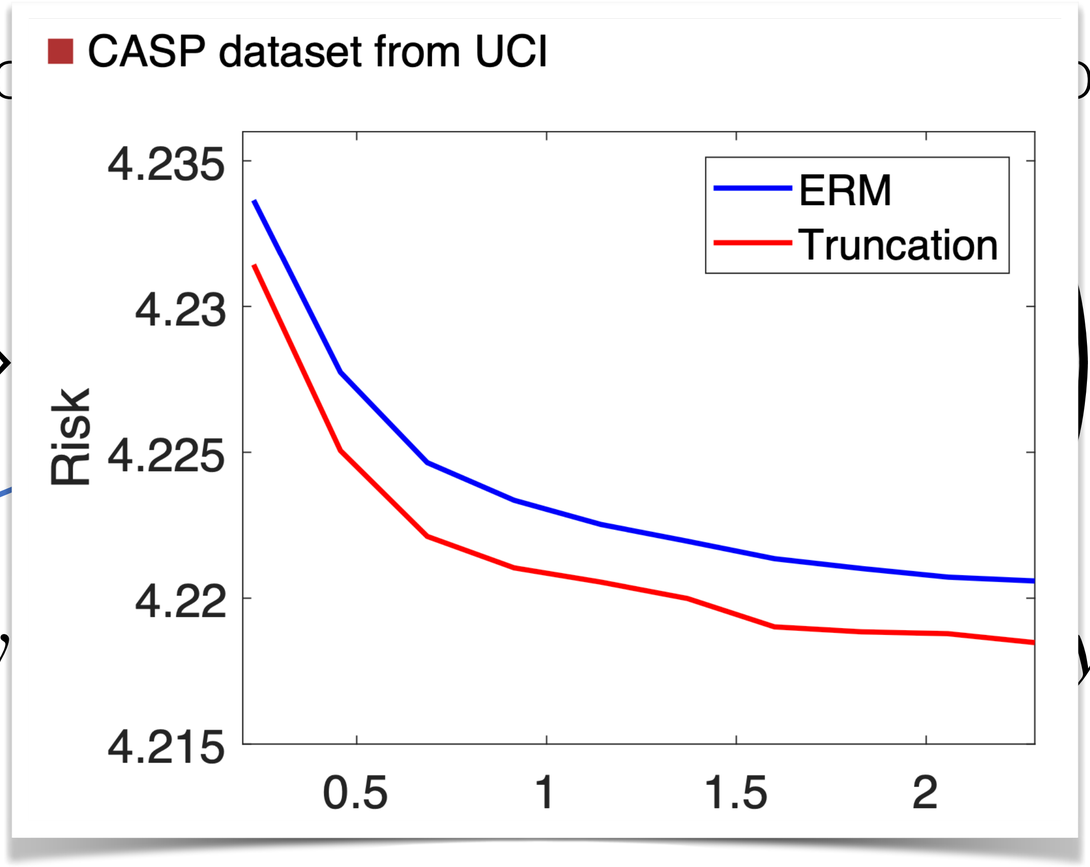
$$\mathbb{P}(X > x) \sim x^{-\alpha} \quad \Rightarrow \quad \left|\widehat{R}_\alpha(h) - R(h)\right| = O\left(\sqrt{\frac{d}{n}}\right)$$

$$\underbrace{\widehat{R}_\alpha(h)}_{\text{Truncated Risk}} = \frac{1}{n\alpha}\sum_{i=1}^{n} \psi(\alpha \ell(h(\mathbf{x}_i), y_i)) \qquad \underbrace{R(h)}_{\text{Risk}} = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(h(\mathbf{x}), y)]$$

ℓ_1-regression with Heavy-Tailed Distributions, Zhang & Zhou, ***NeurIPS*** 2018

Heavy-Tail and Truncated Risk Minimization





ℓ_1-regression with Heavy-Tailed Distributions, Zhang & Zhou, ***NeurIPS*** 2018

Application 1
Bandit with Heavy-Tail Reward

Related Works

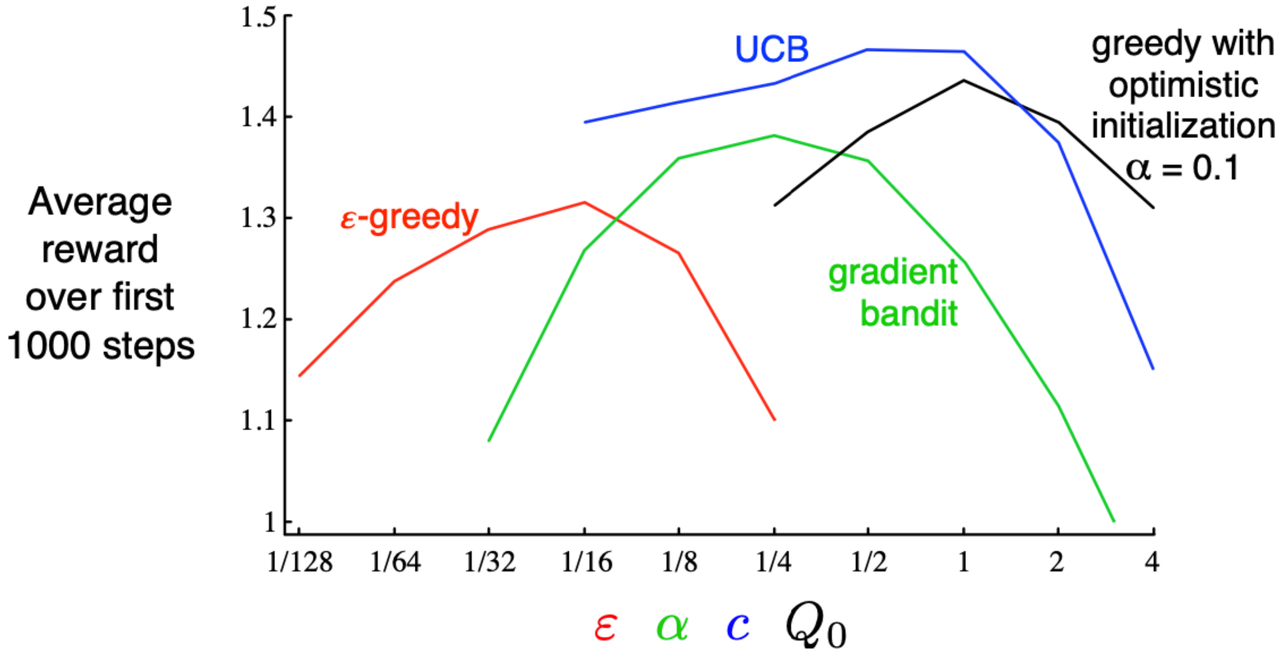
Optimal Algorithms for Stochastic Multi-Armed Bandits with Heavy-Tailed Rewards, Lee et al., ***NeurIPS*** (2020)

Minimax Optimal Bandits for Heavy Tail Rewards, Lee & Lim, ***TNNLS*** 2022

Bandit with Heavy-Tail Reward

- Upper Confidence Bound (UCB)
- Bubeck et al. (2013) proposes Robust UCB for bandit with heavy-tailed rewards
  - requires robust estimator to obtain the UCB of the action
  - truncated or median of mean
  - Catoni's estimator

$$A_t \doteq \arg\max_a \left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right]$$



Bandits with Heavy Tail, S. Bubeck et al., ***Transactions on Information Theory*** (2013)

Sub-optimality of Robust UCB

- Surprisingly, robust UCB is a ***sub-optimal*** algorithm
  - there is a counterexample for which the lower bound of the regret of robust UCB has a sub-optimal factor $\ln(T)^{1-\frac{1}{p}}$

Theorem 1. *There exists a K-armed stochastic bandit problem for which the regret of robust UCB has the following lower bound, for $T > 25$,*

$$\mathbb{E}[\mathcal{R}_T] \geq \Omega\left((K\ln(T))^{1-1/p}\, T^{1/p}\right). \tag{2}$$

Optimal Algorithms for Stochastic Multi-Armed Bandits with Heavy-Tailed Rewards, ***NeurIPS*** (2020)

Joint work with K. Lee (SNU), H. Yang (UNIST), S. Oh (SNU)

p-Robust estimator

$$b_p = \left(2\left(\frac{2-p}{p-1}\right)^{1-2/p} + \left(\frac{2-p}{p-1}\right)^{2-2/p}\right)^{-p/2}$$

- In general, we can deal with $1 < p \leq 2$

$$-\log\left(1 - x + b_p|x|^p\right) \leq \psi_p(x) \leq \log\left(1 + x + b_p|x|^p\right)$$

$$\hat{Y}_n := \frac{c}{n^{1-\frac{1}{p}}} \sum_{i=1}^{n} \psi_p\left(\frac{y_i}{cn^{1/p}}\right) \quad \Rightarrow \quad \mathbb{P}\left(|\bar{y} - \hat{Y}_n| > \epsilon\right) \leq \exp\left(-\frac{n^{\frac{p-1}{p}}\epsilon}{c} + \frac{b_p m_p}{c^p}\right)$$

- The above result recover Cesa's estimator when $p = 2$.

Optimal Algorithms for Stochastic Multi-Armed Bandits with Heavy-Tailed Rewards, ***NeurIPS*** (2020)

Joint work with K. Lee (SNU), H. Yang (UNIST), S. Oh (SNU)

APE^2 with a p-robust estimator

- FTPL (Follow The Perturbed Leader) method provides an optimal exploration strategy under ***heavy-tail*** assumption
  - for sub-Gaussian, see Kim & Tewari (***NeurIPS*** 2019)
- We employ FTPL instead of the UCB strategies

this is more robust than UCB strategy

Algorithm 1 Adaptively Perturbed Exploration with a p-robust estimator (APE^2)

Require: c, T, and $F^{-1}(y)$

Initialize $\{\hat{r}_{0,a}=0, n_{0,a}=0\}$, select $a_1, \cdots, a_K$ and receive $\mathbf{R}_{1,a_1}, \cdots, \mathbf{R}_{K,a_K}$ once

for $t = K+1, \cdots, T$ **do**

  for $\forall a \in \mathcal{A}$ **do**

    $\beta_{t-1,a} \leftarrow c/\left(n_{t,a}\right)^{1-1/p}$ and $G_{t,a} \leftarrow F^{-1}(u)$ with $u \sim \text{Uniform}(0,1)$

    $\hat{r}_{t-1,a} \leftarrow c/\left(n_{t,a}\right)^{1-1/p} \cdot \sum_{k=1}^{t-1} \mathbb{I}\left[a_k = a\right] \psi_p\left(\mathbf{R}_{k,a}/\left(c \cdot \left(n_{t,a}\right)^{1/p}\right)\right)$

  end for

  Choose $a_t = \arg\max_{a \in \mathcal{A}} \{\hat{r}_{t-1,a} + \beta_{t-1,a} G_{t,a}\}$ and receive $\mathbf{R}_{t,a_t}$

end for

Dist. on G	Prob. Dep. Bnd. $O(\cdot)$	Prob. Indep. Bnd. $O(\cdot)$	Low. Bnd. $\Omega(\cdot)$	Opt. Params.	Opt. Bnd. $\Theta(\cdot)$
Weibull	$\sum_{a \neq a^*} A_{c,\lambda,a} \left(\ln\left(B_{c,a}T\right)\right)^{\frac{p}{k(p-1)}}$	$C_{K,T} \ln\left(K\right)^{\frac{1}{k}}$	$C_{K,T} \ln\left(K\right)$	$k = 1, \lambda \geq 1$	$K^{1-1/p} T^{1/p} \ln\left(K\right)$
Gamma	$\sum_{a \neq a^*} A_{c,\lambda,a} \alpha^{p/(p-1)} \ln\left(B_{c,a}T\right)^{p/(p-1)}$	$C_{K,T} \frac{\ln\left(\alpha K^{1+p/(p-1)}\right)^{p/(p-1)}}{\ln(K)^{\frac{1}{p-1}}}$	$C_{K,T} \ln\left(K\right)$	$\alpha = 1, \lambda \geq 1$	
GEV	$\sum_{a \neq a^*} A_{c,\lambda,a} \ln_\zeta\left(B_{c,a}T\right)^{p/(p-1)}$	$C_{K,T} \frac{\ln_\zeta\left(K^{\frac{2p-1}{p-1}}\right)^{p/(p-1)}}{\ln_\zeta(K)^{\frac{1}{p-1}}}$	$C_{K,T} \ln_\zeta\left(K\right)$	$\zeta = 0, \lambda \geq 1$	
Pareto	$\sum_{a \neq a^*} A_{c,\lambda,a} \left[B_{c,a}T\right]^{\frac{p}{\alpha(p-1)}}$	$C_{K,T} \alpha^{1+\frac{p^2}{\alpha(p-1)^2}} K^{\frac{1}{\alpha(p-1)}}$	$C_{K,T} \alpha K^{\frac{1}{\alpha}}$	$\alpha = \lambda = \ln(K)$	
Fréchet	$\sum_{a \neq a^*} A_{c,\lambda,a} \left[B_{c,a}T\right]^{\frac{p}{\alpha(p-1)}}$	$C_{K,T} \alpha^{1+\frac{p^2}{\alpha(p-1)^2}} K^{\frac{1}{\alpha(p-1)}}$	$C_{K,T} \alpha K^{\frac{1}{\alpha}}$	$\alpha = \lambda = \ln(K)$	

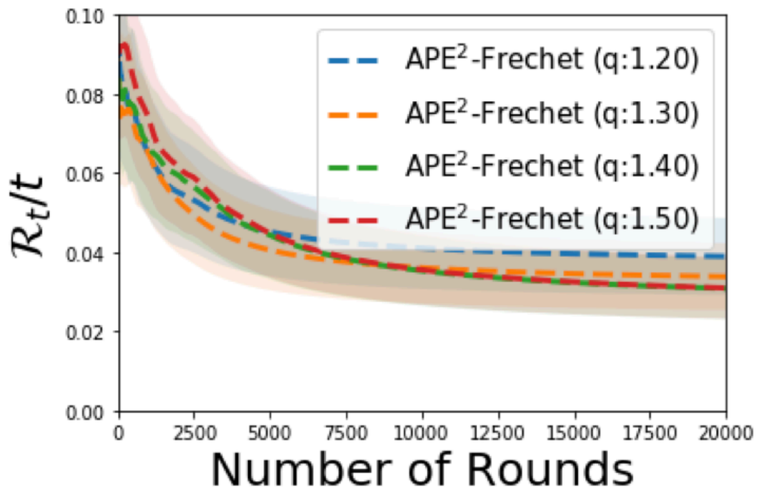
$A_{c,\lambda,a} := \left((3c\lambda)^p/\Delta_a\right)^{\frac{1}{p-1}}$, $B_{c,a} := \left(\Delta_a/c\right)^{p/(p-1)}$, and $C_{K,T} := K^{1-1/p}T^{1/p}$

Optimal Algorithms for Stochastic Multi-Armed Bandits with Heavy-Tailed Rewards, ***NeurIPS*** (2020)

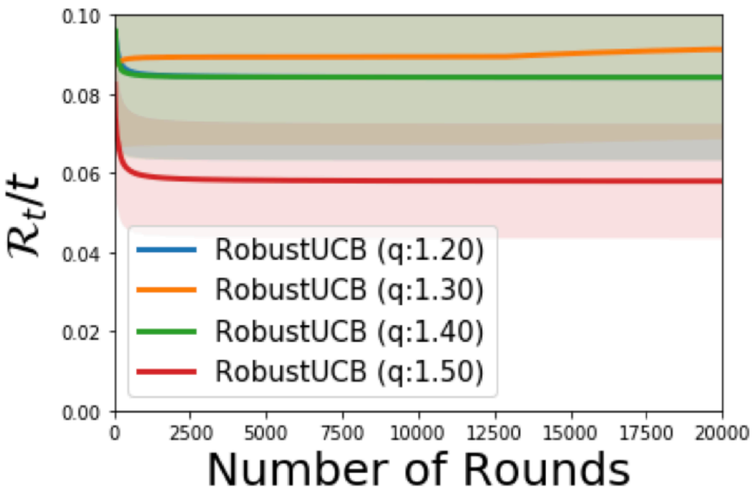
Joint work with K. Lee (SNU), H. Yang (UNIST), S. Oh (SNU)

APE^2 with a p-robust estimator

- FTF
  exp
  - f
- We



(a) APE^2 with Frechet



(b) Robust UCB

Fig. 4. $\mathcal{R}_t/t$ plot with $p = 1.5, \Delta = 0.1$. (a) APE^2 with Frechet perturbation shows a robust performance while (b) the robust UCB is sensitive depending on the choice of q, the moment parameter for the algorithm. Other perturbations show similar tendency.

Algorith

Require:
1: Initial
2: **for** t =
3: **for**
4:
5:
6: **end**
7: Ch
8: **end fo**

re robust
strategy

Opt. Bnd. $\Theta(\cdot)$
$K^{1-1/p}T^{1/p}\ln(K)$
$K^{1-1/p}T^{1/p}$

Optimal Algorithms for Stochastic Multi-Armed Bandits with Heavy-Tailed Rewards, ***NeurIPS*** (2020)

Joint work with K. Lee (SNU), H. Yang (UNIST), S. Oh (SNU)

More Reads

- Robust linear least squares regression, Audibert & Catoni, ***Annals of Statistics***, 2011
- Empirical risk minimization for stochastic convex optimization, Zhang et al., ***COLT***, 2017
- ℓ_1-regression with Heavy-tailed Distributions, Zhang & Zhou, ***NeurIPS*** 2018
- Optimal Algorithms for Stochastic Multi-Armed Bandits with Heavy-Tailed Rewards, Lee et al., ***NeurIPS*** 2020
- No-Regret Reinforcement Learning with Heavy-Tailed Rewards, Zhuang & Sui, ***AISTATS***, 2021
- Minimax Optimal Bandits for Heavy Tail Rewards, Lee & Lim, ***TNNLS*** 2022
- Nearly Optimal Catoni's M-estimator for Infinite Variance, Bhatt et al, ***ICML*** 2022

This paper's result is quite similar to Lee et al., *NeurIPS* 2020

Application 2
Heavy-Tail Noise in Diffusion Model

Related Works

Score-Based Generative Models with Lévy Processes, Yoon et al., ***NeurIPS Workshop on Score-Based Methods*** (2022)

The Era of Diffusion Models

Sprouts in the shape of text 'Imagen' coming out of a fairytale book.

A photo of a Shiba Inu dog with a backpack riding a bike. It is wearing sunglasses and a beach hat.

A high contrast portrait of a very happy fuzzy panda dressed as a chef in a high end kitchen making dough. There is a painting of flowers on the wall behind him.

Teddy bears swimming at the Olympics 400m Butterfly event.

A cute corgi lives in a house made out of sushi.

A cute sloth holding a small treasure chest. A bright golden glow is coming from the chest.

text2image (Imagen, Google)

An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy.

"I won, and I didn't break any rules," the artwork's creator says.

Give this article 1.5K

Théâtre D'opéra Spatial (Stable Diffusion, J. Allen)

Too Ideal Properties of Gaussian Distribution

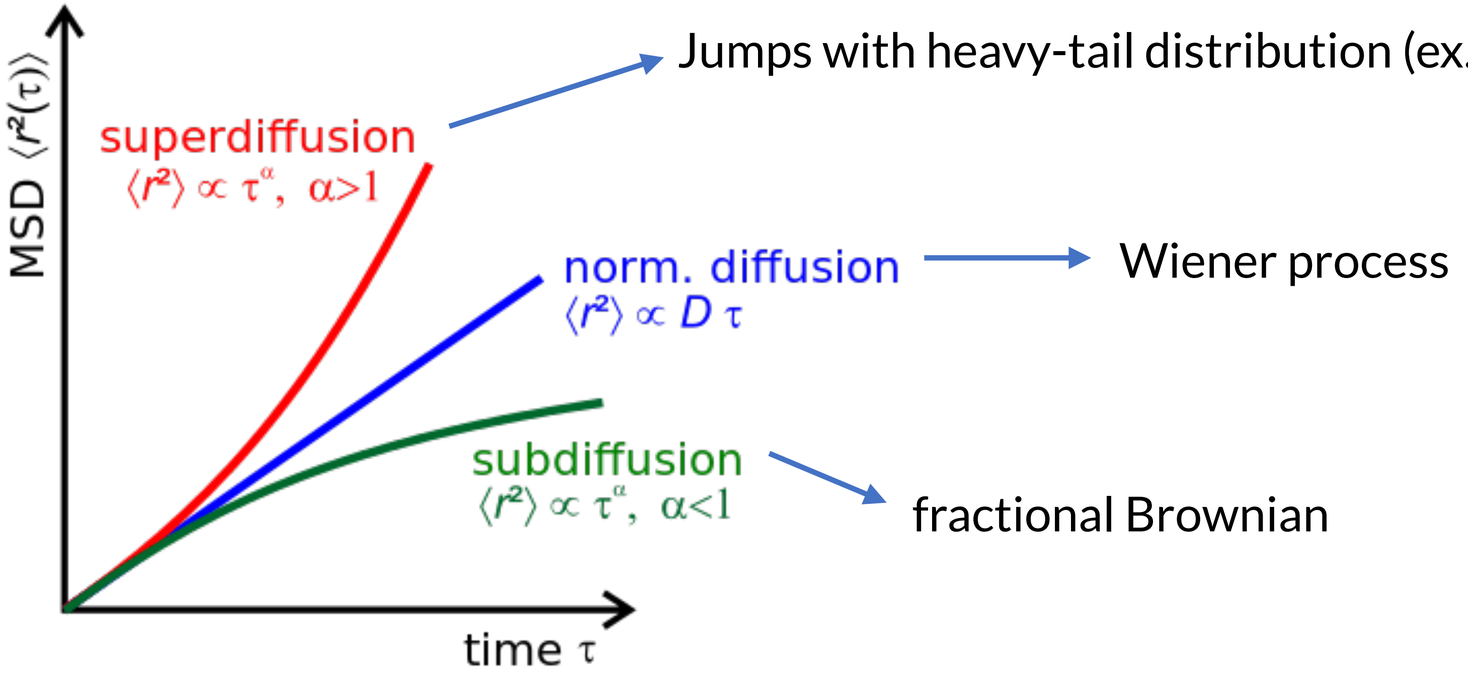
- Density function
  - Affine & Fourier transform of Gaussian
  - Exponential decay at tail → cannot model ***heavy-tail*** behavior
- Wiener process (or Brownian motion)
  - Markov property
  - Continuous trajectory → cannot model ***jump*** movement
  - Fokker-Planck Kolmogorov equations
  - Closed under time reversal

too ideal property
→ does not fit real data!

Anomalous Diffusion

anomalous diffusion is a **wide** class
which overlaps heavy-tail noise

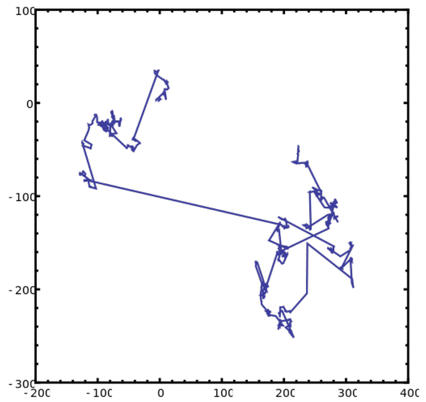
- A diffusion process with a nonlinear mean-square displacement
  - usually described by a power law $\langle r^2(\tau)\rangle = K_\alpha \tau^\alpha$



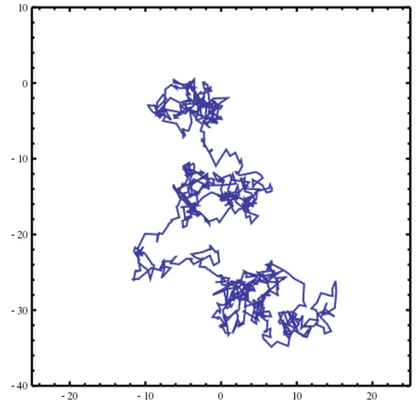
Jumps with heavy-tail distribution (ex. Lévy process)

Wiener process

fractional Brownian

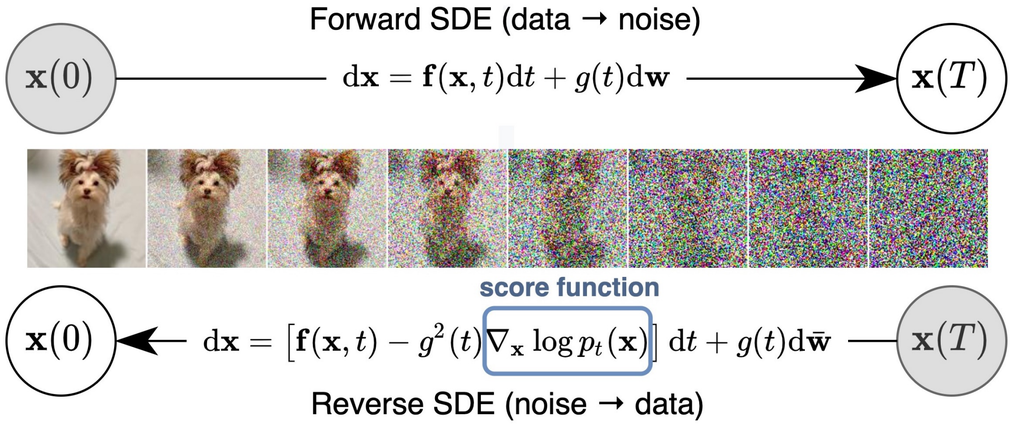


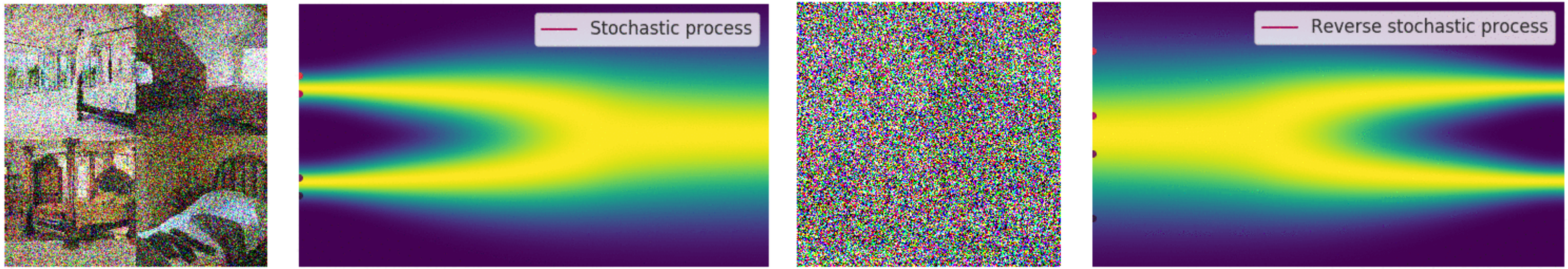
Lévy process



Wiener process

Time Reversal of Diffusion Processes



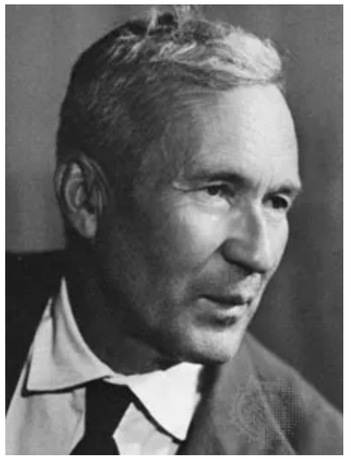


Forward Process

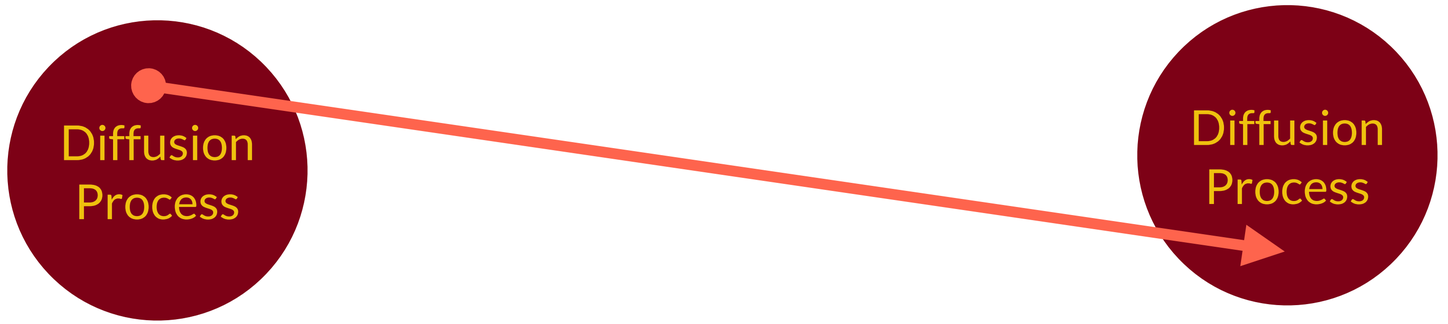
Reverse Process

Score-Based Generative Modeling Through Stochastic Differential Equations, Song et al., ***ICLR*** (2021)

Time Reversal of Diffusion Processes



Andrey N. Kolmogorov

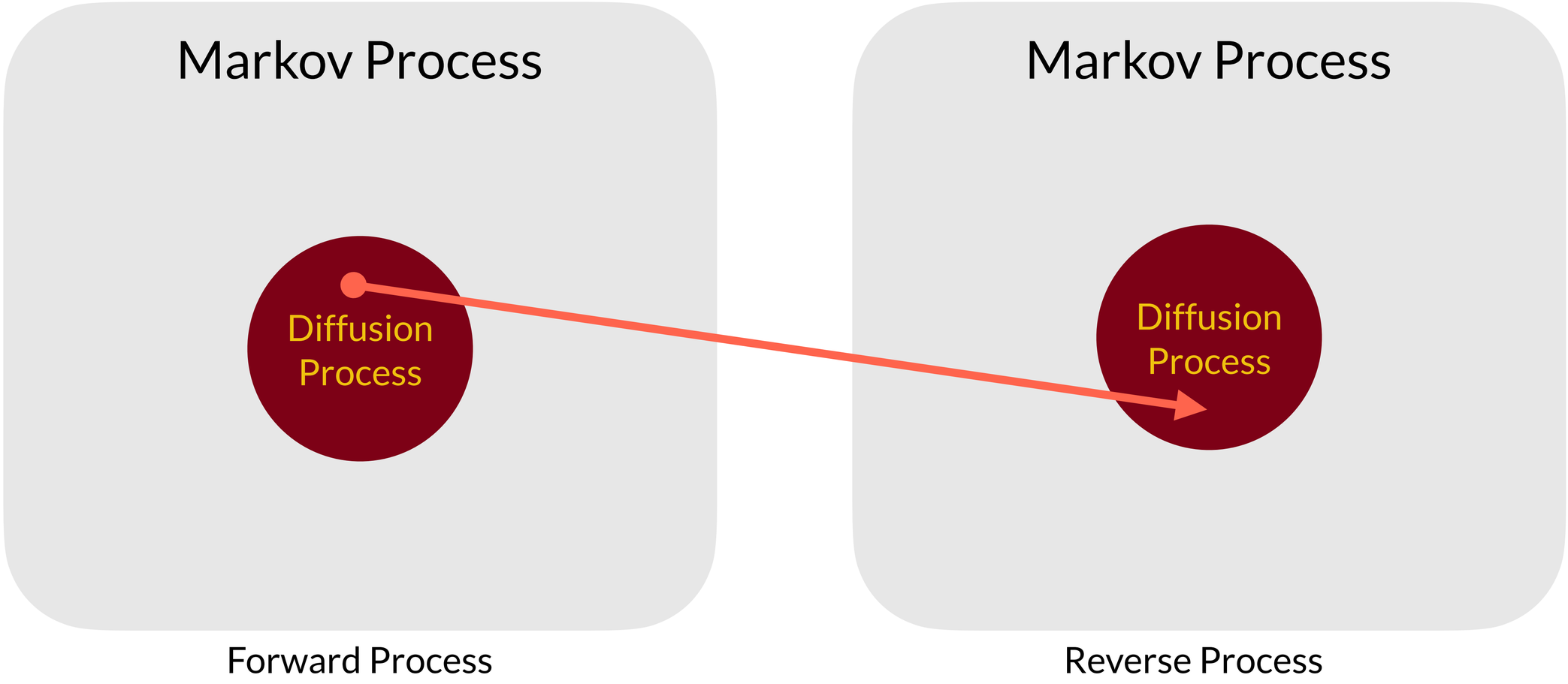


Forward Process

Reverse Process

Zur Unkehrbarkeit der statistischen, A. N. Kolmogorov, ***Math. Ann.*** (1937)

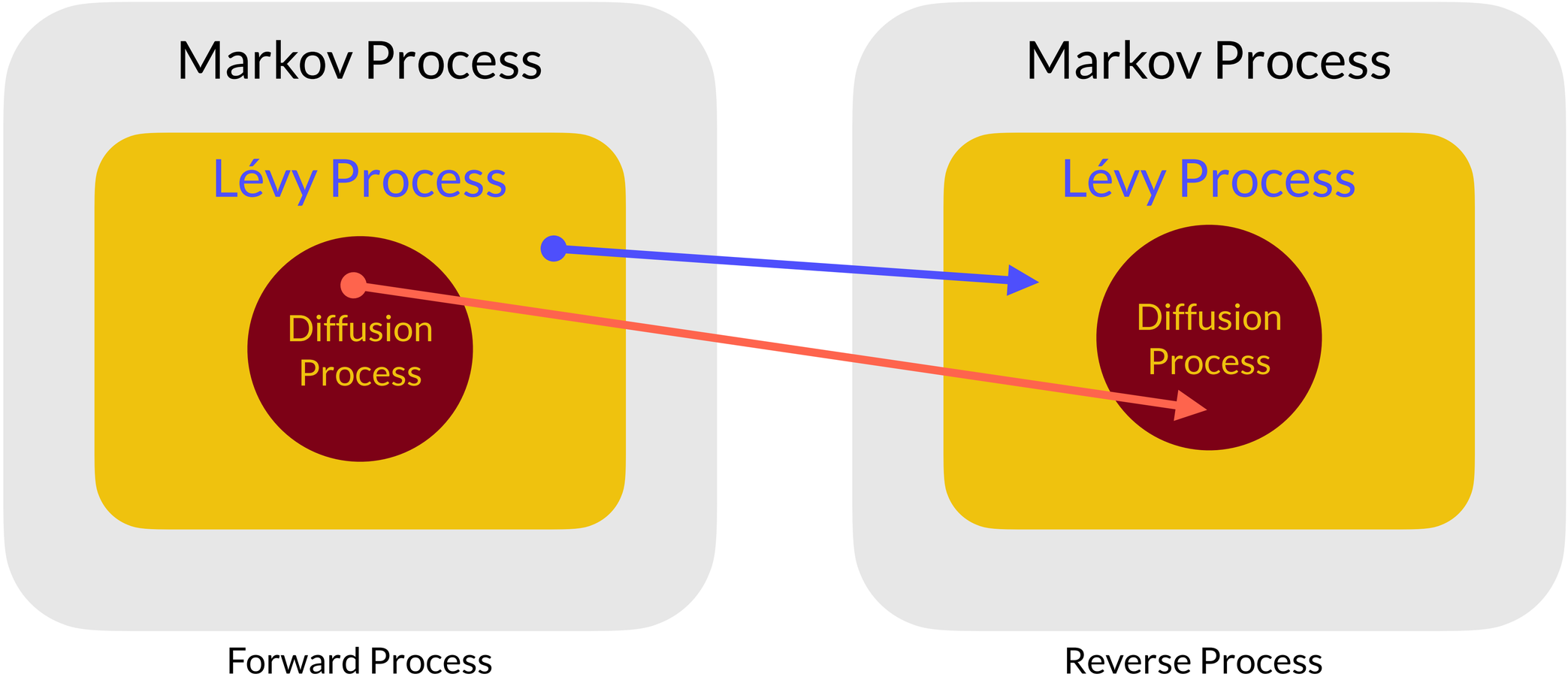
Time Reversal of General Markov Processes



Zur Theorie der Markoffschen Ketten, A. N. Kolmogorov, ***Math. Ann.*** (1936)

Time Reversal of Lévy Processes

Lévy = Diffusion + Jump



Time Reversal on Lévy Processes, Jacod & Protter, ***Annals of Probability*** (1988)

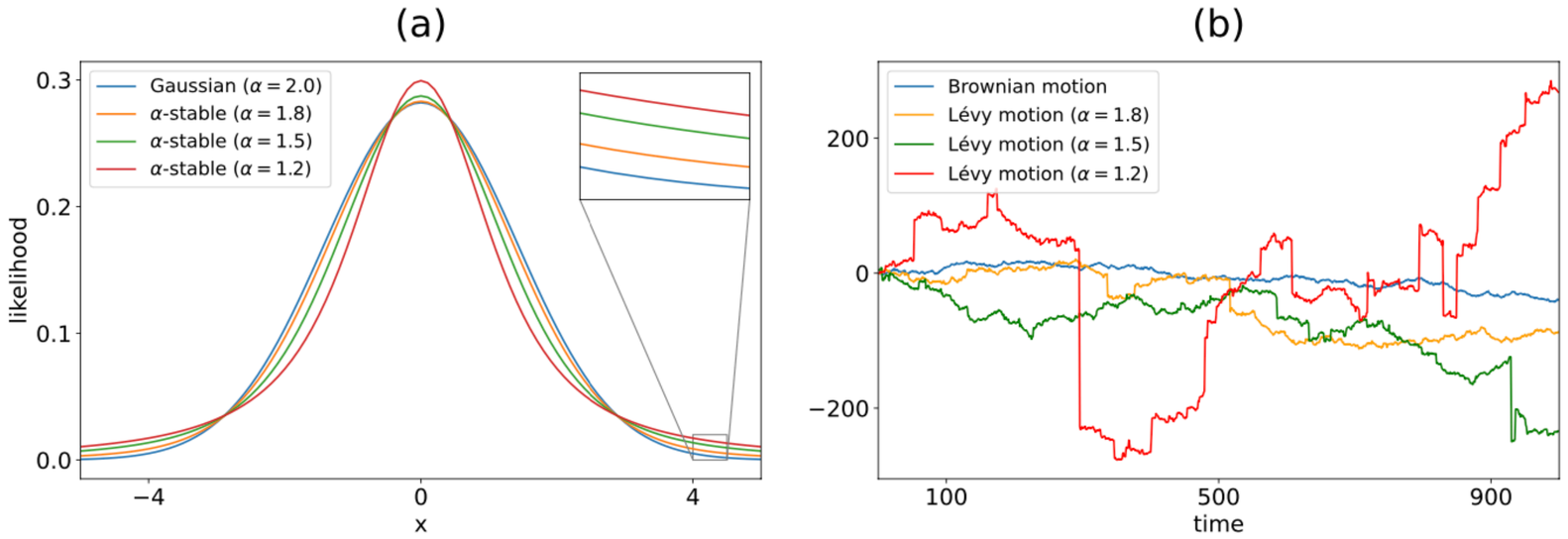
α-Stable Lévy process

Definition

A stochastic process L_t is called Lévy process if it satisfies (i) independent increments, (ii) stationary increments, and (iii) stochastic continuous

$$\mathbb{P}(X > x) \sim x^{-\alpha}$$

α determines the jump behavior of processes



Fractional Langevin Monte Carlo

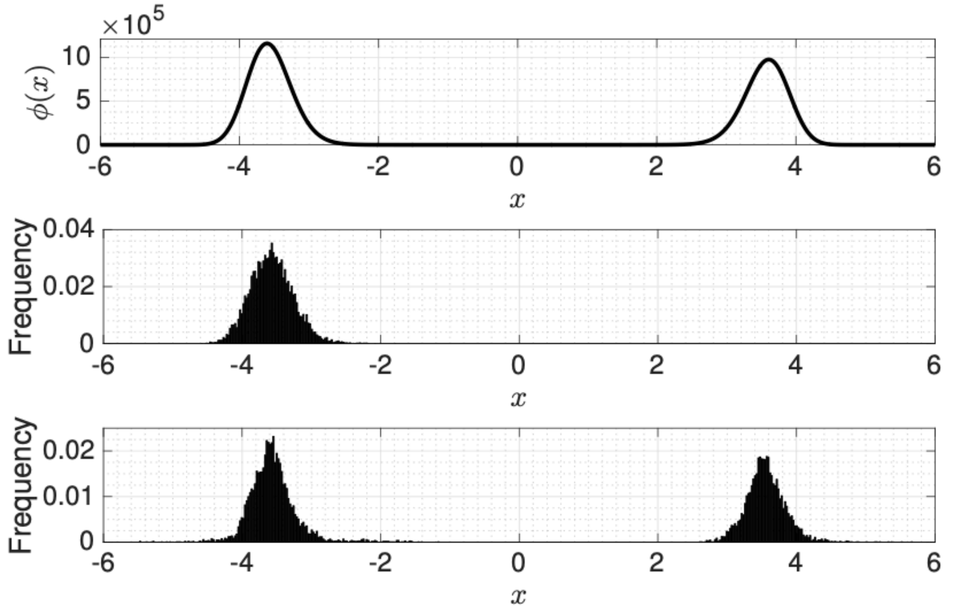
SDEs driven by symmetric stable Lévy processes : In this study, we are interested in SDEs driven by symmetric α-stable Lévy processes, which are defined as follows:

$$dX_t = b(X_{t-}, \alpha)dt + dL_t^\alpha, \tag{3}$$

Theorem 1. *Consider the SDE (3), where b is defined as:*

$$b(x, \alpha) \triangleq \left(\mathcal{D}^{\alpha-2} f_\pi(x)\right)/\phi(x). \tag{5}$$

Here, $f_\pi(x) \triangleq -\phi(x)\partial_x U(x)$ and $\mathcal{D}^\gamma$ is defined in (4). Then, π is an invariant measure of the Markov process $(X_t)_{t\geq 0}$ that is a càdlàg solution of the SDE given in (3). Furthermore, if b is Lipschitz continuous, then π is the unique invariant measure of the process $(X_t)_{t\geq 0}$.



Gaussian

Lévy noise

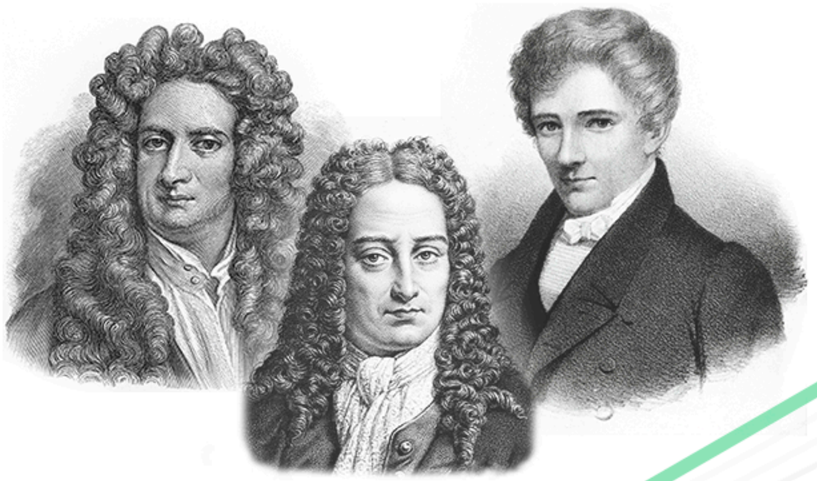
$$\mathcal{D}^\gamma f = \mathcal{F}^{-1}\{|\xi|^\gamma \hat{f}(\xi)\}$$

fractional Riesz derivative

$$\approx \frac{1}{h^\gamma} \sum_{k\in\mathbb{Z}} \frac{(-1)^k \Gamma(\gamma+1)}{\Gamma(\frac{\gamma}{2}-k+1)\Gamma(\frac{\gamma}{2}+k+1)} f(x-kh)$$

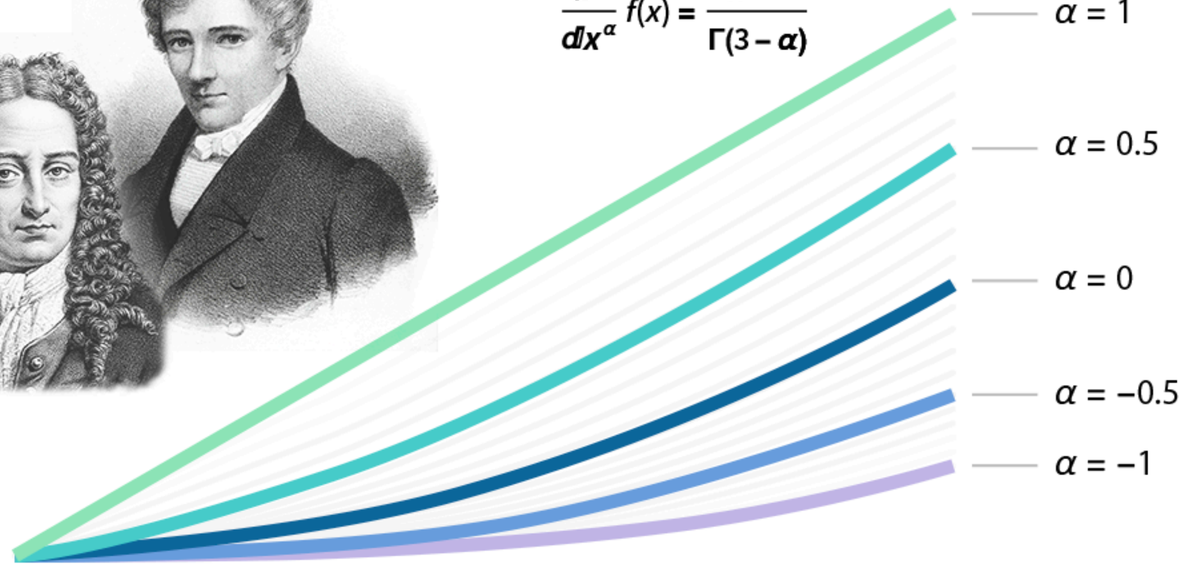
Fractional Langevin Monte Carlo, Umut Simsekli, ***ICML*** (2017)

Fractional Calculus



$$f(x) = \frac{x^2}{2}$$

$$\frac{d^{\alpha}}{dx^{\alpha}} f(x) = \frac{x^{2-\alpha}}{\Gamma(3-\alpha)}$$



Tigran Ishkhanyan, Wolfram Blog

Tautochrone Problem

Time-Reversal of SDEs driven by Lévy Process

- We can derive a time-reversal of SDEs driven by α-stable Lévy processes

$$d\vec{X}_t = b(t, \vec{X}_t)dt + \sigma_B(t)dB_t + \sigma_L(t)dL_t^\alpha,$$

Forward Process

Dual

$$d\overleftarrow{X}_t = \left(b(t, \overleftarrow{X}_t) - \sigma_B^2(t)\nabla_x \log p_t(\overleftarrow{X}_t) - \alpha \cdot \sigma_L^\alpha(t) \frac{\partial_{|x|}^{\alpha-2}\partial_x p_t(\overleftarrow{X}_t)}{p_t(\overleftarrow{X}_t)}\right)dt + \sigma_B(t)d\bar{B}_t + \sigma_L(t)d\bar{L}_t^\alpha.$$

Reverse Process

fractional Riesz derivative

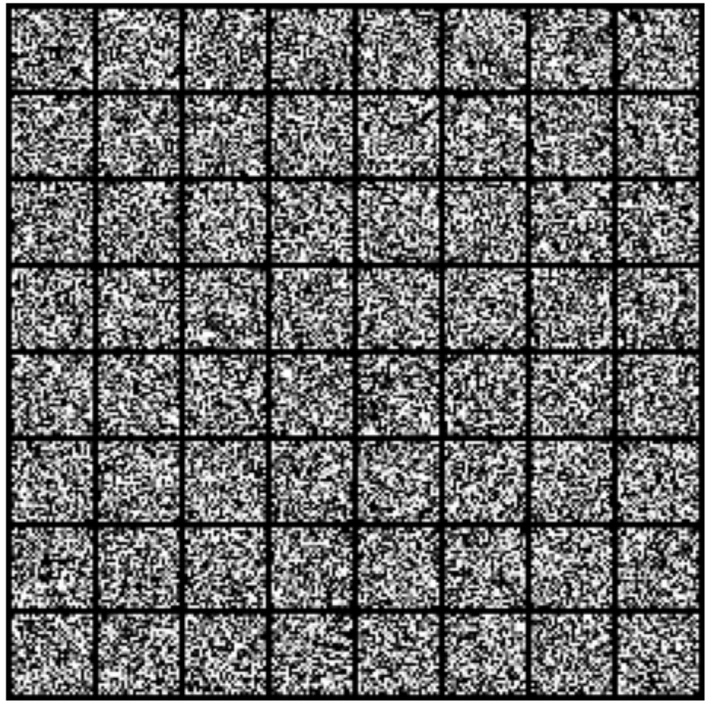
score function

we can also obtain probability ODE

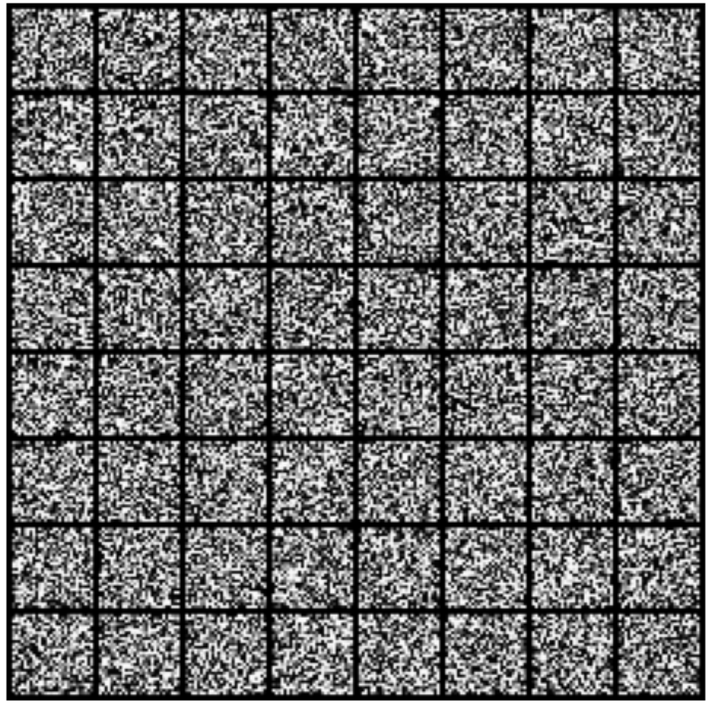
Score-Based Generative Models with Lévy Processes, ***NeurIPS Workshop on Score-Based Methods*** (2022)

Joint work with E. Yoon (UNIST), K. Park (UNIST), J. Kim (UNIST)

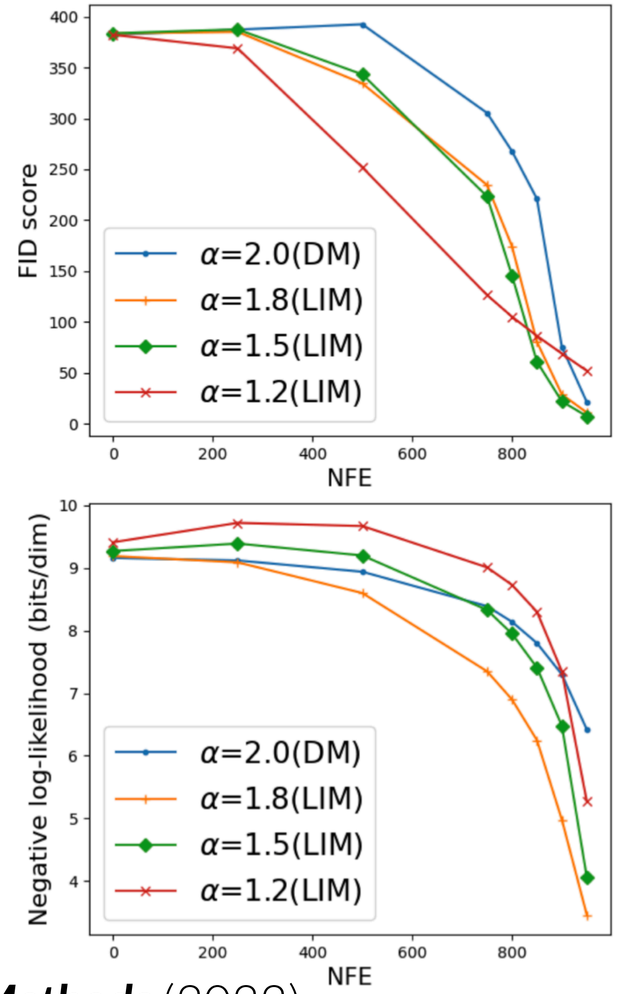
Diffusion Model vs Lévy-Itō Model



Diffusion Model (Song et al., 2021)



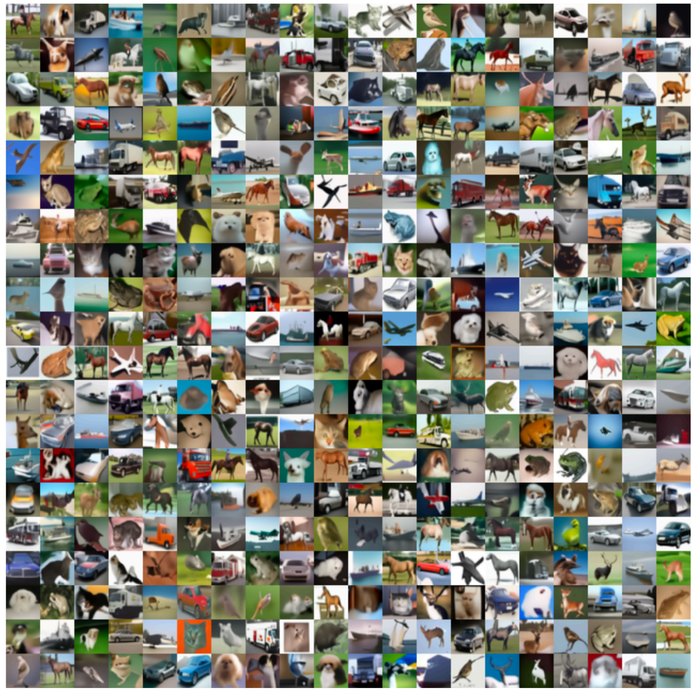
Lévy-Itō Model (Ours)



Score-Based Generative Models with Lévy Processes, ***NeurIPS Workshop on Score-Based Methods*** (2022)

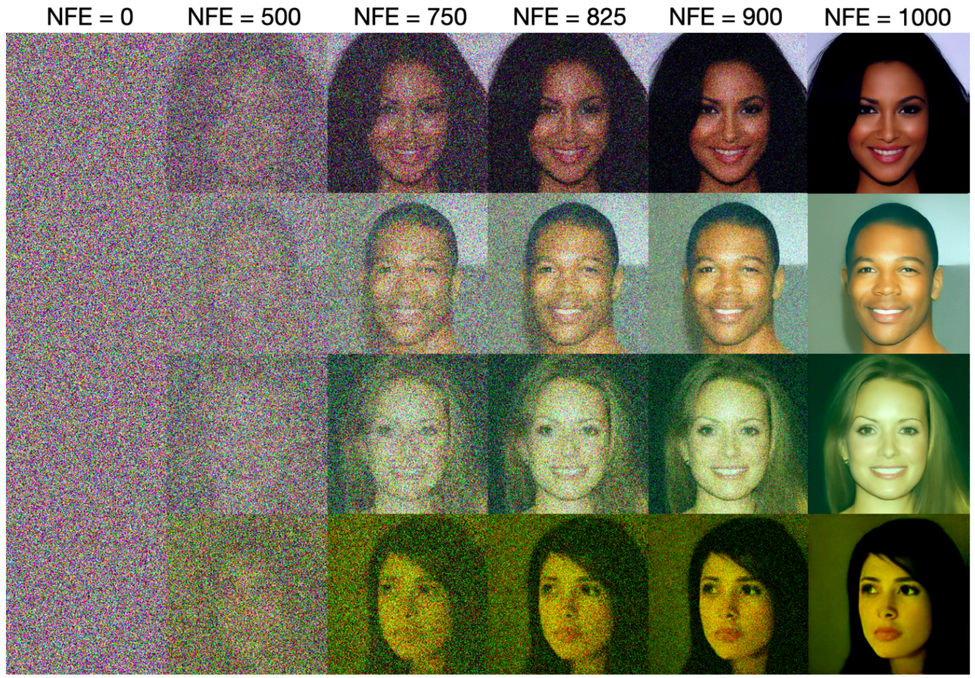
Joint work with E. Yoon (UNIST), K. Park (UNIST), J. Kim (UNIST)

Generated Images by LIM



CIFAR10 (32x32)

CelebA (64x64)



CelebA-HQ (256x256)

Score-Based Generative Models with Lévy Processes, ***NeurIPS Workshop on Score-Based Methods*** (2022)

Joint work with E. Yoon (UNIST), K. Park (UNIST), J. Kim (UNIST)

More Reads

- Time Reversal on Lévy Processes, Jacod & Protter, ***Annals of Probability*** (1988)
- Asymptotic Behaviors of Fundamental Solution and its Derivatives related to Space-Time Fractional Differential Equations, Kim & Lim, ***Journal of KMS***, 2016
- Fractional Langevin Monte Carlo: Exploring Lévy Driven Stochastic Differential Equations for Markov Chain Monte Carlo, Umut Şimşekli, ***ICML*** (2017)
- Heavy-Tailed Denoising Score Matching, Deasy et al., ***arXiv*** (2021)
- First Hitting Diffusion Models for Generating Manifold, Graph and Categorical Data, Ye at al., ***NeurIPS*** 2022
- Score-Based Generative Models with Lévy Processes, Yoon et al., ***NeurIPS Workshop on Score-Based Methods*** (2022)

Summary

- Heavy-tail distributions have interesting behaviors compared to Gaussian
- Special function theory is necessary to deal with heavy-tail
  - theoretically, we need to apply different mathematical tools
  - empirically, we need to implement additional numerical tools
- We can deal with heavy-tail distributions in machine learning problems
  - multi-armed bandits
  - reinforcement learning
  - diffusion models with Lévy processes

Q & A