Heavy Tails in Generative Modeling

HTGAN and Stochastic Interpolants

GAME 2026 Workshop

Jean Pachebat

Based on PhD work at École Polytechnique

12 June 2026

Outline

Generative models

We observe data drawn from an unknown distribution.

$$X_1, \dots, X_N \in \mathbb{R}^d, \qquad \text{i.i.d. samples of } X \sim p_{\text{data}}.$$

Generative models

We observe data drawn from an unknown distribution.

$$X_1, \dots, X_N \in \mathbb{R}^d, \qquad \text{i.i.d. samples of } X \sim p_{\text{data}}.$$

Goal: Build a model that generates new data from the same distribution. Transform simple random noise via a map G_θ:

$$z \sim p_z \quad \text{(noise source, typically gaussian or uniform)},$$
$$X_\theta = G_\theta(z) \quad \text{(generated data)}; \quad X_\theta \sim p_\theta.$$

Match the generated distribution to the data distribution:

$$\min_\theta \ \mathcal{D}\big(p_{\text{data}} \,\|\, p_\theta\big).$$

Generative models

We observe data drawn from an unknown distribution.

$$X_1, \dots, X_N \in \mathbb{R}^d, \qquad \text{i.i.d. samples of } X \sim p_{\text{data}}.$$

Goal: Build a model that generates new data from the same distribution. Transform simple random noise via a map G_θ:

$$z \sim p_z \quad \text{(noise source, typically gaussian or uniform)},$$
$$X_\theta = G_\theta(z) \quad \text{(generated data)}; \quad X_\theta \sim p_\theta.$$

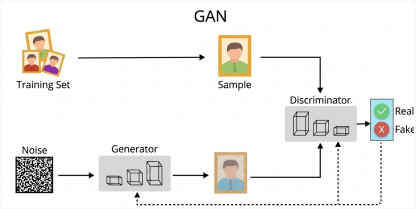
Match the generated distribution to the data distribution:

$$\min_\theta \ \mathcal{D}\big(p_{\text{data}} \,\|\, p_\theta\big).$$

Question: Can generative models (GMs) reproduce marginals and dependence structure for heavy-tailed targets?

Two GMs: GANs and Diffusion

Generative Adversarial Networks (GANs)

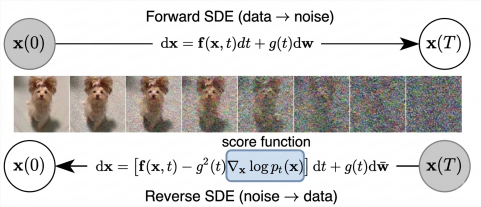


G_θ and D_ϕ play an adversarial game.

Objective: $\min_\theta \max_\phi \mathbb{E}_{x\sim p_{\text{data}}}[\log D_\phi(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D_\phi(G_\theta(z)))]$

Generation: noise $z \sim p_z$ passed through the generator G_θ.

Diffusion Models



Reconstruct the data from noise via the score function.

Objective: learn the score $s_\theta(x_t, t) \approx \nabla_{x_t} \log p_t(x_t)$.

Generation: discretization of the reverse SDE (noise → data).

Linked to the general framework of Stochastic Interpolants:

$$X_t = \alpha_t X_0 + \beta_t X_1. \tag{1}$$

Contributions

1. HTGAN

Question: *With appropriate heavy tailed noise, GANs can reproduce tail behaviour. Can they reproduce extremal dependence with approximation guarantees?*

Contribution: Fréchet noise at the GAN input; approximation properties of the dependence structure.

HTGAN: heavy-tail GAN for multivariate dependent extremes via latent-dimensional control, S. Girard, E. Gobet, J. Pachebat. *IJCM*, 2025.

Contributions

1. HTGAN

Question: *With appropriate heavy tailed noise, GANs can reproduce tail behaviour. Can they reproduce extremal dependence with approximation guarantees?*

Contribution: Fréchet noise at the GAN input; approximation properties of the dependence structure.

HTGAN: heavy-tail GAN for multivariate dependent extremes via latent-dimensional control, S. Girard, E. Gobet, J. Pachebat. *IJCM*, 2025.

2. Log-FM

Question: *Can we adapt diffusion / flow models to produce heavy-tailed data?*

Contribution: flow matching in log-space (soft-log transform) can be seen as continuous tail annealing.

Tail Annealing for Heavy-Tailed Flow Matching, J. Pachebat. *ICML 2026*.

Outline

HTGAN: heavy-tailed GAN

GAN [Goodfellow et al., 2014 ; Arjovsky et al., 2017]**.** A generator G_θ (MLP) transforms noise z into data:

$$X_\theta = G_\theta(z), \qquad G_\theta(z) = W_L \cdot \sigma\big(W_{L-1} \cdot \sigma\big(\cdots \sigma(W_1 z)\big)\big),\ \sigma(x) = x_+ = \max\{0, x\}$$

Observation [Wiese et al., 2020]**.** For large values, $G_\theta(z) \approx \sigma(Az)$. With $z \sim \mathcal{N}(0, I)$: outputs are sub-Gaussian $\Rightarrow$ light-tailed margins.

HTGAN: heavy-tailed GAN

GAN [Goodfellow et al., 2014 ; Arjovsky et al., 2017]**.** A generator G_θ (MLP) transforms noise z into data:

$$X_\theta = G_\theta(z), \qquad G_\theta(z) = W_L \cdot \sigma\big(W_{L-1} \cdot \sigma\big(\cdots \sigma(W_1 z)\big)\big),\ \sigma(x) = x_+ = \max\{0, x\}$$

Observation [Wiese et al., 2020]**.** For large values, $G_\theta(z) \approx \sigma(Az)$. With $z \sim \mathcal{N}(0, I)$: outputs are sub-Gaussian $\Rightarrow$ light-tailed margins.

HTGAN. We do not change the architecture G_θ; we change z:

LTGAN (standard GAN): $z \sim \mathcal{N}(0, I)$ (light tail)

HTGAN: $z = (z_j)_{j=1}^N \overset{\text{iid}}{\sim} \text{Fréchet}(\alpha)$ (heavy tail)

With suitable α, we recover the target marginal tail index.

HTGAN: heavy-tailed GAN

GAN [Goodfellow et al., 2014 ; Arjovsky et al., 2017]**.** A generator G_θ (MLP) transforms noise z into data:

$$X_\theta = G_\theta(z), \qquad G_\theta(z) = W_L \cdot \sigma\big(W_{L-1} \cdot \sigma\big(\cdots \sigma(W_1 z)\big)\big),\ \sigma(x) = x_+ = \max\{0, x\}$$

Observation [Wiese et al., 2020]**.** For large values, $G_\theta(z) \approx \sigma(Az)$. With $z \sim \mathcal{N}(0, I)$: outputs are sub-Gaussian $\Rightarrow$ light-tailed margins.

HTGAN. We do not change the architecture G_θ; we change z:

LTGAN (standard GAN): $z \sim \mathcal{N}(0, I)$ (light tail)

HTGAN: $z = (z_j)_{j=1}^N \overset{\text{iid}}{\sim} \text{Fréchet}(\alpha)$ (heavy tail)

With suitable α, we recover the target marginal tail index.

Question: is extremal dependence structure recovered?

HTGAN: stdf, asymptotic angular measure and D-norm

[Falk, 2019 ; Hüsler & Reiss, 1989]

Definition (Stable tail dependence function). For a RV $X \in \mathbb{R}^d$ with unit Fréchet margins, define:

$$\ell(\mathbf{x}) \;=\; \lim_{t\to\infty} t\,\mathbb{P}\left(\bigcup_{j=1}^{d}\{X_j > t/x_j\}\right), \qquad \mathbf{x}\in\mathbb{R}_+^d.$$

HTGAN: stdf, asymptotic angular measure and D-norm

[Falk, 2019 ; Hüsler & Reiss, 1989]

Definition (Stable tail dependence function). For a RV $X \in \mathbb{R}^d$ with unit Fréchet margins, define:

$$\ell(\mathbf{x}) \;=\; \lim_{t\to\infty} \; t\,\mathbb{P}\left(\bigcup_{j=1}^{d}\{X_j > t/x_j\}\right), \qquad \mathbf{x} \in \mathbb{R}_+^d.$$

Spectral representation of the stdf [de Haan & Ferreira, 2006, Remark 6.1.16]:

$$\ell(\mathbf{x}) \;=\; \int_{\Delta_{d-1}} \max_j(w_j\, x_j)\; \mathrm{d}\Phi(w)$$

HTGAN: stdf, asymptotic angular measure and D-norm

[Falk, 2019 ; Hüsler & Reiss, 1989]

Definition (Stable tail dependence function). For a RV $X \in \mathbb{R}^d$ with unit Fréchet margins, define:

$$\ell(\mathbf{x}) \;=\; \lim_{t\to\infty} \; t\,\mathbb{P}\left(\bigcup_{j=1}^{d}\{X_j > t/x_j\}\right), \qquad \mathbf{x} \in \mathbb{R}_+^d.$$

Spectral representation of the stdf [de Haan & Ferreira, 2006, Remark 6.1.16]:

$$\ell(\mathbf{x}) \;=\; \int_{\Delta_{d-1}} \max_j(w_j\, x_j)\; \mathrm{d}\Phi(w)$$

D-norm (equivalent formulation):

$$\ell(\mathbf{x}) \;=\; \|\mathbf{x}\|_D \;:=\; \mathbb{E}[\max_j |x_j|\, V_j]$$

where $V \in \mathbb{R}_+^d$ is a *generator* of the D-norm, $V_j \geq 0$ a.s., $\mathbb{E}[V_j] = 1$.

HTGAN: stdf, asymptotic angular measure and D-norm

[Falk, 2019 ; Hüsler & Reiss, 1989]

Definition (Stable tail dependence function). For a RV $X \in \mathbb{R}^d$ with unit Fréchet margins, define:

$$\ell(\mathbf{x}) \;=\; \lim_{t\to\infty} \; t\,\mathbb{P}\left(\bigcup_{j=1}^{d}\{X_j > t/x_j\}\right), \qquad \mathbf{x} \in \mathbb{R}_+^d.$$

Spectral representation of the stdf [de Haan & Ferreira, 2006, Remark 6.1.16]:

$$\ell(\mathbf{x}) \;=\; \int_{\Delta_{d-1}} \max_j(w_j\, x_j)\; \mathrm{d}\Phi(w)$$

D-norm (equivalent formulation):

$$\ell(\mathbf{x}) \;=\; \|\mathbf{x}\|_D \;:=\; \mathbb{E}[\max_j |x_j|\, V_j]$$

where $V \in \mathbb{R}_+^d$ is a *generator* of the D-norm, $V_j \geq 0$ a.s., $\mathbb{E}[V_j] = 1$.

Consequence: approximating the spectral measure Φ / D-norm generator $\Rightarrow$ approximating the stdf.

HTGAN: formalized problem

Set-up. A generator with Fréchet noise input:

$$G_\theta : \mathbb{R}^N \to \mathbb{R}^d, \qquad z = (z_1, \dots, z_N), \;\; z_j \overset{\text{iid}}{\sim} \text{Fréchet}(\alpha)$$

The output $G_\theta(z) \in \mathbb{R}^d$ is a random vector whose *extremal dependence* is characterised by its *induced* stdf:

$$\ell_{G_\theta(z)}(\mathbf{x}) \;=\; \int_{\Delta_{d-1}} \max_j \big(w_j \, x_j\big) \, \mathrm{d}\Phi_{G_\theta(z)}(w)$$

HTGAN: formalized problem

Set-up. A generator with Fréchet noise input:

$$G_\theta : \mathbb{R}^N \to \mathbb{R}^d, \qquad z = (z_1, \dots, z_N), \;\; z_j \overset{\text{iid}}{\sim} \text{Fréchet}(\alpha)$$

The output $G_\theta(z) \in \mathbb{R}^d$ is a random vector whose *extremal dependence* is characterised by its *induced* stdf:

$$\ell_{G_\theta(z)}(\mathbf{x}) \;=\; \int_{\Delta_{d-1}} \max_j \big(w_j \, x_j\big) \, \mathrm{d}\Phi_{G_\theta(z)}(w)$$

Target. A RV $X \in \mathbb{R}^d$, with stdf ℓ_X.

HTGAN: formalized problem

Set-up. A generator with Fréchet noise input:

$$G_\theta : \mathbb{R}^N \to \mathbb{R}^d, \qquad z = (z_1, \dots, z_N), \;\; z_j \overset{\text{iid}}{\sim} \text{Fréchet}(\alpha)$$

The output $G_\theta(z) \in \mathbb{R}^d$ is a random vector whose *extremal dependence* is characterised by its *induced* stdf:

$$\ell_{G_\theta(z)}(\mathbf{x}) \;=\; \int_{\Delta_{d-1}} \max_j \big(w_j \, x_j\big) \, \mathrm{d}\Phi_{G_\theta(z)}(w)$$

Target. A RV $X \in \mathbb{R}^d$, with stdf ℓ_X.

Question. Can we choose θ and N ($= \dim(z)$) such that $\ell_{G_\theta(z)}$ approximates ℓ_X? With what error, and what rate in N?

$$\inf_\theta \sup_{\mathbf{x}} \frac{|\ell_X(\mathbf{x}) - \ell_{G_\theta}(\mathbf{x})|}{\|\mathbf{x}\|_\infty} \;\leq ?\; (N,\, d)$$

HTGAN: approximation results

Proposition. $z \in \mathbb{R}^N$, $z_j \overset{\text{iid}}{\sim} \text{Fréchet}(\alpha)$, G_θ ReLU $\implies \Phi_{G_\theta(z)} = \sum_{k=1}^{N} c_k \, \delta_{w_k}$.

HTGAN: approximation results

Proposition. $z \in \mathbb{R}^N$, $z_j \overset{\text{iid}}{\sim}$ Fréchet(α), G_θ ReLU $\implies \Phi_{G_\theta(z)} = \sum_{k=1}^{N} c_k \, \delta_{w_k}$.

Proof sketch.

- i.i.d. heavy-tailed noise has a **discrete** asymptotic angular measure (N atoms, concentrated on the axes);
- $G_\theta(z)$ is asymptotically linear, so its spectral measure is supported on at most N atoms.

HTGAN: approximation results

Proposition. $z \in \mathbb{R}^N$, $z_j \overset{\text{iid}}{\sim}$ Fréchet(α), G_θ ReLU $\implies \Phi_{G_\theta(z)} = \sum_{k=1}^{N} c_k \, \delta_{w_k}$.

Proof sketch.

- i.i.d. heavy-tailed noise has a **discrete** asymptotic angular measure (N atoms, concentrated on the axes);
- $G_\theta(z)$ is asymptotically linear, so its spectral measure is supported on at most N atoms.

Theorem. For every stdf ℓ_X in dimension d and every $N \in \mathbb{N}^*$ ($N = \dim(z)$):

$$\inf_\theta \sup_{\mathbf{x}} \frac{|\ell_X(\mathbf{x}) - \ell_{G_\theta}(\mathbf{x})|}{\|\mathbf{x}\|_\infty} \lesssim \frac{d^{1/(d-1)}}{2} N^{-1/(d-1)} \underset{d\to\infty}{\sim} \frac{1}{2} N^{-1/(d-1)}.$$

HTGAN: approximation results

Proposition. $z \in \mathbb{R}^N$, $z_j \overset{\text{iid}}{\sim}$ Fréchet(α), G_θ ReLU $\implies \Phi_{G_\theta(z)} = \sum_{k=1}^{N} c_k \, \delta_{w_k}$.

Proof sketch.

- i.i.d. heavy-tailed noise has a **discrete** asymptotic angular measure (N atoms, concentrated on the axes);
- $G_\theta(z)$ is asymptotically linear, so its spectral measure is supported on at most N atoms.

Theorem. For every stdf ℓ_X in dimension d and every $N \in \mathbb{N}^*$ $(N = \dim(z))$:

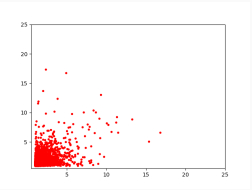
$$\inf_\theta \sup_{\mathbf{x}} \frac{|\ell_X(\mathbf{x}) - \ell_{G_\theta}(\mathbf{x})|}{\|\mathbf{x}\|_\infty} \lesssim \frac{d^{1/(d-1)}}{2} N^{-1/(d-1)} \underset{d\to\infty}{\sim} \frac{1}{2} N^{-1/(d-1)}.$$

Proof sketch.

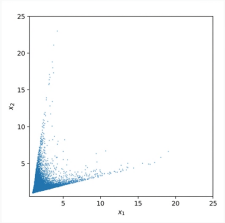
- every stdf is characterised by its asymptotic angular measure on the simplex Δ_{d-1};
- approximating the stdf $\equiv$ **quantizing** this measure by a discrete measure with N atoms [Graf & Luschgy, 2000];
- the rate $N^{-1/(d-1)}$ is the optimal rate of a **covering** of Δ_{d-1} by N points.

HTGAN: approximation, scatter plots

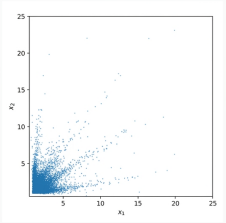
Increasing the latent noise dimension N improves the approximation of the extremal dependence.



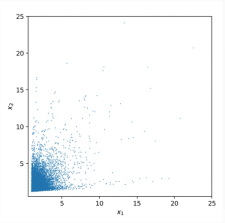
real data



$N = 2$



$N = 10$



$N = 50$

Samples of trained G_θ for different latent noise dimensions $N = \dim(z)$.

HTGAN: experimental set-up

Comparison

HTGAN (Fréchet) vs LTGAN (Gaussian)
Same architecture; only the noise differs.

Metrics (thresholds $\xi \in \{90, 95, 99\}\%$)

AKE dependence
SWD marginal quality

Baselines

LTGAN: Gaussian noise
HTGAN: Fréchet noise

Synthetic data

Copulas: Gumbel ($\beta \in \{4/3, 2, 4\}$), Hüsler-Reiß, Gaussian
(β: dependence parameter, $\beta \uparrow \Rightarrow$ stronger dep.)
$d \in \{2, 5, 10, 20, 50\}$, $\alpha \in \{1.5, 2.0, 2.5\}$

Real data

S&P 500, Financials ($d=27$) and Utilities ($d=57$) sectors.
Normalised returns: $r_t = (s_t - s_{t-1})/s_t$ (s_t: price). Supposed iid.

HTGAN: synthetic results (Gumbel copula)

% of parametrizations where **HTGAN outperforms LTGAN**, averaged over $d \in \{2, 5, 10, 20, 50\}$.

	$\beta = 4/3$			$\beta = 2$			$\beta = 4$		
α	1.5	2.0	2.5	1.5	2.0	2.5	1.5	2.0	2.5
AKE_{95}	65.6	69.5	66.4	79.7	79.4	69.2	87.8	87.2	75.1
SWD_{95}	84.5	74.3	51.1	76.9	66.7	54.3	62.0	52.2	39.5

AKE_{95}, SWD_{95} : threshold 95%. β : Gumbel dependence parameter ($\beta \uparrow \Rightarrow$ stronger dependence). α : Pareto tail index ($\alpha \downarrow \Rightarrow$ heavier tail).

HTGAN: synthetic results (Gumbel copula)

% of parametrizations where **HTGAN outperforms LTGAN**, averaged over $d \in \{2, 5, 10, 20, 50\}$.

	$\beta = 4/3$			$\beta = 2$			$\beta = 4$		
α	1.5	2.0	2.5	1.5	2.0	2.5	1.5	2.0	2.5
AKE_{95}	65.6	69.5	66.4	79.7	79.4	69.2	87.8	87.2	75.1
SWD_{95}	84.5	74.3	51.1	76.9	66.7	54.3	62.0	52.2	39.5

AKE_{95}, SWD_{95} : threshold 95%. β : Gumbel dependence parameter ($\beta \uparrow \Rightarrow$ stronger dependence). α : Pareto tail index ($\alpha \downarrow \Rightarrow$ heavier tail).

HTGAN dominates on heavy tails ($\alpha = 1.5$) and strong dependence ($\beta = 4$): up to $\sim \mathbf{88}\%$ win rate on AKE. Advantage is weaker for lighter tails.

HTGAN: influence of the data dimension

Question. Does the optimal latent dimension $N^\star$ scale with the data dimension d ?

HTGAN: influence of the data dimension

Question. Does the optimal latent dimension $N^\star$ scale with the data dimension d ?

Set-up. Mean N over the top 5% HTGAN runs (by SWD_{90}), averaged over $\beta \in \{4/3, 2, 4\}$.

d	$\alpha = 1.5$ (heaviest tails)	$\alpha = 2.0$	$\alpha = 2.5$ (lightest tails)
2	76	41	39
5	91	56	50
10	91	57	57
20	91	65	73
50	108	69	58

HTGAN: influence of the data dimension

Question. Does the optimal latent dimension $N^\star$ scale with the data dimension d ?

Set-up. Mean N over the top 5% HTGAN runs (by SWD_{90}), averaged over $\beta \in \{4/3, 2, 4\}$.

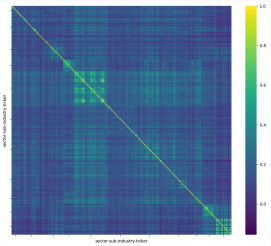
d	$\alpha = 1.5$ (heaviest tails)	$\alpha = 2.0$	$\alpha = 2.5$ (lightest tails)
2	76	41	39
5	91	56	50
10	91	57	57
20	91	65	73
50	108	69	58

Observation. $N^\star$ **grows with** d, particularly for heavy tails ($\alpha = 1.5$: $76 \to 108$).

HTGAN: S&P 500 results

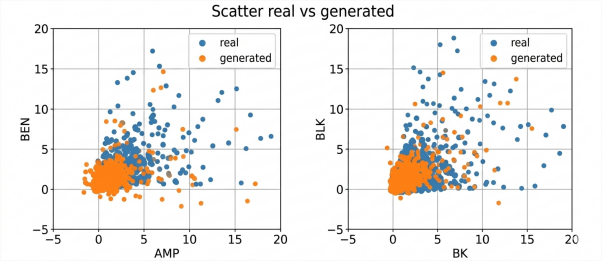
S&P 500, Financials ($d=27$) and Utilities ($d=57$) sectors.

Kendall τ



	Financials	Utilities
AKE (%)	**78**	65
SWD (%)	71	**89**

% of configurations where HTGAN > LTGAN



HTGAN: Summary

Theory. i.i.d. Fréchet noise $(z_i)_{i=1}^N$ at the input of a ReLU GAN $\Rightarrow$ approximation of the target stdf at rate $N^{-1/(d-1)}$.

HTGAN: Summary

Theory. i.i.d. Fréchet noise $(z_i)_{i=1}^N$ at the input of a ReLU GAN $\Rightarrow$ approximation of the target stdf at rate $N^{-1/(d-1)}$.

Experiments. HTGAN outperforms LTGAN on AKE (up to 89% of configurations). Validated on S&P 500.

HTGAN: Summary

Theory. i.i.d. Fréchet noise $(z_i)_{i=1}^N$ at the input of a ReLU GAN $\Rightarrow$ approximation of the target stdf at rate $N^{-1/(d-1)}$.

Experiments. HTGAN outperforms LTGAN on AKE (up to 89% of configurations). Validated on S&P 500.

Limitation. Curse of dimensionality: for a fixed stdf error ε, $N \sim \varepsilon^{-(d-1)}$.

HTGAN: Summary

Theory. i.i.d. Fréchet noise $(z_i)_{i=1}^N$ at the input of a ReLU GAN $\Rightarrow$ approximation of the target stdf at rate $N^{-1/(d-1)}$.

Experiments. HTGAN outperforms LTGAN on AKE (up to 89% of configurations). Validated on S&P 500.

Limitation. Curse of dimensionality: for a fixed stdf error ε, $N \sim \varepsilon^{-(d-1)}$.

Perspectives.

- *Replicate methodology in other models*: induced stdf approximation for heavy-tailed diffusion [Shariatian+24; Pandey+24], Wasserstein-Aitchison GAN [Lhaut, Rootzén & Segers, 2026].
- *Extend the i.i.d. setting*: sequential structure for autocorrelated returns [TimeGAN, Yoon et al., 2019; Quant GAN, Wiese et al., 2020].

Outline

Stochastic interpolants & flow matching

Interpolant [Albergo & Vanden-Eijnden, 2023 ; Lipman et al., 2023]**.** Bridge data X_0 and Gaussian noise $X_1 \sim \mathcal{N}(0, I_d)$, a modern view of diffusion models:

$$X_t = \alpha_t X_0 + \beta_t X_1;$$
$$(\alpha_0, \beta_0) = (1, 0), (\alpha_1, \beta_1) = (0, 1), \quad \alpha_t \text{ decreasing, } \beta_t \text{ increasing.}$$

Stochastic interpolants & flow matching

Interpolant [Albergo & Vanden-Eijnden, 2023 ; Lipman et al., 2023]**.** Bridge data X_0 and Gaussian noise $X_1 \sim \mathcal{N}(0, I_d)$, a modern view of diffusion models:

$$X_t = \alpha_t X_0 + \beta_t X_1;$$
$$(\alpha_0, \beta_0) = (1, 0), (\alpha_1, \beta_1) = (0, 1), \quad \alpha_t \text{ decreasing, } \beta_t \text{ increasing.}$$

Transport. Defining the interpolant fixes the marginals p_t of X_t; they are realised by the probability-flow ODE

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = v_t(X_t), \qquad v_t(x) := \mathbb{E}\big[\dot{X}_t \mid X_t = x\big] = \mathbb{E}\big[\dot{\alpha}_t X_0 + \dot{\beta}_t X_1 \mid X_t = x\big].$$

Stochastic interpolants & flow matching

Interpolant [Albergo & Vanden-Eijnden, 2023 ; Lipman et al., 2023]**.** Bridge data X_0 and Gaussian noise $X_1 \sim \mathcal{N}(0, I_d)$, a modern view of diffusion models:

$$X_t = \alpha_t X_0 + \beta_t X_1;$$
$$(\alpha_0, \beta_0) = (1, 0), (\alpha_1, \beta_1) = (0, 1), \quad \alpha_t \text{ decreasing, } \beta_t \text{ increasing.}$$

Transport. Defining the interpolant fixes the marginals p_t of X_t; they are realised by the probability-flow ODE

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = v_t(X_t), \qquad v_t(x) := \mathbb{E}\big[\dot{X}_t \mid X_t = x\big] = \mathbb{E}\big[\dot{\alpha}_t X_0 + \dot{\beta}_t X_1 \mid X_t = x\big].$$

Training. v_t is intractable, but it is the minimiser of the regression

$$\mathcal{L}(\theta) = \mathbb{E}_{t,\, X_0,\, X_1} \left\| v_\theta(X_t, t) - (\dot{\alpha}_t X_0 + \dot{\beta}_t X_1) \right\|^2.$$

This objective function is trained by gradient descent on the network parameters.

Stochastic interpolants & flow matching

Interpolant [Albergo & Vanden-Eijnden, 2023 ; Lipman et al., 2023]**.** Bridge data X_0 and Gaussian noise $X_1 \sim \mathcal{N}(0, I_d)$, a modern view of diffusion models:

$$X_t = \alpha_t X_0 + \beta_t X_1;$$
$$(\alpha_0, \beta_0) = (1, 0), (\alpha_1, \beta_1) = (0, 1), \quad \alpha_t \text{ decreasing, } \beta_t \text{ increasing.}$$

Transport. Defining the interpolant fixes the marginals p_t of X_t; they are realised by the probability-flow ODE

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = v_t(X_t), \qquad v_t(x) := \mathbb{E}\big[\dot{X}_t \mid X_t = x\big] = \mathbb{E}\big[\dot{\alpha}_t X_0 + \dot{\beta}_t X_1 \mid X_t = x\big].$$

Training. v_t is intractable, but it is the minimiser of the regression

$$\mathcal{L}(\theta) = \mathbb{E}_{t,\, X_0,\, X_1} \left\| v_\theta(X_t, t) - (\dot{\alpha}_t X_0 + \dot{\beta}_t X_1) \right\|^2.$$

This objective function is trained by gradient descent on the network parameters.
Sampling. Draw $X_1 \sim \mathcal{N}(0, I_d)$ and integrate $\dot{X}_t = v_\theta(X_t, t)$ backward, $t : 1 \to 0$.

Why log-space?

Problem.

$$X_t = \alpha_t X_0 + \beta_t X_1$$

Tail index is that of the heaviest component. X_t keeps X_0's heavy tail for **every** $\alpha_t > 0$, and is light only at $t = 1$:

$$\text{tail index of } X_t \;=\; \begin{cases} \text{heavy } (X_0\text{'s}) & \text{for } \alpha_t > 0, \\ \text{light } (X_1\text{'s}) & \text{for } \alpha_t = 0 \end{cases} \quad \Rightarrow \quad \text{discontinuous in } t.$$

Why log-space?

Problem.

$$X_t = \alpha_t X_0 + \beta_t X_1$$

Tail index is that of the heaviest component. X_t keeps X_0's heavy tail for **every** $\alpha_t > 0$, and is light only at $t = 1$:

$$\text{tail index of } X_t \;=\; \begin{cases} \text{heavy } (X_0\text{'s}) & \text{for } \alpha_t > 0, \\ \text{light } (X_1\text{'s}) & \text{for } \alpha_t = 0 \end{cases} \quad \Rightarrow \quad \text{discontinuous in } t.$$

Fix. Interpolate in *log-space* via the soft-log $\phi(x) = \operatorname{sign}(x)\log(1+|x|)$, which sends heavy tails to exponential (light): $\mathbb{P}(\phi(X) > y) \approx e^{-y/\gamma}$.

$$\tilde{X}_t = \alpha_t \phi(X_0) + \beta_t \phi(X_1);$$
$$X_t = \phi^{-1}(\tilde{X}_t).$$

Why it works: tail annealing

Arithmetic interpolation in log-space = geometric in real space. With $\tilde{X}_0 = \phi(X_0) \approx \log X_0$ and $\tilde{X}_t = \alpha_t \tilde{X}_0 + \beta_t \tilde{X}_1$, exponentiating gives

$$X_t = \phi^{-1}(\tilde{X}_t) \;\approx\; \underbrace{X_0^{\alpha_t}}_{\text{power of data}} \cdot \underbrace{e^{\beta_t \tilde{X}_1}}_{\text{log-normal noise}} .$$

Why it works: tail annealing

Arithmetic interpolation in log-space = geometric in real space. With $\tilde{X}_0 = \phi(X_0) \approx \log X_0$ and $\tilde{X}_t = \alpha_t \tilde{X}_0 + \beta_t \tilde{X}_1$, exponentiating gives

$$X_t = \phi^{-1}(\tilde{X}_t) \;\approx\; \underbrace{X_0^{\alpha_t}}_{\text{power of data}} \cdot \underbrace{e^{\beta_t \tilde{X}_1}}_{\text{log-normal noise}} .$$

Power transform. $X_0 \sim \mathrm{Pareto}(\gamma) \Rightarrow X_0^{\alpha_t} \sim \mathrm{Pareto}(\gamma\alpha_t)$, tail index $\gamma\alpha_t$. The log-normal factor (extreme-value index 0) adds *no* tail factor. X_t has tail index $\gamma\alpha_t$.

Why it works: tail annealing

Arithmetic interpolation in log-space = geometric in real space. With $\tilde{X}_0 = \phi(X_0) \approx \log X_0$ and $\tilde{X}_t = \alpha_t \tilde{X}_0 + \beta_t \tilde{X}_1$, exponentiating gives

$$X_t = \phi^{-1}(\tilde{X}_t) \;\approx\; \underbrace{X_0^{\alpha_t}}_{\text{power of data}} \cdot \underbrace{e^{\beta_t \tilde{X}_1}}_{\text{log-normal noise}} .$$

Power transform. $X_0 \sim \mathrm{Pareto}(\gamma) \Rightarrow X_0^{\alpha_t} \sim \mathrm{Pareto}(\gamma\alpha_t)$, tail index $\gamma\alpha_t$. The log-normal factor (extreme-value index 0) adds *no* tail factor. X_t has tail index $\gamma\alpha_t$.

Tail annealing. As $\alpha_t : 1 \to 0$, the tail index $\gamma\alpha_t$ decreases *continuously* from γ (heavy) to 0 (light).

Bounded score

Second consequence of working in log-space: the score the network is well behaved.

Original space. The Pareto score

$$s_X(x) = \nabla_x \log p_X(x) = -\frac{1/\gamma + 1}{x}$$

diverges as $x \to 0^+$ and decays as $1/x$ in the tails: hard to regress.

Bounded score

Second consequence of working in log-space: the score the network is well behaved.

Original space. The Pareto score

$$s_X(x) = \nabla_x \log p_X(x) = -\frac{1/\gamma + 1}{x}$$

diverges as $x \to 0^+$ and decays as $1/x$ in the tails: hard to regress.

Log-space. Since $\tilde{X} \approx \log X \sim \mathrm{Exp}(1/\gamma)$,

$$\nabla_{\tilde{x}} \log p_{\tilde{X}}(\tilde{x}) \;\approx\; -\frac{1}{\gamma} \qquad \text{(bounded, constant in the tails).}$$

Bounded score

Second consequence of working in log-space: the score the network is well behaved.

Original space. The Pareto score

$$s_X(x) = \nabla_x \log p_X(x) = -\frac{1/\gamma + 1}{x}$$

diverges as $x \to 0^+$ and decays as $1/x$ in the tails: hard to regress.

Log-space. Since $\tilde{X} \approx \log X \sim \mathrm{Exp}(1/\gamma)$,

$$\nabla_{\tilde{x}} \log p_{\tilde{X}}(\tilde{x}) \;\approx\; -\frac{1}{\gamma} \qquad \text{(bounded, constant in the tails).}$$

$\Rightarrow$ the velocity field is Lipschitz; a **standard MLP suffices**. No architectural change for heavy tails.

Log-FM: experimental set-up

Experiments set-up

Known margins *and* known copulas.

Copulas	Gaussian, Gumbel, Hüsler–Reiß
Margins	70% Pareto $\alpha \in \{1.5, 1.75, 2, 2.5\}$ + 30% standard Gaussian
Dims	$d \in \{10, 20, 50, 100\}$

144 configs $\times$ 5 methods $\times$ 20 reps

Baselines

TTF, TTFfix [Hickling+25]
gTAF [Jaini+20]
Arcsinh-FM (ablation)

Metrics

W_1^P	Average W_1 on Pareto margins
CVaR$_{99}$, $Q_{99.9}$	risk / deep tail
angular W_2, AKE	dependence

Parameter-free: no estimation of tail index.

Log-FM: tail-metric results

Median W_1^P across Gumbel + Gaussian copulas (lower is better). Average over 20 rep.

Method	d=10	d=20	d=50	d=100
$\alpha = 1.5$ *(heaviest)*				
Log-FM	**0.522**	**0.451**	0.534	**0.525**
Arcsinh-FM	0.561	0.550	**0.534**	0.592
TTFfix	1.149	0.963	0.867	1.046
TTF	0.758	0.938	0.758	0.919
gTAF	0.534	0.583	0.761	3.320
$\alpha = 2.0$				
Log-FM	**0.153**	**0.146**	0.195	**0.196**
Arcsinh-FM	0.166	0.162	**0.174**	0.212
TTFfix	0.235	0.222	0.220	0.294
TTF	0.182	0.203	0.207	0.252
gTAF	0.171	0.174	0.198	0.219

Log-FM: experiments at a glance

Metrics computed (medians over 144 configs, 20 reps each):

Marginal tail	**Light margins**	**Dependence**
W_1^P, Hill index,	W_1^G	AKE, angular W_2,
VaR_{99}, CVaR_{99}, $Q_{99.5}$, $Q_{99.9}$	(Gaussian coordinates)	sliced-W, energy distance

Summary. Best on tail & risk metrics, by far the most stable (no divergence for $\alpha \geq 1.75$), dependence preserved.

Log-FM: Summary

Theory. The soft-log transform compresses heavy tails to exponential. Bounded log-space score $\Rightarrow$ standard MLP is sufficient. Induced *tail annealing* $X_t \approx X_0^{\alpha_t} e^{\beta_t \tilde{X}_1}$: tail index decreases continuously.

Log-FM: Summary

Theory. The soft-log transform compresses heavy tails to exponential. Bounded log-space score $\Rightarrow$ standard MLP is sufficient. Induced *tail annealing* $X_t \approx X_0^{\alpha_t} e^{\beta_t \tilde{X}_1}$: tail index decreases continuously.

Experiments.

- Best on every tail / risk metric (W_1^P, CVaR$_{99}$, $Q_{99.9}$);
- Uniquely stable (no severe divergence for $\alpha \geq 1.75$);
- Parameter-free: no estimation of tail index.

Log-FM: Summary

Theory. The soft-log transform compresses heavy tails to exponential. Bounded log-space score $\Rightarrow$ standard MLP is sufficient. Induced *tail annealing* $X_t \approx X_0^{\alpha_t} e^{\beta_t \tilde{X}_1}$: tail index decreases continuously.

Experiments.

- Best on every tail / risk metric (W_1^P, CVaR$_{99}$, $Q_{99.9}$);
- Uniquely stable (no severe divergence for $\alpha \geq 1.75$);
- Parameter-free: no estimation of tail index.

Limitation. Exponential ϕ^{-1} amplifies deep-tail error for $\gamma > 1$.

Log-FM: Summary

Theory. The soft-log transform compresses heavy tails to exponential. Bounded log-space score $\Rightarrow$ standard MLP is sufficient. Induced *tail annealing* $X_t \approx X_0^{\alpha_t} e^{\beta_t \tilde{X}_1}$: tail index decreases continuously.

Experiments.

- Best on every tail / risk metric (W_1^P, CVaR$_{99}$, $Q_{99.9}$);
- Uniquely stable (no severe divergence for $\alpha \geq 1.75$);
- Parameter-free: no estimation of tail index.

Limitation. Exponential ϕ^{-1} amplifies deep-tail error for $\gamma > 1$.

Perspectives.

- Tighter dependence preservation at high d (small angular-W_2 gap vs. TTF at $d \geq 50$).
- Schedule design (α_t, β_t) as an explicit tail-annealing rate.

Perspectives: recap

HTGAN

- Beyond i.i.d.: time series [TimeGAN, Quant GAN].
- Replicate methodology in other models (heavy-tailed diffusion, W-A GAN).

Log-FM

- Tighter dependence preservation at high d (angular-W_2 gap vs. TTF).
- Schedule (α_t, β_t) as an explicit tail-annealing rate.
- Beyond i.i.d.: extremal dependence in time series.

Perspectives: recap

HTGAN

- Beyond i.i.d.: time series [TimeGAN, Quant GAN].
- Replicate methodology in other models (heavy-tailed diffusion, W-A GAN).

Log-FM

- Tighter dependence preservation at high d (angular-W_2 gap vs. TTF).
- Schedule (α_t, β_t) as an explicit tail-annealing rate.
- Beyond i.i.d.: extremal dependence in time series.

Common thread: reproduce heavy-tailed extremal dependence, via adversarial (HTGAN) and diffusion / flow (Log-FM) models.

Thank you!

Questions?

HTGAN paper

Log-FM paper