Marginal Tail-Adaptive Normalizing Flows

mTAF: Capturing Heavy-Tailed Distributions with Mixed Marginal Tail Behavior

Juyeong Hwang

Outline

Motivation: Why Heavy Tails Matter

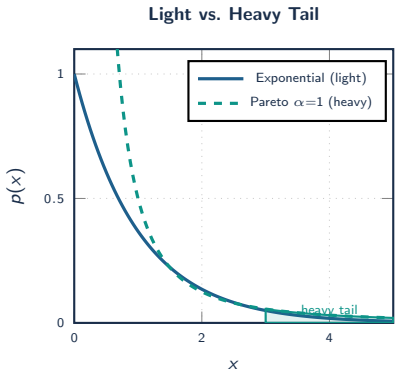
Heavy-tailed phenomena are everywhere:

- **Finance:** Stock returns, loss distributions, VaR
- **Climate:** Extreme weather events, flood levels
- **Biology:** Protein sequence lengths (power law)
- **Networks:** Degree distributions, file sizes

Underestimating tails $\Rightarrow$ catastrophic failure:

- Miscalculated financial risk $\rightarrow$ 2008-style crises
- Underestimated flood levels $\rightarrow$ infrastructure collapse

▲ **Problem:** Deep generative models (incl. Normalizing Flows) concentrate on the *body* of the distribution and fail to capture the **tail behavior**.



Background: Heavy-Tailed Distributions

Definition 2.1 — Heavy-Tailed Random Variable

A random variable $x \in \mathbb{R}$ is **heavy-tailed** if and only if

$$\forall \lambda > 0 : \quad \mathbb{E}\left[e^{\lambda x}\right] = \infty.$$

The function $m_p(\lambda) := \mathbb{E}[e^{\lambda x}]$ is the *moment-generating function*. Variables that are NOT heavy-tailed are called **light-tailed**.

Definition 2.2 — Tail Index α

x has **tail index** α if

$$\mathbb{E}\left[|x|^{\beta}\right] \begin{cases} < \infty & \text{if } \beta < \alpha, \\ = \infty & \text{if } \beta > \alpha. \end{cases}$$

Larger $\alpha \Rightarrow$ lighter tail. **Example:** t_ν distribution has tail index ν.

Estimation of Tail Index (EVT):

- Hill estimator (only regularly varying)
- Moments estimator (Dekkers et al. 1989)
- Kernel-based estimator (Csorgo et al. 1985)
- All require threshold selection k (tricky!)

Multivariate Notion (ℓ_2-heavy-tailed):

$\mathbf{x} \in \mathbb{R}^D$ is ℓ_2-heavy-tailed if $\|\mathbf{x}\|$ is heavy-tailed.

But this is too coarse — it cannot distinguish fully vs. mixed-tailed distributions!

Background: Normalizing Flows

Core idea: Learn invertible transformation $T_\theta : \mathbb{R}^D \to \mathbb{R}^D$ from tractable base $\mathbf{z}$ to target $\mathbf{x}$.

Theorem 2.4 — Change-of-Variables Formula

If $\mathbf{x} = T(\mathbf{z})$ where T is a diffeomorphism, then

$$p(\mathbf{x}) \;=\; q\Big(T^{-1}(\mathbf{x})\Big) \cdot \left|\det J_{T^{-1}}(\mathbf{x})\right|,$$

where $J_{T^{-1}}$ is the Jacobian of T^{-1}.

Masked Autoregressive Flow (MAF) — triangular structure:

$$T_j(\mathbf{z}) \;=\; \mu_j(\mathbf{z}_{<j}) + \exp\big(\sigma_j(\mathbf{z}_{<j})\big) \cdot z_j, \quad j = 1, \ldots, D,$$

where μ_j, σ_j are neural networks taking the first $j-1$ components as input.

Advantages:

- Triangular Jacobian $\Rightarrow$ efficient det
- Exact density evaluation & sampling
- Composable: $T = T^{(L)} \circ \cdots \circ T^{(1)}$

Linear layers between flows:

- Simple: random **permutation** $P^{(l)}$
- Better: **LU-decomposition** layer (learnable linear mixing)

Background: Tail-Adaptive Flows (TAF) — Prior Work

Theorem 2.5 — Jaini et al. (2020)

Let $\mathbf{z}$ be ℓ_2-light-tailed and T be an **affine triangular flow**:

$$T_j(\mathbf{z}_{\leq j}) = \mu_j(\mathbf{z}_{<j}) + \sigma_j(\mathbf{z}_{<j})\, z_j.$$

If σ_j is *bounded above* and μ_j is *Lipschitz* for all j, then $\mathbf{x} = T(\mathbf{z})$ is also ℓ_2-light-tailed.

Consequence: To model heavy-tailed targets, we need a **heavy-tailed base distribution**.
Solution by Jaini et al. (2020) — Tail-Adaptive Flow (TAF):

$$\mathbf{z} \sim t_{\hat{\nu}}(\mathbf{0}, I) \quad \text{(multivariate } t \text{ with one learnable d.o.f. } \hat{\nu})$$

▲ **Critical Limitation:** Since all marginals of $t_\nu(\mathbf{0}, I)$ share the same tail index ν, **TAF** can only produce distributions where **ALL** marginals are heavy-tailed OR **ALL** are light-tailed.

⇒ **Cannot handle mixed-tailed distributions!**

Problem: Mixed Marginal Tail Behavior

Definition 3.1 — Marginal Tail Behavior (New)

- $\mathbf{x}$ is j-**heavy-tailed** if its j-th marginal x_j is heavy-tailed.
- $\mathbf{x}$ is **mixed-tailed** if $\exists\, j_1, j_2$ such that x_{j_1} is heavy- and x_{j_2} is light-tailed.
- $\mathbf{x}$ is **fully heavy-tailed** if x_j is heavy-tailed for all j.
- $\mathbf{x}$ and $\mathbf{z}$ have **equal tail behavior** if for all j:

$$x \text{ is } j\text{-heavy-tailed} \iff \mathbf{z} \text{ is } j\text{-heavy-tailed}.$$

Proposition 3.2

j-heavy-tailedness $\Rightarrow$ ℓ_2-heavy-tailedness.
The new notion is **strictly finer** than ℓ_2-heavy-tailedness: it distinguishes fully vs. mixed-tailed variables.

Example — 4-dim mixed-tailed:

$x_1 \sim \mathcal{N}$	light-tailed
$x_2 \sim t_2$	heavy-tailed
$x_3 \sim \mathcal{N}$	light-tailed
$x_4 \sim$ **Pareto**	heavy-tailed

Theoretical Results (1/2): Impossibility

Proposition 3.3 — Triangular Affine Maps Preserve Full Heavy-Tailedness

Let $\mathbf{z}$ be **fully heavy-tailed** (Assumption A.7 on copula density) and T a triangular affine map

$$T_j(\mathbf{z}_j, \mathbf{z}_{<j}) = \mu_j(\mathbf{z}_{<j}) + \sigma_j(\mathbf{z}_{<j})\, z_j, \quad \sigma_j > 0.$$

Then $T(\mathbf{z})$ is also **fully heavy-tailed**.

Proof sketch: Examine $m_{x_j}(\lambda)$ for $j > 1$:

$$m_{x_j}(\lambda) = \int_{\mathbb{R}^j} e^{\lambda[\mu(\mathbf{z}_{<j}) + \sigma(\mathbf{z}_{<j}) z_j]}\, q_{\leq j}(\mathbf{z}_{\leq j})\, d\mathbf{z}_{\leq j}.$$

Using **Sklar's theorem** to factor $q_{\leq j} = c_j(F_1(z_1), \ldots, F_j(z_j)) \prod_{i<j} q_i(z_i)$ and Assumption A.7 to bound the copula density from below, the inner integral over z_j is $= \infty$ because z_j is heavy-tailed.

▲ **Implication:** Any flow with a *fully heavy-tailed* base (incl. **TAF**) **cannot** produce a mixed-tailed target. Permutation layers do not change marginal tailedness, so this applies to all layers.

Proof Detail (1/3): Why Copula Appears — Sklar's Theorem

Problem: After one flow step, $z_1, \ldots, z_j$ are **no longer independent**. Their joint density $q_{\leq j}$ contains a complex dependency structure.

Sklar's Theorem — Any Joint Density Decomposes As

For any joint density $q_{\leq j}$ with marginal CDFs $F_1, \ldots, F_j$, there exists a **copula density** $c_j : [0,1]^j \to \mathbb{R}_{\geq 0}$ such that

$$q_{\leq j}(z_1, \ldots, z_j) = \underbrace{c_j(F_1(z_1), \ldots, F_j(z_j))}_{\text{dependency structure}} \cdot \underbrace{\prod_{i=1}^{j} q_i(z_i)}_{\text{marginals}}.$$

The copula c_j captures *only* how the variables co-vary — independent of each marginal.

Why this decomposition helps: We can now analyze the MGF integral separately:

$$m_{x_j}(\lambda) = \int_{\mathbf{z}_{<j}} e^{\lambda \mu(\mathbf{z}_{<j})} q_{<j}(\mathbf{z}_{<j}) \underbrace{\int_{z_j} e^{\lambda \sigma(\mathbf{z}_{<j}) z_j} \cdot c_j(\cdots) \cdot q_j(z_j)\, dz_j}_{A(\mathbf{z}_{<j})} d\mathbf{z}_{<j}$$

If $c_j \to 0$ too fast as $z_j \to \infty$, it could **cancel** the divergence from $e^{\lambda \sigma z_j}$, making $A(\mathbf{z}_{<j}) < \infty$ and destroying heavy-tailedness.

$\Rightarrow$ We need to **bound** c_j **from below** — that is exactly Assumption A.7.

Proof Detail (2/3): Assumption A.7 — Bounding the Copula

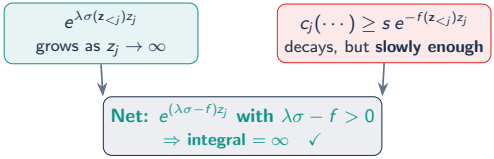
Goal: Prevent the copula density from decaying so fast that it cancels the heavy-tail divergence.

Assumption A.7

There exist a compact set $S \subset \mathbb{R}^{j-1}$, constants $z_j^* > 0$, $s > 0$, and a function $f(\mathbf{z}_{<j}) < \lambda\sigma(\mathbf{z}_{<j})$ such that

$$c_j(F_1(z_1), \ldots, F_j(z_j)) \;\geq\; s\, e^{-f(\mathbf{z}_{<j})\, z_j} \quad \text{for } z_j > z_j^*,\ \mathbf{z}_{<j} \in S.$$

Intuition — a race between growth and decay:



When is A.7 satisfied?

- **Independence copula** ($c_j \equiv 1$): trivially — mTAF base is designed this way
- **Gaussian copula**: verified via $\Phi^{-1}(1-y) \sim \sqrt{-2\log y}$ asymptotics
- **Any copula bounded below** ($c_j \geq a > 0$): trivially

Theoretical Results (2/2): Main Theorem

Theorem 3.4 — Learning the Correct Tail Behavior (Main Result)

Let $\mathbf{z} \in \mathbb{R}^D$ satisfy:

$$z_j \text{ is } \textbf{light-tailed} \text{ for } j \in \{1, \ldots, d_l\}, \quad z_j \text{ is } \textbf{heavy-tailed} \text{ for } j \in \{d_l+1, \ldots, D\}.$$

Under the same conditions as Theorem 2.5 and Proposition 3.3, a triangular affine flow T satisfies:

$$\boxed{\mathbf{z} \text{ and } \mathbf{x} = T(\mathbf{z}) \text{ have } \textbf{equal marginal tail behavior.}}$$

Proof sketch:

- *Light-tailed marginals* ($j \leq d_l$): The MGF $m_{x_j}(\lambda)$ is bounded by the argument of Theorem 2.5 (Jaini et al.), because the conditioning on $\mathbf{z}_{<j}$ involves only light-tailed components.
- *Heavy-tailed marginals* ($j > d_l$): The proof of Prop. 3.3 shows that heavy-tailedness of z_j propagates *regardless* of the tailedness of $\mathbf{z}_{<j}$. Hence $x_j = T_j(\mathbf{z}_{\leq j})$ is heavy-tailed.

Corollary 3.5 — mTAF is Marginally Tail-Adaptive

Under the same assumptions, $\mathbf{z}$ and $\mathbf{x} = T(\mathbf{z})$ have **identical marginal tail behavior**.

Proof Detail (3/3): Heavy-Tail Propagation Mechanism

Key question: Does z_j being heavy-tailed guarantee $x_j = \mu_j(\mathbf{z}_{<j}) + \sigma_j(\mathbf{z}_{<j})z_j$ is heavy-tailed, *regardless* of the tailedness of $\mathbf{z}_{<j}$?

Answer: Yes. Here is why. Focus on the inner integral $A(\mathbf{z}_{<j})$:

$$A(\mathbf{z}_{<j}) = \int_{z_j > z_j^*} e^{\lambda\sigma(\mathbf{z}_{<j})z_j} \cdot c_j(\cdots) \cdot q_j(z_j)\, dz_j \overset{\text{A.7}}{\geq} s \int_{z_j > z_j^*} e^{\underbrace{(\lambda\sigma(\mathbf{z}_{<j}) - f(\mathbf{z}_{<j}))}_{>0 \text{ by A.7}} z_j} q_j(z_j)\, dz_j.$$

The Crucial Step

Define $\lambda' := \lambda\sigma(\mathbf{z}_{<j}) - f(\mathbf{z}_{<j}) > 0$. Since z_j is **heavy-tailed**, by Definition 2.1:

$$\int_{z_j > z_j^*} e^{\lambda' z_j} q_j(z_j)\, dz_j = \infty \quad \text{for all } \lambda' > 0.$$

Therefore $A(\mathbf{z}_{<j}) = \infty$ for **every** $\mathbf{z}_{<j} \in S$, regardless of its own tail behavior.

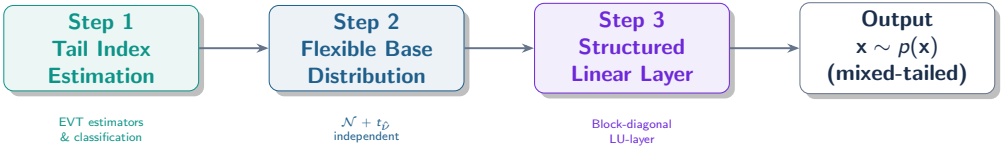
Why $\mathbf{z}_{<j}$ doesn't matter:

- $\mathbf{z}_{<j}$ only enters through $\sigma(\mathbf{z}_{<j})$ and $f(\mathbf{z}_{<j})$
- A.7 ensures $\lambda\sigma(\mathbf{z}_{<j}) - f(\mathbf{z}_{<j}) > 0$ for all $\mathbf{z}_{<j} \in S$
- So $\lambda' > 0$ is guaranteed **no matter what** $\mathbf{z}_{<j}$ is
- The divergence $A(\mathbf{z}_{<j}) = \infty$ is driven purely by z_j**'s heavy-tailedness**

Consequence for Theorem 3.4

For $j > d_l$: z_j is heavy-tailed $\Rightarrow$ x_j is heavy-tailed, even if $z_1, \ldots, z_{d_l}$ are light-tailed. Mixed base $\Rightarrow$ mixed target. ✓

mTAF: Overview



Step 1

Estimate marginal tail indices; classify as heavy or light-tailed.

Step 2

Build flexible base with $\mathcal{N}$ and $t_{\hat{\nu}}$ marginals; reorder components.

Step 3

Replace permutations with block-diagonal LU-layers; train end-to-end.

Step 1: Tail Index Estimation

Goal: For each marginal x_j, decide heavy- or light-tailed and estimate $\hat{\nu}_j$.

Challenge: Tail behavior depends only on $|x| > C$ — fitting a full model gives biased estimates!

Estimators used (Algorithm 1):

1. **Moments double-bootstrap estimator** (Draisma et al. 1999) — consistent for all max-domains
2. **Kernel-type double-bootstrap estimator** (Groeneboom et al. 2003)
3. If both predict light-tailed $\Rightarrow$ set $q_j = \mathcal{N}(0, 1)$
4. Otherwise $\Rightarrow$ run **Hill double-bootstrap** (Danielsson et al. 2001) for $\hat{\nu}_j$
5. **Clip**: if $\hat{\nu}_j > 10$, treat as light-tailed (very mild heaviness $\approx$ Gaussian)

Why clip at 10? Without clipping, nearly all marginals would be classified as heavy-tailed (Hill can misfire on Gaussians). Clipping prevents over-restriction of the permutation scheme.

Learnable refinement: Initialize $\nu_j \leftarrow \hat{\nu}_j$, then make it a **trainable parameter** alongside flow weights — accounts for estimation errors.

Step 2: Flexible Base Distribution

Construction (mean-field / independent marginals):

$$q(\mathbf{z}) = \prod_{j=1}^{D} q_j(z_j), \quad q_j = \begin{cases} \mathcal{N}(0,1) & j \in \{1, \ldots, d_l\} \text{ (light-tailed)} \\ t_{\hat{\nu}_j}(0,1) & j \in \{d_l{+}1, \ldots, D\} \text{ (heavy-tailed)} \end{cases}$$

Reordering: Sort marginals so light-tailed come first (required by Thm. 3.4). Apply the same permutation to the data.

Training objective (maximize log-likelihood):

$$\mathcal{L}(\hat{\theta}, \hat{\boldsymbol{\nu}};\, \mathcal{X}) = \sum_{n=1}^{N} \left(\sum_{j=1}^{d_l} \log \phi\Big(T_{\hat{\theta}}^{-1}(\mathbf{x}^{(n)})_j \Big) + \sum_{j=d_l+1}^{D} \log t_{\hat{\nu}_j}\Big(T_{\hat{\theta}}^{-1}(\mathbf{x}^{(n)})_j \Big) - \log \Big| \det J_{T_{\hat{\theta}}}\Big(\mathbf{x}^{(n)}\Big) \Big| \right)$$

where ϕ = standard Gaussian PDF, $t_{\hat{\nu}_j}$ = t-PDF with $\hat{\nu}_j$ d.o.f.

Why mean-field? Independence satisfies Assumption A.7 automatically (independence copula: $c \equiv 1$). Dependency structure is learned by the flow layers.

Step 3: Structured Linear Layers

Problem with standard LU-layers: They mix all components freely, potentially mapping a light-tailed component into a heavy-tailed one — violating the ordering required by Theorem 3.4.

Theorem 3.6 — Block-Diagonal Linear Layers Preserve Marginal Tailedness

Let $\mathbf{z}$ be j-light-tailed for $j \leq d_l$ and j-heavy-tailed for $j > d_l$. Consider the block-diagonal matrix

$$W = \begin{pmatrix} A & \mathbf{0} \\ B & C \end{pmatrix}, \quad A \in \mathbb{R}^{d_l \times d_l},\ B \in \mathbb{R}^{d_h \times d_l},\ C \in \mathbb{R}^{d_h \times d_h},$$

where $\mathbf{0}$ is a $d_l \times d_h$ zero block. Then $W\mathbf{z}$ and $\mathbf{z}$ have **equal tail behavior**.

Why does this work?

- $(W\mathbf{z})_j$ for $j \leq d_l$: linear combo of $z_1, \ldots, z_{d_l}$ only $\Rightarrow$ light-tailed.
- $(W\mathbf{z})_j$ for $j > d_l$: involves at least one heavy-tailed z_k $(k > d_l) \Rightarrow$ heavy-tailed.

Efficient implementation via LU-decomposition:

$$\log|\det W| = \log|\det A| + \log|\det C|, \qquad W^{-1} = \begin{pmatrix} A^{-1} & \mathbf{0} \\ -C^{-1}BA^{-1} & C^{-1} \end{pmatrix}.$$

Parameterize A and C via LU-decomposition (efficient inverse & log-det); B unconstrained.

Extension 1: Neural Spline Flows (NSF)

NSF (Durkan et al. 2019) replace affine layers with **rational-quadratic splines**:

$$T_j^{\text{NSF}}(z_j \mid \mathbf{z}_{<j}) = \begin{cases} \text{rational quadratic spline} & z_j \in [-b, b] \\ z_j + \text{const} & |z_j| > b \quad \text{(linear in the tails!)} \end{cases}$$

Key Observation

Because NSF layers are **affine linear outside** $[-b, b]$, all theory from Jaini et al. (2020) and Theorem 3.4 applies directly to NSF layers as well.

$\Rightarrow$ Simply replacing standard LU-layers with **block-diagonal LU-layers** makes NSF marginally tail-adaptive.

Subtlety: Even though each autoregressive NSF layer is linear outside $[-b, b]$, the *composition* of layers is NOT globally linear, because block-diagonal LU-layers can map tail values back into $[-b, b]$ on subsequent layers. This gives the flow expressive power in the tails.

Extension 2: Generalized Tail-Adaptive Flow (gTAF)

Motivation: mTAF has strong guarantees but restricts the linear layer structure. Can we get the best of both worlds?

gTAF construction:

- Drop structural restriction on linearities (use standard LU-layers)
- Set **all** base marginals to t_ν with **independently learnable** ν_j: $z_j \sim t_{\nu_j}(0,1)$, $j = 1, \ldots, D$
- As $\nu_j \to \infty$: $t_{\nu_j} \to \mathcal{N}(0,1)$ — so gTAF approximates light-tailed marginals

Theorem 3.7 — gTAF Converges to mTAF

Let $T : \mathbb{R}^D \to \mathbb{R}^D$ be a.s. continuous. Let $\mathbf{z}_\nu$ be the gTAF base with $z_{\nu,j} \sim t_\nu$ for $j \leq d_l$. Then $T(\mathbf{z}_\nu) \xrightarrow{D} T(\mathbf{z})$ as $\nu \to \infty$ (gTAF converges in distribution to the mTAF solution).

Summary — Trade-offs:

	Vanilla	TAF	gTAF	mTAF
Mixed-tail theory	✗	✗	approx.	✓
Flexible linearities	✓	✓	✓	partial
# tail params	0	1	D	d_h

Experiments: Synthetic Data Setup

Data: 8-dim Gaussian copula; $d_h \in \{1, 4\}$ heavy-tailed marginals ($\nu = 2$).

- Light-tailed marginals: Gaussian / mixture of Gaussians
- Heavy-tailed marginals: mixture of two t_2-distributions
- Copula with random correlation structure ($R_{ij} = 0.25$ for 16 random pairs)

Models compared: Vanilla NSF | **TAF** | **gTAF** | **mTAF** **Oracle:** Gaussian copula fit

Evaluation Metrics (all lower is better):

Neg. log-likelihood $\mathcal{L}$

Overall distributional fit

Area (log-log)

$$\sum_i \left|\log \bar{F}^{-1}_{\text{data}} \tfrac{i}{n} - \log \bar{F}^{-1}_{\text{flow}} \tfrac{i}{n}\right| \log \tfrac{i+1}{i}$$

Reweighted tail fit

tVaR (Expected Shortfall)

$\mathrm{tVaR}_\alpha = \frac{1}{1-\alpha} \int_\alpha^1 F^{-1}(u)\, du$

Abs. difference at $\alpha = 0.95$

25 runs $\times$ 3 distributions $=$ 75 runs total. Training: 5k steps, Adam, lr $= 10^{-5}$.

Experiments: Quantitative Results

Results ($\nu = 2$, lower is better):

$d_h = 1$ **(1 heavy-tailed marginal)**

	$\mathcal{L}$	Area_l	Area_h	tVaR_l	tVaR_h
Vanilla	10.25	0.25	4.19	0.60	29.56
TAF	10.15	0.37	3.55	0.78	4.21
gTAF	10.12	0.55	3.24	1.16	2.67
mTAF	**10.11**	**0.26**	**2.22**	**0.59**	**2.94**
Copula	*9.75*	*0.20*	*1.23*	*0.45*	*2.22*

$d_h = 4$ **(4 heavy-tailed marginals)**

	$\mathcal{L}$	Area_l	Area_h	tVaR_l	tVaR_h
Vanilla	8.98	0.25	4.30	0.54	28.55
TAF	8.69	0.42	4.05	0.89	4.36
gTAF	8.57	0.50	3.38	0.98	5.55
mTAF	**8.55**	**0.25**	**2.60**	**0.57**	**6.74**
Copula	*9.75*	*0.19*	*1.43*	*0.46*	*3.49*

Key observations:

- Vanilla: good on Area_l but completely fails on tVaR_h (28–29 $\gg$ oracle)
- **TAF** ($d_h = 1$): only 1 shared ν for 7 light + 1 heavy $\Rightarrow$ poor balance
- **mTAF**: best balance between light- and heavy-tailed metrics; closest to oracle on Area_h
- **gTAF**: strong on tVaR_h, but sacrifices light-tail accuracy

Experiments: Tail Recovery (Confusion Matrices)

Test: Generate samples from each flow; classify each marginal as heavy- or light-tailed.

Vanilla

	Act L	Act H
Gen L	97.7%	95.3%
Gen H	2.3%	4.7%

TAF

	90.0%	86.7%
	10.0%	13.3%

gTAF

	77.3%	1.6%
	22.7%	98.4%

mTAF ★

	96.7%	1.3%
	3.3%	98.7%

- **Vanilla:** All generated as light-tailed — zero heavy-tail recall
- **TAF:** Slight improvement, but still mostly light-tailed
- **gTAF:** 98.4% heavy-tail recall, but 22.7% false heavy-tail rate
- **mTAF: Near-perfect: 96.7% L-accuracy AND 98.7% H-accuracy**

Experiments: Climate Application

Dataset: EUMETSAT NWP-SAF — 25,000 atmospheric profiles; **412 dimensions**

- 3 quantities $\times$ 137 atmospheric levels: dry-bulb temperature (K), atmospheric pressure (hPa), cloud optical depth
- Heavy-tailed components identified manually (Table 7): e.g. temperature levels 80–137, pressure 100–137, optical depth 58–137

Architecture: 5-layer NSF, LU-linearities, ResNet conditioner (2 hidden layers, 100 neurons), 3 spline bins, tail-bound $= 2.5$, batch norm, cosine annealing, 20k steps.

Quantitative results (neg. log-likelihood, 25 trials):

	Vanilla	TAF	gTAF	**mTAF**
$\mathcal{L}$	-2101.91 ± 9.44	-2110.56 ± 7.87	-2113.48 ± 7.93	$-\mathbf{2121.38} \pm 10.91$

1D random projection analysis: Project onto 100 random directions $\mathbf{w}_j \sim \mathcal{U}([0,1]^{412})$; compare mean, std, and 1%-quantile.

- All methods match the mean well
- Only **mTAF** correctly captures the **standard deviation and extreme quantiles**

Limitations & Future Directions

▲ Asymmetric Tail Behavior

Current framework treats both tails equally. Many real variables (e.g. optical depth ≥ 0) have only one heavy tail.

Future: Composite body+tail base distributions (COMET Flows, McDonald et al. 2022).

⇛ Tail Dependencies

mTAF uses an independent (mean-field) base — ignores co-tail structure $\lambda_{i,j} = \lim_{u \to 1} P(x_j > F_j^{-1}(u) \mid x_i > F_i^{-1}(u))$.

Future: Replace with copula base $q(\mathbf{z}) = c(F_1(z_1), \ldots) \prod q_j(z_j)$.

Summary & Contributions

01 **Problem identified:** **TAF** cannot model mixed-tailed distributions (Prop. 3.3).
- A fully heavy-tailed base always yields a fully heavy-tailed target.

02 **Theory:** Theorem 3.4 — a flow with a mixed base ($\mathcal{N} + t_\nu$ marginals) and block-diagonal linear layers preserves marginal tail behavior.

03 **mTAF proposed:** Tail estimation (EVT) $\rightarrow$ flexible base $\rightarrow$ block-diagonal LU-layers.
- Marginally tail-adaptive by construction (Corollary 3.5).

04 **NSF extended:** NSF tails are affine-linear, so all theory applies; modified LU-layers preserve full marginal tailedness.

05 **gTAF:** Flexible relaxation with D learnable degrees of freedom; converges to mTAF as light-tailed $\nu_j \rightarrow \infty$ (Theorem 3.7).

06 **Experiments:** Near-perfect tail recovery on synthetic data; best NLL on 412-dim climate dataset.

Thank You

Questions & Discussion

Juyeong Hwang