

U1.T2 — The Zoo and Its Limits

Flow-Based Generative Models · UFRJ · 2026.2

01 | Glow: the coupling family at image scale

Three debts, on the board

Left open	By	Paid in
The depth question — how expressive is a stack of coupling layers?	U1.T1	the universality block
The fine print — a flow is a diffeomorphism; what can one <i>not</i> do?	U1.T1	the topology block
The checkerboard gaps — 29.8% of the mass where the data density is zero	U1.L1	the topology block

Before the debts: the rest of the zoo (the Glow and autoregressive blocks), read through one table.

The two costs, version 1

THE TWO COSTS (V1) – U1.T1

Operation	Direction	Needs, per layer	Cost driver
Sampling	forward, $z \rightarrow x$	evaluate ψ_k	one pass per layer
Density / training	inverse, $x \rightarrow z$	ψ_k^{-1} and $\log \det J_{\psi_k} $	the inverse, and the log-det

The unit's thesis: the zoo is a list of tricks to make both rows cheap at once.

Today the table gets one row per family (v2). Convention unchanged: $\psi: z \rightarrow x$ generates; a library's $f: x \rightarrow z$ is $f = \psi^{-1}$. U1.T1 wrote the same two rows as $x_0 \rightarrow x_1$ and $x_1 \rightarrow x_0$; these are U1.L1's letters.

Glow is RealNVP, engineered

Affine coupling at 32×32 and 256×256 (Kingma and Dhariwal 2018, sec. 3), with two extra layer types whose log-determinants are as cheap as coupling's:

- **actnorm**: a per-channel scale and bias, in place of batch normalization;
- **invertible 1×1 convolution**: a *learned* permutation of the channels;

and a **multi-scale** arrangement of the stack. No new theorem.

On an $h \times w \times c$ feature map, with $s, b \in \mathbb{R}^c$, at every position (i, j) :

$$y_{ij} = s \odot x_{ij} + b, \quad x_{ij} = (y_{ij} - b)/s,$$
$$\log|\det J| = h w \sum_{j=1}^c \log|s_j|.$$

Example, a $32 \times 32 \times 12$ map: 24 parameters; $\log|\det J| = 1024 \sum_{j=1}^{12} \log|s_j|$ — twelve logarithms, one sum.

Actnorm is not batch normalization

- s, b are *initialized* from the first minibatch: zero mean, unit variance per channel (Kingma and Dhariwal 2018, sec. 3.1).
- After that they are ordinary parameters: no batch statistics, no train/eval split, one function of one input.
- Invertible when every $s_j \neq 0$.

Data-dependent initialization chooses the starting point. It is not a normalization layer.

The invertible 1×1 convolution

One matrix $W \in \mathbb{R}^{c \times c}$ at every position; a permutation is the case $W = P$:

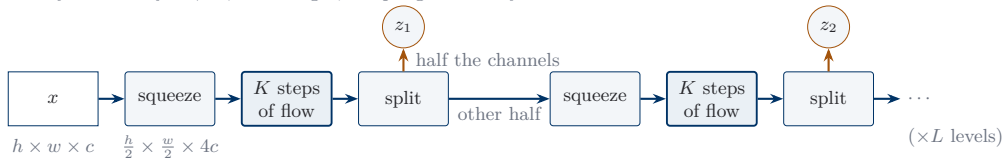
$$\begin{aligned} y_{ij} &= Wx_{ij}, & \log|\det J| &= h w \log|\det W| \\ &= h w \sum_{i=1}^c \log|s_i| & \text{for } W = PL(U + \text{diag } s). \end{aligned}$$

- P fixed permutation, L unit lower-triangular, U strictly upper, s learned (Kingma and Dhariwal 2018, sec. 3.2). Invertible when every $s_i \neq 0$.
- Determinant bookkeeping: $O(c)$. **Applying the mixing: still $O(hwc^2)$.**
- Same $32 \times 32 \times 12$ map: 144 parameters, 1024×144 multiply-adds to apply, twelve logarithms for the log-det.

Multi-scale, in one picture

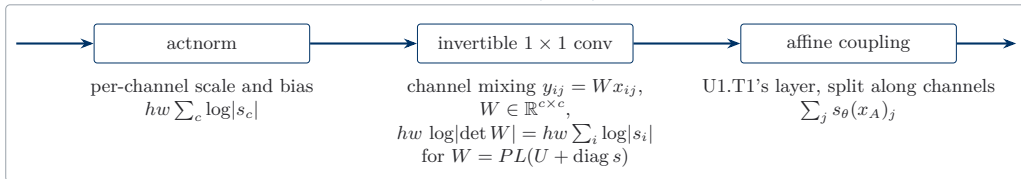
multi-scale: squeeze $\rightarrow K$ steps of flow $\rightarrow$ split, over L levels

density direction $f = \psi^{-1}$, left to right; sampling runs every arrow backwards



$z = (z_1, \dots, z_L)$ has the same total size as x . Squeeze is a reshape (log-det 0); split is a partition of the state, not a transform.

one step of flow, and what each layer adds to $\log|\det J|$



$\times K$ per level. Every layer has an analytic inverse, so the whole stack sits in the coupling row of the two-costs table.

Glow on the table

GLOW, PLACED

	Density / likelihood	Sampling
Glow	exact: coupling sums + actnorm sums + PLU sums	direct: every layer has an analytic inverse

The coupling row, engineered for images. It removes no limitation that the universality and topology blocks will state.

02 | Autoregressive flows: MAF and IAF

The autoregressive factorization

Any density on $\mathbb{R}^d$ factorizes by the chain rule. With coordinates $x = (\xi_1, \dots, \xi_d)$:

$$p(x) = \prod_{i=1}^d p(\xi_i \mid \xi_{<i}), \quad \xi_{<i} = (\xi_1, \dots, \xi_{i-1}).$$

A network that produces each factor from earlier coordinates only has an **exact likelihood** and a fixed ordering. That is the whole **autoregressive (AR)** family.

Coordinates are ξ_i (of x) and ζ_i (of z): in this course x_1 is a data point, never a coordinate.

MAF: the density transform

Parameterize $f = \psi^{-1}: x \rightarrow z$ coordinate by coordinate (Papamakarios, Pavlakou, and Murray 2017, sec. 3):

$$\zeta_i = \frac{\xi_i - \mu_i(\xi_{<i})}{\sigma_i(\xi_{<i})}, \quad \log|\det J_\psi(z)| = \sum_{i=1}^d \log \sigma_i(\xi_{<i}).$$

- **Density: one pass.** Every ξ_i is known, so one masked network (its masks let output i read only $\xi_{<i}$) gives all (μ_i, σ_i) at once.
- **Sampling: sequential.** $\xi_i = \mu_i(\xi_{<i}) + \sigma_i(\xi_{<i}) \zeta_i$ needs $\xi_{<i}$ first.

IAF: the generative transform

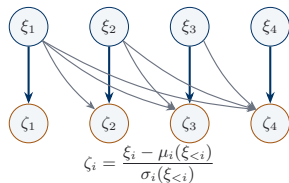
Parameterize $\psi: z \rightarrow x$ instead, with the conditioner reading the **source** coordinates (Kingma et al. 2016, secs. 3–4):

$$\xi_i = \mu_i(\zeta_{<i}) + \sigma_i(\zeta_{<i}) \zeta_i, \quad \log|\det J_\psi(z)| = \sum_{i=1}^d \log \sigma_i(\zeta_{<i}).$$

- **Sampling: one pass.** Every ζ_i is known.
- **Density: sequential.** $\zeta_i = (\xi_i - \mu_i(\zeta_{<i}))/\sigma_i(\zeta_{<i})$ needs $\zeta_{<i}$ first.

Duality: the direction of a triangle

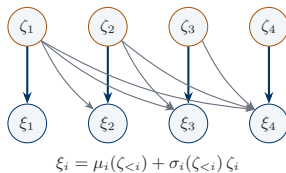
MAF: density transform $f = \psi^{-1}: x \rightarrow z$



density: all ξ known $\Rightarrow$ every ζ_i at once, one pass

sampling: ξ_i needs $\xi_{<i}$ first $\Rightarrow$ sequential

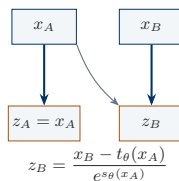
IAF: generative transform $\psi: z \rightarrow x$



sampling: all ζ known $\Rightarrow$ every ξ_i at once, one pass

density: ζ_i needs $\zeta_{<i}$ first $\Rightarrow$ sequential

coupling: two blocks



both directions: one pass. The conditioner never reads what it moves.

Same triangular Jacobian in every panel: $\log|\det J|$ is a sum of $\log \sigma_i$ (or of s_θ). What the direction of the gray edges decides is *which* of the two costs is the sequential one.

MAF and IAF on the table

MAF AND IAF, PLACED

Family	Density / likelihood	Sampling
MAF	parallel, one pass	sequential inverse (event dimension)
IAF	sequential inverse (event dimension)	parallel, one pass

“ d passes” is an intuition for affine scalar autoregression. The implementation-level statement is **sequential in the event dimension** — along the d coordinates of one point — not a promise about a count of network calls.

Coupling is block-autoregressive

Two blocks in place of d coordinates: $z_A = x_A$ copied, $z_B = (x_B - t_\theta(x_A)) \odot e^{-s_\theta(x_A)}$ conditioned on the first.

- U1.T1's affine coupling layer is an autoregressive map with two blocks (Papamakarios et al. 2021, sec. 3.1.2).
- The conditioner **never reads what it moves**, so *both* directions are one pass.
- The price: one coupling layer cannot represent an arbitrary autoregressive map. Expressivity is paid in **depth**, with permutations between layers (U1.T1 §6).

Coupling, MAF and IAF: one triangular idea, at different block sizes, in different directions.

03 | Universality, and its fine print

The depth question, first half: a negative result

NICE's additive layers, and any stack of them, are **volume-preserving**: $|\det J| = 1$.

THEOREM (DRAXLER ET AL. 2024, THM. 4.2 AND PROP. 4.3)

Flows with constant Jacobian determinant are **not** universal distribution approximators under KL. A flow of constant-Jacobian diffeomorphisms has the **same number of modes** as its base distribution.

A constant factor on the density cannot move a maximum. **NICE's final scaling layer does not help**: its determinant is a constant, so the theorem applies with the layer as without it — the answer to U1.T1's Exercise 2. What removes the obstruction is a determinant that **varies with the input**: RealNVP's $e^{s_\theta(x_A)}$ (Draxler et al. 2024, sec. 4).

The depth question, second half: a positive result

Constant-Jacobian stacks are capped. Does depth in a richer family do better? Yes, with fine print:

THEOREM (DRAXLER ET AL. 2024, THM. 5.4)

For every **continuous** density p on $\mathbb{R}^d$ with **finite first and second moments** and **infinite support**, there is a sequence of flows f_n of n affine coupling blocks, each L -bi-Lipschitz (distances stretched or compressed by at most a factor L , in either direction), with $(f_n)_\# p \rightarrow \mathcal{N}(\mathbf{o}, I_d)$ as $n \rightarrow \infty$ — in the paper's convergence measure Δ_{affine} (the KL improvement one more block could buy), which KL-convergence implies.

Why the affine case needed its own theorem

The classical argument (Papamakarios et al. 2021, sec. 2.2, §3.1.2):

- For any everywhere-positive target with differentiable conditional CDFs, the **conditional-CDF map** $\zeta_i = \Pr(\xi'_i \leq \xi_i \mid \xi_{<i})$ is a triangular diffeomorphism onto the uniform on $(0, 1)^d$.
- An autoregressive flow represents it when conditioners *and transformers* are flexible enough; d coupling layers with the right splits represent any autoregressive map.
- **An affine transformer is not flexible enough:** for fixed $\xi_{<i}$ it is affine in ξ_i , and an affine function is not a conditional CDF.

So the classical existence argument says nothing about RealNVP. The theorem on the previous slide is what settles the affine case.

The fine print, read slowly

1. **Existence over a growing family.** For each n a good flow exists; nothing says training finds it, and no bound on n is given.
2. **The hypotheses on the target matter.** Continuity, finite moments, **infinite support**. The gapped checkerboard fails two of three.
3. **Nothing about finite depth, trainability, or the conditioning of the composition.** Blocks are L -bi-Lipschitz; the *stack* is not bounded independently of n .
4. **Approximating a distribution is not approximating a map.** Many maps push p_0 near the same target.

The pivot

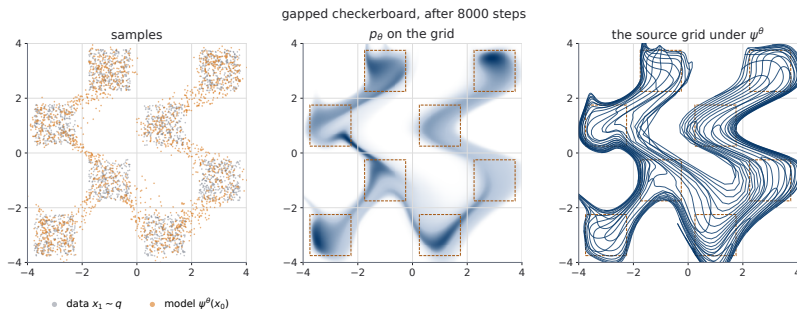
Universality holds, with fine print, for targets of infinite support.

What happens when the target has literal hard-zero gaps — a support in several pieces — that a flow from a Gaussian source cannot reproduce exactly?

That is the checkerboard. The answer is a theorem about supports.

04 | Exact support: the topology obstruction

Replay: the gapped checkerboard



Eight separated squares, **exactly zero density between them**; 0 of 100,000 generated points outside. The model: 29.8% of its mass in the gaps, NLL 0.50 nats above log 18.

The precision rule

“Topology obstruction” is a statement about supports and globally invertible maps.

It is **not** a statement about density level sets. A diffeomorphism can make a bimodal density from a unimodal source — U1.L1’s two-moons model is the proof by example.

Recall: the **support** of P is the closed set of points all of whose open neighbourhoods have positive measure. $\text{supp } \mathcal{N}(o, I_d) = \mathbb{R}^d$. The checkerboard’s support is eight closed squares: **disconnected**.

The support proposition

PROPOSITION

Let P have support S and let $\psi: \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a homeomorphism (onto, continuous, continuous inverse). Then

$$\text{supp}(\psi_{\#}P) = \psi(S).$$

If S is connected, so is the support of the pushforward.

Proof. ($\supseteq$) V open around $\psi(s)$, $s \in S$: $\psi^{-1}(V)$ is open around s , so $P(\psi^{-1}(V)) > 0$.
($\subseteq$) y in the support, $s = \psi^{-1}(y)$, U open around s : $\psi(U)$ is open around y , so $P(U) = \psi_{\#}P(\psi(U)) > 0$, hence $s \in S$. Connectedness: continuous image of a connected set. ■

Every flow of this unit is onto $\mathbb{R}^d$

Layer	Why it is a bijection of $\mathbb{R}^d$ onto $\mathbb{R}^d$
affine coupling	analytic inverse, defined for every x
actnorm, 1×1 conv	affine and linear, $s_j \neq 0$, $\det W \neq 0$
MAF, IAF, $\sigma_i > 0$	inverse recursion: one affine equation per coordinate
permutation, squeeze	reshapes and swaps
residual, $\text{Lip}(g) < 1$	$x + g(x) = z$ has one solution (residual-flow block)

$$\text{supp } p_\theta = \psi^\theta(\mathbb{R}^d) = \mathbb{R}^d \quad \text{for every } \theta.$$

The verdict, and what it is not

A flow of this unit **cannot exactly equal**:

1. a target whose support is **disconnected** — the checkerboard: $\mathbb{R}^2$ is not eight squares, for any θ ;
2. a target whose support has **empty interior** — a distribution on a curve or a surface.

Not capacity, not depth, not optimization.

Not about level sets. The model's superlevel sets $\{p_\theta > c\}$ *can* be disconnected: $|\det J|$ varies from point to point. A **constant** Jacobian guarantees the mode count (the universality block); a variable one guarantees nothing, and two-moons shows the count changing. Do not write “level-set topology is preserved” — it is false.

Approaching a mismatched target

The proposition rules out equality. What does *closeness* cost? Define the **bi-Lipschitz constant**

$$\text{BiLip}(f) = \inf \left\{ M \in [1, \infty] : M^{-1} \|z - z'\| \leq \|f(z) - f(z')\| \leq M \|z - z'\| \text{ for all } z, z' \right\}.$$

$\text{BiLip}(f) = \infty$ when no finite M works. A large constant means that somewhere the map stretches or compresses enormously, and its inverse does the opposite: the constant measures how *numerically* invertible f is.

Cornish's theorem

THEOREM (CORNISH ET AL. 2020, THM. 2.1)

If $\text{supp } P_Z$ and $\text{supp } P^*$ are **not homeomorphic**, then for any measurable f_n :

$$(f_n)_\# P_Z \rightarrow P^* \text{ weakly} \quad \Rightarrow \quad \text{BiLip}(f_n) \rightarrow \infty.$$

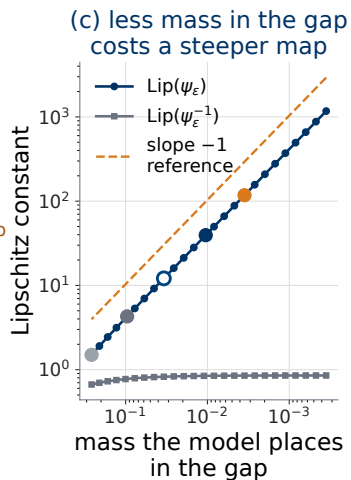
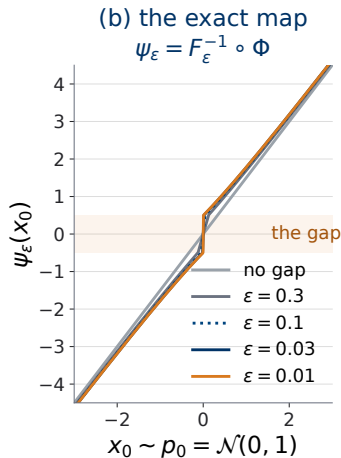
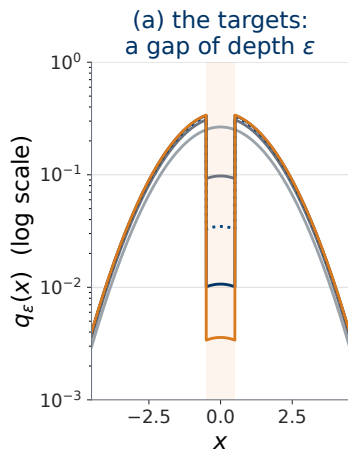
Weak convergence — expectations of bounded continuous functions converge — is implied by minimizing KL. So: **approaching the checkerboard forces the flow to become arbitrarily badly conditioned** (Cornish et al. 2020, sec. 2).

Cornish's corollary, and who pays first

Corollary 2.2. For any weak-topology metric ρ there is $M(\varepsilon) \rightarrow \infty$ as $\varepsilon \rightarrow 0$ such that: target within ε of a mismatched P^* , model within ε of the target $\Rightarrow \text{BiLip}(f) \geq M(\varepsilon)$.

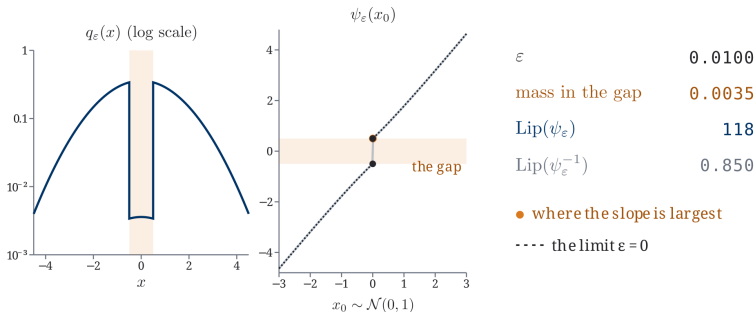
- The realistic case: a *nearly* hard-gapped target costs almost as much as a hard-gapped one.
- Universality says the approximation exists; this says what it costs.
- A residual flow of K blocks with $\text{Lip}(g_k) \leq \kappa < 1$ has $\text{BiLip} \leq \max(1 + \kappa, (1 - \kappa)^{-1})^K$ (Cornish et al. 2020, eq. 6): **bounded by construction**, against a theorem that demands unbounded.

The price of a gap, in closed form



The price of a gap, as a sweep

The price of a gap: less mass, a steeper map



At $\epsilon = 0$ the map jumps from -0.5 to +0.5 at $x_0 = 0$:
no longer a homeomorphism. The support proposition takes over.

Animated in the web deck — final frame shown.

Reading the checkerboard correctly

DIAGNOSTICS: EXPLAINED, AND NOT

Observation on a trained flow	Status
nonzero model density in hard-zero gaps	structural (full-support source + diffeomorphism)
extreme Jacobian conditioning as the mismatch is pushed	supported by the theorem
inverse round-trip degradation, instability	plausible numerical consequence; measure it
NLL plateau, poor samples	non-specific : capacity or optimization

What 29.8% is a number about

- **Structural:** every finite, regular, Gaussian-based flow puts nonzero density in the gaps.
- **Not structural:** the *amount*. 29.8% is a fact about one six-layer model, trained 8000 steps at a bounded log-scale — not about diffeomorphisms.
- The bridge can be pushed toward zero; the price is conditioning (the theorem, and the one-dimensional family).
- The half nat above $\log 18$ is $\text{KL}(q \parallel p_\theta)$: never exactly zero for a flow of this unit, but with no positive floor — it can be driven toward zero at the conditioning cost. 0.50 is what this architecture paid.

Topology is a hypothesis to test, not a verdict every bridge proves.

Where the obstruction goes next

- **U2.** A flow of a well-posed differential equation is *still* a diffeomorphism of $\mathbb{R}^d$ onto itself. Continuous time changes the parameterization and the costs, not this constraint. U2 adds a separate restriction of its own; the two stay apart.
- **U3.** A hard-support target forces a choice at the data end of the path from p_0 to q : a limiting endpoint, or a smoothed one.
- **U6.** Manifold data: choose a source and a map that live on the right space; on that space a regular flow is still a diffeomorphism, and the proposition applies unchanged.

05

**Residual flows:
invertibility
without
triangularity**

The dialect, declared once

Everything so far bought a cheap log-det by forcing a **triangular** Jacobian. Residual flows give that up.

The literature writes the map the way a residual network computes it — data to source. That is the density transform:

$$f = \psi^{-1}, \quad f(x) = x + g(x), \quad g: \mathbb{R}^d \rightarrow \mathbb{R}^d \text{ a network.}$$

The identity plus a perturbation. **If the perturbation is small enough, the identity wins.** “Small enough” is a Lipschitz condition (Behrmann et al. 2019, sec. 2).

Invertibility by Banach

Fix z ; $x + g(x) = z$ is the fixed-point equation $x = z - g(x)$. Iterate:

$$x^{(m+1)} = z - g(x^{(m)}), \quad x^{(0)} = z.$$

LEMMA (THE ITERATION MAP IS A CONTRACTION)

If $\text{Lip}(g) = L < 1$, then $T_z(x) = z - g(x)$ satisfies

$$\|T_z(x) - T_z(x')\| = \|g(x') - g(x)\| \leq L \|x - x'\|.$$

Banach on the complete space $\mathbb{R}^d$: a unique fixed point, hence a unique solution of $f(x) = z - f$ **is a bijection of $\mathbb{R}^d$ onto itself**, and the iteration converges to $f^{-1}(z)$ from anywhere (Behrmann et al. 2019, Thm. 1). On the *Measured, on one block* slide: $L = 0.5$ reaches 10^{-10} in 27 iterations; $L = 0.95$ has not after 60.

PS1.5 — residual-flow invertibility via Banach

With $\text{Lip}(g) = L < 1$ and $f(x) = x + g(x)$:

- **(a)** $T_z(x) = z - g(x)$ is an L -contraction, with the definitions written out.
- **(b)** Banach on $\mathbb{R}^d$ (say why it is complete): f is a bijection of $\mathbb{R}^d$ onto itself and the iteration converges to $f^{-1}(z)$ from every start.
- **(c)** $\text{Lip}(f) \leq 1 + L$ and $\text{Lip}(f^{-1}) \leq 1/(1 - L)$: f is a homeomorphism, and a diffeomorphism when g is C^1 .
- **(d) Bonus.** $\|x^{(m)} - x^*\| \leq L^m(1 - L)^{-1}\|x^{(1)} - x^{(0)}\|$; iterations for tolerance δ ; the cost of one sample from K blocks.

Enforcing $L < 1$: spectral normalization — divide each weight matrix by (a bound on) its largest singular value (Behrmann et al. 2019, sec. 2.1). This is the bounded constant of the topology block.

The log-determinant, as a series

$J_f = I + J_g$ is **dense**. Under $\|J_g\|_2 \leq L < 1$: every eigenvalue of J_f has positive real part, so $\det J_f > 0$; and $\log \det = \text{tr} \log$ has a power series:

$$\log \det (I + J_g) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} \text{tr} (J_g^k),$$

valid when $\rho(J_g) < 1$.

The condition is part of the formula — the Maclaurin series of $\log(1 + a)$ with a matrix for a , convergent for the same reason. Truncation after n terms: error $\leq d \sum_{k>n} L^k/k$, geometric at rate L (Behrmann et al. 2019, sec. 3, Thm. 3).

Hutchinson's estimator

Infinitely many traces of matrices that are never formed. The device:

HUTCHINSON'S TRACE ESTIMATOR (V1)

For $A \in \mathbb{R}^{d \times d}$ and a random v with $\mathbb{E}_v[vv^\top] = I_d$ (Rademacher ± 1 , or Gaussian):

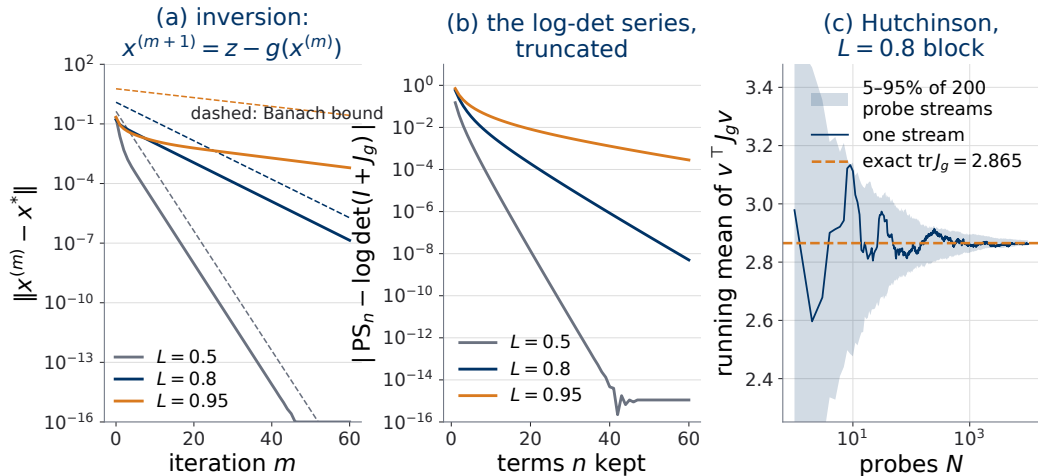
$$\text{tr}(A) = \mathbb{E}_v[v^\top A v].$$

Usage. $v^\top J_g^k v$ by k vector-Jacobian products: $w^\top \leftarrow v^\top$, then $w^\top \leftarrow w^\top J_g$ k times, then $w^\top v$. One backward pass of g per product; J_g never materialized. One probe serves every term.

Deferred. Proof and variance: U2.T2, where the estimator carries the weight.

Named only: the truncated series is biased; the Russian-roulette estimator of (Chen et al. 2019) removes the bias by a random truncation with reweighting.

Measured, on one block



Residual flows on the table

RESIDUAL FLOWS, PLACED

	Density / likelihood	Sampling
Residual flow	$f(x) = x + g(x)$ direct ; $\log \det J $ an infinite series, estimated stochastically	$\psi = f^{-1}$ by fixed-point iteration , several evaluations of g per block

One row stochastic and approximate, the other iterative and exact to any tolerance. **Not** “both directions are stochastic”.

06 | Depth to infinity: the cliffhanger

A residual stack with a step size

Put a step $h = 1/K$ in front of each residual branch and let the branch depend on a time $t_k = kh$:

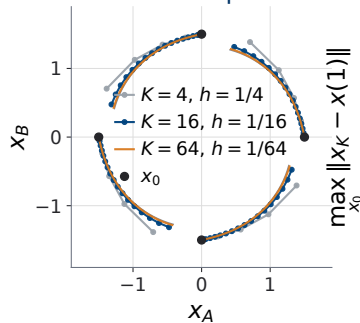
$$x_{k+1} = x_k + h f(t_k, x_k), \quad k = 0, \dots, K-1.$$

The letter f is now the **right-hand side of a differential equation**, time first — the course's $f(t, x)$; the residual branch of the residual-flow block is $g_k = hf(t_k, \cdot)$. This is the **forward-Euler** step, and as $h \rightarrow 0$ the stack's output converges, for a regular enough f , to the solution of

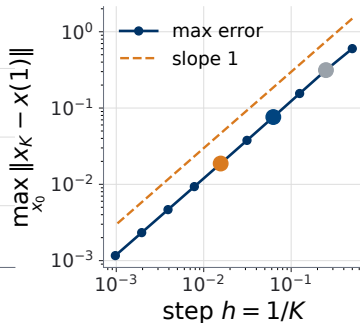
$$\frac{dx}{dt} = f(t, x), \quad t \in [0, 1].$$

Measured: the stack converges as $h \rightarrow 0$

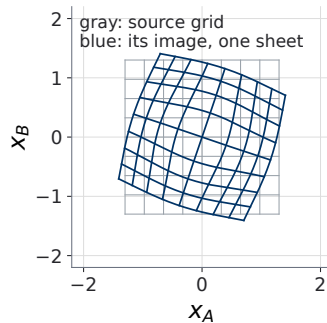
(a) $x_{k+1} = x_k + h f(t_k, x_k)$,
three depths



(b) halving h
halves the error

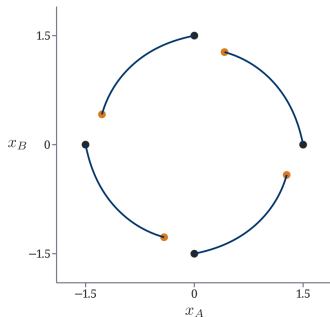


(c) the source grid
under the $K=64$ stack



The stack, one doubling at a time

Depth to infinity: the residual stack refines into a curve



$$x_{k+1} = x_k + h f(t_k, x_k), \quad h = 1/K$$

$K = 256$

$$h = 1/256$$

```
max endpoint error    0.0047
```

$$\text{error}(K/2) / \text{error}(K) \quad 2.003$$

— the solution of $dx/dt = f(t, x)$

- the stack, K blocks

- where the stack ends

The log-determinant of the stack is a sum of K terms.

What does that sum become as $K \rightarrow \infty$? (U2)

Animated in the web deck — final frame shown.

The promise, without overclaiming

- Discrete residual invertibility needed a **strong layerwise condition**: every block a contraction. That capped the bi-Lipschitz constant (the topology block) and is what spectral normalization enforces.
- A well-posed equation is invertible on its interval under existence-and-uniqueness and no-blow-up conditions: run it backwards. **The field still needs regularity** — a Lipschitz-type assumption does not disappear — but the architecture is no longer forced to make every finite block a contraction.
- The stack's log-det is $\sum_k \log \det(I + hJ_f)$. As $h \rightarrow 0$ it wants to become an integral. **What it becomes is not derived here.**

Left on the board for U2

What does $\log|\det J|$ become in continuous time?

Not answered today. U2.T1 sets up the flows of differential equations and the cost of a solver inside every training step; U2.T2 answers this question, with Hutchinson's estimator on the critical path.

07 | Unit close: the two costs, version 2

The two costs, version 2

THE TWO COSTS (V2)

Family	Density	Sampling	Trick, and its price
NICE / RealNVP	exact, triangular	direct	coupling; depth
Glow	exact	direct	mixing, multi-scale
MAF	parallel	sequential	AR, data side
IAF	sequential	parallel	AR, source side
Residual flow	direct; series log-det	iterative	contraction, traces

Version 3 in U2.T2, where the cells change. U3.L1 touches neither column.

Reading the unit

- **Exactness** was achieved in U1.T1, by construction.
- **Architecture choices expose computational dualities:** a triangle read in either direction; the direction decides which cost is sequential.
- **Support topology and conditioning reveal a real limitation**, not a capacity gap — with a calibrated way to tell the two apart.
- **Residual flows** dropped triangularity, paid with a series and an iteration, and pointed at continuous time as the escape from *architectural bookkeeping* — not from mathematics.

Reading the board

On the board	Status
The depth question	Answered twice (universality)
The fine print	Examined (topology): support, not level sets
The checkerboard gaps	Explained and calibrated (topology)
$\log \det J $ in continuous time	Open. U2

Assigned today: **PS1.5**. Next session: the lab (U1.L2), multi-scale flows on images. The theory session after it opens U2 on the question above.

References I

- Behrmann, Jens, Will Grathwohl, Ricky T. Q. Chen, David Duvenaud, and Jörn-Henrik Jacobsen. 2019. "Invertible Residual Networks." In *International Conference on Machine Learning*. <https://arxiv.org/abs/1811.00995>.
- Chen, Ricky T. Q., Jens Behrmann, David Duvenaud, and Jörn-Henrik Jacobsen. 2019. "Residual Flows for Invertible Generative Modeling." In *Advances in Neural Information Processing Systems*. <https://arxiv.org/abs/1906.02735>.
- Cornish, Rob, Anthony Caterini, George Deligiannidis, and Arnaud Doucet. 2020. "Relaxing Bijectivity Constraints with Continuously Indexed Normalising Flows." In *International Conference on Machine Learning*. <https://arxiv.org/abs/1909.13833>.

References II

- Draxler, Felix, Stefan Wahl, Christoph Schnörr, and Ullrich Köthe. 2024. “On the Universality of Volume-Preserving and Coupling-Based Normalizing Flows.” In *International Conference on Machine Learning*.
<https://arxiv.org/abs/2402.06578>.
- Kingma, Diederik P., and Prafulla Dhariwal. 2018. “Glow: Generative Flow with Invertible 1×1 Convolutions.” In *Advances in Neural Information Processing Systems*. <https://arxiv.org/abs/1807.03039>.
- Kingma, Diederik P., Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. 2016. “Improved Variational Inference with Inverse Autoregressive Flow.” In *Advances in Neural Information Processing Systems*.
<https://arxiv.org/abs/1606.04934>.

References III

- Papamakarios, George, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. 2021. "Normalizing Flows for Probabilistic Modeling and Inference." *Journal of Machine Learning Research* 22 (57): 1–64. <https://arxiv.org/abs/1912.02762>.
- Papamakarios, George, Theo Pavlakou, and Iain Murray. 2017. "Masked Autoregressive Flow for Density Estimation." In *Advances in Neural Information Processing Systems*. <https://arxiv.org/abs/1705.07057>.