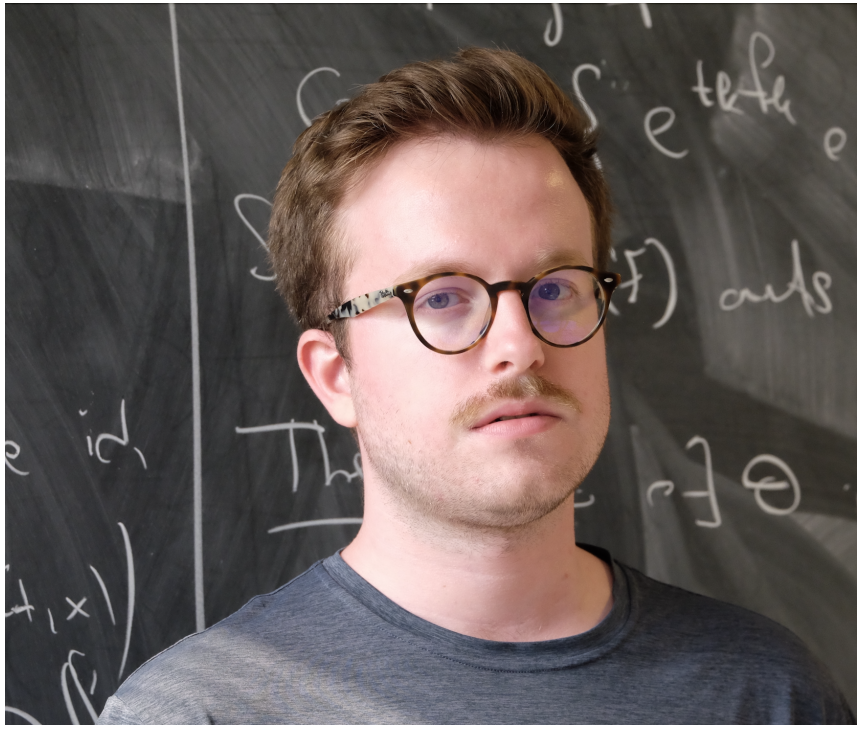
Entropy-driven control of the continuity equation for normalizing flows

Antonio Álvarez-López

FAU Erlangen–Nürnberg
Universidad Autónoma de Madrid

September 29, 2025

Joint work with...



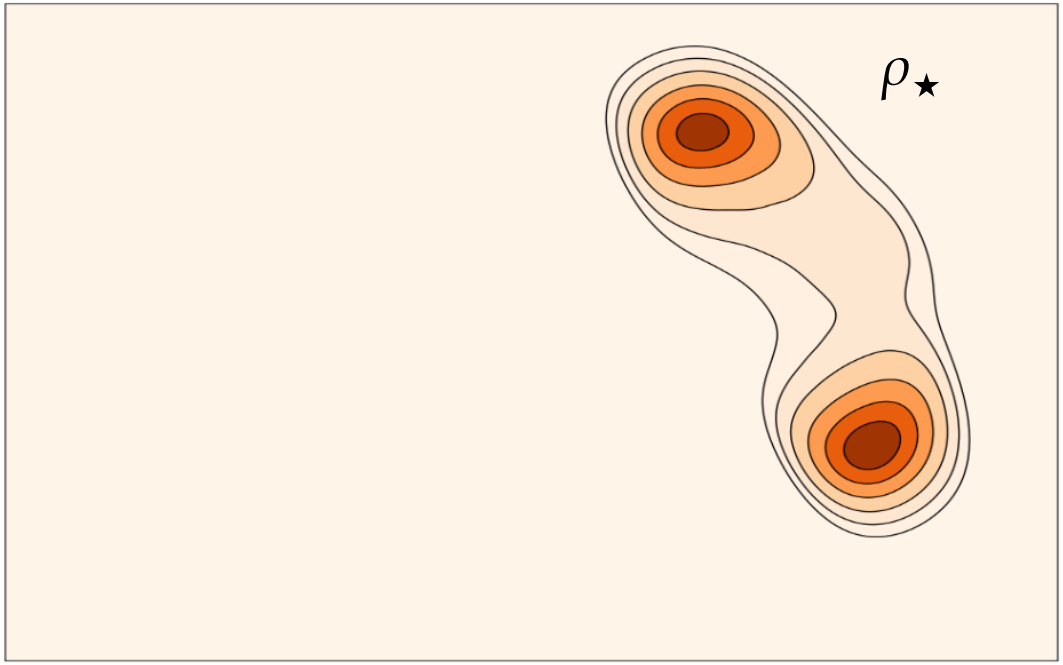
Borjan Geshkovski
Inria Paris — Sorbonne Université (LJLL).

Domènec Ruiz-Balet
Universidad de Barcelona.

Outline

Generative modeling

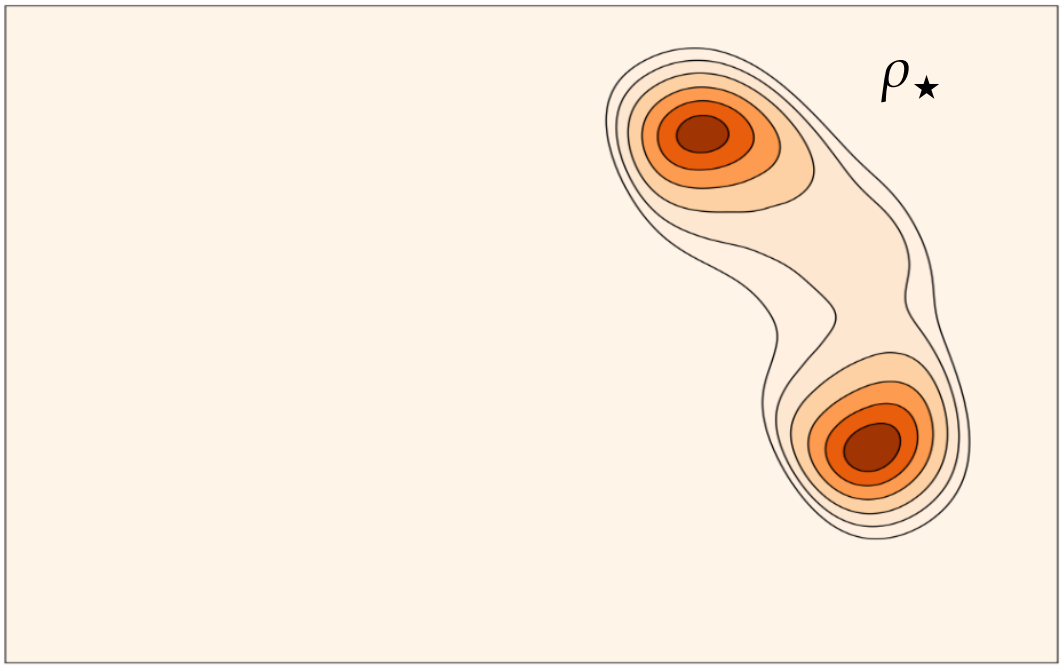
What is generative modeling?



Generative modeling

What is generative modeling?

Generate samples that follow a law... (complex and high dimension!)
by transforming samples from a base law via a learned map T:



Generative models

Flow-based generative models:

$$\mathsf{T} \;=\; \text{Flow map of an ODE/SDE.}$$

- **Continuous Normalizing Flows (CNFs).** Invertible ODE flows trained by maximum likelihood. *Chen et al., 2018; Grathwohl et al., 2019.*
- **Diffusion Models.** Forward noising and reverse SDE/ODE denoiser using scores. *Ho et al., 2020; Song et al., 2021.*
- **Flow Matching.** Regress a time–dependent field to prescribed probability paths for fast few–step sampling. *Lipman et al., 2023.*

Common thread:

$$\underbrace{\mathcal{N}(0, \mathrm{Id})}_{\text{easy to sample}} \longrightarrow \underbrace{\rho_\star}_{\text{hard to sample}}$$

Generative models

Flow-based generative models:

$$\mathsf{T} \;=\; \text{Flow map of an ODE/SDE.}$$

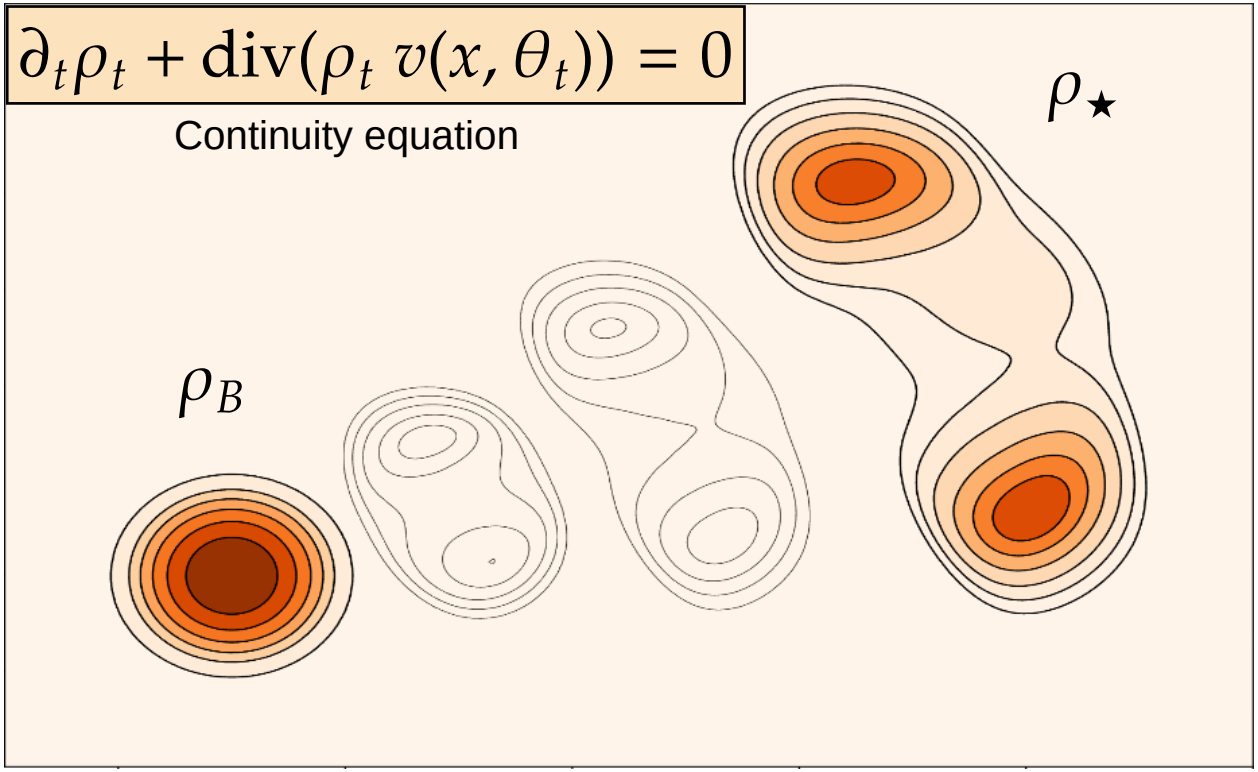
- **Continuous Normalizing Flows (CNFs).** Invertible ODE flows trained by maximum likelihood. *Chen et al., 2018; Grathwohl et al., 2019.*
- **Diffusion Models.** Forward noising and reverse SDE/ODE denoiser using scores. *Ho et al., 2020; Song et al., 2021.*
- **Flow Matching.** Regress a time–dependent field to prescribed probability paths for fast few–step sampling. *Lipman et al., 2023.*

Common thread:

$$\underbrace{\mathcal{N}(0, \mathrm{Id})}_{\text{easy to sample}} \longrightarrow \underbrace{\rho_\star}_{\text{hard to sample}}$$

Continuous Normalizing Flows

- Base $\rho_B \equiv \mathcal{N}(0, \mathrm{Id}) \mapsto$ Target $\rho_\star$ known or estimated from data $x_1, \dots, x_N \sim \rho_\star$.



- Fix Lipschitz control family $\{v(\cdot, \theta)\}_\theta$. Find θ and $T > 0$ such that $\rho_T^\theta \approx \rho_\star$.
- Sample $x \sim \rho_B$ follows the flow: $\begin{cases} \frac{d}{dt}\Phi_t^\theta(x) & = v(\Phi_t^\theta(x), \theta_t) \\ \Phi_0^\theta(x) & = x \end{cases}$

Continuous Normalizing Flows

- **Optimization**:

$$\min_{\theta} \ \mathrm{KL}(\mu_\star \| \mu_T^\theta)$$

$$\mu_T^\theta = (\Phi_T^\theta)_{\#}\mu_{\mathrm{B}}$$
$$\mathrm{d}\mu_\star = \rho_\star \, dx$$
$$\mathrm{d}\mu_T^\theta = \rho_T^\theta \, dx$$

where the Kullback-Leibler divergence/relative entropy is

$$\mathrm{KL}(\mu_\star \| \mu_T^\theta) := \begin{cases} \mathbb{E}_{\mu_\star}\left[\log\left(\frac{\mathrm{d}\mu_\star}{\mathrm{d}\mu_T^\theta}\right)\right] = \int_{\mathbb{R}^d} \log\left(\frac{\rho_\star}{\rho_T^\theta}\right)\rho_\star, & \mu_\star \ll \mu_T^\theta, \\ +\infty, & \text{otherwise.} \end{cases}$$

Continuous Normalizing Flows

- **Optimization**:

$$\min_{\theta} \ \mathrm{KL}(\mu_\star \| \mu_T^\theta)$$

$$\mu_T^\theta = (\Phi_T^\theta)_{\#}\mu_B$$
$$\mathrm{d}\mu_\star = \rho_\star \, dx$$
$$\mathrm{d}\mu_T^\theta = \rho_T^\theta \, dx$$

where the Kullback-Leibler divergence/relative entropy is

$$\mathrm{KL}(\mu_\star \| \mu_T^\theta) := \begin{cases} \mathbb{E}_{\mu_\star}\left[\log\left(\frac{\mathrm{d}\mu_\star}{\mathrm{d}\mu_T^\theta}\right)\right] = \int_{\mathbb{R}^d} \log\left(\frac{\rho_\star}{\rho_T^\theta}\right)\rho_\star, & \mu_\star \ll \mu_T^\theta, \\ +\infty, & \text{otherwise.} \end{cases}$$

- **Why KL?** Max. likelihood $\iff$ Min. KL divergence:

Empirical MLE: $\max_{\theta} \ \mathbb{P}_\theta(x_1, \ldots, x_n) \overset{\text{i.i.d.}}{=} \max_{\theta} \prod_{i=1}^n \rho_T^\theta(x_i) \iff \max_{\theta} \frac{1}{n}\sum_{i=1}^n \log \rho_T^\theta(x_i)$
$(x_1,\ldots,x_n \sim \mu_\star)$

Population objective: $\mathrm{KL}(\mu_\star \| \mu_T^\theta) = \underbrace{\mathbb{E}_{\mu_\star}[\log \rho_\star]}_{\text{indep. of } \theta} - \mathbb{E}_{\mu_\star}[\log \rho_T^\theta].$

Continuous Normalizing Flows

We can view it as an **approximate controllability problem**:

Given $\varepsilon > 0$, find $\theta = (\theta_t)_{t\in[0,T]}$ such that

$$\mathrm{KL}(\mu_\star \| \mu_T^\theta) < \varepsilon, \quad \text{s.t.} \quad \begin{cases} \partial_t \rho_t^\theta(x) + \mathrm{div}\big(\rho_t^\theta(x)\, v(x;\theta_t)\big) = 0, \\ \rho_0(x) = \rho_{\mathrm{B}}(x). \end{cases}$$

Continuous Normalizing Flows

We can view it as an **approximate controllability problem**:

Given $\varepsilon > 0$, find $\theta = (\theta_t)_{t\in[0,T]}$ such that

$$\mathrm{KL}(\mu_\star \| \mu_T^\theta) < \varepsilon, \quad \text{s.t.} \quad \begin{cases} \partial_t \rho_t^\theta(x) + \mathrm{div}\big(\rho_t^\theta(x)\, v(x;\theta_t)\big) = 0, \\ \rho_0(x) = \rho_{\mathrm{B}}(x). \end{cases}$$

- **Difficulty 1: Not a distance**

  KL is non-symmetric $+$ no triangle inequality!

Continuous Normalizing Flows

We can view it as an **approximate controllability problem**:

Given $\varepsilon > 0$, find $\theta = (\theta_t)_{t\in[0,T]}$ such that

$$\mathrm{KL}(\mu_\star \| \mu_T^\theta) < \varepsilon, \quad \text{s.t.} \quad \begin{cases} \partial_t \rho_t^\theta(x) + \mathrm{div}\big(\rho_t^\theta(x)\, v(x;\theta_t)\big) = 0, \\ \rho_0(x) = \rho_{\mathrm{B}}(x). \end{cases}$$

- **Difficulty 1: Not a distance**

  KL is non-symmetric + no triangle inequality!

- **Difficulty 2: Blow-up**. We have $\mathrm{KL} \to +\infty$ whenever
  - **Smaller support.** $\exists A \subset \mathbb{R}^d$ with $\mu_\star(A) > 0$ and $\mu_T^\theta(A) = 0$.
  - **Relative decay.** $\mathbb{E}_{\mu_\star}\big[\,|\log \rho_T^\theta(X)|\,\big] \to +\infty$.

(Example) $\rho_\star(x) = \mathbf{1}_{(0,1)}(x)$, $\rho_T^\theta(x) = \dfrac{e^{-1/x}}{Z}\,\mathbf{1}_{(0,1)}(x)$, $\mathrm{KL} = \log Z + \displaystyle\int_0^1 \frac{\mathrm{d}x}{x} = +\infty$

Continuous Normalizing Flows

We can view it as an **approximate controllability problem**:

Given $\varepsilon > 0$, find $\theta = (\theta_t)_{t\in[0,T]}$ such that

$$\mathrm{KL}(\mu_\star \| \mu_T^\theta) < \varepsilon \quad \text{s.t.} \quad \begin{cases} \partial_t \rho_t^\theta(x) + \mathrm{div}\big(\rho_t^\theta(x)\, v_t(x;\theta_t)\big) = 0, \\ \rho_0(x) = \rho_{\mathrm{B}}(x). \end{cases}$$

$\Longrightarrow$ We need control over the full support and over the relative tail decay!

- Our controllability result is valid for densities supported on $\mathbb{R}^d$.
- But $\mu_\star$ is an empirical measure built with n samples...

  In practice: build new target $\widehat{\mu}_{\star,n} \approx \mu_\star$ via (Gaussian) KDE.

Controllability of continuity/Liouville equation:

Brief review

- ***Moser (1965); Dacorogna–Moser (1990)*** — Mass-preserving diffeomorphisms; baseline density-to-density flows.
- ***Duprez–Morancey–Rossi (2019)*** — Continuity eq. with localized control; approx. controllability (crossing); exact with low regularity.
- ***Duprez–Morancey–Rossi (2020)*** — Minimal time for localized control; sharp mass-through-control-region condition.
- ***Raginsky (2024)*** — Liouville controllability; implements broad classes of diffeomorphisms for linear systems; OT links.
- ***Brockett (notes)*** — Ensemble controllability framework; state-dependent feedback viewpoint.
- **OT/OC link** — *Benamou–Brenier (2000)* dynamic OT; *Elamvazhuthi (2023)* dynamical OT for control-affine systems.

Control family

Neural ODEs (ReLU, one-neuron)

$$v(x, \theta_t) = w(t)(\langle a(t), x\rangle + b(t))_+, \qquad \theta \in L^\infty\left(0, T; \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}\right)$$

- **Simple** control but **high** expressivity![1]

[1] *Ruiz-Balet--Zuazua, 2023., Á-L.--Orive--Illera--Zuazua, 2024.*

Control family

Neural ODEs (ReLU, one-neuron)

$$v(x,\theta_t) = w(t)(\langle a(t), x\rangle + b(t))_+, \qquad \theta \in L^\infty\left(0, T; \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}\right)$$

- **Simple** control but **high** expressivity![1]
- The solution of

$$\partial_t \rho_t^\theta(x) + \operatorname{div}\left(\rho_t^\theta(x) w(\langle a, x\rangle + b)_+\right) = 0, \qquad \rho_0 = \rho_{\mathrm{B}}$$

has **closed-form** (by the method of characteristics):

$$\rho_t^\theta(x) = \begin{cases} \rho_{\mathrm{B}}(x), & x \in H^- := \{\langle a, x\rangle + b \leq 0\}, \\ e^{-t\langle w,a\rangle}\, \rho_{\mathrm{B}}(\mathrm{A}_t(x)), & x \in H^+ := \{\langle a, x\rangle + b > 0\}, \end{cases}$$

with

$$\mathrm{A}_t(x) = e^{-t\, wa^\top} x - \frac{b}{\langle w, a\rangle}\left(1 - e^{-t\langle w,a\rangle}\right) w \quad \to \quad \textbf{Affine} \text{ in } x.$$

[1] *Ruiz-Balet--Zuazua, 2023., Á-L.--Orive--Illera--Zuazua, 2024.*

Control family

Neural ODEs (ReLU, one-neuron)

$$v(x,\theta_t) = w(t)(\langle a(t), x\rangle + b(t))_+, \qquad \theta \in L^\infty\left(0, T; \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}\right)$$

- **Simple** control but **high** expressivity![1]
- The solution of

$$\partial_t \rho_t^\theta(x) + \operatorname{div}\left(\rho_t^\theta(x) w(\langle a, x\rangle + b)_+\right) = 0, \qquad \rho_0 = \rho_{\mathrm{B}}$$

has **closed-form** (by the method of characteristics):

$$\rho_t^\theta(x) = \begin{cases} \rho_{\mathrm{B}}(x), & x \in H^- := \{\langle a, x\rangle + b \le 0\}, \\ e^{-t\langle w,a\rangle}\, \rho_{\mathrm{B}}(\mathrm{A}_t(x)), & x \in H^+ := \{\langle a, x\rangle + b > 0\}, \end{cases}$$

with

$$\mathrm{A}_t(x) = e^{-t\, wa^\top} x - \frac{b}{\langle w, a\rangle}\left(1 - e^{-t\langle w,a\rangle}\right) w \quad \to \quad \textbf{Affine} \text{ in } x.$$

Degenerate case. If $\langle w, a\rangle = 0$, then $\mathrm{A}_t(x) = (I - t\, wa^\top)x - t\, b\, w$.

[1] *Ruiz-Balet--Zuazua, 2023., Á-L.--Orive--Illera--Zuazua, 2024.*

Main result

Theorem

Let $\mu_{\mathrm{B}} = \mathcal{N}(m_{\mathrm{B}}, \Sigma_{\mathrm{B}})$ *and a probability density* $\rho_\star$ *s.t. for some* $M, \sigma_\bullet > 0$

$$\rho_\star(x) \leq \rho_\bullet(x) := \frac{1}{(2\pi\sigma_\bullet^2)^{d/2}} e^{-\|x\|^2/(2\sigma_\bullet^2)} \quad (\|x\| \geq M).$$

Then for all $T > 0, \varepsilon > 0$ *there exists a piecewise-constant* θ *such that*

$$\mathrm{KL}\left(\mu_\star \,\|\, \mu_T^\theta\right) = \int \rho_\star \log \frac{\rho_\star}{\rho_T^\theta} < \varepsilon.$$

Main result

Theorem

Let $\mu_{\mathrm{B}} = \mathcal{N}(m_{\mathrm{B}}, \Sigma_{\mathrm{B}})$ *and a probability density* $\rho_\star$ *s.t. for some* $M, \sigma_\bullet > 0$

$$\rho_\star(x) \leq \rho_\bullet(x) := \frac{1}{(2\pi\sigma_\bullet^2)^{d/2}} e^{-\|x\|^2/(2\sigma_\bullet^2)} \quad (\|x\| \geq M).$$

Then for all $T > 0, \varepsilon > 0$ *there exists a piecewise-constant* θ *such that*

$$\mathrm{KL}\left(\mu_\star \,\|\, \mu_T^\theta\right) = \int \rho_\star \log \frac{\rho_\star}{\rho_T^\theta} < \varepsilon.$$

Number of discontinuities: We can choose

$$N_{\mathrm{d}}(\varepsilon) \;\leq\; (d+10)\left\lceil \tfrac{2R_\varepsilon}{h_\varepsilon} \right\rceil^d + 2d$$

where $[-R_\varepsilon, R_\varepsilon]^d$ contains $\geq 1-\varepsilon$ mass of both μ_{B} and $\mu_\star$; and h_ε mesh size of uniform grid on which piecewise-constant interpolants have L^1 error $\leq \varepsilon$.

Main result

Theorem

Let $\mu_{\mathrm{B}} = \mathcal{N}(m_{\mathrm{B}}, \Sigma_{\mathrm{B}})$ *and a probability density* $\rho_\star$ *s.t. for some* $M, \sigma_\bullet > 0$

$$\rho_\star(x) \le \rho_\bullet(x) := \frac{1}{(2\pi\sigma_\bullet^2)^{d/2}} e^{-\|x\|^2/(2\sigma_\bullet^2)} \quad (\|x\| \ge M).$$

Then for all $T > 0, \varepsilon > 0$ *there exists a piecewise-constant* θ *such that*

$$\mathrm{KL}\left(\mu_\star \,\|\, \mu_T^\theta\right) = \int \rho_\star \log \frac{\rho_\star}{\rho_T^\theta} < \varepsilon.$$

Number of discontinuities: We can choose

$$N_{\mathrm{d}}(\varepsilon) \;\le\; (d+10) \underbrace{\left\lceil \tfrac{2R_\varepsilon}{h_\varepsilon} \right\rceil^d}_{\text{Number of cells}} + 2d$$

where $[-R_\varepsilon, R_\varepsilon]^d$ contains $\ge 1-\varepsilon$ mass of both μ_{B} and $\mu_\star$; $h_\varepsilon \equiv$ mesh size of a uniform grid on which Riemann approximation has L^1-error $\le \varepsilon$.

Proof idea: (i) + (ii) + (iii)

(i) **TV-controllability** (Prior work, *Ruiz-Balet--Zuazua, 2024*):

$$[0, T/2]: \qquad \left|\mu^{\theta}_{T/2} - \mu_\star\right|_{\mathrm{TV}} = \frac{1}{2}\|\rho^{\theta}_{T/2} - \rho_\star\|_{L^1} < \varepsilon \quad \text{(constructive } \theta\text{)}.$$

Proof idea: (i) + (ii) + (iii)

(i) **TV-controllability** (Prior work, *Ruiz-Balet--Zuazua, 2024*):

$$[0, T/2]: \qquad \left|\mu^{\theta}_{T/2} - \mu_\star\right|_{\mathrm{TV}} = \frac{1}{2}\|\rho^{\theta}_{T/2} - \rho_\star\|_{L^1} < \varepsilon \quad \text{(constructive } \theta\text{)}.$$

(ii) **Positivity/Tail control:**

$$[T/2, T]: \qquad \rho^{\theta}_T > 0 \quad \text{and} \quad \lim_{\|x\|\to\infty} \frac{\rho_\star(x)}{\rho^{\theta}_T(x)} < \infty \quad \text{(bounded Radon–Nikodym deriva}$$

Proof idea: (i) + (ii) + (iii)

(i) **TV-controllability** (Prior work, *Ruiz-Balet--Zuazua, 2024*):

$$[0, T/2]: \qquad \left|\mu^\theta_{T/2} - \mu_\star\right|_{\mathrm{TV}} = \frac{1}{2}\|\rho^\theta_{T/2} - \rho_\star\|_{L^1} < \varepsilon \quad \text{(constructive } \theta\text{)}.$$

(ii) **Positivity/Tail control:**

$$[T/2, T]: \qquad \rho^\theta_T > 0 \quad \text{and} \quad \lim_{\|x\|\to\infty} \frac{\rho_\star(x)}{\rho^\theta_T(x)} < \infty \quad \text{(bounded Radon–Nikodym deriva}$$

(iii) Pinsker inequality: $|\mu_\star - \mu^\theta_T|_{\mathrm{TV}} \le \sqrt{2 \cdot \mathrm{KL}(\mu_\star \,\|\, \mu^\theta_T)}$ useless here.

Proof idea: (i) + (ii) + (iii)

(i) **TV-controllability** (Prior work, *Ruiz-Balet--Zuazua, 2024*):

$$[0, T/2]: \qquad \left|\mu^\theta_{T/2} - \mu_\star\right|_{\mathrm{TV}} = \frac{1}{2}\|\rho^\theta_{T/2} - \rho_\star\|_{L^1} < \varepsilon \quad \text{(constructive } \theta\text{)}.$$

(ii) **Positivity/Tail control:**

$$[T/2, T]: \qquad \rho^\theta_T > 0 \quad \text{and} \quad \lim_{\|x\|\to\infty} \frac{\rho_\star(x)}{\rho^\theta_T(x)} < \infty \quad \text{(bounded Radon–Nikodym deriva}$$

(iii) Pinsker inequality: $|\mu_\star - \mu^\theta_T|_{\mathrm{TV}} \le \sqrt{2 \cdot \mathrm{KL}(\mu_\star \,\|\, \mu^\theta_T)}$ useless here.

Reverse Pinsker inequality lifts TV $\to$ KL:

$$\mathrm{KL}(\mu_\star\|\mu^\theta_T) \le \frac{1}{2} \cdot \frac{\log \left\| \frac{\mathrm{d}\mu_\star}{\mathrm{d}\mu^\theta_T} \right\|_\infty}{1 - \left\| \frac{\mathrm{d}\mu_\star}{\mathrm{d}\mu^\theta_T} \right\|_\infty^{-1}} \cdot |\mu_\star - \mu^\theta_T|_{\mathrm{TV}} = \frac{1}{2} \cdot \underbrace{\frac{\log \left\| \frac{\rho_\star}{\rho^\theta_T} \right\|_\infty}{1 - \left\| \frac{\rho_\star}{\rho^\theta_T} \right\|_\infty^{-1}}}_{\text{Finite thanks to (ii)}} \cdot \|\rho_\star - \rho^\theta_T\|_{L^1}$$

Tool I: Contraction in L^1

- (**Lemma**). Let $\varepsilon > 0$ and consider any perturbation

$$\rho_{\mathrm{B}}(\cdot) + \varepsilon(\cdot) \qquad \text{with} \qquad \|\varepsilon(\cdot)\|_{L^1(\mathbb{R}^d)} < \varepsilon.$$

Then, for any $v \in L^\infty(0, T; \mathrm{Lip}(\mathbb{R}^d; \mathbb{R}^d))$,

$$\partial_t \rho_t^\varepsilon(x) + \mathrm{div}\,(\rho_t^\varepsilon(x) v(t,x)) = 0, \qquad \rho_0^\varepsilon(\cdot) = \rho_{\mathrm{B}}(\cdot) + \varepsilon(\cdot)$$

satisfies

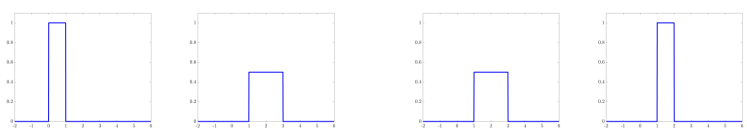
$$\|\rho_t^0 - \rho_t^\varepsilon\|_{L^1(\mathbb{R}^d)} < \varepsilon \qquad \text{for all } t > 0.$$

Step 1: Control in L^1 (Basic transformations)

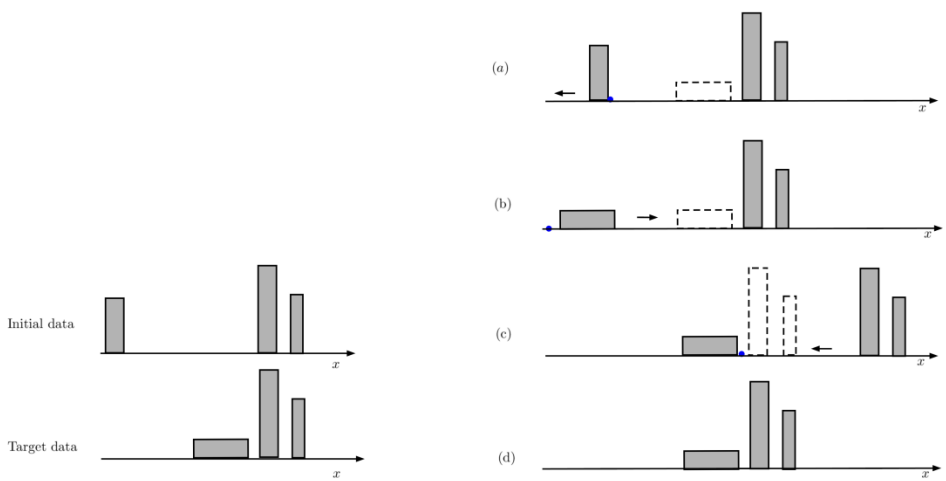
$$\partial_t \rho + \partial_x \big(w\, \sigma(x-b)\rho \big) = 0, \qquad \mathrm{d}\rho_{\mathrm{B}} = \eta\, \mathbb{1}_{(x_0, x_0+1/\eta)}$$

- $\underline{b = x_0}$: $\rho_t^\theta = \eta e^{-wt}\, \mathbb{1}_{(x_0,\, b+(x_0+\frac{1}{\eta}-b)e^{wt})}$ dilation/compression.
- $\underline{b < x_0}$: $\rho_t^\theta = \eta e^{-wt}\, \mathbb{1}_{(b+(x_0-b)e^{wt},\, b+(x_0+\frac{1}{\eta}-b)e^{wt})}$ translation + dilation.
- $\underline{x_0 < b < x_0 + \frac{1}{\eta}}$: $\rho_t^\theta = \eta\, \mathbb{1}_{(x_0,\, b)} + \eta e^{-wt}\, \mathbb{1}_{(b,\, b+(x_0+\frac{1}{\eta}-b)e^{wt})}$ split

STEP 1: Translation and dilation STEP 2: Compression/dilation

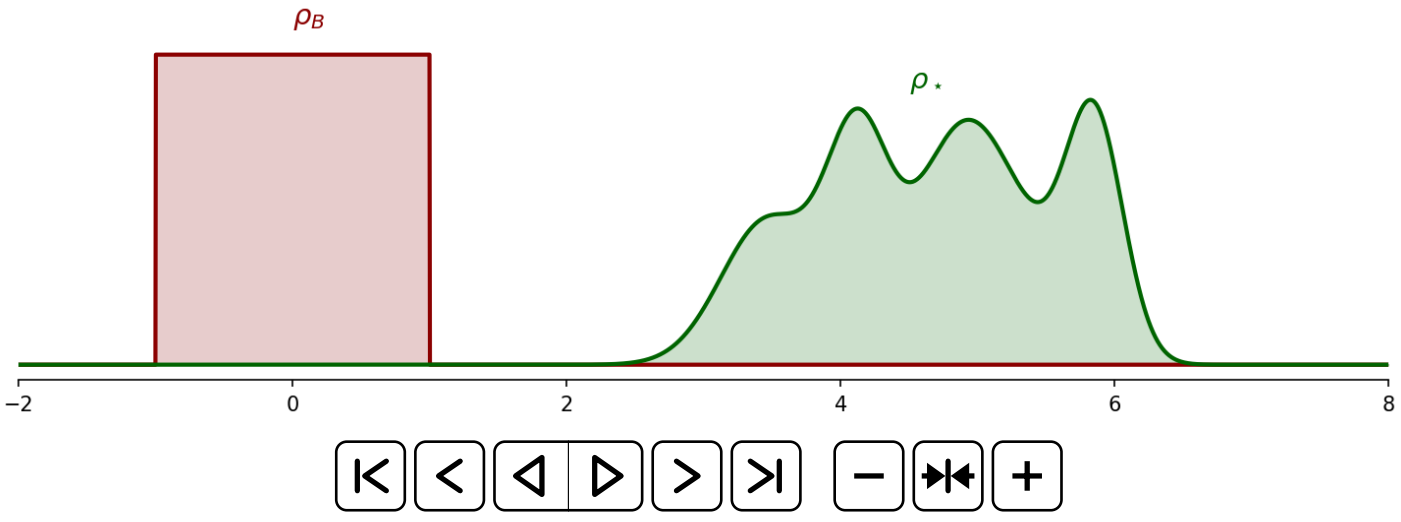


(a) Translation as composition of two steps.



(b) Localized translation.

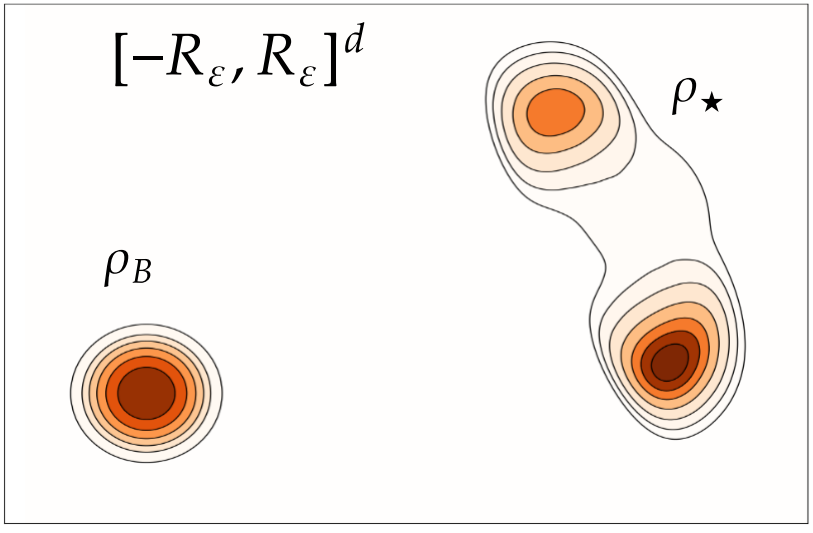
Step 1: Control in L^1 ($d = 1$)



Mechanism: Approximation by K rectangles + inductive exact control

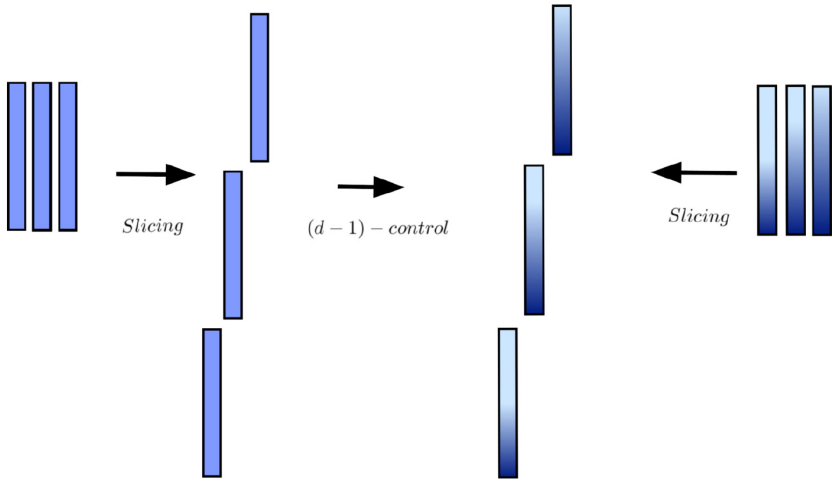
1. Set $\rho_{\mathrm{B}} = 1_{[x_0-1,x_0]}$ with $x_0 < \inf \operatorname{supp} \rho_\star$.
2. For $\varepsilon > 0$ choose $h = h(\varepsilon)$ and ε-Riemann approximations $\rho_\star^h$ and ρ_{B}^h in L^1 consisting of $K(\varepsilon)$ disjoint blocks. Each ordered pair must contain the same mass.
3. Inductively (i) compress/dilate and (ii) translate locally the rectangles, then $\rho_{\mathrm{B}}^h \to \rho_\star^h$ exactly. By L^1-contraction, it follows that $\|\rho_\star - \rho_{T/2}^\theta\|_{L^1} < \varepsilon$.

Step 1: Control in L^1 ($d > 1$)



1. **Discretization.** Let $[-R_\varepsilon, R_\varepsilon]^d$ contain $1-\varepsilon$ mass of ρ_{B} and $\rho_\star$. Grid it with step h, remove tiny strips, and take ε-Riemann approximations for $\rho_{\mathrm{B}}^h, \rho_\star^h$ in L^1.
2. **Simplified goal.** Transform any Riemann approximation on a grid into a constant function on the same grid.

Step 1: Control in L^1 ($d > 1$)



Iterative process to reduce control to $d = 1$.

1. **Discretization.** Let $[-R_\varepsilon, R_\varepsilon]^d$ contain $1-\varepsilon$ mass of ρ_{B} and $\rho_\star$. Grid it with step h, remove tiny strips, and take ε-Riemann approximations for $\rho_{\mathrm{B}}, \rho_\star^h$ in L^1.
2. **Simplified goal.** Transform any Riemann approximation on a grid into a constant density with the same mass and defined on the same grid.
3. **Dimension reduction.** Translate "mass columns" until they are ordered in coordinate d. Project onto the first $d-1$ coords.
4. **Iterate to $d = 1$ and apply Step 1.**

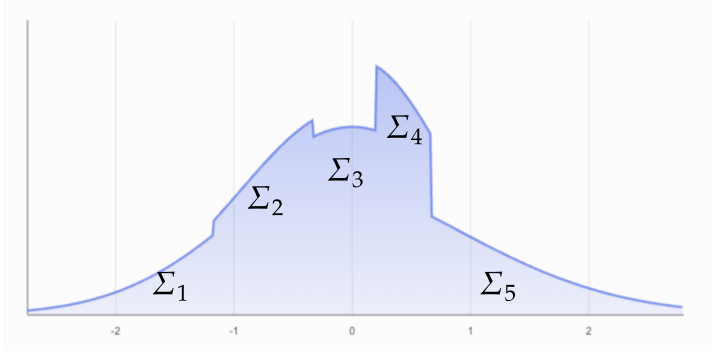
Step 2: Tail control

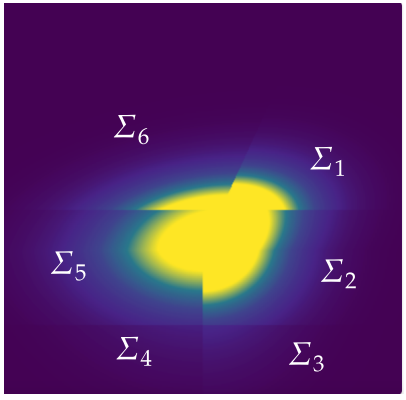
- Recall: Under a constant control (w, a, b) if $\langle a, x\rangle + b > 0$ then

$$x \mapsto \rho_t^{\theta}(x) = e^{-t\langle w,a\rangle}\rho_{\mathrm{B}}\left(e^{-t\,wa^{\top}}x + \text{shift}\right).$$

- $\Longrightarrow$ Step I transforms the initial Gaussian tail into a **piecewise Gaussian tail** with covariance matrices depending on each $(\langle a, x\rangle + b)_+$:

$$\Sigma \longmapsto \Sigma_i(t) := \left(e^{-t\,wa^{\top}}\right)\Sigma\left(e^{-t\,wa^{\top}}\right)^{\top} \qquad \text{or} \qquad \Sigma \longmapsto \Sigma_i(t) := \Sigma.$$





Evolution of Gaussian after piecewise constant controls.

Tool II: Variance/tail domination

- (**Lemma**). Suppose

$$\langle(\Sigma_2 - \Sigma_1)e_k, e_k\rangle > 0, \qquad \text{for some } k \in [d].$$

  Then, for any $M > 0$, there exists $M(k) > 0$ such that

$$\rho_1(x) < \rho_2(x) \qquad \text{in } \left\{x \in \mathbb{R}^d : |x_k| > M(k) \text{ and } |x_j| < M \text{ for } j \neq k\right\}.$$

- (**Lemma**). Suppose $\Sigma_1 \prec \Sigma_2$, then there exists some $\overline{M} > 0$ such that

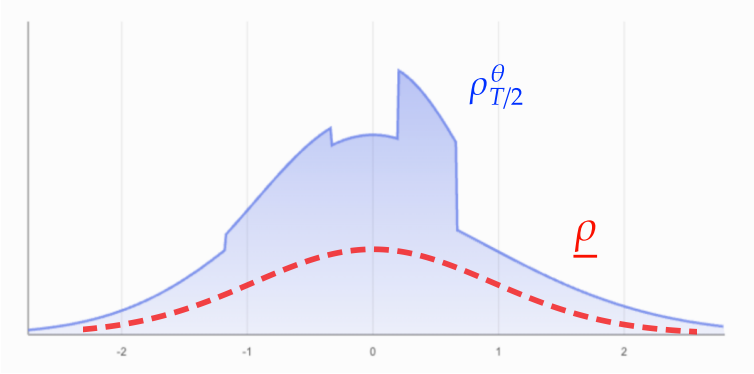
$$\rho_1(x) < \rho_2(x) \qquad \text{for all } \|x\| > \overline{M}.$$

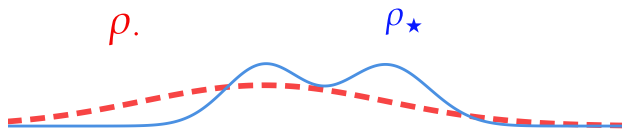
Step 2: Tail control

1. New base $\underline{\rho} = \alpha \exp\left(-\|x\|^2/2\underline{\sigma}^2\right)$ with $\underline{\sigma}^2 I_d \prec \Sigma_i \ \forall i$, and $\alpha > 0$ so that $\underline{\rho} < \rho^{\theta}_{T/2}$.
2. New target $\rho_\bullet := (2\pi\sigma_\bullet^2)^{-d/2} e^{-\|x\|^2/(2\sigma_\bullet^2)}$ (recall $\rho_\star(x) \le \rho_\bullet(x)$ for $\|x\| \ge M$).
3. New goal:

$$\text{Obtain} \begin{cases} \left\langle \left(\Sigma_i(T) - \sigma_\bullet^2 I_d\right) e_k, e_k \right\rangle > 0 & \text{on the } k-th \text{ face of } [-R_\varepsilon, R_\varepsilon]^d \\ \sigma_\bullet^2 I_d \prec \Sigma_i(T) & \text{on each corner.} \end{cases}$$

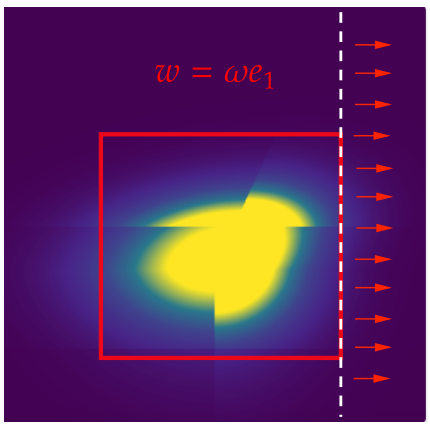
$\Big($ This induces full tail control since $\rho_0 < \tilde{\rho}_0 \implies \rho_t < \tilde{\rho}_t$ for all $t \in [0, T]\Big)$.

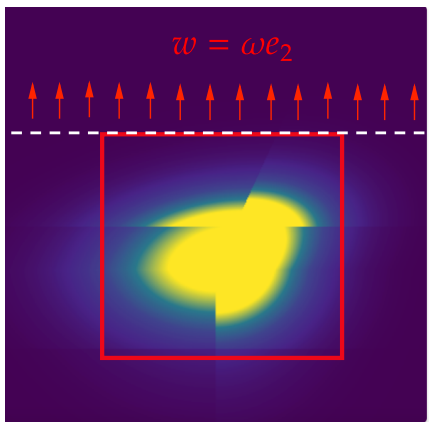




Step 2: Tail control

1. Choose (a, b) so that $H := \{a^\top x + b = 0\} \equiv e_k$-face of $[-R_\varepsilon, R_\varepsilon]^d$ and $[-R_\varepsilon, R_\varepsilon]^d \subset H^-$. Then the density will remain fixed inside $[-R_\varepsilon, R_\varepsilon]^d$.
2. Take $w = \omega e_k$ so in H^+. For $\omega > \frac{2d}{T} \log\left(\frac{\sigma_\bullet^2}{\sigma^2}\right)$, the goal is achieved at face k.
3. Repeat for each of the $2d$ faces of the cube. Note that in the corner regions d controls intervene so we obtain $\Sigma_\star \prec \Sigma_i(T)$.





Step 3: Combine both ingredients

- (Step I) We achieve
$$\|\rho_{T/2} - \rho_\star\|_{L^1} < \varepsilon_0.$$
- (Step II) We achieve:
  1. $\rho_T^\theta \geq \rho_\star$ on $\mathbb{R}^d \setminus [-S_\varepsilon, S_\varepsilon]$ for some $S_\varepsilon > R_\varepsilon$.
  2. $\rho_T = \rho_{T/2}$ on $[-R_\varepsilon, R_\varepsilon]^d$ containing $1-\varepsilon$ mass of $\rho_\star$ and ρ_T^θ.
- Positivity of ρ_{B} is preserved to ρ_T^θ by the continuity equation.
- Then $\sup_{\mathbb{R}^d} \frac{\rho_\star}{\rho_T^\theta} < +\infty$, and moreover

$$\left\|\rho_T^\theta - \rho_\star\right\|_{L^1} \leq \int_{[-S_\varepsilon, S_\varepsilon]^d} \left|\rho_{T/2} - \rho_\star\right| + \int_{\mathbb{R}^d \setminus [-S_\varepsilon, S_\varepsilon]^d} \rho_T^\theta + \int_{\mathbb{R}^d \setminus [-S_\varepsilon, S_\varepsilon]^d} \rho_\star < 3\varepsilon_0.$$

By Reverse Pinsker Inequality, choosing ε_0 small enough we conclude

$$\mathrm{KL}\left(\mu_\star \,\|\, \mu_T^\theta\right) < \varepsilon. \quad \square$$

Reverse KL (variational inference)

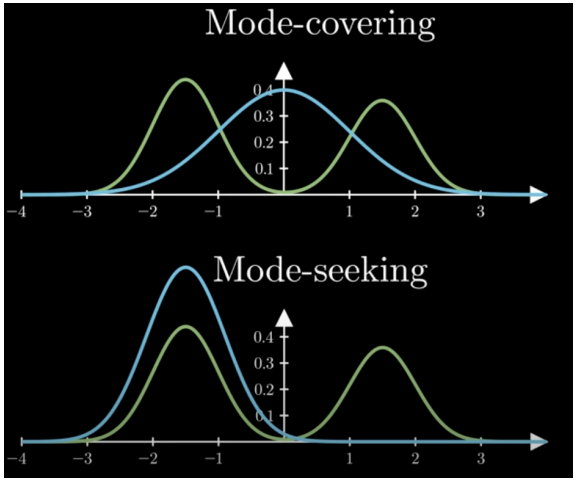
Theorem 2

Assume $\rho_\star > 0$, $\rho_\star \log \rho_\star \in L^1$ *and, for some* $M, \sigma_\bullet > 0$,

$$\rho_\star(x) \geq \rho_\bullet(x) = \frac{1}{(2\pi\sigma_\bullet^2)^{d/2}} e^{-\|x\|^2/(2\sigma_\bullet^2)} \quad (\|x\| \geq M).$$

Then $\forall\, T, \varepsilon > 0$ *there is a piecewise-constant* θ *with finitely many switches s.t.*

$$\mathrm{KL}\left(\mu_T^\theta \,\|\, \mu_\star\right) \leq \varepsilon.$$



Beyond Gaussians

- Write
$$\rho_{\mathrm{B}}(x) = e^{-U_{\mathrm{B}}(x)}, \quad \rho_\star(x) = e^{-U_\star(x)}.$$
- **Assumptions:**
  1. Linear growth:
  $$U_\star(x) \gtrsim \|x\| \qquad \text{for } \|x\| \gg 1.$$
  2. Convex envelope condition:
  Given $A \in M_{d\times d}(\mathbb{R})$ with $\mathrm{spec}(A) \subset (0, +\infty)$, there exists $\lambda_A > 0$ s.t.
  $$\frac{\mathrm{conv}\, U_\star(\lambda_A x)}{U_{\mathrm{B}}(Ax)} \to +\infty.$$
- **Idea.** The convex envelope condition guarantees that, after affine transport, the tails of $\rho_T^\theta > \rho_\star$ eventually dominate some λ_A rescale of $\rho_\star$'s tails.
- Then $\forall\, T, \varepsilon > 0$ there is piecewise-constant θ with finitely many switches s.t.
$$\mathrm{KL}\left(\mu_T^\theta \,\|\, \mu_\star\right) \leq \varepsilon.$$

Beyond KL

Csiszár *f*-divergences. For a convex $f : (0, \infty) \to \mathbb{R}$ with $f(1) = 0$,

$$\mathsf{D}_f(\mu_1, \mu_2) := \int f\left(\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}\right) \mathrm{d}\mu_2.$$

- Examples.
  - $f(t) = t \ln t \;\Rightarrow$ KL divergence.
  - $f(t) = |t - 1| \;\Rightarrow$ Total Variation.
  - $f(t) = (\sqrt{t} - 1)^2 \;\Rightarrow$ Squared Hellinger distance.
  - $f(t) = (t - 1)^2 \;\Rightarrow \chi^2$-divergence.
- Our main result extends as:

$$\sup_{t>0} \frac{f(t)}{t \ln t} < \infty \quad \Longrightarrow \quad \mathsf{D}_f(\mu_\star \,\|\, \mu_T^\theta) < \varepsilon.$$

Beyond KL

Rényi divergences. For $\lambda > 0$, $\lambda \neq 1$ ($\lambda \to 1$: recovers KL)

$$\mathsf{D}_\lambda(\mu_1, \mu_2) := \tfrac{1}{\lambda - 1} \log \int \left(\tfrac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}\right)^\lambda \mathrm{d}\mu_2.$$

- Our main result extends as:

$$0 \leq \lambda \leq 1 \quad \Longrightarrow \quad \mathsf{D}_\lambda(\mu_\star \,\|\, \mu_T^\theta) < \varepsilon.$$

Beyond Lipschitz

Motivation. Can we match densities with **different tail regimes**?

It is possible by breaking global Lipschitz constraint of the vector field!

Beyond Lipschitz

- Take the locally Lipschitz activation $\sigma(x) = x \log x \, 1_{\{x>1\}}$:

$$\partial_t \rho + \partial_x(-\sigma(x)\rho) = 0, \quad \rho_{\mathrm{B}}(x) \propto e^{-|x|^p}.$$

- Explicit solution:

$$\Phi_t^\theta(x) = \begin{cases} x, & x \leq 1, \\ x^{e^{-t}}, & x > 1 \end{cases} \qquad \rho_t^\theta(x) = \begin{cases} \rho_B(x), & x \leq 1, \\ e^t \, \rho_B\left(x^{e^t}\right) \, x^{e^t - 1}, & x > 1. \end{cases}$$

- **Result.** For each $q > 0$, there exists[2] $T = T(p, q) \geq 0$ such that

$$\lim_{x \to \infty} \frac{\rho_T(x)}{e^{-|x|^q}} = \lim_{x \to \infty} e^{|x|^q - |x|^{p e^t}} = 0.$$

- **Idea.** Non-Lipschitz $x \log x$ accelerates decay in the tails while preserving mass.

  Characteristics increase exponentially but don't blow up!

[2] Take any $T > \log(q/p)$ if $q > p$ and $T = 0$ if $q \leq p$.

Thank you for your attention!

The article.

Takeaways

1. **One-neuron ReLU perceptrons suffice** to train a CNF by minimizing KL (under tail-compatibility).
2. **Explicit construction** $\Rightarrow$ blueprint for architectures.
3. Extensions to variational inference and other metrics.
4. Locally-Lipschitz activation $\sigma(x) = x \log x \, 1_{\{x>1\}}$ enables matching between **different tail decays**.

KL Divergence

- **Interpretation**: Expected excess surprise when using ν as a model for the real distribution μ.
- (Basic) **properties**:
  - **Non-symmetric**: In general, $\mathrm{KL}(\mu||\nu) \neq \mathrm{KL}(\nu||\mu)$. This implies KL is not a distance.
  - **Non-negative**: $\mathrm{KL}(\mu||\nu) \geq 0$ for all μ, ν and $\mathrm{KL}(\mu||\nu) = 0 \iff \mu = \nu$.[3]
  - **Jointly convex**: For any $(\mu_1, \nu_1), (\mu_2, \nu_2) \in \mathcal{P}(\mathbb{R}^d)$, $\lambda \in [0, T]$

$$\mathrm{KL}(\lambda\mu_1 + (1-\lambda)\mu_2||\lambda\nu_1 + (1-\lambda)\nu_2) \leq \lambda\mathrm{KL}(\mu_1||\nu_1) + (1-\lambda)\mathrm{KL}(\mu_2||\nu_2).$$

[3] Proof:

$$\mathrm{KL}(\mu||\nu) = \mathbb{E}_\mu \underbrace{\left[-\log\left(\frac{d\nu}{d\mu}\right)\right]}_{\text{convex}} \overset{\text{(Jensen ineq.)}}{\geq} -\log\left(\mathbb{E}_\mu\left[\frac{d\nu}{d\mu}\right]\right) = \log 1 = 0$$

Basic properties of divergences

- **More (basic) properties:**
  - **Bounds.** For any convex $f : (0, \infty) \to \mathbb{R}$ with $f(1) = 0$,

$$D_f(\mu_1 \| \mu_2) \;:=\; \int f\left(\frac{d\mu_1}{d\mu_2}\right) d\mu_2.$$

    If there exists $C > 0$ such that

$$f(u) \;\le\; C\,(u \log u - u + 1) \quad \forall u > 0,$$

    then (whenever $\mu_1 \ll \mu_2$)

$$D_f(\mu_1 \| \mu_2) \;\le\; C\,\mathrm{KL}(\mu_1 \| \mu_2).$$

  - **Rényi divergence (order $\lambda \in (0, 1)$).**

$$D_\lambda(\mu_1 \| \mu_2) \;:=\; \frac{1}{\lambda - 1} \log \int \left(\frac{d\mu_1}{d\mu_2}\right)^\lambda d\mu_2 \;\le\; \mathrm{KL}(\mu_1 \| \mu_2),$$

    using the monotonicity of D_α in α and $D_1 = \mathrm{KL}$.
  - **Unbalanced mass.** If $\mu(\mathbb{R}^d) = \alpha \neq \beta = \nu(\mathbb{R}^d)$ and $\mu \ll \nu$, write $\bar{\mu} := \mu/\alpha$, $\bar{\nu} := \nu/\beta$ (probability measures). Then

$$\mathrm{KL}(\mu \| \nu) \;=\; \alpha \log\left(\frac{\alpha}{\beta}\right) + \alpha\,\mathrm{KL}(\bar{\mu} \,\|\, \bar{\nu}).$$

Proof sketch: 1D simultaneous control

Step A (approximate target): Riemann (or P_0 FEM) approximation

$$\rho_T^h = \sum_{k=1}^{K} \rho_{T,k}\, \mathbb{1}_{(x_{k-1},\, x_k-\delta)}, \qquad \|\rho_T - \rho_T^h\|_{L^1} < \varepsilon.$$

Step B (approximate source): Split ρ_{B} into K disjoint intervals with the *same* masses as the K target blocks, introducing tiny gaps of total mass $\tau \le \varepsilon$:

$$\rho_{\mathrm{B}}^h = \sum_{k=1}^{K-1} \mathbb{1}_{(y_{k-1},\, y_k-\tau/(K-1))} + \mathbb{1}_{(y_{K-1},y_K)}, \qquad \|\rho_{\mathrm{B}} - \rho_{\mathrm{B}}^h\|_{L^1} \le \varepsilon.$$

Step C (simultaneous control of blocks):

$$\partial_t \rho^{(k)} + \partial_x\big(w(t)\sigma(a(t)x + b(t))\, \rho^{(k)}\big) = 0,$$

$$\rho_0^{(k)} = \mathbb{1}_{I_k} \;\longrightarrow\; \rho_T^{(k)} = \rho_{T,k}\, \mathbb{1}_{J_k}, \qquad k = 1, \dots, K.$$

Mechanism: Induction over k using *(i)* local compression/dilation and *(ii)* local translation , with choices of (a, b) that make the field vanish on protected regions.

Conclusion: $\rho_{\mathrm{B}}^h \to \rho_T^h$ exactly; by L^1-contraction $\|\rho_\star - \rho_T\|_{L^1} \le 2\varepsilon$.

Proof sketch: parallel translation in $d \geq 2$

Half-space transport with ReLU: Choose $a \perp w$ so $\operatorname{div}(w\,\sigma(\langle a, x\rangle + b)) = 0$ and the field vanishes on one side of the hyperplane $\{\langle a, x\rangle + b = 0\}$.

Two-step composition: Pick b_1, b_2 and apply

$$\begin{cases} (a, w) = ((1, 0, \ldots, 0),\ -(0, 1, 0, \ldots, 0)),\ b = -b_1 \\ (a, w) = ((1, 0, \ldots, 0),\ +(0, 1, 0, \ldots, 0)),\ b = -b_2 \end{cases}$$

to obtain a net *parallel translation* of selected mass along e_2, leaving the complementary half-space invariant.

Use: Enables independent movement of separated parts of the support in $\mathbb{R}^d$.

Proof sketch: Dimension descent

(1) Truncate & grid: pick cube $\mathcal{R} = [-R, R]^d$, mesh size h, hyperplanes $H_{k\ell} : \; x^{(k)} = c_{k,\ell}$.

$$\rho_{0,T} \approx \sum_j m_{j,0/T} \, \mathbb{1}_{\square_{j+h(1,\ldots,1)/2,\; h-\delta}},$$

removing strips Ω_δ around hyperplanes.
(2) Align along $x^{(d)}$: By parallel translations slab-by-slab so projected boxes in $(x^{(1)}, \ldots, x^{(d-1)})$ are *disjoint* and independent.
(3) Recurse: the new densities are constant in $x^{(d)}$ on their support $\Rightarrow$ reduce to a $(d-1)$-D problem. Iterate down to 1-D.
(4) Stability: L^1-contraction propagates the ε-approximations from truncation and meshing.