Constructive approximate transport maps with normalizing flows

Antonio Álvarez-López, Borjan Geshkovski, Domènec Ruiz-Balet

Constructive approximate transport maps with normalizing flows

Antonio Álvarez-López
Universidad Autónoma de Madrid

Borjan Geshkovski
Inria & Sorbonne Université

Domènec Ruiz-Balet
Université Paris Dauphine

January 23, 2025

Abstract

We study an approximate controllability problem for the continuity equation and its application to constructing transport maps with normalizing flows. Specifically, we construct time-dependent controls $\theta = (w, a, b)$ in the vector field $x \mapsto w(a^\top x + b)_+$ to approximately transport a known base density ρ_{B} to a target density ρ_*. The approximation error is measured in relative entropy, and θ are constructed piecewise constant, with bounds on the number of switches being provided. Our main result relies on an assumption on the relative tail decay of ρ_* and ρ_{B}, and provides hints on characterizing the reachable space of the continuity equation in relative entropy.

Keywords. Normalizing flows, approximate controllability, optimal transport, Pinsker inequality.

AMS classification. 35Q49, 93C15, 68T07, 93C20, 46N10, 49Q22.

Contents

1 Introduction

By viewing discrete layers as a continuous time variable, the framework of *neural ODEs*—proposed by [E17] and made feasible in computing practice by [HZRS16, CRBD18]—has significantly influenced the study and application of neural networks. They are transparent objects: each (of many) data-point $x \in \mathbb{R}^d$ is propagated through the flow of the Cauchy problem

$$\begin{cases} \dot{x}(t) = v(x(t), \theta(t)) & \text{for } t \in [0, T] \\ x(0) = x; \end{cases} \tag{1.1}$$

the velocity field is, canonically, a perceptron

$$v(x, \theta) \coloneqq w \left(a^\top x + b \right)_+, \tag{1.2}$$

with $\theta = (w, a, b) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R} \cong \mathbb{R}^{2d+1}$ being the parameters of the network. (Natural generalizations of (1.2) can be devised, by replacing the inner product by a matrix multiplication or discrete convolution for example.) We will consistently assume that $\theta(\cdot) \in L^\infty(0, T; \mathbb{R}^{2d+1})$, ensuring the applicability of the Cauchy-Lipschitz theorem. A neural network can thus be interpreted as the flow map $\Phi_\theta^T : \mathbb{R}^d \to \mathbb{R}^d$ of (1.1) at time $t = T$, yielding a representation $x(T)$ of each data point x. This representation is then fed into a last layer that corresponds to a classical machine learning task such as linear or logistic regression, in order to predict the corresponding label.

Besides serving as a useful abstraction of discrete neural networks, this formalism has had tremendous impact on density estimation through the guise of *normalizing flows* [GCB+18]. This is a popular methodology in machine learning as evidenced by [PNR+21, KPB20, LCBH+22, ABVE23, GRVE22, PBHDE+23] and the references therein, and goes back at least as [TVE10, TT13, DKB14, TW16]. To estimate an unknown probability density ρ_*, of which samples are known, one takes a simple initial probability density ρ_{B}, and matches the solution of the continuity (or conservative transport) equation

$$\begin{cases} \partial_t \rho(t, x) + \operatorname{div}\Big(v(x, \theta(t)) \rho(t, x) \Big) = 0 & \text{in } [0, T] \times \mathbb{R}^d \\ \rho(0, x) = \rho_{\mathrm{B}}(x) & \text{on } \mathbb{R}^d \end{cases} \tag{1.3}$$

to the known samples. Specifically, given n iid samples $x_1, \dots, x_n$ of the unknown density ρ_*, one solves

$$\max_\theta \frac{1}{n} \sum_{i=1}^n \log \rho_{\mathrm{B}}(\Phi_\theta^{-T}(x_i)) + \log |\det \nabla \Phi_\theta^{-T}(x_i)|. \tag{1.4}$$

(We shall keep this presentation formal, and do not detail what constraints are imposed on θ to ensure existence of maximizers.) This is the problem of *maximum (log-)likelihood estimation* on the data. Here Φ_θ^{-T} designates the end-time flow map of (1.1) ran backwards in time, from T to 0; the eponym "normalizing" then stems from the fact that ρ_{B} is most often a Gaussian probability density in practice.

If ρ_* were to be known, then (1.4) is the empirical version of

$$\max_\theta \int \rho_*(x) \left(\log \rho_{\mathrm{B}}(\Phi_\theta^{-T}(x)) + \log |\det \nabla \Phi_\theta^{-T}(x)| \right) \mathrm{d}x. \tag{1.5}$$

By the change of variable theorem, we have $\rho_{\mathrm{B}}(\Phi_\theta^{-T}(x))|\det \nabla \Phi_\theta^{-T}(x)| = \rho(T, x)$, and thus

$$\begin{aligned} (1.5) &= \max_\theta \int \rho_*(x) \log \rho(T, x) \, \mathrm{d}x \\ &= \min_\theta \int \rho_*(x) \log \frac{\rho_*(x)}{\rho(T, x)} \, \mathrm{d}x. \end{aligned} \tag{1.6}$$

This is the question of minimizing the *relative entropy* between ρ_* and $\rho(T)$, or equivalently, the *Kullback-Leibler divergence* between the measures μ_* and $\Phi_{\theta\#}^T \mu_{\mathrm{B}}$, henceforth denoted as

$$\mathsf{KL}\left(\mu_* \,\|\, \Phi_{\theta\#}^T \mu_{\mathrm{B}}\right) := \int \rho_*(x) \log \frac{\rho_*(x)}{\rho(T, x)} \, \mathrm{d}x,$$

where $\rho_* = \frac{\mathrm{d}\mu_*}{\mathrm{d}x}$. It is a problem of general interest in Bayesian inference and sampling—all we do here is to parametrize it by densities given as solutions to (1.3).

Our focus in this paper is the feasibility problem associated to (1.6)—namely, given an arbitrarily small $\varepsilon > 0$, we wish to find θ (depending on ε) such that

$$\int \rho_*(x) \log \frac{\rho_*(x)}{\rho(T, x)} \, \mathrm{d}x \leqslant \varepsilon.$$

This constitutes an *approximate controllability* or *approximate transportation* problem, which is *nonlinear* in the controls θ. Results for (1.3) are only known in weaker[1] topologies—such as L^1, L^2 or Wasserstein distances [RBZ23, RBZ24, MRWZ24, ÂLHSZ24, GRRB24]—and, again in weaker topologies, also for continuity equations with vector fields that are linear in the controls θ [ABBG$^+$12, Rag24,

[1] see (1.8) and Remark 1.7.

DMR19, SF23]. In addition to addressing the strong topology imposed by the relative entropy, we will require θ to be piecewise constant, with the aim of counting the number of discontinuities (switches). This is of interest, as these switches can be interpreted as analogous to discrete layers. In the setting of such θ, given an initial density $\rho_{\mathrm{B}} \in \mathscr{C}^0(\mathbb{R}^d)$, (1.3) has a unique solution $\rho \in \mathscr{C}^0([0,T]; L^1 \cap L^\infty(\mathbb{R}^d))$ which is explicit and found by the method of characteristics (Lemma 2.3).

1.1 Main result

Our main result is as follows.

Theorem 1.1. *Suppose $\mu_{\mathrm{B}} = \mathcal{N}(m_{\mathrm{B}}, \Sigma_{\mathrm{B}})$ for an arbitrary mean $m_{\mathrm{B}} \in \mathbb{R}^d$ and covariance matrix $\Sigma_{\mathrm{B}} \in \mathcal{S}_d^+(\mathbb{R})$. Let ρ_* be any probability density on $\mathbb{R}^d$ such that $\rho_* \log \rho_* \in L^1(\mathbb{R}^d)$ and for some $M > 0$ and $\sigma_\bullet > 0$,*

$$\rho_*(x) \leqslant \frac{1}{(2\pi\sigma_\bullet^2)^{\frac{d}{2}}} e^{-\frac{\|x\|^2}{2\sigma_\bullet^2}} =: \rho_\bullet(x) \qquad \text{for all } \|x\| \geqslant M. \tag{1.7}$$

Then for any $T > 0$ and $\varepsilon > 0$, there exists $\theta = (w, a, b) : [0,T] \to \mathbb{R}^{2d+1}$, piecewise constant with finitely many switches, such that the corresponding solution ρ to (1.3) with data $\rho_{\mathrm{B}} = \frac{\mathrm{d}\mu_{\mathrm{B}}}{\mathrm{d}x}$ satisfies

$$\int \rho_*(x) \log \frac{\rho_*(x)}{\rho(T,x)} \,\mathrm{d}x \leqslant \varepsilon.$$

Equivalently,

$$\mathsf{KL}\left(\mu_* \,\|\, \Phi_{\theta\#}^T \mu_{\mathrm{B}}\right) \leqslant \varepsilon,$$

where $\rho_* = \frac{\mathrm{d}\mu_*}{\mathrm{d}x}$ and $\Phi_\theta^T : \mathbb{R}^d \to \mathbb{R}^d$ is the flow map of (1.1) at time $t = T$.

We defer a detailed discussion on the setup and extensions of the above result to **Section 1.2**, and instead briefly motivate the proof. Our starting point is [RBZ24], in which L^1-approximate controllability is shown:

$$\|\rho(T) - \rho_*\|_{L^1(\mathbb{R}^d)} = \left|\Phi_{\theta\#}^T \mu_{\mathrm{B}} - \mu_*\right|_{\mathsf{TV}} \leqslant \varepsilon.$$

The proof is entirely constructive and hinges on performing exact transportation between piecewise constant approximations of the initial and target densities, along with a continuity argument. This is achieved by moving and deforming each segment of the approximated initial density to match the approximated target density using a piecewise constant control, θ. Consequently, the number of switches in θ corresponds to the number of segments in the approximation. In higher dimensions, the proof extends by employing piecewise constant approximations over grids, which result in an exponential number of cells (and thus segments) in the dimension d. The argument proceeds via induction on the dimension. Since the number of cells grows exponentially with d, the number of switches in θ similarly scales exponentially with d.

One way to make use of [RBZ24] is to employ a functional inequality linking KL and TV. A well-known example is the (Csizár-Kullback-)*Pinsker* inequality

$$|\mu_2 - \mu_1|_{\mathsf{TV}} \leqslant \sqrt{2 \cdot \mathsf{KL}(\mu_2 \,\|\, \mu_1)} \tag{1.8}$$

which holds for arbitrary probability measures μ_1, μ_2. While the inequality sign in (1.8) is not in the desired direction so that we can straightforwardly apply it to conclude, it turns out that it can be reversed under stronger conditions on the tails of the measures (see [Ver14, Sas15, SV16]):

Lemma 1.2 (Reverse Pinsker inequality)**.** *Let μ_1, μ_2 be probability measures on $\mathbb{R}^d$ such that $\mu_2 \ll \mu_1$. Then*

$$\mathsf{KL}(\mu_2 \,\|\, \mu_1) \leqslant \frac{1}{2} \cdot \frac{\log \left\| \frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_1} \right\|_{L^\infty(\mathbb{R}^d)}}{1 - \frac{1}{\left\| \frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_1} \right\|_{L^\infty(\mathbb{R}^d)}}} \cdot |\mu_2 - \mu_1|_{\mathsf{TV}}. \tag{1.9}$$

Recall that the Radon-Nikodym theorem only ensures that $\frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_1}$ is measurable. Yet for (1.9) to be non-void we need

$$\left\| \frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_1} \right\|_{L^\infty(\mathbb{R}^d)} < +\infty.$$

(Note that this quantity is always $\geqslant 1$ by definition.) Suppose instead that μ_1 and μ_2 are absolutely continuous with respect to the Lebesgue measure. We are asking for

$$\left\| \frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_1} \right\|_{L^\infty(\mathbb{R}^d)} = \operatorname*{esssup}_{x \in \mathbb{R}^d} \frac{\rho_2(x)}{\rho_1(x)} < +\infty.$$

Therefore, to convert the TV approximation result of [RBZ24] to a KL one, we also ought to ensure that

$$\mathsf{L}(\rho_*, \rho(T)) \coloneqq \lim_{\|x\| \to \infty} \frac{\rho_*(x)}{\rho(T, x)} < +\infty, \tag{1.10}$$

as well as $\rho(T, x) > 0$ for all $x \in \mathbb{R}^d$. Ensuring this condition is in essence the main novelty of our study. The complete proof can be found in **Section 3.1**.

At this point, we can also observe that KL is not symmetric with respect to its inputs, whence Theorem 1.1 does not directly entail a result when the densities are swapped. Interestingly, minimizing the reverse-KL is a problem of interest in its own right and is referred to as *variational inference* (see [WJ08, BKM17, RM15, JCP23] and the references therein for works on the topic). The following result turns out to be a simple modification of the proof of Theorem 1.1.

Theorem 1.3. *Suppose $\mu_{\mathrm{B}} = \mathcal{N}(m_{\mathrm{B}}, \Sigma_{\mathrm{B}})$ for an arbitrary mean $m_{\mathrm{B}} \in \mathbb{R}^d$ and covariance matrix $\Sigma_{\mathrm{B}} \in \mathcal{S}_d^+(\mathbb{R})$. Let ρ_* be any probability density on $\mathbb{R}^d$ such that $\rho_* \log \rho_* \in L^1(\mathbb{R}^d)$ and for some $M > 0$ and $\sigma_\bullet > 0$,*

$$\rho_*(x) \geqslant \frac{1}{(2\pi\sigma_\bullet^2)^{\frac{d}{2}}} e^{-\frac{\|x\|^2}{2\sigma_\bullet^2}} =: \rho_\bullet(x) \qquad \text{for all } \|x\| \geqslant M. \tag{1.11}$$

Then for any $T > 0$ and $\varepsilon > 0$, there exists $\theta = (w, a, b) : [0, T] \to \mathbb{R}^{2d+1}$, piecewise constant with finitely many switches, such that the corresponding solution ρ to (1.3) satisfies

$$\int \rho(T, x) \log \frac{\rho(T, x)}{\rho_*(x)} \, \mathrm{d}x \leqslant \varepsilon.$$

Equivalently

$$\mathsf{KL}\left(\Phi_{\theta\#}^T \mu_{\mathrm{B}} \,\|\, \mu_*\right) \leqslant \varepsilon,$$

where $\rho_* = \frac{\mathrm{d}\mu_*}{\mathrm{d}x}$ and $\Phi_\theta^T : \mathbb{R}^d \to \mathbb{R}^d$ is the flow map of (1.1) at time $t = T$.

We provide the proof in **Section 3.2**.

1.2 Discussion

We turn to commenting the various assumptions made in Theorems 1.1 and 1.3. In **Section 1.2.1**, we discuss estimates on the number of switches (Remark 1.4), the assumption of a Gaussian initial density (Remark 1.5) and the possible removal of (1.7) (Proposition 1.6), considering different metrics (Remark 1.7), and estimation from finitely many samples (Remark 1.8). In **Section 1.2.2**, we discuss changing the model (1.2) to the one considered in [SF23], which is *linear* in θ.

1.2.1 Generalities

Remark 1.4 (Number of switches)**.** *The controls θ from Theorem 1.1 have at most*

$$\left\lceil \frac{2R(\varepsilon)}{h(\varepsilon)} \right\rceil^d (d + 10) + 2d$$

switches, where $R(\varepsilon), h(\varepsilon) > 0$ are such that

$$\int_{[-R(\varepsilon), R(\varepsilon)]^d} \rho_{\mathrm{B}}(x) \, \mathrm{d}x \wedge \int_{[-R(\varepsilon), R(\varepsilon)]^d} \rho_*(x) \, \mathrm{d}x > 1 - \varepsilon$$

and

$$\int_{[-R(\varepsilon), R(\varepsilon)]^d} \left|\rho_{\mathrm{B}}^{h_\varepsilon}(x) - \rho_{\mathrm{B}}(x)\right| \, \mathrm{d}x \vee \int_{[-R(\varepsilon), R(\varepsilon)]^d} \left|\rho_*^{h_\varepsilon}(x) - \rho_*(x)\right| \, \mathrm{d}x < \varepsilon,$$

where $\rho_{\mathrm{B}}^{h_\varepsilon}$ and $\rho_*^{h_\varepsilon}$ are piecewise constant interpolants of ρ_{B} and ρ_* on a regular grid of the hypercube $[-R(\varepsilon), R(\varepsilon)]^d$ with cells of spacing $h(\varepsilon)$. The exponential growth

with d of the number of switches is due to [RBZ24, Theorem 1], where the error scales inversely with the cell-sizeof a multidimensional grid.

In Theorem 1.3, θ has at most

$$\left\lceil \frac{2R(\varepsilon)}{h(\varepsilon)} \right\rceil^d (d+10) + 2d$$

switches, where $R(\varepsilon) > 0$ and $h(\varepsilon) > 0$ are as in Theorem 1.1.

Remark 1.5 (Beyond Gaussians)**.** *Theorems 1.1 and 1.3 can be generalized beyond the case of Gaussian initial densities. For Theorem 1.1, it suffices to replace (1.7) by the following three conditions:*

- *The densities ρ_{B} and ρ_* have full support[2]. Thus, $\rho_{\mathrm{B}}(x) = e^{-U_{\mathrm{B}}(x)}$ as well as $\rho_*(x) = e^{-U_*(x)}$.*
- *There exists $M_* > 0$ such that*

$$U_*(x) \gtrsim \|x\| \qquad \text{for all } \|x\| > M_*. \tag{1.12}$$

- *Given $A \in \mathscr{M}_{d\times d}(\mathbb{R})$ with $\operatorname{spec}(A) \subset (0, +\infty)$, there exists $\lambda_A > 0$ such that the lower convex envelope, defined for all $x \in \mathbb{R}^d$ by*

$$\operatorname{conv} U_*(x) = \sup_{\substack{U \leqslant U_* \\ U \text{convex}}} U(x)$$

satisfies

$$\lim_{\|x\|\to\infty} \frac{\operatorname{conv} U_*(\lambda_A x)}{U_{\mathrm{B}}(Ax)} = +\infty. \tag{1.13}$$

Similarly, to extend Theorem 1.3 we replace (1.11) with these conditions and instead consider $U_\bullet$ as the upper convex envelope. Furthermore, we invert the quotient in (1.13). We provide the arguments of the proofs in **Section 3.3**.

To build on Remark 1.5, one can inquire whether (1.2)–(1.3) can be used to couple two densities with tails of different nature. To this end, we suspect that the globally Lipschitz character of (1.2) needs to be violated, in which case we could obtain an affirmative answer as illustrated by the following partial result.

Proposition 1.6. *Fix $p > 0$. Consider*

$$\begin{cases} \partial_t \rho(t,x) + \partial_x((|x| \log x_+)_+ \rho(t,x)) = 0 & \text{in } \mathbb{R}_{>0} \times \mathbb{R} \\ \rho(0,x) = \rho_{\mathrm{B},p}(x) & \text{in } \mathbb{R} \end{cases} \tag{1.14}$$

where $\rho_{\mathrm{B},p} \in \mathscr{C}^0(\mathbb{R}^d)$ with $\rho_{\mathrm{B},p} \propto \exp(-|x|^p)$. Then, for every $q \in (0,p]$ there exist a time $T_q > 0$ such that the unique weak solution ρ to (1.14) satisfies

$$\lim_{x\to+\infty} \frac{\rho(T_q,x)}{\rho_{\mathrm{B},q}(x)} < +\infty.$$

[2] If $t \mapsto \rho_*(ut)$ is compactly supported for some $u \in \mathbb{S}^{d-1}$, then $\lim_{|t|\to\infty} \rho_*(tu)/\rho(T,tu) = 0$ trivially. This means that we would only have to care about the directions in which the support is unbounded.

We provide the proof in **Section 3.4**. Proposition 1.6 asserts that one can avoid assuming (1.7) in Theorem 1.1 by sacrificing the globally Lipschitz character of the nonlinearity and, in turn, use (1.3) to transport a Gaussian to a Laplace density, for example.

Remark 1.7 (Metric). *We discuss extensions of Theorem 1.1 to other divergences which are common in statistics and information theory.*

- *Consider the squared* Hellinger distance*:*

$$\mathsf{H}^2(\mu_1,\mu_2) \coloneqq \int \left(\sqrt{\mu_1(\,\mathrm{d}x)} - \sqrt{\mu_2(\,\mathrm{d}x)}\right)^2.$$

  One has

$$\mathsf{H}^2(\mu_1,\mu_2) \leqslant |\mu_1-\mu_2|_{\mathsf{TV}(\mathbb{R}^d)} \leqslant \sqrt{2}\cdot \mathsf{H}(\mu_1,\mu_2).$$

  Thus a direct application of [RBZ24, Theorem 1] gives

$$\mathsf{H}^2\left(\Phi_{\theta\#}^T\mu_{\mathrm{B}},\mu_*\right) \leqslant \varepsilon.$$

- *More generally one could consider any* f-divergence *between* $\mu_1 \ll \mu_2$*. For a convex function* $f:(0,+\infty)\to\mathbb{R}$ *with* $f(1)=0$ *and* $\lim_{t\to 0^+} f(t) \in [0,+\infty]$*, the* f*-divergence [PW24] is defined as*

$$\mathsf{D}_f(\mu_1,\mu_2) = \int f\left(\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}\right)\mathrm{d}\mu_2 \tag{1.15}$$

  For instance, $f(t)=t\ln t$ *corresponds to* KL*,* $f(t)=|t-1|$ *corresponds to* TV*,* $f(t)=(\sqrt{t}-1)^2$ *corresponds to* H^2*, and* $f(t)=|t-1|^2$ *corresponds to* χ^2*. From* (1.15) *one sees that* $f_1 \lesssim f_2$ *is equivalent to* $\mathsf{D}_{f_1}(\mu_1,\mu_2) \lesssim \mathsf{D}_{f_2}(\mu_1,\mu_2)$ *for all* μ_1,μ_2 *with* $\mu_1 \ll \mu_2$*. Thus by virtue of Theorem 1.1 we deduce that for any* $\varepsilon>0$ *and* f *as above,*

$$\textit{if}\quad \sup_{t>0}\frac{f(t)}{t\ln t} < +\infty \qquad \textit{then}\quad \mathsf{D}_f\left(\mu_* \,\|\, \Phi_{\theta\#}^T\mu_{\mathrm{B}}\right) \leqslant \varepsilon.$$

  The condition is satisfied by $f(t)=|t-1|$ *and* $f(t)=(\sqrt{t}-1)^2$*.*

- *One could also consider the* Rényi entropy *[PW24]: for any* $\lambda>0$ *with* $\lambda\neq 1$*, the Rényi entropy of order* λ *between* $\mu_1 \ll \mu_2$*, is defined as*

$$\mathsf{D}_\lambda(\mu_1,\mu_2) \coloneqq \frac{1}{\lambda-1}\log\int_{\mathbb{R}^d}\left(\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}\right)^\lambda \mathrm{d}\mu_2.$$

  By continuity $\mathsf{D}_1 = \mathsf{KL}$*, and similarly* $\mathsf{D}_\infty(\mu_1,\mu_2) = \log \operatorname{esssup}\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}$*. The latter is known as* worst-case regret*. If* μ_1 *and* μ_2 *have continuous densities denoted* ρ_1 *and* ρ_2 *respectively, then* $\mathsf{D}_\infty(\mu_1,\mu_2)\leqslant\varepsilon$ *implies*

$$\|\rho_1-\rho_2\|_{L^\infty(\mathbb{R}^d)} \lesssim \varepsilon. \tag{1.16}$$

Our methods do not extend to the case of worst-case regret since we rely on the $L^1(\mathbb{R}^d)$–approximation of densities of [RBZ24], which doesn't imply (1.16). On the other hand, for any fixed $\mu_1 \ll \mu_2$, the map $\lambda \in [0, +\infty] \mapsto \mathsf{D}_\lambda(\mu_1, \mu_2)$ is non-decreasing [VEH14]. Theorem 1.1 thus implies that

$$\mathsf{D}_\lambda \left(\mu_* \,\|\, \Phi^T_{\theta\#} \mu_{\mathrm{B}} \right) \leqslant \varepsilon$$

if $0 \leqslant \lambda \leqslant 1$.

Remark 1.8 (Statistics). *In practice one only has access to n samples of the target density ρ_*. Much like [RBZ24], we can use Theorem 1.1 to obtain sample complexity bounds regarding the explicit construction of θ. To this end, one first employs a non-parametric estimator $\widehat{\rho}_{*,n}$ of ρ_* using the n samples (e.g., via a kernel density estimator), and then uses Theorem 1.1 to approximate $\widehat{\rho}_{*,n}$ to any desired accuracy ε. In doing so, sample complexity-wise, the specific construction of Theorem 1.1 can do as well as any non-parametric estimator of ρ_*, but at the additional cost that is the number of switches, which may be exponential in the dimension d.*

1.2.2 Linear vector field

The related work [SF23] considers the simpler vector field which is linear in θ:

$$v(x, \theta) = w\sigma(x) + b, \qquad \theta = (w, b) \in \mathscr{M}_{d\times d}(\mathbb{R}) \times \mathbb{R}^d, \tag{1.17}$$

with $\sigma : \mathbb{R} \to \mathbb{R}$ assumed Lipschitz and smooth, and defined component-wise.

When $\sigma(x) = (x)_+$, (1.17) cannot yield (1.10) in general, because $\sigma|_{\mathbb{R}_{\leqslant 0}} = 0$. Indeed for $R > 0$ large enough (depending on b) and $x \in (\mathbb{R}_{<0})^d \cap \{x \colon \|x\| \geqslant R\}$, the solution to (1.3) reads

$$\rho(t, x) = \rho_{\mathrm{B}} \left(x - \int_0^t b(s)\, \mathrm{d}s \right), \qquad \text{for } x \in (\mathbb{R}_{<0})^d \cap \{x \colon \|x\| \geqslant R\}.$$

For x away from the origin, $\rho(t, x)$ is simply a translation of the initial data ρ_B. Thus, the tails will not be generally be of the same order. Much like (1.2) however (see [LLS22, RBZ23, CLLS23]), one can use (1.1)–(1.17) to transport an empirical measure $\mu_0 = \frac{1}{n} \sum_{i\in[n]} \delta_{x^{(i)}}$ to a target $\mu_1 = \frac{1}{n} \sum_{i\in[n]} \delta_{y^{(i)}}$. In the case where σ is smooth, this can be done by using techniques from geometric control theory based on computing iterated Lie brackets, in the spirit of the celebrated Chow-Rashevskii theorem [AS22, CLT20, Sca23, TG22, EGBO22]. The result is not known for (1.17) with $\sigma(x) = (x)_+$.

We consider

$$\left(x^{(i)}, y^{(i)}\right) \in \mathbb{R}^d \times \mathbb{R}^d, \qquad \text{for } i \in [n],$$

and, to avoid unnecessary technical details, work under the benign assumption that $x^{(i)} \neq x^{(j)}$ and $y^{(i)} \neq y^{(j)}$ for $i \neq j$. The following holds.

Proposition 1.9. *Let $T > 0$ and suppose $\sigma(x) \equiv (x)_+$ in (1.17). Then there exists $\theta = (w, b) : [0, T] \to \mathscr{M}_{d\times d}(\mathbb{R}) \times \mathbb{R}^d$, piecewise constant with at most $4n + 2$ switches, such that the flow map $\Phi_\theta^t : \mathbb{R}^d \to \mathbb{R}^d$ ($t \in [0, T]$) of (1.17) satisfies*

$$\Phi_\theta^T\left(x^{(i)}\right) = y^{(i)}, \qquad \text{for all } i \in [n].$$

The proof may be found in **Section 3.5** and relies on several modifications of that presented in [RBZ23]. One may then ask at what level (1.1) differs from (1.17). As it turns out, one important deviation is in the stability of θ with respect to perturbations in the data (estimate (1.19)). This can be derived easily in the context of (1.1) as done in [BP24]. In the context of (1.17) however, the most we are able to say is the following.

Proposition 1.10. *Suppose $T > 0$, $\sigma \in \mathscr{C}^0(\mathbb{R})$ and $d \geqslant n$. Let $y^{(1)}, \dots, y^{(n)} \in \mathbb{R}^d$ be such that*

$$\operatorname{span}\left\{\sigma(y^{(1)}), \dots, \sigma(y^{(n)})\right\} = \mathbb{R}^n. \tag{1.18}$$

Then there exist $\varepsilon > 0$ and $C > 0$ such that for all $x^{(1)}, \dots, x^{(n)} \in \mathbb{R}^d$ with

$$\max_{i\in[n]} \|x^{(i)} - y^{(i)}\|_1 \leqslant \varepsilon,$$

there exist $w \in \mathscr{C}^0([0, T]; \mathscr{M}_{d\times d}(\mathbb{R}))$ and $(x_i(\cdot))_{i\in[n]} \in \mathscr{C}^1([0, T]; \mathbb{R}^{nd})$ satisfying

$$\begin{cases} \dot{x}_i(t) = w(t)\sigma(x_i(t)) & \text{for } t \in [0, T] \\ x_i(0) = x^{(i)} \\ x_i(T) = y^{(i)} \end{cases}$$

for all $i \in [n]$. Moreover,

$$\|w\|_{\mathscr{C}^0([0,T];\mathscr{M}_{d\times d}(\mathbb{R}))} \leqslant \frac{C}{T} \max_{i\in[n]} \|x^{(i)} - y^{(i)}\|_1. \tag{1.19}$$

Remark 1.11 (Genericity)**.** *Suppose $\sigma : \mathbb{R} \to \mathbb{R}$ is Lipschitz and strictly increasing, and $d \geqslant n$. Then (1.18) holds with probability 1 for n random points $y^{(i)}$ sampled iid from any probability density $\rho = \frac{\mathrm{d}\mu}{\mathrm{d}x}$ on $\mathbb{R}^d$. (For instance, this holds if $\sigma(\cdot) \equiv (\cdot)_+$ and ρ is supported on $(\mathbb{R}_{>0})^d$.)*

This is seen by induction on the number of points n. The base case $n = 1$ trivially holds since for the element-wise extension $\sigma : \mathbb{R}^d \to \mathbb{R}^d$ of σ, $\sigma_\#\mu$ preserves the zero-measure sets of μ. Suppose that $\sigma(y^{(1)}), \dots, \sigma(y^{(n)})$ are almost surely linearly independent for some $1 \leqslant n < d$. By absolute continuity of $\sigma_\#\mu$, the subspace spanned by these points has measure zero, so the next vector $\sigma(y^{(n+1)})$ lies in the complement with probability 1, as desired.

Remark 1.12. • *Extending Proposition 1.10 to account for a bias $b(t)$ as in (1.17) is straightforward: we simply write $\dot{z}(t) = \overline{w}(t)\sigma(z(t))$ where*

$$z \coloneqq \begin{bmatrix} x \\ 1 \end{bmatrix} \in \mathbb{R}^{d+1} \quad \text{and} \quad \overline{w} \coloneqq \begin{bmatrix} w & b \\ 0 & 0 \end{bmatrix} \in \mathscr{M}_{(d+1)\times(d+1)}(\mathbb{R}).$$

We now rather need $d \geqslant n - 1$.

- *With regard to Remark 1.11, when $\sigma(\cdot) \equiv (\cdot)_+$ and ρ is supported on $\mathbb{R}^d$, one could simply translate the inputs to $(\mathbb{R}_{>0})^d$ using $b(t)$, but then a bound on $b(t)$ as in (1.19) cannot hold.*

1.3 Notation

We use $[n] := \{1, \dots, n\}$, $x \wedge y = \min\{x, y\}$, $x \vee y = \max\{x, y\}$. We use $f(x) \lesssim g(x)$ if there exists a finite constant $C > 0$ such that $f(x) \leqslant Cg(x)$, and $f(x) \propto g(x)$ if $f(x) = Cg(x)$.

We denote by $\mathcal{S}_d^+(\mathbb{R})$ the space of positive definite matrices. On $\mathcal{M}_{d\times d}(\mathbb{R})$, we consider the partial order $P \prec Q$ iff $Q - P$ is positive definite. We also alternate between $a^\top b$ and $\langle a, b\rangle$ depending on presentational convenience.

Given two measures μ, ν on $\mathbb{R}^d$, we write $\mu \ll \nu$ to say that μ is absolutely continuous with respect to ν. If $\mathsf{T} : \mathbb{R}^d \to \mathbb{R}^d$ is measurable, then $\mathsf{T}_{\#}\mu$ denotes the pushforward of μ, defined by $\mathsf{T}_{\#}\mu(A) = \mu(\mathsf{T}^{-1}(A))$ for any measurable A. The Gaussian measure with mean $m \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathcal{S}_d^+(\mathbb{R})$ is denoted by $\mathcal{N}(m, \Sigma)$. Finally we write $\mathfrak{L}_\mu = \mu(\mathbb{R}^d)$.

Funding

This work was initiated during AÁ and DRB's visit to the Department of Mathematics at MIT in the summer of 2023. AÁ has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 777822. DRB acknowledges "France 2030" support managed by the Agence Nationale de la Recherche, under the reference ANR-23-PEIA-0004.

2 Preliminary lemmas

We begin with a brief discussion on when $\mathsf{L}(\rho_*, \rho(T)) = \lim_{\|x\|\to\infty} \rho_*(x)/\rho(T, x)$ can be made finite when $d = 1$, purely for illustrative purposes. Recall

Lemma 2.1. *Take $T > 0$, a probability density $\rho_{\mathrm{B}} \in \mathscr{C}^0(\mathbb{R})$, and*

$$(w, a, b)(t) = \sum_{k=1}^{K} (w_k, a_k, b_k) 1_{[t_{k-1}, t_k)}(t) \qquad \text{for } t \in [0, T],$$

where $w_k, a_k, b_k \in \mathbb{R}$ with $a_k \neq 0$ and $0 = t_0 < t_1 < \ldots < t_K = T$. Then the unique solution to

$$\begin{cases} \partial_t \rho(t, x) + \partial_x \left(w(t)(a(t)x + b(t))_+ \rho(t, x)\right) = 0 & \text{on } [0, T] \times \mathbb{R}, \\ \rho(0, x) = \rho_{\mathrm{B}}(x) & \text{on } \mathbb{R} \end{cases}$$

is given by

$$\rho(t,x) = \rho_{\mathrm{B}}(x)1_{\left\{\cdot\leqslant -\frac{b_1}{a_1}\right\}}(x) + e^{-w_1a_1t}\rho_{\mathrm{B}}\left(\left(x+\frac{b_1}{a_1}\right)e^{-w_1a_1t} - \frac{b_1}{a_1}\right)1_{\left\{\cdot>-\frac{b_1}{a_1}\right\}}(x)$$

for $t \in [0,t_1]$ *and*

$$\rho(t,x) = \rho(t_{k-1},x)1_{\left\{\cdot\leqslant -\frac{b_k}{a_k}\right\}}(x) + e^{-w_ka_kt}\rho\left(t_{k-1},\left(x+\frac{b_k}{a_k}\right)e^{-w_ka_kt} - \frac{b_k}{a_k}\right)1_{\left\{\cdot>-\frac{b_k}{a_k}\right\}}(x)$$

for $t \in [t_{k-1},t_k]$, *for all* $x \in \mathbb{R}$.

Proof. The derivation is straightforward: for $t \in [0,t_1)$, $\rho(t,\cdot)$ has a single discontinuity at $x = -b_1/a_1$, and clearly $\rho(t,x) = \rho_{\mathrm{B}}(x)$ for $x \leqslant -b_1/a_1$. In the case $x > -b_1/a_1$, note that $t \mapsto \rho(t,x(t))$ is constant along the characteristics which solve $\dot{x}(t) = w_1(a_1x(t)+b_1)$, whereupon we deduce the first formula. The computation is easily repeated for $k \geqslant 1$. □

Remark 2.2. *If* $a_k = 0$, *then* $\rho(t,x) = \rho(t_{k-1}, x - w_k(b_k)_+t)$ *for all* $t \in [t_{k-1},t_k]$ *and* $x \in \mathbb{R}$, *again by the method of characteristics.*

From Lemma 2.1 we readily gather that

$$\rho(T,x) = \begin{cases} \alpha^+\rho_{\mathrm{B}}(\alpha^+x+\beta^+), & \text{if } x \geqslant M, \\ \alpha^-\rho_{\mathrm{B}}(\alpha^-x+\beta^-), & \text{if } x \leqslant -M \end{cases}$$

for some $\alpha^\pm > 0$, $\beta^\pm \in \mathbb{R}$ and $M > 0$, all explicitly depending on (w,a,b)—in other words, one can choose $\alpha^\pm, \beta^\pm$ by choosing (w,a,b). Now consider ρ_{B} and ρ_* in Gibbs form:

$$\rho_{\mathrm{B}}(x) \propto e^{-U_{\mathrm{B}}(x)}, \qquad \rho_*(x) \propto e^{-U_*(x)},$$

with $U_{\mathrm{B}}, U_* \in \mathscr{C}^0(\mathbb{R};\mathbb{R}_{\geqslant 0})$. We see that

$$\frac{\rho_*(x)}{\rho(T,x)} \propto \exp\left(U_{\mathrm{B}}(\alpha^\pm x+\beta^\pm) - U_*(x)\right)$$

for all $\|x\| \geqslant M$. So $\mathsf{L}(\rho_*,\rho(T))$ defined in (1.10) is finite if and only if

$$\lim_{\|x\|\to\infty} U_{\mathrm{B}}(\alpha^\pm x+\beta^\pm) - U_*(x) \in [-\infty,+\infty). \tag{2.1}$$

One sees the constraints that (2.1) entails. For instance, one can consider as input the Laplace distribution—corresponding to $U_{\mathrm{B}}(x) = |x-m|/\sigma$ with parameters $m_{\mathrm{B}} \in \mathbb{R}$ and $\sigma_{\mathrm{B}} > 0$—, and as target the Gaussian—corresponding to $U_*(x) = (x-m_*)^2/2\sigma_*^2$ with $m_* \in \mathbb{R}$ and $\sigma_* > 0$—, but not the converse.

Lemma 2.1 can be generalized to the case $d \geqslant 1$ without difficulty.

Lemma 2.3. *Take $T > 0$, a probability density $\rho_{\mathrm{B}} \in \mathscr{C}^0(\mathbb{R}^d)$, and*

$$(w, a, b)(t) = \sum_{k=1}^{K} (w_k, a_k, b_k) 1_{[t_{k-1}, t_k)}(t) \qquad \text{for } t \in [0, T],$$

where $w_k, a_k \in \mathbb{R}^d$, $b_k \in \mathbb{R}$, and $0 = t_0 < t_1 < \ldots < t_K = T$. Then the unique solution to

$$\begin{cases} \partial_t \rho(t,x) + \operatorname{div}\left(w(t) \left(a(t)^\top x + b(t) \right)_+ \rho(t,x) \right) = 0 & \text{on } [0,T] \times \mathbb{R}^d, \\ \rho(0,x) = \rho_{\mathrm{B}}(x) & \text{on } \mathbb{R}^d \end{cases} \tag{2.2}$$

is given by

$$\begin{aligned} \rho(t,x) = \rho(t_{k-1}, x) 1_{\left\{ \langle a_k, \cdot \rangle + b_k \leqslant 0 \right\}}(x) \\ + e^{-w_k^\top a_k (t - t_{k-1})} \rho\left(t_{k-1}, \mathsf{A}_k(x) \right) 1_{\left\{ \langle a_k, \cdot \rangle + b_k > 0 \right\}}(x) \end{aligned}$$

if $w_k^\top a_k \neq 0$, where

$$\mathsf{A}_k(x) = e^{-(t - t_{k-1}) w_k a_k^\top} x - b_k w_k^\top a_k \left(1 - e^{-(t - t_{k-1}) w_k^\top a_k} \right) w_k,$$

and

$$\rho(t,x) = \rho(t_{k-1}, x) 1_{\left\{ \langle a_k, \cdot \rangle + b_k \leqslant 0 \right\}}(x) + \rho(t_{k-1}, \mathsf{B}_k(x)) 1_{\left\{ \langle a_k, \cdot \rangle + b_k > 0 \right\}}(x)$$

if $\langle w_k, a_k \rangle = 0$, where

$$\mathsf{B}_k(x) = e^{-(t - t_{k-1}) w_k a_k^\top} x - e^{-(t - t_{k-1}) w_k a_k^\top} (t - t_{k-1}) b_k w_k,$$

for all $t \in [t_{k-1}, t_k]$ and for all $x \in \mathbb{R}^d$.

Proof. We again start by $t \in [0, t_1)$. For $x \in \{ \langle a_1, \cdot \rangle + b_1 \leqslant 0 \}$, we clearly have $\rho(t,x) = \rho_{\mathrm{B}}(x)$. On $\{ \langle a_1, \cdot \rangle + b_1 > 0 \}$, $t \mapsto \rho(t, x(t))$ is constant along the characteristic surfaces parametrized by solutions to

$$\dot{x}(t) = w_1 a_1^\top x(t) + w_1 b_1.$$

Suppose $w_1^\top a_1 \neq 0$. Since

$$e^{t w_1 a_1^\top} = I_d + \frac{(e^{t w_1^\top a_1} - 1)}{w_1^\top a_1} w_1 a_1^\top,$$

we have

$$\begin{aligned} & e^{t w_1 a_1^\top} \left(\int_0^t e^{-s w_1 a_1^\top} \, \mathrm{d}s \right) w_1 b_1 \\ & \quad = e^{t w_1 a_1^\top} b_1 \left(t + \left(1 - e^{-t w_1^\top a_1} - \frac{t}{w_1^\top a_1} \right) w_1 a_1^\top \right) w_1 \\ & \quad = b_1 \langle w_1, a_1 \rangle \left(1 - e^{-t \langle w_1, a_1 \rangle} \right) e^{t w_1 a_1^\top} w_1. \end{aligned}$$

Thus

$$x(t) = e^{tw_1a_1^\top} x(0) + b_1 w_1^\top a_1 \left(1 - e^{-tw_1^\top a_1}\right) e^{tw_1a_1^\top} w_1,$$

and so

$$\rho(t,x) = \rho_{\mathrm{B}}\left(e^{-tw_1a_1^\top} x - b_1 w_1^\top a_1 \left(1 - e^{-tw_1^\top a_1}\right) w_1\right)$$

for $x \in \{\langle a_1, \cdot\rangle + b_1 > 0\}$ when $w_1^\top a_1 \neq 0$.

Now suppose $w_1^\top a_1 = 0$. Then $w_1 a_1^\top$ is nilpotent, so $e^{tw_1a_1^\top} = I_d + tw_1a_1^\top$. Thus

$$x(t) = e^{tw_1a_1^\top} x(0) + tb_1 w_1,$$

and so

$$\rho(t,x) = \rho_{\mathrm{B}}\left(e^{-tw_1a_1^\top} x - e^{-tw_1a_1^\top} tb_1 w_1\right)$$

for $x \in \{\langle a_1, \cdot\rangle + b_1 > 0\}$.

The computations are easily repeated for $k \geqslant 1$. □

The following comparison principle will be useful.

Lemma 2.4. *Let ρ_1 and ρ_2 be two solutions to (2.2) corresponding to the same θ, and to data $\rho_1(0,\cdot)$ and $\rho_2(0,\cdot)$ respectively. Suppose $\rho_1(0,x) < \rho_2(0,x)$ for all $x \in \mathbb{R}^d$. Then $\rho_1(t,x) < \rho_2(t,x)$ for all $t \in [0,T]$ and $x \in \mathbb{R}^d$.*

Proof. Set $f(t,x) \coloneqq \rho_2(t,x) - \rho_1(t,x)$ and $v(t,x) \coloneqq w(t)(a(t)^\top x + b(t))_+$. Along the characteristics $\dot{x}(t) = v(t,x(t))$,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} f(t,x(t)) &= \partial_t f(t,x(t)) + \nabla_x f(t,x(t))^\top v(t,x(t)) \\ &= -f(t,x(t)) \operatorname{div}_x v(t,x(t)).\end{aligned}$$

Thus

$$f(t,x(t)) = f(0,x(0)) \exp\left(-\int_0^t \operatorname{div}_x v(s,x(s))\,\mathrm{d}s\right)$$

for all $t \in [0,T]$ and $x(0) \in \mathbb{R}$. We may conclude. □

We also use the following elementary lemma.

Lemma 2.5. *Let ρ_1 and ρ_2 be two nonnegative and integrable functions that are proportional to Gaussian densities with covariance matrices Σ_1 and Σ_2 respectively. Suppose*

$$\left\langle (\Sigma_2^{-1} - \Sigma_1^{-1}) e_i, e_i \right\rangle < 0, \qquad \text{for some } i \in [d].$$

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

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

If $\Sigma_2^{-1} \prec \Sigma_1^{-1}$, then there exists some $\overline{M} > 0$ such that

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

Proof. Write

$$\frac{\rho_2(x)}{\rho_1(x)} \propto e^{-\frac{1}{2}\left\langle\left(\Sigma_2^{-1}-\Sigma_1^{-1}\right)x,x\right\rangle+O(\|x\|)}.$$

The proof follows. □

3 Proofs

3.1 Proof of Theorem 1.1

Proof of Theorem 1.1. Fix $\varepsilon_0 > 0$; $\theta = (w, a, b)$ takes the form

$$\theta(t) = \bar{\theta}(t)1_{\left[0,\frac{T}{2}\right]}(t) + \sum_{k=1}^{2d} \theta_k 1_{[T_{k-1},T_k]}(t),$$

where $T_k := \frac{T}{2} + \frac{kT}{4d}$, and $\bar{\theta}$ is piecewise constant with at most $\left\lceil \frac{2R(\varepsilon)}{h(\varepsilon)} \right\rceil^d (d+10)$ switches and such that

$$\left\| \rho\left(\frac{T}{2}\right) - \rho_* \right\|_{L^1(\mathbb{R}^d)} \leqslant \varepsilon_0. \tag{3.1}$$

Such a $\bar{\theta}$ exists by virtue of [RBZ24, Theorem 1]. We seek to construct $\theta_1, \ldots, \theta_{2d}$ so that

$$\mathsf{L}(\rho_*, \rho(T)) = \lim_{\|x\|\to+\infty} \frac{\rho_*(x)}{\rho(T,x)} = 0.$$

Step 1.

Thanks to Lemma 2.3, there exists a finite $n \geqslant 1$, as well as polyhedra $P_i \subset \mathbb{R}^d$ and coefficients $(\alpha_i, A_i, \beta_i) \in \mathbb{R}_{>0} \times \mathscr{M}_{d\times d}(\mathbb{R}) \times \mathbb{R}$ with $\operatorname{spec}(A_i) \subset (0, +\infty)$ for $i \in [n]$, such that

$$\rho\left(\frac{T}{2}, x\right) = \sum_{i=1}^{n} \alpha_i \rho_{\mathrm{B}}(A_i x + \beta_i) 1_{P_i}(x) \qquad \text{for all } x \in \mathbb{R}^d.$$

Moreover, the polyhedra P_i form a partition of $\mathbb{R}^d$, which implies $\rho(T/2) > 0$. Each $\rho_i := \alpha_i \rho_{\mathrm{B}}(A_i \cdot + \beta_i)$ is a Gaussian function on $\mathbb{R}^d$, and the Gaussian probability density $\mathfrak{Z}_i^{-1}\rho_i$ has covariance matrix

$$\Sigma_i := A_i^{-1}\Sigma_{\mathrm{B}}\left(A_i^{-1}\right)^\top.$$

Take $0 < \underline{\sigma}^2 < \min_{i\in[n]} \operatorname{spec}(\Sigma_i)$ and define

$$\underline{\rho}(x) := \alpha \exp\left(-\frac{\|x\|^2}{2\underline{\sigma}^2}\right) \tag{3.2}$$

for $x \in \mathbb{R}^d$, where $\alpha > 0$ is a constant to be specified. Then $\Sigma_i^{-1} \prec I_d/\underline{\sigma}^2$ for all $i \in [n]$, so by virtue of Lemma 2.5, there exists $M(i) > 0$ such that

$$\underline{\rho}(x) < \rho_i(x) \qquad \text{for all} \quad \|x\| > M(i).$$

Taking $\overline{M} \geqslant M \vee \max_{i\in[n]} M(i)$ it follows that (1.7) and $\underline{\rho}(x) < \rho\,(T/2, x)$ hold for all $\|x\| > \overline{M}$. Choosing α small enough, we can deduce

$$\underline{\rho}(x) < \rho\left(\frac{T}{2}, x\right) \qquad \text{for all } x \in \mathbb{R}^d, \tag{3.3}$$

and $\overline{M}$ large enough to deduce

$$\int_{[-\overline{M},\overline{M}]^d} \rho_*(x)\,\mathrm{d}x \wedge \int_{[-\overline{M},\overline{M}]^d} \rho\left(\frac{T}{2}, x\right)\,\mathrm{d}x > 1 - \varepsilon_0. \tag{3.4}$$

Step 2.

We will choose $\theta_1, \dots, \theta_{2d}$ so that

$$\rho_\bullet(x) < \underline{\rho}(T, x), \qquad \text{in } \mathbb{R}^d \setminus [-\overline{\overline{M}}, \overline{\overline{M}}]^d, \tag{3.5}$$

for some larger $\overline{\overline{M}} \geqslant \overline{M}$, where $\underline{\rho}(t, \cdot)$ is the unique solution to (1.3) in $[T/2, T]$ with data $\underline{\rho}\,(T/2, \cdot) = \underline{\rho}(\cdot)$. To this end, take

$$\begin{aligned} \theta_{2k-1} &\coloneqq (\omega e_k, e_k, -\overline{M}), \\ \theta_{2k} &\coloneqq (-\omega e_k, -e_k, -\overline{M}), \end{aligned}$$

where $\omega > 0$ is to be chosen, for $k \in [d]$. By virtue of Lemma 2.3, the solution reads

$$\begin{aligned} \underline{\rho}(t, x) =& \underline{\rho}(T_{2k-2}, x) 1_{\{x\colon x_k \leqslant \overline{M}\}}(x) \\ &+ e^{-\omega(t-T_{2k-2})} \underline{\rho}\left(T_{2k-2}, \mathsf{A}_{2k-2}(t, x)\right) 1_{\{x\colon x_k > \overline{M}\}}(x) \end{aligned}$$

if $t \in [T_{2k-2}, T_{2k-1}]$, and

$$\begin{aligned} \underline{\rho}(t, x) =& \underline{\rho}(T_{2k-1}, x) 1_{\{x\colon x_k \geqslant -\overline{M}\}}(x) \\ &+ e^{-\omega(t-T_{2k-1})} \underline{\rho}\left(T_{2k-1}, \mathsf{A}_{2k-1}(t, x)\right) 1_{\{x\colon x_k < -\overline{M}\}}(x) \end{aligned}$$

if $t \in [T_{2k-1}, T_{2k}]$, where

$$\mathsf{A}_j(t, x) = e^{-\omega(t-T_j)e_k e_k^\top} x + (-1)^j \overline{M}\omega^2 \left(1 - e^{-\omega(t-T_j)}\right) e_k$$

for $j \in \{2k-2, 2k-1\}$, for all $k \in [d]$ and $x \in \mathbb{R}^d$. It ensues that

$$\begin{aligned} \underline{\rho}(T, x) =& \underline{\rho}\left(\frac{T}{2}, x\right) 1_{[-\overline{M},\overline{M}]^d}(x) \\ &+ \sum_{k=1}^{d} \sum_{\ell=0}^{1} \underline{\rho}_1^{k\ell}(x) 1_{S_{k\ell}}(x) + \sum_{q\in\{0,1\}^d} \underline{\rho}_2^q(x) 1_{\mathbb{Q}_q}(x), \end{aligned} \tag{3.6}$$

where $\underline{\rho}_1^{k\ell}$, $\underline{\rho}_2^k$ are Gaussian functions on $\mathbb{R}^d$, and $\mathbb{Q}_q$, $\mathcal{S}_{k\ell}$ are defined by

$$\mathbb{Q}_q := \left\{ x \in \mathbb{R}^d : (-1)^{q_j} x_j > \overline{M} \quad \text{for all } j \in [d] \right\}$$

and

$$\mathcal{S}_{k\ell} := \left\{ (-1)^\ell x_k > \overline{M} \right\} \setminus \bigcup_{q \in \{0,1\}^d} \mathbb{Q}_q.$$

The Gaussian density $\mathfrak{T}^{-1}\underline{\rho}_1^{k\ell}$ has covariance

$$\begin{aligned}
\Sigma_{1,k\ell} &= \left(e^{-\frac{\omega T}{4d}} e_k e_k^\top \right)^{-1} \underline{\sigma}^2 I_d \left(\left(e^{-\frac{\omega T}{4d}} e_k e_k^\top \right)^{-1} \right)^\top \\
&= \underline{\sigma}^2 \left(I_d + (e^{-\frac{\omega T}{4d}} - 1) e_k e_k^\top \right)^{-1} \left(I_d + (e^{-\frac{\omega T}{4d}} - 1) e_k e_k^\top \right)^{-1} \\
&= \underline{\sigma}^2 \left(I_d + (e^{\frac{\omega T}{4d}} - 1) e_k e_k^\top \right) \left(I_d + (e^{\frac{\omega T}{4d}} - 1) e_k e_k^\top \right) \\
&= \underline{\sigma}^2 \left(I_d + (e^{\frac{\omega T}{2d}} - 1) e_k e_k^\top \right),
\end{aligned}$$

Taking

$$\omega > \frac{2d}{T} \log \left(\frac{\sigma_\bullet^2}{\underline{\sigma}^2} \right), \tag{3.7}$$

we find for each $k \in [d]$, $\ell \in \{0,1\}$, that

$$\left\langle \left(\Sigma_{1,k\ell}^{-1} - \frac{1}{\sigma_\bullet^2} I_d \right) e_k, e_k \right\rangle = \frac{e^{-\frac{\omega T}{2d}}}{\underline{\sigma}^2} - \frac{1}{\sigma_\bullet^2} < 0.$$

A similar computation for the covariance of $\mathfrak{T}^{-1}\underline{\rho}_2^q$ gives $\Sigma_{2,q} = \underline{\sigma}^2 e^{\frac{\omega T}{2d}} I_d$. We see that $\Sigma_{2,q}^{-1} \prec I_d/\sigma_\bullet^2$ for all $q \in \{0,1\}^d$ when ω satisfies (3.7). By Lemma 2.5 there exists some $M(0) \geqslant \overline{M}$ such that

$$\rho_\bullet(x) < \underline{\rho}_2^q(x) \qquad \text{for all } \|x\| > M(0),$$

for $q \in \{0,1\}^d$. Also by Lemma 2.5, for all $k \in [d]$ and $\ell \in \{0,1\}$ there exists $M(k) > 0$ such that

$$\rho_\bullet(x) < \underline{\rho}_1^{k\ell}(x) \qquad \text{in } \left\{ x \in \mathbb{R}^d \colon |x_k| > M(k) \text{ and } |x_j| < M(0) \text{ for } j \neq k \right\}.$$

Thanks to (3.6), taking $\overline{\overline{M}} := \max_{k \in \{0,\dots,d\}} M(k)$ ensures (3.5).

Step 3.

Combining (1.7), (3.3), (3.5) and Lemma 2.4, we find

$$\rho_*(x) < \rho(T,x), \qquad \text{in } \mathbb{R}^d \setminus [-\overline{\overline{M}}, \overline{\overline{M}}]^d.$$

In particular, this implies $\mathsf{L}(\rho_*, \rho(T)) = 0$, where L is defined in (1.10). We moreover have

$$\min_{x \in [-\overline{\overline{M}}, \overline{\overline{M}}]^d} \rho\,(T,x) > \min_{x \in [-\overline{\overline{M}}, \overline{\overline{M}}]^d} \underline{\rho}\,(T,x) > 0$$

by (3.2), (3.3) and (3.6), so we conclude that

$$\sup_{x \in \mathbb{R}^d} \frac{\rho_*(x)}{\rho(T,x)} < +\infty.$$

On the other hand, $\rho(T,x) = \rho(T/2,x)$ for $x \in [-\overline{M}, \overline{M}]^d$, which, combined with (3.1) and (3.4), yields

$$\begin{aligned} \|\rho(T) - \rho_*\|_{L^1(\mathbb{R}^d)} \leqslant & \int_{[-\overline{M}, \overline{M}]^d} \left| \rho\left(\frac{T}{2}, x\right) - \rho_*\,(x) \right| \mathrm{d}x \\ & + \int_{\mathbb{R}^d \setminus [-\overline{M}, \overline{M}]^d} \rho(T,x)\,\mathrm{d}x \\ & + \int_{\mathbb{R}^d \setminus [-\overline{M}, \overline{M}]^d} \rho_*(x)\,\mathrm{d}x < 3\varepsilon_0. \end{aligned}$$

Taking ε_0 small enough and applying Lemma 1.2, we conclude. □

3.2 Proof of Theorem 1.3

Proof of Theorem 1.3. The proof is almost identical to that of Theorem 1.1–we only discuss the differences.

In Step 1, we take $\overline{\sigma}^2 > \max_{i \in [n]} \operatorname{spec}(\Sigma_i)$ instead of $\underline{\sigma}^2$, defining $\overline{\rho}(x)$ as $\underline{\rho}(x)$ in (3.2), as to have $\overline{\rho}(x) > \rho(T/2,x)$ for all $x \in \mathbb{R}^d$.

In Step 2, we take

$$\begin{aligned} \theta_{2k-1} &= (-\omega e_k, e_k, -\overline{M}), \\ \theta_{2k} &= (\omega e_k, -e_k, -\overline{M}), \end{aligned}$$

for $k \in [d]$, with $\omega > 2d \log\left(\sigma_\bullet^2/\overline{\sigma}^2\right)/T$. Then the same argument as in Step 2 yields $\overline{\rho}(T,x) < \rho_\bullet(x)$ for all $x \in \mathbb{R}^d \setminus [-\overline{\overline{M}}, \overline{\overline{M}}]^d$, for some $\overline{\overline{M}} \geqslant \overline{M}$.

In Step 3, Lemma 2.4 then yields $\rho(T,x) < \rho_*(x)$ for all $x \in \mathbb{R}^d \setminus [-\overline{\overline{M}}, \overline{\overline{M}}]^d$. This implies $\sup_{x \in \mathbb{R}^d} \frac{\rho_*(x)}{\rho(T,x)} < +\infty$ because $\rho_* > 0$ in $\mathbb{R}^d$. The conclusion follows thereupon. □

3.3 Beyond Gaussians

In this section we comment on how the extension discussed in Remark 1.5 can be accounted for in the above proofs. Fix $\varepsilon_0 > 0$. From [Aza13] we deduce the existence of a convex $U_\bullet \in \mathscr{C}^2(\mathbb{R}^d)$ with $U_\bullet \leqslant \operatorname{conv} U_*$ such that

$$\|U_\bullet - \operatorname{conv} U_*\|_{\mathscr{C}^0(\mathbb{R}^d)} < \varepsilon_0.$$

For any $A \in \mathscr{M}_{d\times d}(\mathbb{R})$ with $\mathrm{spec}(A) \subset (0,+\infty)$, the $\lambda_A > 0$ from (1.13) gives

$$\begin{aligned}&\lim_{\|x\|\to+\infty} \frac{U_\bullet(\lambda_A x)}{U_{\mathrm{B}}(Ax)}\\ &= \lim_{\|x\|\to+\infty} \frac{U_\bullet(\lambda_A x) - \operatorname{conv} U_*(\lambda_A x)}{U_{\mathrm{B}}(Ax)} + \frac{\operatorname{conv} U_*(\lambda_A x)}{U_{\mathrm{B}}(Ax)} = +\infty.\end{aligned}$$

We now show that $e^{-U_\bullet}$ satisfies a comparison principle similar to Lemma 2.5. Namely, given any P, Q diagonal with positive entries and $\beta \in \mathbb{R}^d$, we prove that

$$\lim_{|x_i|\to+\infty} U_\bullet(Px+\beta) - U_\bullet(Qx) = +\infty \qquad \text{if } P_i > Q_i, \tag{3.8}$$

and

$$\lim_{\|x\|\to+\infty} U_\bullet(Px+\beta) - U_\bullet(Qx) = +\infty \qquad \text{if } P \succ Q. \tag{3.9}$$

If $\varepsilon_0 > 0$ is small enough, thanks to (1.12) we have

$$U_\bullet(x) \gtrsim \|x\| \qquad \text{for all } \|x\| > M_*.$$

By virtue of $U_\bullet \in \mathscr{C}^2$ and convexity, there exists $M_\bullet > 0$ such that for all $i \in [d]$,

$$\partial_{x_i} U_\bullet(x)|x_i| \gtrsim x_i \qquad \text{for all } \|x\| > M_\bullet.$$

Take $\|x\| > M_\bullet$ and, without loss of generality, suppose $x \in (\mathbb{R}_{>0})^d$, so that

$$\partial_{x_i} U_\bullet(x) \gtrsim 1$$

for all $i \in [d]$. We now decompose

$$\begin{aligned}U_\bullet(Px) - U_\bullet(Qx) &= U_\bullet(Px) - U_\bullet\left(P_1x_1, \dots, P_{d-1}x_{d-1}, Q_dx_d\right)\\ &\quad + \cdots + U_\bullet\left(P_1x_1, Q_2x_2, \dots, Q_dx_d\right) - U_\bullet\left(Qx\right)\\ &= \sum_{i=1}^d \int_{(Qx)_i}^{(Px)_i} \partial_{x_i} U_\bullet((Px)_{<i}, t, (Qx)_{>i})dt,\end{aligned}$$

where

$$((Px)_{<i}, t, (Qx)_{>i}) = ((Px)_1, \dots, (Px)_{i-1}, t, (Qx)_{i+1}, \dots, (Qx)_n).$$

Suppose $P_i > Q_i$ for all i. We deduce:

$$U_\bullet(Px) - U_\bullet(Qx) \gtrsim \sum_{i=1}^d (P_i - Q_i)x_i \gtrsim \|x\|.$$

Now suppose $P_i > Q_i$ for some $i \in [d]$. If we fix x_j for all $j \neq i$, we deduce

$$U_\bullet(Px) - U_\bullet(Qx) \gtrsim (P_i - Q_i)x_i - 1 \gtrsim x_i - 1.$$

These inequalities give (3.8) and (3.9).

It remains to outline the adjustments in the proof of Theorem 1.1 required to generalize the result. The generalization of Theorem 1.3 is completely analogous.

In Step 1, since $\beta_i \in \mathbb{R}^d$ and $\alpha_i > 0$ are fixed for all $i \in [n]$, thanks to (1.13) we can find $\underline{\lambda} > \max_{i\in[n]} \lambda_{A_i}$ and $\overline{M} > 0$ such that

$$U_\bullet(\underline{\lambda}x) > \max_{i\in[n]} U_{\mathrm{B}}(A_i x + \beta_i) - \log \alpha_i \qquad \text{for all } \|x\| > \overline{M}.$$

We define $\underline{\rho}(x) = \alpha e^{-U_\bullet(\underline{\lambda}x)}$ for all $x \in \mathbb{R}^d$, which satisfies

$$\underline{\rho}(x) < \min_{i\in[n]} \alpha_i \rho_{\mathrm{B}}(A_i x + \beta_i) \leqslant \rho\left(\frac{T}{2}, x\right) \qquad \text{for all} \quad \|x\| > \overline{M}.$$

Choosing α small enough, we can deduce $\underline{\rho}(x) < \rho\,(T/2, x)$ for all $x \in \mathbb{R}^d$.

In Step 2, we take the same controls $\theta_1, \dots, \theta_{2d}$ that give (3.6), now with

$$\underline{\rho}_1^{k\ell}(x) = e^{-\frac{\omega T}{4d}} \underline{\rho}\left(\left(I_d + (e^{-\frac{\omega T}{4d}} - 1)e_k e_k^\top\right) x + \beta_{1,k\ell}\right),$$
$$\underline{\rho}_2^{q}(x) = e^{-\frac{\omega T}{4}} \underline{\rho}\left(e^{-\frac{\omega T}{4d}} x + \beta_{2,q}\right)$$

for some $\beta_{1,k\ell}, \beta_{2,q} \in \mathbb{R}^d$. Take $\omega > 4d \log \underline{\lambda}/T$, which gives $\underline{\rho}(T, x) > e^{-U_\bullet(x)}$ for $\|x\|$ large enough, thanks to (3.8) and (3.9). Since $U_\bullet \leqslant U_*$ by construction, Lemma 2.4 ensures the existence of $\overline{\overline{M}} \geqslant \overline{M} \vee M_\bullet$ such that

$$\rho_*(x) < \underline{\rho}(T, x) < \rho(T, x), \qquad \text{in } \mathbb{R}^d \setminus [-\overline{\overline{M}}, \overline{\overline{M}}]^d.$$

3.4 Proof of Proposition 1.6

Proof of Proposition 1.6. For $x \in \mathbb{R}$ and $t \in \mathbb{R}$, consider

$$x(t) = \begin{cases} x^{e^t} & \text{if } x \geqslant 1 \\ x & \text{if } x \leqslant 1. \end{cases}$$

This function solves

$$\begin{cases} \dot{x}(t) = (|x(t)| \log x(t)_+)_+ & t \in \mathbb{R} \\ x(0) = x. \end{cases}$$

It is in fact the unique solution. Indeed, the local Lipschitz condition guarantees uniqueness on the maximal interval of existence, and $x(t)$ is defined for all $t \in \mathbb{R}$.

One can thus define the solution $\rho(t, x)$ to (1.14) uniquely in the weak sense. We focus on the behavior of the solution $\rho(t, \cdot)$ on $\{x\colon x > 1\}$. By the method of characteristics, we can write

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho(t, x(t)) = -(\log x(t) + 1)\rho(t, x(t)),$$

which translates into

$$\int_0^t \frac{\mathrm{d}}{\mathrm{d}s}\rho(s,x(s))\rho(s,x(s))^{-1}\,\mathrm{d}s = \int_0^t -(\log(x)e^s+1)\,\mathrm{d}s.$$

Whereupon,

$$\log\rho(t,x(t)) - \log\rho_{\mathrm{B}}(x) = -\left(\log(x)(e^t-1)+t\right),$$

whence

$$\rho(t,x(t)) = \rho_{\mathrm{B}}(x)\exp\left(-\log x^{(e^t-1)} - t\right)$$

and

$$\rho(t,x^{e^t}) = e^{-t}\rho_{\mathrm{B}}(x)x^{-(e^t-1)}.$$

Taking $y = x^{e^t}$, we obtain $y^{\frac{1}{e^t}} = x$ and

$$\rho(t,y) = e^{-t}\rho_{\mathrm{B}}\left(y^{\frac{1}{e^t}}\right)y^{1-\frac{1}{e^t}}.$$

Upon considering the case $x < 1$, we conclude

$$\rho(t,x) = \rho_{\mathrm{B}}(x)1_{\{x\colon x<1\}}(x) + e^{-t}\rho_{\mathrm{B}}\left(x^{\frac{1}{e^t}}\right)x^{1-\frac{1}{e^t}}1_{\{x\colon x\geqslant 1\}}(x).$$

We thus see that

$$\lim_{x\to+\infty}\frac{\rho(t,x)}{\rho_{\mathrm{B},q}(x)} \propto e^{-t}\lim_{x\to+\infty}\frac{\exp\left(-x^{\frac{p}{e^t}}\right)x^{1-\frac{1}{e^t}}}{\exp(-x^q)}.$$

It therefore suffices to take

$$T_q > \log\left(\frac{p}{q}\right),$$

which is positive if $p > q$, to ensure

$$\lim_{x\to+\infty}\frac{\rho(T_q,x)}{\rho_{\mathrm{B},q}(x)} < +\infty. \qquad \square$$

Remark 3.1. *Observe that by simply changing the sign of the vector field (or reversing time), we obtain the same result for $q \geqslant p$.*

3.5 Proof of Proposition 1.9

Ultimately, $\theta = (w,b)$ takes the form

$$\theta(t) = \sum_{k=1}^{4n+3}\theta_k 1_{[T_{k-1},T_k]}(t),$$

where $\theta_k = (w_k,b_k) \in \mathscr{M}_{d\times d}(\mathbb{R})\times\mathbb{R}^d$ are constant and determined throughout the proof, whereas $T_k \coloneqq \frac{kT}{4n+3}$.

Step 1.

We begin by taking $\theta_1 = (w_1, b_1)$ with

$$w_1 = 0, \qquad b_1 = \beta_1 \mathbf{1}$$

for $\beta_1 > 0$. Then for all $i \in [n]$,

$$x^{(i)}(t) = x^{(i)}(0) + t\beta_1 \, \mathbf{1} \qquad \text{for } t \in [0, T_1].$$

Taking β_1 large enough, we find

$$x_k^{(i)}(T_1) > 0 \qquad \text{for all } k \in [d].$$

We now take $\theta_2 = (w_2, b_2)$ with

$$w_2 = \alpha_1 \begin{bmatrix} 0 & \dots & 0 & \dots & 0 \\ v_1 & \dots & v_k & \dots & v_d \\ 0 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 0 \end{bmatrix}, \qquad b_2 = 0,$$

where $\alpha_1 > 0$ and $v = (v_1, \dots, v_d) \in (\mathbb{R}_{>0})^d$ is such that v and $x^{(i)}(T_1) - x^{(j)}(T_1)$ are not orthogonal for all $i \neq j$. Then

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(x_2^{(i)} - x_2^{(j)}\right)(t) = \alpha_1 \left\langle v, \left(x^{(i)} - x^{(j)}\right)(t) \right\rangle \qquad \text{for } t \in [T_1, T_2],$$

hence $x_2^{(i)}(t) - x_2^{(j)}(t)$ has a sign whenever $t - T_1 \ll 1$, for all $i \neq j$. Taking α_1 sufficiently small and rescaling time appropriately, we deduce $x_2^{(i)}(T_2) \neq x_2^{(j)}(T_2)$ for all $i \neq j$, or equivalently

$$\Phi_{\theta_2}^{T_2}\left(\Phi_{\theta_1}^{T_1}\left(x^{(i)}\right)\right)_2 \neq \Phi_{\theta_2}^{T_2}\left(\Phi_{\theta_1}^{T_1}\left(x^{(j)}\right)\right)_2, \qquad \text{for all } i \neq j. \tag{3.10}$$

Consider the backward equation

$$\begin{cases} \dot{y}(t) = \overline{w}(t)(y(t))_+ + \overline{b}(t) & t \in [T_{4n+1}, T], \\ y(T) = y_0, \end{cases}$$

and note that y also solves the forward equation (1.17) for $\theta(t) = -\overline{\theta}(T - t)$.

Take $\theta_{4n+3} = (w_{4n+3}, b_{4n+3})$ with

$$w_{4n+3} = 0, \qquad b_{4n+3} = -\beta_2 \mathbf{1}$$

for $\beta_2 > 0$ large enough to ensure

$$y_k^{(i)}(T_{4n+2}) > 0 \qquad \text{for all } (k, i) \in [d] \times [n].$$

Take $\theta_{4n+2} = (w_{4n+2}, b_{4n+2})$ with

$$w_{4n+2} = -\alpha_2 \begin{bmatrix} u_1 & \dots & u_k & \dots & u_d \\ 0 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 0 \end{bmatrix}, \qquad b_{4n+2} = 0$$

where $\alpha_2 > 0$ and $u = (u_1, \dots, u_d) \in (\mathbb{R}_{>0})^d$ is such that u and $y^{(i)} - y^{(j)}$ are not orthogonal for all $i \neq j$. Setting

$$\overline{y}^{(i)} := \Phi_{-\theta_{4n+2}}^{T_{4n+1}} \left(\Phi_{-\theta_{4n+3}}^{T_{4n+2}} \left(y^{(i)} \right) \right)$$

for $i \in [n]$, arguing just as above, by choosing α_2 sufficiently small, it ensues

$$\overline{y}_1^{(i)} \neq \overline{y}_1^{(j)} \qquad \text{for all } i \neq j.$$

Step 2.

Due to (3.10), we can relabel all points as to have

$$x_2^{(i)}(T_2) < x_2^{(i+1)}(T_2) \qquad \text{for all } i \in [n-1]. \tag{3.11}$$

We now show that

$$x_1^{(i)}(T_{2n+1}) = \overline{y}_1^{(i)} + c_1 \tag{3.12}$$

for some $c_1 \in \mathbb{R}$ and all $i \in [n]$. We argue by induction. The base case $i = 1$ trivially holds with $\theta_3 \equiv 0$ and $c_1 = x_1^{(1)} - \overline{y}_1^{(1)}$, at time T_3. Of course, since the points don't move, the order (3.11) is propagated up to time T_3. Assume (3.11) and (3.12) hold at time T_{2n-1} for all $i \in [n-1]$. Take $\theta_{2n} = (w_{2n}, b_{2n})$ with

$$w_{2n} = 0, \qquad b_{2n} = -\alpha_3 \operatorname{sign} \left(x_2^{(n-1)}(T_{2n-1}) \right) e_2,$$

where $\alpha_3 > 0$. Then for all $i \in [n]$,

$$\begin{aligned} x_2^{(i)}(t) &= x_2^{(i)}(T_{2n-1}) - \alpha_3 t \operatorname{sign} \left(x_2^{(n-1)}(T_{2n-1}) \right), \\ x_k^{(i)}(t) &= x_k^{(i)}(T_{2n-1}) \qquad\qquad\qquad \text{for } k \neq 2, \end{aligned} \qquad \text{for } t \in [T_{2n-1}, T_{2n}].$$

By virtue of the order (3.11) which holds at time T_{2n-1} by heredity, we can choose α_3 as to ensure

$$x_2^{(1)}(T_{2n}) < \ldots < x_2^{(n-1)}(T_{2n}) < 0 < x_2^{(n)}(T_{2n}). \tag{3.13}$$

All the while,

$$x_k^{(i)}(T_{2n}) = x_k^{(i)}(T_{2n-1})$$

for all $k \neq 2$ and $i \in [n]$. Now take $\theta_{2n+1} = (w_{2n+1}, b_{2n+1})$ with

$$w_{2n+1} = \frac{4n+3}{Tx_2^{(n)}(T_{2n})} \begin{bmatrix} 0 & \omega & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}, \qquad b_{2n+1} = 0,$$

where

$$\omega = \overline{y}_1^{(n)} - x_1^{(n)}(T_{2n}) + c_1.$$

Because of (3.13) we see that $x_1^{(n)}(T_{2n+1}) = \overline{y}_1^{(n)} + c_1$ holds, and we moreover have $x_1^{(i)}(T_{2n+1}) = x_1^{(i)}(T_{2n}) = \overline{y}_1^{(i)} + c_1$ for $i \in [n-1]$ because of the heredity assumption. All in all, we conclude that (3.12) holds.

Step 3.

We will now apply this argument simultaneously to all coordinates. We relabel all points anew as to have

$$x_1^{(i)}(T_{2n+1}) < x_1^{(i+1)}(T_{2n+1}) \tag{3.14}$$

for all $i \in [n-1]$. We show that

$$x^{(i)}(T_{4n}) = \overline{y}^{(i)} + c \tag{3.15}$$

for some $c \in \mathbb{R}^d$ and all $i \in [n]$. We again argue by induction. The base case trivially holds at time T_{2n+2} by taking $\theta_{2n+2} \equiv 0$ and $c_k = x_k^{(1)}(T_{2n+1}) - \overline{y}_k^{(1)}$ for all $k \in [d]$. Again, since the points don't move, the order (3.14) is propagated up to time T_{2n+2}. Assume that (3.14) and (3.15) hold at time T_{4n-2} for all $i \in [n-1]$. Take $\theta_{4n-1} = (w_{4n-1}, b_{4n-1})$ with

$$w_{4n-1} = 0, \qquad b_{4n-1} = -\alpha_4 \operatorname{sign}\left(x_1^{(n-1)}(T_{4n-2})\right) e_1$$

for $\alpha_4 > 0$. Then for all $i \in [n]$,

$$\begin{aligned} x_1^{(i)}(t) &= x_1^{(i)}(T_{4n-2}) - \alpha_4 t \operatorname{sign}\left(x_1^{(n-1)}(T_{4n-2})\right), \\ x_k^{(i)}(t) &= x_k^{(i)}(T_{4n-2}), \qquad \text{for } k \neq 1, \end{aligned} \qquad \text{for } t \in [T_{4n-2}, T_{4n-1}].$$

By virtue of (3.14) which holds at time T_{4n-2} by heredity, we can choose α_4 as to ensure

$$x_1^{(1)}(T_{4n-1}) < \ldots < x_1^{(n-1)}(T_{4n-1}) < 0 < x_1^{(n)}(T_{4n-1}). \tag{3.16}$$

All the while

$$x_k^{(i)}(T_{4n-1}) = x_k^{(i)}(T_{4n-2})$$

for all $k \neq 1$ and $i \in [n]$. Now take $\theta_{4n} = (w_{4n}, b_{4n})$ as

$$w_{4n} = \frac{4n+3}{Tx_1^{(n)}(T_{4n-1})} \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \omega_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{d-1} & 0 & \cdots & 0 \end{bmatrix}, \qquad b_{4n} = 0,$$

where

$$\omega_k = \overline{y}_k^{(n)} - x_k^{(n)}(T_{4n-1}) + c_k$$

for $k \in [d-1]$. Because of (3.16), we see that $x^{(n)}(T_{4n}) = \overline{y}^{(n)} + c$, and we moreover have $x^{(i)}(T_{4n}) = x^{(i)}(T_{4n-1}) = \overline{y}^{(i)} + c$ for $i \in [n-1]$ because of the heredity assumption. All in all, we conclude that (3.15) holds.

We can conclude the proof by taking $\theta_{4n+1} = (w_{4n+1}, b_{4n+1})$ with

$$w_{4n+1} = 0, \qquad b_{4n+1} = -c. \qquad \square$$

3.6 Proof of Proposition 1.10

Let $Y \in \mathcal{M}_{d\times n}(\mathbb{R})$ be the matrix whose i-th column, for $i \in [n]$, is $y^{(i)} \in \mathbb{R}^d$. We define $\gamma : [0,1] \times \mathcal{M}_{d\times n}(\mathbb{R}) \to \mathcal{M}_{d\times n}(\mathbb{R})$ by

$$\gamma(s, X) = (1-s)X + sY.$$

There exists $\varepsilon > 0$ such for all $X \in \mathcal{M}_{d\times n}(\mathbb{R})$ with $\|Y - X\|_1 \leqslant \varepsilon$, we have

$$\operatorname{rank}(\sigma(\gamma(s, X))) = n \tag{3.17}$$

for all $s \in [0,1]$. Take any $X \in \mathcal{M}_{d\times n}(\mathbb{R})$ as above, and then consider the map $\mathcal{G} : [0,1] \times \mathcal{M}_{d\times d}(\mathbb{R}) \to \mathcal{M}_{d\times n}(\mathbb{R})$ defined by

$$\mathcal{G}(s, w) = w\sigma(\gamma(s, X)),$$

with σ being applied element-wise to the matrix. Observe that $\mathcal{G}(s, \cdot)$ is a linear map for any $s \in [0,1]$, and (3.17) ensures that $\mathcal{G}(s, \cdot)$ is surjective. Therefore $\mathcal{G}(s, \cdot)$ has a right inverse $\mathcal{F}(s, \cdot) : \mathcal{M}_{d\times n}(\mathbb{R}) \to \mathcal{M}_{d\times d}(\mathbb{R})$, also a linear map, which takes the form

$$\mathcal{F}(s, v) = \operatorname*{argmin}_{w \in \ker(\mathcal{G}(s,\cdot) - v)} \|w\|_2.$$

The family of linear operators $(\mathcal{F}(s, \cdot))_{s\in[0,1]}$ is uniformly bounded in operator norm, that is to say,

$$\max_{s\in[0,1]} \|\mathcal{F}(s, \cdot)\|_{\mathcal{L}(\mathcal{M}_{d\times n}(\mathbb{R}); \mathcal{M}_{d\times d}(\mathbb{R}))} \leqslant C \tag{3.18}$$

for some constant $C > 0$. For $t \in [0, T]$ set

$$w(t) := \mathcal{F}\left(\frac{t}{T}, \frac{Y - X}{T}\right).$$

For $i \in [n]$, the i-th column $x^i(t) \in \mathbb{R}^d$ of

$$X(t) \coloneqq \gamma\left(\frac{t}{T}, X\right) = \left(1 - \frac{t}{T}\right) X + \frac{t}{T} Y$$

satisfies

$$\begin{cases} \dot{x}^i(t) = w(t)\sigma(x^i(t)) & \text{for } t \in [0,T], \\ x^i(0) = x^{(i)}, \\ x^i(T) = y^{(i)}. \end{cases}$$

By virtue of (3.18),

$$\|w(t)\|_2 = \left\| \mathscr{F}\left(\frac{t}{T}, \frac{Y-X}{T}\right) \right\|_2 \lesssim \frac{C}{T} \|Y - X\|_1,$$

as desired. □

References

[ABBG+12] Fatiha Alabau-Boussouira, Roger Brockett, Olivier Glass, Jérôme Le Rousseau, Enrique Zuazua, and Roger Brockett. Notes on the control of the liouville equation. *Control of Partial Differential Equations: Cetraro, Italy 2010, Editors: Piermarco Cannarsa, Jean-Michel Coron*, pages 101–129, 2012.

[ABVE23] Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

[ÁLHSZ24] Antonio Álvarez-López, Arselane Hadj Slimane, and Enrique Zuazua. Interplay between depth and width for interpolation in neural odes. *Neural Networks*, 180:106640, 2024.

[AS22] Andrei Agrachev and Andrey Sarychev. Control on the manifolds of mappings with a view to the deep learning. *Journal of Dynamical and Control Systems*, 28(4):989–1008, 2022.

[Aza13] Daniel Azagra. Global and fine approximation of convex functions. *Proceedings of the London Mathematical Society*, 107(4):799–824, October 2013.

[BKM17] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[BP24] Jon Asier Bárcena-Petisco. Optimal control for neural ode in a long time horizon and applications to the classification and simultaneous controllability problems. 2024.

[CLLS23] Jingpu Cheng, Qianxiao Li, Ting Lin, and Zuowei Shen. Interpolation, approximation and controllability of deep neural networks. *arXiv preprint arXiv:2309.06015*, 2023.

[CLT20] Christa Cuchiero, Martin Larsson, and Josef Teichmann. Deep neural networks, generic universal interpolation, and controlled odes. *SIAM Journal on Mathematics of Data Science*, 2(3):901–919, 2020.

[CRBD18] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 31, 2018.

[DKB14] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

[DMR19] Michel Duprez, Morgan Morancey, and Francesco Rossi. Approximate and exact controllability of the continuity equation with a localized vector field. *SIAM Journal on Control and Optimization*, 57(2):1284–1311, 2019.

[E17] Weinan E. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 1(5):1–11, 2017.

[EGBO22] Karthik Elamvazhuthi, Bahman Gharesifard, Andrea L Bertozzi, and Stanley Osher. Neural ode control for trajectory approximation of continuity equation. *IEEE Control Systems Letters*, 6:3152–3157, 2022.

[GCB+18] Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*, 2018.

[GRRB24] Borjan Geshkovski, Philippe Rigollet, and Domènec Ruiz-Balet. Measure-to-measure interpolation using transformers. *arXiv preprint arXiv:2411.04551*, 2024.

[GRVE22] Marylou Gabrié, Grant M Rotskoff, and Eric Vanden-Eijnden. Adaptive monte carlo augmented with normalizing flows. *Proceedings of the National Academy of Sciences*, 119(10):e2109420119, 2022.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[JCP23] Yiheng Jiang, Sinho Chewi, and Aram-Alexandre Pooladian. Algorithms for mean-field variational inference via polyhedral optimization in the wasserstein space. *arXiv preprint arXiv:2312.02849*, 2023.

[KPB20] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979, 2020.

[LCBH+22] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[LLS22] Qianxiao Li, Ting Lin, and Zuowei Shen. Deep learning via dynamical systems: An approximation perspective. *Journal of the European Mathematical Society*, 25(5):1671–1709, 2022.

[MRWZ24] Youssef Marzouk, Zhi Robert Ren, Sven Wang, and Jakob Zech. Distribution learning via neural differential equations: a nonparametric statistical perspective. *Journal of Machine Learning Research*, 25(232):1–61, 2024.

[PBHDE+23] Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky TQ Chen. Multisample flow matching: straightening flows with minibatch couplings. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28100–28127, 2023.

[PNR+21] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

[PW24] Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2024.

[Rag24] Maxim Raginsky. Some Remarks on Controllability of the Liouville Equation. *arXiv preprint arXiv:2404.14683*, 2024.

[RBZ23] Domenec Ruiz-Balet and Enrique Zuazua. Neural ode control for classification, approximation, and transport. *SIAM Review*, 65(3):735–773, 2023.

[RBZ24] Domènec Ruiz-Balet and Enrique Zuazua. Control of neural transport for normalising flows. *Journal de Mathématiques Pures et Appliquées*, 181:58–90, 2024.

[RM15] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[Sas15] Igal Sason. On Reverse Pinsker Inequalities. *arXiv preprint arXiv:1503.07118*, 2015.

[Sca23] Alessandro Scagliotti. Deep learning approximation of diffeomorphisms via linear-control systems. *Mathematical Control and Related Fields*, 13(3):1226–1257, 2023.

[SF23] Alessandro Scagliotti and Sara Farinelli. Normalizing flows as approximations of optimal transport maps via linear-control neural odes. *arXiv preprint arXiv:2311.01404*, 2023.

[SV16] Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

[TG22] Paulo Tabuada and Bahman Gharesifard. Universal approximation power of deep residual neural networks through the lens of control. *IEEE Transactions on Automatic Control*, 2022.

[TT13] Esteban G Tabak and Cristina V Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

[TVE10] Esteban G. Tabak and Eric Vanden-Eijnden. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217 – 233, 2010.

[TW16] Jakub M Tomczak and Max Welling. Improving variational auto-encoders using Householder flow. *arXiv preprint arXiv:1611.09630*, 2016.

[VEH14] Tim Van Erven and Peter Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[Ver14] Sergio Verdú. Total variation distance and the distribution of relative information. In *2014 information theory and applications workshop (ITA)*, pages 1–3. IEEE, 2014.

[WJ08] Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Antonio Álvarez-López
Departamento de Matemáticas
Universidad Autónoma de Madrid
C/ Francisco Tomás y Valiente 7,
28049 Madrid, Spain
e-mail: antonio.alvarezl@uam.es

Borjan Geshkovski
Inria & Laboratoire Jacques-Louis Lions
Sorbonne Université
4 Place Jussieu
75005 Paris, France
e-mail: borjan.geshkovski@inria.fr

Domènec Ruiz-Balet
CEREMADE, UMR CNRS 7534
Université Paris-Dauphine, Université PSL
Pl. du Maréchal de Lattre de Tassigny
75016 Paris, France
e-mail: domenec.ruiz-i-balet@dauphine.psl.eu