# Score Operator Newton Transport

**Nisha Chandramoorthy**
Georgia Tech

**Florian Schäfer**
Georgia Tech

**Youssef Marzouk**
MIT

## Abstract

We propose a new approach for sampling and Bayesian computation that uses the score of the target distribution to construct a transport from a given reference distribution to the target. Our approach is an infinite-dimensional Newton method, involving an elliptic PDE, for finding a zero of a "score-residual" operator. We prove sufficient conditions for convergence to a valid transport map. Our Newton iterates can be computed by exploiting fast solvers for elliptic PDEs, resulting in new algorithms for Bayesian inference and other sampling tasks. We identify elementary settings where score operator Newton transport achieves fast convergence while avoiding mode collapse.

## 1 INTRODUCTION

Generating samples from a complex (e.g., non-Gaussian, high-dimensional) probability distribution is a core computational challenge in diverse applications, ranging from computational statistics and machine learning to molecular simulation. A recurring setting is where the density $\rho$ of the target distribution is specified up to a normalizing constant—for example, in Bayesian modeling, where $\rho$ represents the posterior density. Here, evaluations of the *score* $\nabla \log \rho$ are often available as well, even for complex statistical models (Villa et al., 2021). Alternatively, many new methods enable effective score estimation from data, without explicit density estimation; examples include score estimation from time series observations in chaotic dynamical systems (Chandramoorthy and Wang, 2022; Ni, 2020) and score-based modeling of image distributions (Song et al., 2020b,a).

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. 

In these settings, transport or "flow"-driven algorithms for generating samples have seen extensive success. The central idea is to construct a transport map from a simple, prescribed source distribution to the target distribution of interest. One class of transport approaches, e.g., as represented by variational inference with normalizing flows, involves constructing a *parametric* class of invertible maps and minimizing some statistical divergence between the pushforward (see Section 2) of the source by a member of this class and the target. A different, essentially nonparametric, class of transport approaches are based on particle systems, e.g., Stein variational gradient descent (SVGD) (Liu and Wang, 2016) and its many variants (Li et al., 2020a; Chen and Ghattas, 2020; Detommaso et al., 2018). These methods can be interpreted as gradient flows (Jordan et al., 1998) of some functional on the space of probability measures, for different choices of geometry (Chewi et al., 2020; Duncan et al., 2019). Such methods yield implicit representations of transport maps, through the paths taken by sample trajectories (Han and Liu, 2017). Yet another class of transport methods involve *prescribing* continuous paths (Masrani et al., 2021; Albergo et al., 2023) between the source and target distributions, and approximating these paths with particle systems or learned velocity fields.

Of course, none of these approaches is without drawbacks. Parametric representations of transport maps often involve ad hoc choices of parametric class, where bias must be balanced against complexity of the representation; moreover, the optimization problems that must be solved over such classes seldom have guarantees. On the other hand, gradient flow approaches, as exemplified by SVGD or more generally standard Langevin dynamics, may be slow to converge and quite sensitive to the geometry and dimensionality of the target distribution. Because the transport map is not explicitly available in these approaches, it is generally difficult to update the map, e.g., for perturbed scores, or to reuse the map for downstream computations. Continuous-time "homotopy" approaches require *a priori* selection of a path through the space of probability measures, and may involve solving equa-

tions that depend explicitly on estimates of the density at the current time (Reich, 2011) or otherwise resorting to cruder approximations (Iglesias and Yang, 2021).

This paper introduces a new sampling approach based on different ingredients: an infinite-dimensional score matching principle and discrete-time dynamics. Specifically, we construct a transport map as the zero of a *score residual* operator via an infinite-dimensional *Newton method* for root finding, which is typically called the Newton–Raphson method in finite dimensions. The transport is a *composition* of maps found, at each step, via solution of a linear elliptic partial differential equation (PDE). We harness regularity theory for elliptic PDEs to prove existence of such a map and to prove convergence of the iterations. The resulting *score-operator Newton* (SCONE) transport construction is illustrated in Figure 1. It applies to any sampling problem where the scores of the source and target measures can be evaluated. Several desirable features of our approach are as follows:

- Newton methods are *efficient*: we will show, empirically and in simple analytical examples (see Appendix B), that very few iterations may be required. The Newton construction also permits an existing map to be updated or fine-tuned, e.g., for perturbed scores; this is useful for applications such as Bayesian filtering.
- Unlike the nonlinear Monge-Ampère equation, which describes optimal transport maps, our construction involves a sequence of *linear* PDEs, which are more amenable to analysis, fast computation, and dimension reduction.
- Elliptic differential operators instantaneously propagate information throughout the domain. Hence, our transport updates, which use elliptic PDE solutions, are intrinsically *global* (Evans, 2022). As evidenced by our numerical results, our construction thus tends to avoid mode collapse, since transport updates are influenced by score values over the entire support of the distributions, including the tails.

We summarize our main contributions as follows: We define a score transformation operator that maps an input score and a transport map to the transported score. We prove the existence of transport maps that are fixed points of an operator based on the score-transformation operator and the target score. Our existence proof is constructive and leads to a Newton method on Banach spaces. Our construction yields a transport map and defines a new notion of score-matching in infinite dimensions. Convergence of transport maps, and scores, is established in classical Hölder

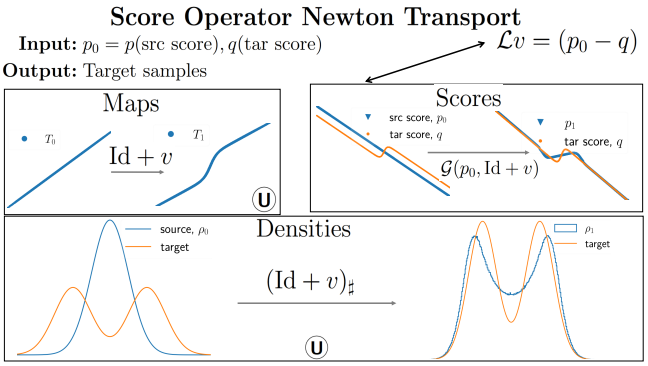


Figure 1: A graphical overview of the construction of transport maps. The *score* of the source and target densities are given as inputs. The method outputs samples from the target distribution. Each iteration involves solving an elliptic PDE that gradually transports the score of the source to that of the target. The PDE solutions across iterations are combined via a simple composition operation to obtain the transport map from the source to the target.

norms, unlike in dual norms typically used for variational inference methods.

In this paper, we establish the theoretical foundations of score operator Newton transport and provide proof-of-concept numerics. Our construction and theory are developed in the infinite-dimensional setting, enabling flexible representations of the transport map. Hence, the construction facilitates the development of many new sampling algorithms, based on kernel methods, deep neural networks, and other discretizations of the underlying linear elliptic PDEs. Developing such scalable algorithms for the Newton updates will be a subject of our future work.

## 2 INFINITE-DIMENSIONAL SCORE MATCHING

Suppose $\nu$ is our unknown target probability measure on $\mathbb{R}^d$ with associated density $\rho^\nu$. We define the *score* of the target to be the vector-valued function $q := \nabla \log \rho^\nu : \mathbb{R}^d \to \mathbb{R}^d$. In our setting, the target score $q(x)$ is available at every $x \in \mathbb{R}^d$.

Let $\mu$ be a source or reference probability measure on $\mathbb{R}^d$ with density $\rho^\mu$. The source is chosen to be easy to sample from, e.g., a Gaussian in $\mathbb{R}^d$. Its vector-valued score function is defined as $p := \nabla \log \rho^\mu$. Here $\nabla$ is the gradient operator on Euclidean space. The main contribution of this work is a new transport map to sample from the target $\nu$ using the source samples

from $\mu$ and the source and target scores,

$$p(x) := \nabla \log \rho^{\mu}(x), \quad q(x) := \nabla \log \rho^{\nu}(x).$$

Our transport map is defined as the solution of an infinite-dimensional root finding problem for the score operator. Next we define both these objects: transport maps and the score operator.

**Transport maps:** Given two probability measures $\mu$ and $\nu$ on $\mathbb{R}^d$, we say that a measurable function $T : \mathbb{R}^d \to \mathbb{R}^d$ is a *transport map* from $\mu$ to $\nu$ if

$$\nu(A) = (T_\sharp \mu)(A) := \mu(T^{-1}(A)), \tag{1}$$

for every measurable set $A$; here $\sharp$ denotes the pushforward operation. In high-dimensional inference problems, the target $\nu$ that we wish to sample from is a often a complicated "intractable" measure. Our goal is to find an invertible map $T$ that transforms samples of $\mu$ to samples distributed according to $\nu$. Recall that samples from the source $\mu$ are easily obtained.

In general, the transport map $T$ between the source and target measures is not unique. The optimal transport map is one useful and canonical choice, but in the context of Bayesian inference—where our main goal is to simulate $\nu$—it suffices to find any transport $T$ that is computationally feasible. Now we provide a constructive approximation of a map $T$ that satisfies (1) by exploiting the scores, $p = \nabla \log \rho^{\mu}$ and $q = \nabla \log \rho^{\nu}$, associated with the source $\mu$ and target $\nu$, respectively. To do this, we define a *score operator* on the product space of functions representing scores and transport maps.

**Score operator:** We define a *score operator*, denoted by $\mathcal{G}$, that takes as arguments a source score and an invertible transport map and returns the score of the resulting pushforward distribution. That is, if $T$ is an invertible map such that $T_\sharp \mu = \nu$, the operator $\mathcal{G}$ is defined such that

$$\mathcal{G}(p, T) = q. \tag{2}$$

Recall the change of variables formula for probability densities,

$$\rho^{\nu} = \frac{\rho^{\mu} \circ T^{-1}}{|\det \nabla T| \circ T^{-1}}. \tag{3}$$

Taking logarithms and differentiating the above formula, we obtain a definition for the score operator $\mathcal{G}$,

$$\mathcal{G}(s, U) = \left(s(\nabla U)^{-1} - \nabla \log |\det \nabla U| (\nabla U)^{-1}\right) \circ U^{-1} \tag{4}$$

$$= \Big(s\,(\nabla U)^{-1} - \mathrm{tr}\big((\nabla U)^{-1} \nabla^2 U\big)(\nabla U)^{-1}\Big) \circ U^{-1}, \tag{5}$$

where (5) follows from using Jacobi's formula for the derivative of the determinant. We use $\mathrm{tr}((\nabla U)^{-1}\nabla^2 U)$ for the vector-valued function $[\mathrm{tr}((\nabla U)^{-1}\partial_1 \nabla U), \ldots, \mathrm{tr}((\nabla U)^{-1}\partial_d \nabla U)]^\top$, where $\partial_i$ is the partial derivative in the $i$th coordinate. The above operator takes as arguments a score $s$ associated with a probability measure, say $\pi$, and a $\mathcal{C}^2$-diffeomorphism $U$, to return the score associated with the measure $U_\sharp \pi$. That is, $\mathcal{G}$ expresses the change of variables, or the pushforward operation, on the space of scores. We end this section by listing some properties of $\mathcal{G}$ that will be useful in the sections to follow and can be checked by using the above definitions.

(i) $\mathcal{G}(s, \mathrm{Id}) = s$ for any score $s$, expressing the fact that the identity coupling results in the same probability measure;

(ii) $\mathcal{G}(s, \mathrm{Id} + c) = s \circ (\mathrm{Id} - c)$, where $c$ is a constant function;

(iii) $\mathcal{G}$ has the group property, mimicking the pushforward operator. That is,

$$\mathcal{G}(s, \psi_1 \circ \psi_2) = \mathcal{G}(\mathcal{G}(s, \psi_2), \psi_1).$$

Note that the operator $\mathcal{G}$ is not injective in the second argument, and hence not invertible. That is, $\mathcal{G}(s, U_1) = \mathcal{G}(s, U_2)$ does not imply that $U_1 = U_2$. Any solution $T$ to (2) is a valid transport map from $\mu$ to $\nu$. We refer to the problem of finding a $T$ that satisfies (2) as the infinite-dimensional score-matching problem. This is because the solution $T$ is a function, and the score of the pushforward distribution through $T$ matches the target score $q$. Despite its name, our problem is not derived as an infinite-dimensional version of the score-matching problem from (Hyvärinen and Dayan, 2005), which is essentially a density estimation problem. Here our objective is to obtain a transport map given target scores and to use it for sampling. In the next section, we will define a score-residual operator that maps a score and a transport map to the difference between the target score and the score of the pushforward of the source by the transport map. We will then derive a solution to the score-matching problem as a zero of this operator; thus, our solution strategy for transport also deviates from the variational problem involving the typical score-matching objective in the literature (Hyvärinen and Dayan, 2005; Song and Ermon, 2019; Song et al., 2020b; Wibisono and Yang, 2022).

**Remark 1** (Availability of scores)**.** *As mentioned in the introduction, the target score is available in settings beyond Bayesian inference with a tractable likelihood and prior. In other words, in many settings, scores can be approximated well even though an explicit model for*

*the unnormalized density is not available. One such example is the score of an ergodic, invariant measure of certain classes of chaotic systems. Here, fast methods have been developed to evaluate scores at any point on a chaotic orbit (Chandramoorthy and Wang, 2022; Ni, 2020).*

**Remark 2** (Validity of transport maps)**.** *Note that any diffeomorphism that satisfies $\mathcal{G}(p,T) = q$ is a valid transport map in the sense that $T_\sharp \mu = \nu$. By integrating both sides of (5), we obtain that any $T$ that satisfies $\mathcal{G}(p,T) = q$ also satisfies, $\rho^\nu = k(\rho^\mu/|\det \nabla T|) \circ T^{-1}$, for an integration constant $k$. Since the left hand side is a valid density (integrates to 1), we obtain that $k$ must be 1.*

# 3 LEARNING A ZERO OF THE SCORE-RESIDUAL OPERATOR

Fixing $p$ and $q$, we define *the score-residual operator* on the space of $\mathcal{C}^2$-diffeomorphisms on $\mathbb{R}^d$ as

$$\mathcal{R}(T) \coloneqq \mathcal{G}(p,T) - q.$$

A zero of this operator is a transport map between the measures associated with $p$ and $q$. Here we describe an iterative approach for finding a zero of this operator, which is a generalization of Newton's method to infinite-dimensional spaces.

## 3.1 Score operator Newton (SCONE) method

Infinite-dimensional generalizations of the Newton method appear in the analysis of PDEs under the name of Nash–Moser iteration (Berti and Bolle, 2015). They also appear in the context of finding *conjugacies* between nearby dynamical systems, in an approach called the Kolmogorov–Arnold–Moser or KAM method (Moser, 1961). In the next section, we derive sufficient conditions for the convergence of this method.

To develop the SCONE iteration, we expand the score operator, supposing that $p$ is close to $q$ and assuming a linear structure on function space near $(q, \mathrm{Id})$,

$$\begin{aligned}\mathcal{G}(p,T) = \mathcal{G}(q,\mathrm{Id}) + D_1\mathcal{G}(q,\mathrm{Id})(p-q) + \\ D_2\mathcal{G}(q,\mathrm{Id})(T-\mathrm{Id}) + \Delta(p,T),\end{aligned} \tag{6}$$

where Id is the identity function on $\mathbb{R}^d$, $D_1, D_2$ are first-order partial derivatives (Frechét derivatives) of $\mathcal{G}$ in its first and second arguments respectively, and $\Delta(p,T)$ contains nonlinear operators on $p-q$ and $T-\mathrm{Id}$. Analogous to the elementary Newton method, to find the solution to the score-matching problem, we first look for a solution to the *linearized* score-matching problem. That is, we find such a $T$ for which the left hand side of (6) is $q$ and $\Delta(p,T) = 0$. This linearization of the score-matching problem yields,

$$-D_1\mathcal{G}(q,\mathrm{Id})\,(p-q) = D_2\mathcal{G}(q,\mathrm{Id})v, \tag{7}$$

defining the vector field

$$v \coloneqq T - \mathrm{Id}.$$

Since $\mathcal{G}(p,\mathrm{Id}) = p$ for all $p$, we have that $D_1\mathcal{G}(q,\mathrm{Id}) = \mathrm{Id}$. We can explicitly compute the linear operator $D_2\mathcal{G}(q,\mathrm{Id})$ to be the following differential operator, which, for convenience, we define as $\mathcal{L}(q)$,

$$-D_2\mathcal{G}(q,\mathrm{Id})\,v = \mathcal{L}(q)\,v \coloneqq \nabla q\, v + q\,\nabla v + \mathrm{tr}(\nabla^2 v). \tag{8}$$

In the above, the trace $\mathrm{tr}(\nabla^2 v)$ of the tensor $\nabla^2 v$ is a row vector with the $i$th column being $\mathrm{tr}(\partial_i \nabla v)$. Using this, we obtain that $v$ satisfies,

$$(p-q) = \mathcal{L}(q)\,v. \tag{9}$$

We may then iterate this update in the following way. Assuming that $\mathcal{L}(q)$ is invertible, we obtain the solution $v = (\mathcal{L}(q))^{-1}(p-q)$. Using $T = \mathrm{Id} + v$, we obtain $p_1 = \mathcal{G}(p,T)$, and then, we repeat the update (9) with $p_1$ replacing $p$, and solve for $v_1$. We then update the transport map approximation as $T_1 = (\mathrm{Id} + v_1) \circ (\mathrm{Id} + v)$. Proceeding further, at the $n$th iteration, we obtain the function $v_n$ by solving

$$-(q - p_n) = \mathcal{L}(q)v_n = (\nabla q)v_n + q_n(\nabla v_n) + \mathrm{tr}(\nabla^2 v_n). \tag{10}$$

Then, we set,

$$\begin{aligned}T_{n+1} &\leftarrow (\mathrm{Id} + v_n) \circ T_n \\ p_{n+1} &\leftarrow \mathcal{G}(p_n, \mathrm{Id} + v_n).\end{aligned} \tag{11}$$

In the same spirit as the KAM method from the dynamical systems literature, notice that the differential operator $\mathcal{L}(q)$ remains the same across steps, (10). In the next section, we will give sufficient conditions under which the sequence of functions $(T_n)_{n\geq 0}$ converges, in a classical Hölder norm, to a transport map $T$, i.e., an invertible map that satisfies $T_\sharp \mu = \nu$.

## 3.2 SCONE transport algorithm

The steps of the infinite-dimensional Newton algorithm derived above are summarized in Algorithm 1. The input to the algorithm are black-box functions that return the source score $p_0 = p$ and $q$, the target score. In addition, we have $m$ iid samples from the source distribution, say, $\{x_i\}_{i=1}^m$. After $n$ steps of the algorithm, these samples are transformed to $\{T_n(x_i)\}_{i=1}^m$, which represent target samples more accurately as $n$ increases. The algorithm can also return

**Algorithm 1** Score-operator Newton Transport

$T_0(x) = x$, $p_0(x) = \nabla \log \rho^{\mu}(x)$, $x_1, x_2, \cdots, x_m \sim \mu$
**while** $n \leq n_{\max}$ **do**
  $v_n \leftarrow \mathcal{L}(q)^{-1}(p_n - q)$
  $x_i \leftarrow x_i + v_n(x_i)$
  $p_{n+1}(x) \leftarrow \mathcal{G}(p_n, \mathrm{Id} + v_n)(x)$
  $n \to n + 1$
**end while**
Return $\{x_i\}_{1 \leq i \leq m}$

the transport map $T$ as a black-box function that can be evaluated at any point.

In the beginning, we set the initial guess for the transport map $T_0$ (e.g., $T_0(x) = \mathrm{Id}(x) = x$) and $p_0 = p$, the source score. At each iteration, we solve the PDE (10) to obtain the vector field $v_n$. There is extensive literature on approximating solutions of PDEs using neural networks, such as physics-informed neural networks (Raissi et al., 2019), deep Ritz methods (Lu et al., 2021b; Yu et al., 2018) and Fourier neural operators (Li et al., 2020b); as well as kernel methods and Gaussian process approximations (Owhadi and Scovel, 2019; Wendland, 2004; Schaback and Wendland, 2006; Zhang et al., 2000), and finite elements (Ciarlet, 2002; Wang and Ye, 2013). Bespoke methods that exploit the particular structure of $\mathcal{L}(q)$ are deferred to future work (see Section 6).

Using meshfree methods such as PINNs or FNOs, we obtain a black-box solution $v_n$ that can be evaluated at any point. Moreover, these black-box solutions, being neural networks, can also be automatically differentiated. Thus, we can evaluate $v_n(x_i), \nabla v_n(x_i)$ and $\nabla^2 v_n(x_i)$, where $\{x_i\}_{i=1}^m$ are the samples at the $n$th iteration. When we use grid-based methods, we obtain (approximate) evaluations of $v_n$ at the grid points. Then, we use interpolations and finite-differences to obtain $v_n(x_i)$ and its derivatives. Using (5), we update the source score $p_n$ as $p_{n+1} = \mathcal{G}(p_n, \mathrm{Id} + v_n)$. The samples are also transformed as $x_i \to x_i + v_n(x_i)$. Note that this results, at the end of the $n$th iteration, in the samples $\{T_n(x_i)\}$, where $T_n := (\mathrm{Id} + v_n) \circ T_{n-1}$. We repeat this process until the algorithm converges, i.e., $\|v_n\|$ is close to zero or if the maximum number of iterations is reached.

**SCONE complexity:** The computational cost of our method is $\mathcal{O}(n\,(C_{\mathrm{pde}} + C_{\mathrm{u}}))$, where $n$ is the number of Newton iterations, $C_{\mathrm{pde}}$ is the cost of the solving the PDE and $C_{\mathrm{u}}$ is the cost of the update step (11). Typically Newton methods exhibit quadratic convergence (Galántai, 2000), which can be accelerated further for finite-dimensional problems under local smoothness conditions (Gerlach, 1994) (see Section 6). The cost of solving the PDE dominates the per iteration cost and naïve methods typically have a computational complexity that scales exponentially with the problem dimension ($d$). However, in the context of elliptic PDEs, various sparse structures in the solution have been exploited to mitigate the curse of dimensionality, including tensor decompositions (Dahmen et al., 2016), hierarchical low-rank approximations (Boullé and Townsend, 2023), and other notions of model reduction (see Section 6).

### 3.3 Related work

In many variational methods for sampling and Bayesian inference, one seeks transport maps that minimize a divergence or distance functional on a space of probability measures, over a parametric class of maps $\mathcal{U}$. As an example, one could define a score-based distance metric, and seek a $T$ such that

$$T = \arg\min_{U \in \mathcal{U}} \int \|q(x) - \mathcal{G}(p, U)(x)\|^2 \, d\nu(x). \qquad (12)$$

Candidate maps $U \in \mathcal{U}$ are parameterized and an empirical minimization problem for the parameters is solved using a method that is appropriate for the distance functional. Common classes of parametric maps include normalizing flows (Papamakarios et al., 2021), the flows of neural ODEs (Chen et al., 2018), and gradients of input-convex neural networks (Huang et al., 2020). An alternative class of approaches produces *implicit* representations of transport maps through paths taken by sample trajectories of a deterministic or stochastic dynamics. Beginning with the classical work of Jordan et al. (1998), these dynamics are derived such that their mean field limits are gradient flows of the distance functional on the space of probability measures, for appropriate choices of objective and geometry (Chewi et al., 2020; Duncan et al., 2019; Han and Liu, 2017). Methods that fall into this class include SVGD and Langevin dynamics. (See Wibisono (2018) for an overview of the connection between sampling methods and optimization of distance functionals on probability spaces.)

Variational autoencoders (Kingma et al., 2019) or GANs (Goodfellow et al., 2020) also transport a low-dimensional source measure to a target, learning parameterized decoders or generators by minimizing a variety of objectives. These methods have no obvious dynamical interpretation; their stable training and obtaining theoretical guarantees are challenging.

The SCONE method exploits scores but does not make use of parametric spaces to define transport maps. Rather than minimizing a distance functional, we derive a generalized Newton–Raphson method on Banach spaces to construct a zero of the operator $\mathcal{R}(T)$. This approach specifies the transport map

as a composition of functions that is achieved in the limit of a discrete-time dynamical system—as opposed to a continuous-time flow—on function spaces. Our SCONE transport defines, correspondingly, a discrete-time dynamical system on the space of probability densities, with the target being a fixed point.

## 4 CONVERGENCE PROOF

Here we prove the convergence of the SCONE algorithm in 1, establishing a new construction of a transport map. We do not compute explicit bounds on the error norms, $\|p_n - q\|$. We invoke elliptic regularity theory to establish convergence. Our Newton iterates (10) yield second-order linear, elliptic PDEs, as we describe below. Let $M$ be a compact subset of $\mathbb{R}^d$ and $\Omega$ be an open set containing $M$. At each step, we solve $d$ second-order PDEs of the following form,

$$(L(x, D)v)_i = f_i,\ 1 \le i \le d, \tag{13}$$

where the linear differential operator $L$ is given by the $d \times d$ matrix with

$$L(x, D)_{ij} = \sum_{|\alpha| \le 2} a_\alpha^{ij}(x)\, D^\alpha.$$

Here, $\alpha = (\alpha_1, \cdots, \alpha_d)$ is a multi-index, $D^\alpha = \partial_1^{\alpha_1} \cdots \partial_d^{\alpha_d}$, such that $|\alpha| = \alpha_1 + \cdots + \alpha_d \le 2$. Suppose we parameterize the solutions $v_n$ of the system as the gradient $\nabla\phi_n$ of some differentiable function $\phi_n : \Omega \to \mathbb{R}$. Then, substituting into (10), we obtain the following simpler scalar equation for $\phi_n$,

$$\nabla(\nabla \cdot \nabla\phi_n) + \nabla(\nabla\phi_n \cdot q) = p_n - q.$$

Integrating this equation, we find that our Newton iterates, when $v_n = \nabla\phi_n$ satisfies,

$$\nabla \cdot \nabla\phi_n + \nabla\phi_n \cdot q = \log(\rho_n^\mu/\rho^\nu) + C, \tag{14}$$

where $C$ is an integration constant and $\rho_n^\mu$ is the density of the pushforward measure, $T_{n\sharp}\mu$. The operator, $\mathcal{P} = \nabla.\nabla + q \cdot \nabla$, is an *elliptic* operator, which ensures the well-posedness of solutions to (14). We recall the ellipticity condition of an operator, which says that the highest-order (principal) symbol is coercive (see e.g., the textbook of Hörmander (1963) or Dyatlov (2022) for notes and Gilbarg and Trudinger (1977) for Schauder estimates).

**Definition 1** (Hölder space of order $k$ and exponent $\gamma$)**.** *The Hölder space of order $k$ and exponent $\gamma$, denoted $\mathcal{C}^{k,\gamma}(\Omega)$, is a Banach space consisting of functions that have continuous derivatives up to order $k$ and $\gamma$-Hölder continuous $k$th order derivatives. It is complete with respect to the following norm, for $k \ge 1$, $f : \Omega \to \mathbb{R}$,*

$$\|f\|_{k,\gamma} := \|f\|_k + \max_{|\alpha|=k} \sup_{x\in\Omega} \|D^\alpha f\|_{0,\gamma}, \tag{15}$$

*where,*

$$\|f\|_k := \max_{|\alpha|\le k} \sup_{x\in\Omega} |D^\alpha f(x)| \tag{16}$$

$$\|f\|_{0,\gamma} := \|f\|_0 + \sup_{x,y\in\Omega, x\neq y} \frac{|f(x) - f(y)|}{|x - y|^\gamma}. \tag{17}$$

*We define Hölder continuous cotangent vector fields of the form $v = \nabla\phi \in D^*\bar{\Omega}$ for some differentiable function $\phi$, where $D^*\bar{\Omega}$ is used to denote the cotangent bundle (rather than the usual $T^*\bar{\Omega}$, to avoid confusion with transport maps $T$). If $\phi \in \mathcal{C}^{k,\gamma}(\Omega)$, then, we have that the components of $v$ (interpreted as scalar functions) are $\mathcal{C}^{k-1,\gamma}(\Omega)$. Note that, due to the (compact) embedding of Hölder spaces, $\mathcal{C}^{0,\gamma} \to \mathcal{C}^{0,\delta}$ for $\delta < \gamma$, $f \in (\mathcal{C}^{k,\gamma}(\Omega))^d$ with components $f_i \in \mathcal{C}^{0,\gamma_i}(\Omega)$ implies that $\gamma \le \min_i \gamma_i$. In a slight abuse of notation, for a vector-valued function $f \in (\mathcal{C}^{k,\gamma}(\Omega))^d$, including cotangent vector fields, we write,*

$$\|f\|_{k,\gamma}^2 := \sum_{i=1}^d \|f_i\|_{k,\gamma}^2. \tag{18}$$

We use Schauder estimates of the type below (Gilbarg and Trudinger, 1977) from classical elliptic PDE theory.

**Theorem 1.** *Let $\Omega$ be a bounded and open subset of $\mathbb{R}^d$ with a smooth boundary. Let $L(x, D)u = f$ be a second-order strongly elliptic system, with $L(x, D) = \sum_{|\alpha|\le 2} a_\alpha(x)D^\alpha$, and zero Dirichlet boundary conditions. If the coefficients $a_\alpha$ and the right hand side $f$ are in $\mathcal{C}^{s,\gamma}(\bar{\Omega})$, then, $u \in \mathcal{C}^{s+2,\gamma}(\bar{\Omega})$. In particular, for any $s \ge 0$, and $\gamma \in (0, 1)$,*

$$\|u\|_{s+2,\gamma} \le K(\|f\|_{s,\gamma} + \|u\|_s),$$

*where $K$ only depends on $\|a_\alpha\|_{s,\gamma}$ and $d$.*

**Theorem 2** (Score-matching)**.** *Let $\Omega$ be a bounded, open subset of $\mathbb{R}^d$ containing the origin, with a smooth boundary. Let $q \in \mathcal{C}^{s+1,\cdot}(\bar{\Omega})$ be the score of a target density $\rho^\nu \in \mathcal{C}^{s+2,\cdot}(\bar{\Omega})$. Then, for every $\epsilon > 0, s \in \mathbb{N}$ there exists a $\delta > 0$ such that for any reference density $\rho^\mu$ with associated score $p$ such that $\|p - q\|_s \le \epsilon$, there is a transformation $T \in \mathcal{C}^{s+2,\cdot}(M)$ such that (i) $\mathcal{G}(p, T) = q$ and (ii) $\|T - \mathrm{Id}\|_{s+2} \le \delta$.*

*Proof.* Define

$$\mathcal{H}(v) := \mathcal{G}(q, \mathrm{Id} + v) = \Big( q\,(\mathrm{Id} + \nabla v)^{-1} - \mathrm{tr}((\mathrm{Id} + \nabla v)^{-1}\,\nabla^2 v)\,(\mathrm{Id} + \nabla v)^{-1} \Big) \circ (\mathrm{Id} + v)^{-1}. \tag{19}$$

From definition (19), note that $\mathcal{H}(0) = q$. We can explicitly compute the first derivative of $\mathcal{H}$ at 0 to be

$$d\mathcal{H}(0)w = -\nabla q\, w - q\, \nabla w - \mathrm{tr}(\nabla^2 w). \tag{20}$$

We can also deduce that $d\mathcal{H}(0)\nabla\phi$ is equivalent to $-\nabla\mathcal{P}\phi$, where $\mathcal{P} = \nabla\cdot\nabla + q\cdot\nabla$ is an elliptic operator, which is Fredholm on $\mathcal{C}^{s+2,\gamma}(\bar{\Omega})$. Let $C^{s,\gamma}(\bar{\Omega})$ denote the quotient space of $\mathcal{C}^{s,\gamma}(\bar{\Omega})$ corresponding to the equivalence relation $f \sim g$ if $\nabla f(x) = \nabla g(x)$, at all $x \in \bar{\Omega}$. Correspondingly, we define a closed subspace of $\mathcal{C}^s$ cotangent vector fields, $\mathcal{V}^{s,\gamma} := \{x \to v(x) = \nabla\phi(x) \in D_x^*\bar{\Omega}, \phi \in C^{s+1,\gamma}(\bar{\Omega}), x \in \Omega\}$, with norm $\|\nabla\phi\|_* = \|\nabla\phi\|_{s,\gamma}$ and let $B_\theta^{s+2,\gamma}(0)$ be a $\theta$-ball around 0 in $\mathcal{V}^{s+2,\gamma}$. The element, $\phi$, in a sufficiently small set around the constant element in $C^{s+3,\gamma}(\bar{\Omega})$ can identify an element, $\nabla\phi$, of $B_\theta^{s+2,\gamma}(0)$. We note that the operator $\mathcal{H} : B_\theta^{s+2,\gamma}(0) \to (\mathcal{C}^{s,\gamma}(\bar{\Omega}))^d$ is well-defined as a continuous operator.

When $d\mathcal{H}(0)$ is defined on $\mathcal{V}^{s+2,\gamma}$, and using Theorem 1, we know that its kernel only contains the zero element of $\mathcal{V}^{s+2,\gamma}$, which corresponds to $\mathcal{P}\phi = \text{const} \in C^{s,\gamma}(\bar{\Omega})$. Thus, we obtain that $d\mathcal{H}(0) : \mathcal{V}^{s+2,\gamma} \to \mathcal{V}^{s,\gamma}$ is bijective. In particular, for a fixed $q$, the $s$+2-Hölder norm of $w$ that solves $d\mathcal{H}(0)w = f$ for an $f \in \mathcal{V}^{s,\gamma}$ is bounded above, by Theorem 1. Hence, $d\mathcal{H}(0)^{-1}$ is continuous on $\mathcal{V}^{s,\gamma}$.

Thus, we can apply the inverse function theorem for $\mathcal{H}$. There exists an open neighborhood, $B^{s,\gamma}(q,\epsilon)$, of radius, say $\epsilon$, of $q$ in $\mathcal{V}^{s,\gamma}$ and a continuously differentiable map $\mathcal{I} : B^{s,\gamma}(q,\epsilon) \to \mathcal{V}^{s+2,\gamma}$ so that $\mathcal{I}\circ\mathcal{H}(v) = v$. Thus, for any $p = \mathcal{H}(v) \in B_{s,\gamma}(q,\epsilon)$, the map $T = (\mathrm{Id} + v)^{-1}$ is such that $\mathcal{G}(p,T) = q$. This proves $(i)$. From the continuity of $\mathcal{I}$ at $q$ in $B_{s,\gamma}(q,\epsilon)$, we can choose a $\delta_0 > 0$ such that $\|(T - \mathrm{Id})^{-1}\|_{s+2,\gamma} \le \delta_0$. This implies that for some $\delta > 0$, $\|v\|_{s+2,\gamma} \le \delta$, hence proving (ii). □

The above is an existence result for a transport map and is established via the inverse function theorem for the score operator. It is important to note that even though the result is local (that is, for nearby probability densities), uniqueness cannot still be established for the transport map. This is because, in the above proof, the operator $d\mathcal{H}(0)$ has a non-empty kernel on function spaces containing functions of the form, $\nabla\phi$, for some $\phi \in \mathcal{C}^{s+2,\gamma}(\bar{\Omega})$. For any function $\phi$ such that $\mathcal{P}\phi = f + \text{const}$, $v = \nabla\phi$ solves $d\mathcal{H}(0)v = \nabla f$, and therefore $d\mathcal{H}(0)$ is not injective. In other words, we have not shown that there is only one transport map between a given source and target density, even when they are close to each other. We have defined quotient spaces on which to define $d\mathcal{H}(0)$ to make it invertible, using isomorphism theorems for vector spaces.

Proving the inverse function theorem via the Banach fixed point theorem both establishes the existence of the desired map $T$ and also the means to construct $T$ as a fixed point iteration of the contraction map. In the theorem below, we explicitly define such a contraction map whose fixed point is $T$. Further, the fixed point iteration of the map is equivalent to our SCONE iteration. Such an interpretation of the Newton–Raphson method as a fixed point iteration of the linearization of the given map is indeed classical in numerical analysis, when we are interested in finding zeros of a function on a finite-dimensional space. The following theorem extends this idea to infinite-dimensional spaces and serves as the convergence proof of the SCONE method.

**Theorem 3** (SCONE construction of transport)**.** *When $q \in (\mathcal{C}^{s,\gamma}(\bar{\Omega}))^d$, there exists a $\theta > 0$ such that for any $p$ in a $\theta$-neighborhood of $q$, $p_n \to q$ in $(s,\gamma)$-Hölder norm, where,*

$$\begin{aligned} v_n &= \mathcal{L}(q)^{-1}(p_n - q) \\ p_{n+1} &= \mathcal{G}(p_n, \mathrm{Id} + v_n), \quad n \in \mathbb{Z}^+,\ p_0 = p. \end{aligned} \tag{21}$$

*Proof.* Recall the definition of $\mathcal{L}(q)$ from (8). Note that $\mathcal{L}(q)$ is not invertible on $(\mathcal{C}^{s,\gamma}(\bar{\Omega}))^d$, and $v_n$ in the statement of the theorem refers to any cotangent vector field $v_n = \nabla\phi_n$ such that $\mathcal{L}(q)v_n = p_n - q$. We show in the proof of Theorem 2 that $\mathcal{L}(q)^{-1}$ is a homeomorphism between a $\theta$-neighborhood of $q$ in $\mathcal{V}^{s,\gamma}$ (see the proof of 2 for the definition of this space) and a $\mathcal{V}^{s+2,\gamma}$ neighborhood of zero. Thus, choosing $\theta > 0$ sufficiently small, we can combine both the SCONE iteration and update steps ((10) and (11)) to define the operator,

$$\mathcal{J}(v) = \mathcal{L}^{-1}(\mathcal{G}(q + \mathcal{L}v, \mathrm{Id} + v) - q), \tag{22}$$

where we write $\mathcal{L}$ to indicate $\mathcal{L}(q)$ for a fixed target score $q$. It is clear that $\mathcal{J}(0) = 0$ and hence 0 is a fixed point of $\mathcal{J}$. The operator $\mathcal{J}$ is smooth at 0. In particular, through direct computation, we can verify that its first derivative, $d\mathcal{J}(0) = 0$ as an operator from $B_\theta^{s+2,\gamma}(0)$ to $\mathcal{V}^{s+2,\gamma}$. Applying the mean value theorem, we get,

$$\|\mathcal{J}(v_1) - \mathcal{J}(v_2)\| \le \|v_1 - v_2\| \sup_{\eta\in(0,1)} \|d\mathcal{J}(\eta v_1 + (1-\eta)v_2)\|. \tag{23}$$

Since $d\mathcal{J}(0) = 0$ and $d\mathcal{J}$ is continuous, we can choose a $\delta < \theta$ such that for all $v \in B_\delta^{s+2,\gamma}(0)$, $\|d\mathcal{J}(v)\| \le (1/2)$. Thus, we obtain that $\mathcal{J}$ is a contraction on $B_\delta^{s+2,\gamma}(0)$. Thus, the fixed point iteration of $\mathcal{J}$, i.e., $v_{n+1} = \mathcal{J}(v_n)$, starting from any $v_0 \in B_\delta^{s+2,\gamma}(0)$ converges to 0. Note that $\|v_0\| = \|\mathcal{L}^{-1}(p_0 - q)\| \le C(p_0 - q)$, from the continuity of $\mathcal{L}^{-1}$. Thus, if $p_0 \in B_\epsilon(q)$, one can choose $\delta_0 := \min\{\delta, C\epsilon\}$ such

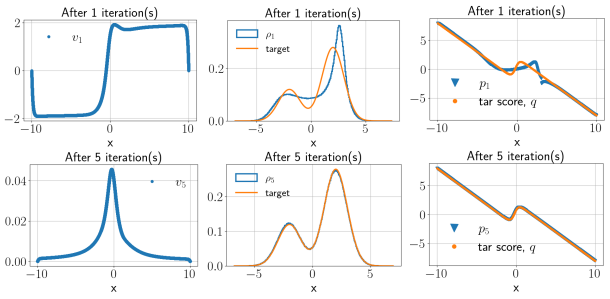


Figure 2: Left: the solution $v_n$ after 1 (top) and 5 (bottom) iterations computed using second order finite difference with O(500) grid points. Center: the transformed empirical density. Right: $p_1$ (top) and $p_5$ (bottom).

that $\mathcal{J}$ is a contraction on $B^{s+2,\gamma}_{\delta_0}(0)$ and hence, the iteration converges.

Now we show that the convergence of $v_n \to 0$ implies the convergence of $T_n$ to a transport $T$. Recall that $T_k = \circ_{n=0}^{k}(\mathrm{Id}+v_n)$, with the compositions being on the left. Hence, $\|T_n - T_{n-1}\| = \|v_n \circ T_{n-1}\| \leq \|v_n\| \to 0$. Finally, to see that the limit $T := \lim_{n\to\infty} T_n$ is a transport, from (21) for a finite $n$, $p_{n+1} = \mathcal{G}(p_n, \mathrm{Id} + v_n) = \mathcal{G}(p_0, T_n)$, by applying the group property of $\mathcal{G}$ iteratively. Taking the limit $n \to \infty$ on both sides, we obtain $q = \mathcal{G}(p_0, T)$. □

**Remark 3.** *Instead of a Newton method, one can also define a different fixed-point iteration on an operator defined using the score-residual operator (see Appendix A). However, under similar smoothness assumptions, these fixed point iteration methods typically show slower convergence (Smale, 1985).*

## 5 NUMERICAL RESULTS

Here we present proof-of-concept numerical results that demonstrate the convergence of our SCONE transport (Algorithm 1). On 1D domains, we find that the SCONE transport map converges to the monotone map or the increasing rearrangement, which is optimal with respect to any convex cost (see Chapter 2 of Santambrogio (2015)). We also demonstrate that SCONE transport maps can effectively tackle multimodality in the target. See Appendix C for details on the numerical methods and additional experiments.

In Figure 2, we show the results of applying the SCONE algorithm to a bimodal target density of the form $w_1\mathcal{N}(m_1, \sigma_1^2) + w_2\mathcal{N}(m_2, \sigma_2^2)$ (shown in orange in the center column). The target score is shown in orange on the right column. We see from the second row of Figure 2 that the transformed scores and densities match those of the target closely after just

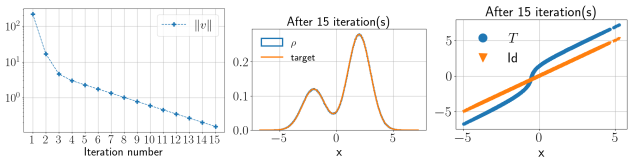


Figure 3: Convergence of SCONE. Left: the convergence of $\|v_n\|$. Center: transformed empirical density after 15 iterations. Right: $T_n$ after 15 iterations.

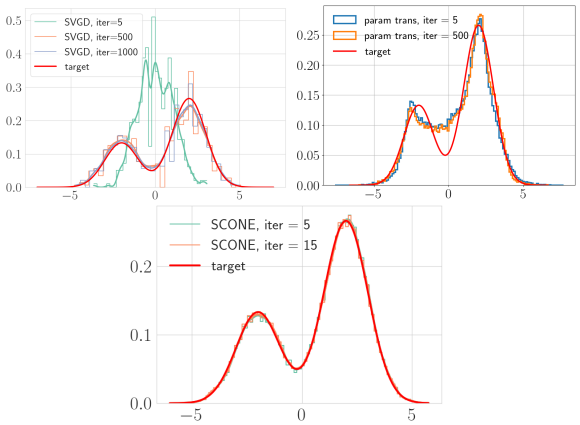


Figure 4: *Left*: SVGD with RBF kernels (median heuristic for bandwidth) and 512 particles. *Center*: parameterized monotone transport map (Parno et al., 2022), with polynomial degree 10, $512^2/11$ samples, optimized using gradient descent + line search. *Right*: SCONE transport with ODE updates solved with 512 grid points.

5 iterations. We observe numerically that solving the SCONE step (10) (which reduces to an ODE in 1D) on a coarser grid (of size $\mathcal{O}(100)$) does not affect fast convergence when we add a small $\ell^2$ regularization. In comparison, SVGD takes $\mathcal{O}(500)$ iterations (Liu and Wang, 2016) with 100 particles for the same problem, and a vanilla GAN implementation (Goodfellow et al., 2020) can lead to unstable training and mode collapse (Thanh-Tung and Tran, 2020; Li et al., 2018). In Figure 3, we show the convergence of the SCONE algorithm to the target. We see that the identified transport map converges to the optimal map (shown in blue in the right column).

**Comparison with other algorithms** In Figure 4, we compare SCONE against SVGD (Liu and Wang, 2016) and parameterized transport maps (Parno et al., 2022) as these are the most widely used classes of algorithms for Bayesian inference. With $N$ samples, an SVGD iteration has $\mathcal{O}(N^2)$ cost. With $G$ grid points, SCONE iterations have $\mathcal{O}(G^2)$ cost, since the linear system to be solved at a SCONE iteration is banded.

With $P$ parameters and $S$ samples to evaluate the variational (KL) objective, the optimization step for a parameterized transport map is $\mathcal{O}(PS)$. We choose $PS = N^2 = G^2 = 512^2$ so that the computational budget *per iteration* remains the same across the different methods. We see in Figure 4 (right) that densities obtained from SCONE most closely match the target (red) after just 5 iterations, while SVGD (left) takes $\mathcal{O}(500)$ iterations to converge to a comparatively worse solution. Parameterized transport (center) converges quickly, with the help of a backtracking line search optimization, but makes a significant approximation error which persists even for higher polynomial degree (we tested up to 20). Figure 4 thus demonstrates that SCONE vastly outperforms SVGD and parameterized transport algorithms for the same computational cost.

## 6 DISCUSSION

We introduce a new notion of infinite-dimensional score-matching that yields a Newton-type method for sampling, and we prove sufficient conditions for its convergence. Our method applies in settings where scores of the source and target measures are easily computed. We comment on theoretical and algorithmic features of this work that will spur further research.

**Learning elliptic PDEs:** Many structure-exploiting, fast, and sample-efficient methods are emerging for learning the solution operators of linear elliptic PDEs; see, e.g., Lu et al. (2021a); Boullé and Townsend (2023); Schäfer and Owhadi (2021). These methods use randomized numerical linear algebra (Boullé and Townsend, 2023), CNN-based encoder-decoder networks (Zhang and Garikipati, 2023), interpolation between deep neural network and Monte Carlo approximations (Nüsken and Richter, 2023), etc. These results suggest that it is possible to develop optimal methods, in terms of computational and sample complexity, for learning our SCONE update operator by exploiting low-rank structure in its solutions. With the same goal, we will also explore particle methods (e.g., from fluid dynamics (Monaghan, 2012; Cottet et al., 2000)), which can also use fast numerical linear algebra and have theoretical guarantees.

**Newton convergence:** Under typical conditions, the classical result of Kantorovich (see Galántai (2000) for a survey) establishes quadratic convergence starting in a ball of sufficiently small radius (as in our Theorem 3). In future work, we will investigate damping and modified SCONE iterations to prevent divergence (Smale, 1985). We will also develop inexact and quasi-Newton variants of SCONE, as a way of further reducing computational cost (Traub and Woźniakowski, 1980) and allowing for errors in $q$. We will derive theoretical convergence guarantees for these modified SCONEs.

**Funding** NC and YMM acknowledge support from NSF grant PHY-2028125. YMM acknowledges support from DOE ASCR award DE-SC0023187, AFOSR award FA9550-20-1-0397, and ONR award N00014-20-1-2595. FS acknowledges support from AFOSR award FA9550-23-1-0668 and ONR award N00014-23-1-2545.

**Acknowledgments** We are greatly indebted to Guillaume Bal for kindly pointing out that the operator $\mathcal{L}$ is not elliptic, as we had assumed in our earlier draft, and generously helping us correct the error. We would also like to thank Daniel Sharp for providing us with an mParT implementation, which we could readily use to generate Figure 4.

### References

Albergo, M. S., Boffi, N. M., and Vanden-Eijnden, E. (2023). Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*.

Berti, M. and Bolle, P. (2015). A Nash-Moser approach to KAM theory. *Hamiltonian partial differential equations and applications*, pages 255–284.

Boullé, N. and Townsend, A. (2023). Learning elliptic partial differential equations with randomized linear algebra. *Foundations of Computational Mathematics*, 23(2):709–739.

Chandramoorthy, N. (2023). https://doi.org/10.5281/zenodo.10307262.

Chandramoorthy, N. and Wang, Q. (2022). Efficient computation of linear response of chaotic attractors with one-dimensional unstable manifolds. *SIAM Journal on Applied Dynamical Systems*, 21(2):735–781.

Chen, P. and Ghattas, O. (2020). Projected Stein variational gradient descent. *Advances in Neural Information Processing Systems*, 33:1947–1958.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. (2018). Neural ordinary differential equations. *Advances in neural information processing systems*, 31.

Chewi, S., Le Gouic, T., Lu, C., Maunu, T., and Rigollet, P. (2020). SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence. *Advances in Neural Information Processing Systems*, 33:2098–2109.

Ciarlet, P. G. (2002). *The finite element method for elliptic problems*. SIAM.

Cottet, G.-H., Koumoutsakos, P. D., et al. (2000). *Vortex methods: theory and practice*, volume 8. Cambridge university press Cambridge.

Dahmen, W., DeVore, R., Grasedyck, L., and Süli, E. (2016). Tensor-sparsity of solutions to high-dimensional elliptic partial differential equations. *Found Comput Math*, 16:813–874.

Detommaso, G., Cui, T., Marzouk, Y., Spantini, A., and Scheichl, R. (2018). A Stein variational Newton method. *Advances in Neural Information Processing Systems*, 31.

Duncan, A., Nüsken, N., and Szpruch, L. (2019). On the geometry of Stein variational gradient descent. *arXiv preprint arXiv:1912.00894*.

Dyatlov, S. (2022). Lecture notes for 18.155: distributions, elliptic regularity, and applications to pdes. https://math.mit.edu/~dyatlov/18.155/155-notes.pdf. [Online; accessed Jan 20th, 2023].

Evans, L. C. (2022). *Partial differential equations*, volume 19. American Mathematical Society.

Galántai, A. (2000). The theory of newton's method. *Journal of Computational and Applied Mathematics*, 124(1):25–44. Numerical Analysis 2000. Vol. IV: Optimization and Nonlinear Equations.

Gerlach, J. (1994). Accelerated convergence in newton's method. *Siam Review*, 36(2):272–276.

Gilbarg, D. and Trudinger, N. S. (1977). *Elliptic partial differential equations of second order*, volume 224. Springer.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Han, J. and Liu, Q. (2017). Stein variational adaptive importance sampling. *arXiv preprint arXiv:1704.05201*.

Hörmander, L. (1963). *Elliptic boundary problems*, pages 242–274. Springer Berlin Heidelberg, Berlin, Heidelberg.

Huang, C.-W., Chen, R. T., Tsirigotis, C., and Courville, A. (2020). Convex potential flows: Universal probability distributions with optimal transport and convex optimization. *arXiv preprint arXiv:2012.05942*.

Hyvärinen, A. and Dayan, P. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4).

Iglesias, M. and Yang, Y. (2021). Adaptive regularisation for ensemble kalman inversion. *Inverse Problems*, 37(2):025008.

Jordan, R., Kinderlehrer, D., and Otto, F. (1998). The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17.

Kingma, D. P., Welling, M., et al. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392.

Li, J., Madry, A., Peebles, J., and Schmidt, L. (2018). On the limitations of first-order approximation in gan dynamics. In *International Conference on Machine Learning*, pages 3005–3013. PMLR.

Li, L., Li, Y., Liu, J.-G., Liu, Z., and Lu, J. (2020a). A stochastic version of Stein variational gradient descent for efficient sampling. *Communications in Applied Mathematics and Computational Science*, 15(1):37–63.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A. (2020b). Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*.

Liu, Q. and Wang, D. (2016). Stein variational gradient descent: A general purpose Bayesian inference algorithm. *Advances in neural information processing systems*, 29.

Lu, Y., Chen, H., Lu, J., Ying, L., and Blanchet, J. (2021a). Machine learning for elliptic pdes: Fast rate generalization bound, neural scaling law and minimax optimality. *arXiv preprint arXiv:2110.06897*.

Lu, Y., Lu, J., and Wang, M. (2021b). A priori generalization analysis of the deep ritz method for solving high dimensional elliptic partial differential equations. In *Conference on learning theory*, pages 3196–3241. PMLR.

Masrani, V., Brekelmans, R., Bui, T., Nielsen, F., Galstyan, A., Ver Steeg, G., and Wood, F. (2021). q-paths: Generalizing the geometric annealing path using power means. In *Uncertainty in Artificial Intelligence*, pages 1938–1947. PMLR.

Monaghan, J. J. (2012). Smoothed particle hydrodynamics and its diverse applications. *Annual Review of Fluid Mechanics*, 44:323–346.

Moser, J. (1961). A new technique for the construction of solutions of nonlinear differential equations. *Proceedings of the National Academy of Sciences*, 47(11):1824–1831.

Ni, A. (2020). Fast linear response algorithm for differentiating chaos.

Nüsken, N. and Richter, L. (2023). Interpolating between bsdes and pinns: deep learning for elliptic and parabolic boundary value problems.

Owhadi, H. and Scovel, C. (2019). *Operator-Adapted Wavelets, Fast Solvers, and Numerical Homogenization: From a Game Theoretic Approach to Numerical Approximation and Algorithm Design*, volume 35. Cambridge University Press.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):2617–2680.

Parno, M., Rubio, P.-B., Sharp, D., Brennan, M., Baptista, R., Bonart, H., and Marzouk, Y. (2022). Mpart: Monotone parameterization toolkit. *Journal of Open Source Software*, 7(80):4843.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707.

Reich, S. (2011). A dynamical systems framework for intermittent data assimilation. *BIT Numerical Mathematics*, 51:235–249.

Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94.

Schaback, R. and Wendland, H. (2006). Kernel techniques: from machine learning to meshless methods. *Acta numerica*, 15:543–639.

Schäfer, F. and Owhadi, H. (2021). Sparse recovery of elliptic solvers from matrix-vector products. *arXiv preprint arXiv:2110.05351*.

Smale, S. (1985). On the efficiency of algorithms of analysis. *Bulletin of the American Mathematical Society*, 13(2):87–121.

Song, J., Meng, C., and Ermon, S. (2020a). Denoising diffusion implicit models. *ArXiv*, abs/2010.02502.

Song, Y. and Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2020b). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Thanh-Tung, H. and Tran, T. (2020). Catastrophic forgetting and mode collapse in gans. In *2020 international joint conference on neural networks (ijcnn)*, pages 1–10. IEEE.

Traub, J. and Woźniakowski, H. (1980). Convergence and complexity of interpolatory—newton iteration in a banach space. *Computers & Mathematics with Applications*, 6(4):385–400.

Villa, U., Petra, N., and Ghattas, O. (2021). HIPPYlib: an extensible software framework for large-scale inverse problems governed by PDEs. *ACM Transactions on Mathematical Software (TOMS)*, 47(2):1–34.

Wang, J. and Ye, X. (2013). A weak galerkin finite element method for second-order elliptic problems. *Journal of Computational and Applied Mathematics*, 241:103–115.

Wendland, H. (2004). *Scattered data approximation*, volume 17. Cambridge university press.

Wibisono, A. (2018). Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, pages 2093–3027. PMLR.

Wibisono, A. and Yang, K. Y. (2022). Convergence in kl divergence of the inexact Langevin algorithm with application to score-based generative models.

Yu, B. et al. (2018). The deep ritz method: a deep learning-based numerical algorithm for solving variational problems. *Communications in Mathematics and Statistics*, 6(1):1–12.

Zhang, X. and Garikipati, K. (2023). Label-free learning of elliptic partial differential equation solvers with generalizability across boundary value problems. *Computer Methods in Applied Mechanics and Engineering*, page 116214.

Zhang, X., Song, K. Z., Lu, M. W., and Liu, X. (2000). Meshless methods based on collocation with radial basis functions. *Computational mechanics*, 26:333–343.

## Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes, in Sections 2 and 3. Our algorithm is also summarized in pseudocode.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes, Sections 3 and 6 include such a discussion. Section 4 is a convergence proof.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes, in the appendix.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes.

(b) Complete proofs of all theoretical results. Yes.

(c) Clear explanations of any assumptions. Yes.

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes. We have included source code for our numerical results in the supplementary material.

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Not applicable.

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes.

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Not applicable.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. Not applicable.

(b) The license information of the assets, if applicable. Not applicable.

(c) New assets either in the supplemental material or as a URL, if applicable. Not applicable.

(d) Information about consent from data providers/curators. Not applicable.

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. Not applicable.

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not applicable.

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not applicable.

# A Other iterations based on the contraction mapping principle

For a fixed $p$, the score operator, $\mathcal{G}(p,\cdot)$, is a map from the space of transformations to scores. We first define a self-map corresponding to the score-residual operator. Then, we define a fixed point iteration for the self-map that converges in a setting where it is a contraction.

For convenience, we denote by $\mathcal{G}_p$ the operator $\mathcal{G}(p,\cdot)$ that returns the transported score, given a transport map. Let $\mathbb{T}$ be a Banach space of functions on $\mathbb{R}^d$ and $\mathbb{S}$ a Banach space of scores. Using the definition of the score operator,

$$\mathcal{G}_p(T)\circ T = p(\nabla T)^{-1} - \mathrm{tr}((\nabla T)^{-1}\,\nabla^2 T)\,(\nabla T)^{-1} = q\circ T, \tag{24}$$

when $T$ is a solution of the score-matching problem. We assume that the target score $q$ is a homeomorphism onto its image. This is indeed a strong condition that multi-modal distributions, for example, do not satisfy. Define a self-map $\mathcal{H}:\mathbb{T}\to\mathbb{T}$

$$\mathcal{H}(T) = q^{-1}\circ\mathcal{G}_p(T)\circ T. \tag{25}$$

Clearly, if $T$ is a solution of the score-matching problem, it is a fixed point of $\mathcal{H}$. Near a fixed point, we can define the following fixed-point iteration of $\mathcal{H}$.

$$T_{n+1} = \mathcal{H}(T_n) = q^{-1}\circ\mathcal{G}_p(T_n)\circ T_n, \tag{26}$$

with an arbitrary map $T_0$. When $\mathcal{H}$ is a contraction near its fixed point, say, $T^*$, then $T_n$ defined in (26) converges to $T^*$, by the contraction mapping principle. Consider one sufficient condition: when the functional derivative $D(\mathcal{G}_p(T)\circ T)(T^*)$ and the gradient $\nabla q^{-1}$ are small in the operator norms and $C^0$ norm respectively. Then, one obtains that $\mathcal{H}$ is contraction, and hence the fixed-point iteration defined above converges to a fixed point or a transport map.

However, this is only a sufficient condition for convergence. When $\mathcal{H}$ is not contractive, the fixed point iterations may not converge, or may exhibit slower than exponential convergence. When $\mathbb{T}$ is large enough to contain multiple fixed points of $\mathcal{H}$, the iterations may oscillate between basins of attraction of multiple fixed points. Note that the sufficient conditions for $\mathcal{H}$ to be a contraction impose restrictions on the behavior of the score operator near the fixed point and on the second-order derivatives of the target, which may preclude multi-modality in the target. The SCONE method, on the other hand, does not explicitly impose such a restriction at the fixed point and is hence more general.

## B SCONE example

As the first example, suppose the target is a univariate Gaussian with mean $m$ and variance $s^2$, while the source distribution is a standard normal in 1D. In this case, $q(x) = -(x-m)/s^2$ and $p(x) = -x$. The SCONE update $v_n$ satisfies the following ODE:

$$p_n - q = v_n'' + qv_n' + q'v_n. \tag{27}$$

At $n = 0$, $p_n = p$. The above ODE is defined on an unbounded domain, without specific decay rates for the solution at the boundaries, $\pm\infty$. This allows for unbounded solutions. Notice that the solution is always affine since, at $n = 0$, the left-hand side is affine. Subsequently, the update equation is satisfied by affine functions $v_n$ at every $n$ and hence $T_n$, which is a composition of affine functions, is affine. By comparing coefficients to solve for the update equations, we can obtain recurrence relationships for the slopes and intercepts of $v_n$, $T_{n+1} := (\mathrm{Id} + v_n) \circ T_n$, and $p_n = (p/T_n') \circ T_n^{-1}$. We can inductively show that all three are affine functions for all $n$. In particular, if $p_n(x) = a_n x + b_n$ and $v_n(x) = (A_n - 1)x + B_n$, we obtain,

$$\begin{aligned} A_n &= -a_n s^2/2 + 1/2 \\ B_n &= -b_n s^2 + m/2 - a_n m s^2/2, \end{aligned}$$

by comparing coefficients in the update (27). Then, the update for the score gives,

$$a_{n+1} = a_n/A_n^2, b_{n+1} = -a_n B_n/A_n^2 + B_n/A_n^2.$$

Considering the set of sequences $\{a_n, b_n, A_n, B_n\}_n$, it is clear from the relationships above that when $A_n \to 1$ and $B_n \to 0$, $a_n \to (-1/s^2)$ and $b_n \to m/s^2$. Thus, when the iterations for $T_n$ converge, or equivalently, when $v_n \to 0$, $p_n \to q$. Moreover, in this case, the limit $T$ is the function $T(x) = sx + m$, which coincides with the increasing rearrangement on $\mathbb{R}$ (and hence the optimal map). The intermediate distributions corresponding to the scores $p_n$ are all Gaussian.

Notice that since this convergence can be established for all $s$ and $m$, it suggests that SCONE transport converges even when in Hölder norm, $\|p - q\|$ is not small. That is, even though the derivation of the is premised on the local expansion of the score operator around $(q, \mathrm{Id})$, the smallness of $p - q$ and $T - \mathrm{Id}$ is not a necessary condition for the convergence of the method.

## C Additional experiments

Considering 1D targets, we first give proof-of-concept numerical results to validate our SCONE construction. We consider two smooth densities supported

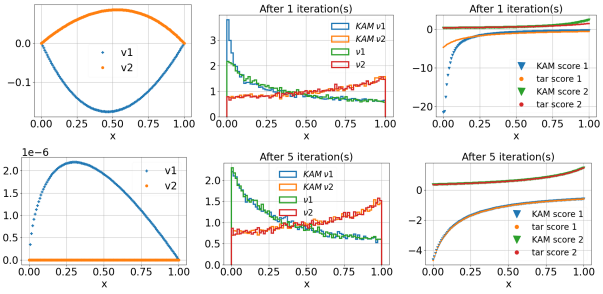


Figure 5: Numerical validation of SCONE transport on 1D densities: the first row depicts results after 1 iteration of our Newton method and the second row after 5 iterations. The scalar field $v$, the histogram approximations of the target density and the target score are plotted for the two different targets described in section C. The results of the computed transformed densities and scores from the SCONE iteration are compared against the target densities and scores in columns 2 and 3.

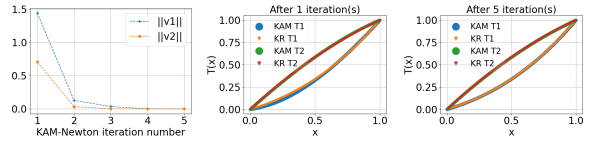


Figure 6: On the left column, we show the convergence of our algorithm ($\|v_n\| \to 0$) for the two different target densities given in section C. The center and right hand side figures compare the optimal map computed analytically against the SCONE transport map computed numerically after one and five iterations respectively.

on $[0,1]$ as our targets: $\rho_{\nu 1} = ((x+1)^3 - 1)/7$ and $\rho_{\nu 2} = 4/3 - (2-x)^2/3$. These densities are shown as histograms in the second column of Figure 5. In 1D, the PDE that needs to be solved at every SCONE iteration is an ODE, which we solve using a finite difference method with 128 grid points and zero Dirichlet boundary conditions on $[0,1]$. Our source density is the uniform (Lebesgue) density on $[0,1]$ whose score is the zero function. As shown in Figure 5 (third column), the transformed score matches the target score within a few SCONE iterations. The solution $v$ also quickly approaches the zero function as confirmed in the first column of Figure 5 and Figure 6(left), thus establishing numerically the convergence of our construction (see Theorem 3 in the main text). On one-dimensional domains, the optimal transport map (for a variety of costs) is simply the increasing rearrangement (see (Santambrogio, 2015), Chapter 2), which can be computed analytically in our setting, and is shown in the second and third columns of Figure 6. As shown, the SCONE construction converges to this transport map.

### C.1 1D unbounded target: bimodal Gaussian with equally weighted modes

Next, we consider a one-dimensional bimodal Gaussian target (shown in Figure 7), $0.5\mathcal{N}(-2,1) + 0.5\mathcal{N}(2,1)$. We again solve the ODE with finite difference on $[-10,10]$. The source is taken to be the standard Gaussian (unimodal distribution). In Figure 7, the first row presents results obtained after 1 iteration of our SCONE algorithm; the second row, after 3 iterations. We see that, after just 1 iteration, the two modes are detected, although the density and scores are not well-approximated. After 3 iterations, the empirical density (shown in blue) of the samples transported by the SCONE transport map match the target density (orange line plot) closely. As shown in Figure 7 (second row), as the SCONE update solution $v$ declines, the density and the scores approach their target values. In Figure 8 (left), we show the convergence of the solution $v$. From the figure, the convergence appears to be exponentially fast, with the rate decreasing after the first 4 iterations. The final transport map (right) is taken after 5 iterations, which accurately models the target density (center). A grid size of 4096 is used for solving the SCONE iteration, but grid size reductions $\mathcal{O}(1000)$ produces similar convergence results. We find that more iterations are needed when the modes are well-separated, e.g., sampling from an equally weighted bimodal distribution with modes centered at -4 and 2 required around 20 iterations.

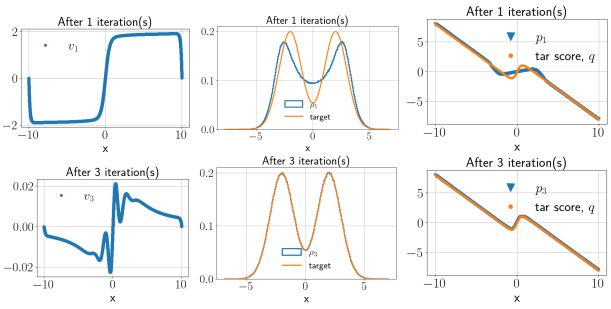


Figure 7: SCONE algorithm results: the target distribution, shown in orange in the center column, is an equally weighted bimodal Gaussian. The application of the SCONE algorithm, as described in section C.1, is shown in the first row after 1 iteration and in the second row after 3 iterations. The scalar field $v$, the histogram approximations of the target density and the target score are plotted in the first, second and third columns, respectively. The results of the computed transformed densities and scores from the SCONE iteration are compared against the target densities and scores in columns 2 and 3.

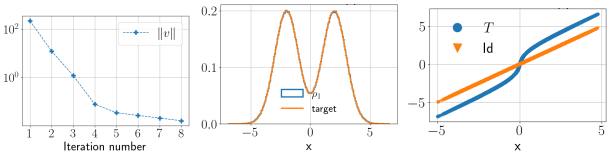


Figure 8: Convergence of SCONE iteration: the target distribution, shown in orange in the center column, is an equally weighted bimodal Gaussian. The application of the SCONE algorithm, as described in section C.1, results in the convergence of $\|v_n\|$, shown in the first column. The final SCONE transport map after 5 iterations is on the right (blue).

### C.2 1D unbounded target: bimodal Gaussian with unequally weighted modes

Previously in section C.1, when the target consisted of equally weighted modes, we used a finite difference method to solve the ODE, and then function interpolation to evaluate $v$ at the samples. Using these interpolated values, the score function, $p$ was updated, and the iterations continued by solving the ODE again. This vanilla scheme leads to numerical blow-up when the target has unequal weights. The reason is that errors in the solution of $v$ leads to the divergence of our Newton-like method. We observe that adding a small regularization term in the finite-difference ODE solution ($\ell^2$ regularization parameter set to 0.01), along with refining the grid near the points where $q'$ is large, induces convergence. The results are in the main text (section 5).

We provide the source code implementing the SCONE algorithm on all the examples above in (Chandramoorthy, 2023). This contains 'oneD.py' that implements the SCONE algorithm on all the examples above. The source file 'oneD_nonUni.py' implements adaptive grid refinement in the finite-difference solver. The unit tests that generate all the figures in this section are in 'tests/test_1D.py'. The code is written in Python and uses numpy and scipy.