A Sampling-Based Optimization Method Using Gradient-Like Generators

Anonymous authors
Paper under double-blind review

Abstract

Many applications require not a single minimizer of an energy function $f(\mathbf{x})$, but a diverse set of states in $\mathbb{R}^D$ whose energies follow a prescribed distribution p^* - a stochastic inverse problem, solved either by sampling an explicitly constructed state-space density, or by training a neural sampler to match the induced push-forward. We instead learn a *curve*: a generator $G : [0, 1] \to \mathbb{R}^D$ anchored at a state $\mathbf{x}_0$ that maps a uniform latent variable U to parts of the state space with energies matching p^*; that is, we learn G such that $f(G(U)) \sim p^*$. Inverse transform sampling turns this requirement into a quantile-matching loss that upper-bounds the squared Wasserstein-2 distance between the law of $f(G(U))$ and p^*, with equality exactly when $f \circ G$ is non-decreasing; the curve then sweeps the level sets of f in order, crossing each one exactly once between its endpoints. We call such generators *gradient-like* and they are not unique. We select the one of least kinetic energy and characterize when it exists. Freely choosing p^* and the law of $\mathbf{x}_0$ gives two dials - which energies the curve visits, and where in state space it is anchored -and our experiments turn both. Concentrating p^* near the optimum turns the sampler into a diverse optimizer: on multi-modal benchmarks p^* sets how deep the curves go and the law of $\mathbf{x}_0$ sets how widely they spread, giving mode coverage comparable to a continuous GFlowNet with additional control over the energy distribution. In image space ($D = 3072$), with the classifier's log-probability of the true class as energy, a curve anchored at a clean image interpolates smoothly from the image to a successful attack.

1 Introduction

Instead of finding the global minimum of an energy function $f : \mathbb{R}^D \to \mathbb{R}$, real-world problems often require exploring different low-energy regions of f and selecting solutions that work best under given constraints. Such situations arise naturally: in engineering, a set of solutions robust to manufacturing tolerances is more useful than one fragile optimum (Beyer & Sendhoff, 2007); in statistical physics and molecular biology, sampling methods are designed to escape energy basins and explore multiple metastable states (Valsson & Parrinello, 2014b; Wei et al., 2016); and in machine learning, distinct minima of a loss landscape can be connected and combined into more robust model ensembles (Garipov et al., 2018). This paper introduces a new, mathematically rigorous optimization method for finding a diverse set of low-energy regions by formulating the search for such a set as a *stochastic inverse problem*.

An inverse problem consists of input variables $\mathbf{x} \in \mathbb{R}^D$ describing a (physical) state and a model F mapping these variables into observations $\mathbf{y} = F(\mathbf{x}) \in \mathbb{R}^m$. The goal is to infer the variables from the observations, or the model from pairs of both (Hansen, 1998; Kaipio & Somersalo, 2005). The statistical Bayesian treatment of inverse problems is the probabilistic counterpart of classical optimization. It treats observations $\mathbf{y} = F(\mathbf{x}) + \varepsilon$ with normally distributed error ε and a Gaussian prior on $\mathbf{x}$, whereas the maximum posterior $\pi(\mathbf{x}|\mathbf{y})$ is the solution of a Tikhonov-regularized least-squares problem. Both classical optimization and statistical Bayes return an answer as a single point or as a belief concentrated around a point. Neither returns a *set* of qualitatively different good solutions.
That requires a distribution with genuine spread over the state space, and sampling the posterior does not provide one: pushing posterior samples forward through F does not in general reproduce the observed distribution (Butler et al., 2018). Finding a state space density π whose pushforward through

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

F equals the observed density p^* formulated as

$$F_{\#}\,\pi = p^* \tag{1}$$

is a structurally different problem and known by different names in different fields of research: *stochastic inverse problem* in mathematics (Breidt et al., 2011), unfolding in physics (Andreassen et al., 2020), population of models in biology (Britton et al., 2013), population inference in astronomy (Mandel et al., 2019), source distribution estimation in Machine Learning (Vetter et al., 2024).

Optimization through sampling by choosing p^*. For the case of scalar observations $m = 1$, Problem (1) is exactly the formalism for finding a diverse-set of low-energy regions. Choosing p^* concentrated near the smallest attainable value of $f = F$, the state space densities π are precisely distributions over near-optimal states: sampling from π produces a diverse set of good candidates rather than a single minimizer.
Different methods exist to learn the state space density π either directly or through variational approximations, see §2. They differ on the selection criterion used to choose among the generally infinitely many solutions to Problem (1). Butler et al. (2018) start with a state-space prior p_0 and *construct* π such that it coincides with p_0 on the level-sets of f. This construction is based on the co-area formula and requires estimating the density of the prior push-forward $f_{\#}p_0$. Sampling from π requires reweighting of prior samples. Vetter et al. (2024) select the distribution of maximum entropy and enforce consistency by incentivizing a small divergence between $F_{\#}\pi$ and p^*. Sampling is amortized and training requires estimating the entropy of π through samples only.

A parameterization by curves in absence of π. We propose a new solution and selection criterion to Problem (1), for the case of scalar observations $m = 1$. The selection criterion is of geometric nature and will be stated in the theoretical results below. Rather than searching over distributions and selecting one by an information-theoretic criterion, we parameterize candidate solutions by *curves through the state space*. Let $G : [0, 1] \to \mathbb{R}^D$ be a curve anchored at a reference state $\mathbf{x}_0$, that is $G(1) = \mathbf{x}_0$, and let $U \sim \mathcal{U}([0, 1])$. Such a curve is a solution to problem (1) whenever

$$f(G(U)) \sim p^*. \tag{2}$$

Inverse transform sampling (Devroye, 2006) tells us that U is transformed into p^* by applying the quantile function P^{-1}, which yields the following consistency loss for G, see Figure 1,

$$\mathcal{L}(G) = \int_0^1 \big(f(G(u)) - P^{-1}(u) \big)^2 \, du. \tag{3}$$

The initial condition $G(1) = \mathbf{x}_0$ can either be explicitly encoded into G or implicitly by adding a residual penalty to loss (3) such as $\lambda_0 ||G(1) - \mathbf{x}_0||_2^2$, where $\lambda_0 \geq 0$ is a hyperparameter.

The restriction to scalar observations is what makes this construction scalable to high-dimensional state-spaces: matching a one-dimensional distribution reduces to matching quantiles pointwise, so (3) requires only pointwise evaluations of f — no density of states, not entropy estimations and no Markov chains, in contrast to existing methods as we discuss in more details in §2.

Our approach comes at the cost of learning the induced state-space density π_G by G only implicitly, and, from a measure theoretic perspective, lies at the opposite extreme from the maximum-entropy solution as the support of π_G is singular, since concentrated on a curve. However, treating the anchor $\mathbf{x}_0$ as random variable with density p_0, the resulting density of $G(U; \mathbf{X}_0)$ is a family of curves filling the state-space. From that perspective, the pair (p_0, G) is our parameterization of solutions, see §3.1.

Theoretical results. In Section 3, we develop the theoretical foundation for this framework. The induced energy density $p_G \sim f(G(U))$ defined by the push-forward $p_G := (f \circ G)_{\#}U$ of the uniform distribution. Theorem 3.2 proves consistency of our approach by estimating the Wasserstein 2-distance from above

$$W_2^2\big(p_G,\ p^*\big) \leq \mathcal{L}(G) \tag{4}$$

where W_2^2 is the 1-dimensional squared Wasserstein 2-distance between the induced p_G and the observed target p^* energy distribution. Minimizing $\mathcal{L}$ therefore matches the density of $f(G(U))$ with the observed target distribution p^* as desired by (2). We show that equality holds if and only if G fulfills the following monotonicity condition

$$\frac{d}{du} f(G(u)) > 0 \tag{5}$$

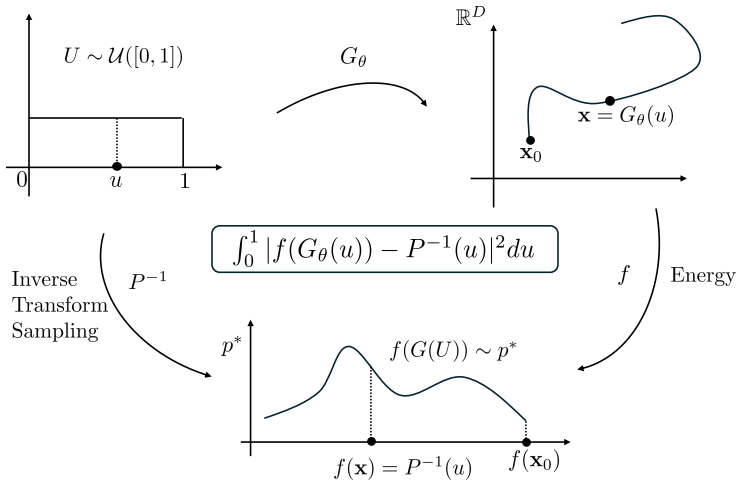


Figure 1: Intuition for solving (2). Starting from a uniform distribution U (top left), we generate samples whose energy values match $P^{-1}(U)$, ensuring $f(G(U)) \sim p^*$ by inverse transform sampling.

for all $u \in [0,1]$. Such curves G are called gradient-like in Morse theory (Milnor, 1965; Smale, 1961). A curve $G : [0,1] \to \mathbb{R}^D$ is a **gradient-like generator** if it satisfies $\frac{d}{du} f(G(u)) > 0$ and $p^* \sim f(G(U))$ simultaneously. It follows from Theorem 3.2 that minimizing our loss is "consistent" with our aim of optimization, meaning the equivalence

$$\left.\begin{array}{l} f(G(U)) \sim p^* \\ \frac{d}{du} f(G(u)) > 0 \end{array}\right\} \quad \Longleftrightarrow \quad \mathcal{L}(G) = 0. \tag{6}$$

Furthermore, we show in Theorem 3.4 that among all gradient-like generators, any critical point of the kinetic energy satisfies $G''(u) = \lambda(u)\, \nabla f(G(u))$ for a scalar function λ. Moreover, when ∇f is an eigenvector of the Hessian $d^2 f$, we can further characterize a local kinetic energy minimizer as the solution to the following ODE

$$G'(u) = \frac{1}{p^*(P^{-1}(u))} \cdot \frac{\nabla f(G(u))}{\|\nabla f(G(u))\|^2}, \qquad G(1) = \mathbf{x}_0. \tag{7}$$

Since (7) is a scaled gradient flow equation (Zhao et al., 2021) on the energy landscape induced by f; therefore, we provide a variational characterization of normalized gradient descent as the solution of a distribution-matching problem. This is our *geometric selection criterion* for G.

2 Related Work

Prescribing a target p^* on the energy values $f(\mathbf{x})$ to solve the stochastic inverse problem (1) is not new. Berg & Neuhaus (1992) and Wang & Landau (2001) both start from a uniform prior p_0 over the state space and target a uniform p^* over the attainable range of f. In neither case is the state-space density π itself learned or represented: only the *density of states* $g \coloneqq f_\# p_0$ is estimated, and samples are drawn from $\pi \propto p_0/(g \circ f)$ via a Markov chain. Butler et al. (2018) generalize this construction to general p_0 and p^*, requiring only that p^* be absolutely continuous with respect to $f_\# p_0$. Arbitrary p^* is likewise admitted in molecular simulation by variationally enhanced sampling (Valsson & Parrinello, 2014a), where p_0 is not a modeling choice but the Boltzmann distribution of the system.

Estimating π directly, e.g. via empirical Bayes or simulation-based inference, is another line of research for solving stochastic inverse problems. We refer to the related work section in Sourcerer Vetter et al. (2024) for an overview. Among all sample-based methods Sourcerer is closest to our work in its training objective: π is learned using neural samplers by maximizing its entropy while minimizing the sliced-Wasserstein distance between $f_\# \pi$ and a target p^* given by samples. Since f is

scalar in our setting, this distance reduces to the Wasserstein-2 distance. SOURCERER differs to our approach in three aspects. First, our loss (3) couples $f(G(u))$ and $P^{-1}(u)$ pointwise in u rather than after sorting; it is therefore not a distance between distributions, and it coincides with the squared Wasserstein-2 distance only for gradient-like curves (Remark 3.3). Second, SOURCERER selects the maximum-entropy solution, which requires an entropy estimator (Berrett et al., 2019; Kozachenko & Leonenko, 1987) in state space. Our selection criterion instead aims to minimize the kinetic energy of G. Third, our generator is anchored at an initial point, $G(1) = \mathbf{x}_0$, which in particular applications serves as a reference design to be improved upon. Treating $\mathbf{x}_0 \sim p_0$ as a random variable §3.1 then turns p_0 into a diversity knob: a concentrated p_0 refines solutions near a given design, while a broad p_0 explores many low-energy regions. Such a design freedom - a prior tied to a reference the curve is required to pass through - is unique to our method.

Diffusion- and control-based samplers (Akhound-Sadegh et al., 2024; Liu & Wang, 2016; Vargas et al., 2023; Zhang & Chen, 2022) pursue a related goal of sampling diverse low-energy solutions without data. We include the continuous GFlowNet (Bengio et al., 2021; Lahlou et al.) as a representative of this family; the resulting energies follow $f_{\#}\pi \propto Rg$ with R the reward and $g = f_{\#}p_0$ the density of states. The same holds for Boltzmann generators (Noé et al., 2019), which learn a normalizing flow targeting $\pi \propto e^{-\beta f}$ in state space. Our method instead prescribes the energy distribution directly, $f_{\#}\pi = p^*$. Moreover, because the target depends on $\mathbf{x}$ only through the scalar $f(\mathbf{x})$, our loss (3) is a direct regression objective. It requires neither a sequential environment nor flow-matching or trajectory-balance constraints, which rely on credit assignment over long trajectories and are known to be difficult to train and prone to mode collapse in continuous spaces (Malkin et al., 2022). The price is reduced generality: GFlowNets can target arbitrary non-negative rewards over $\mathbf{x}$.

3 A MATHEMATICAL FRAMEWORK ADDRESSING OPTIMIZATION PROBLEM (2)

We first show the consistency of our method in Theorem 3.2. We then specify our geometric selection criterion in 3.4 with scaled gradient descent as special case. We end this Section by generalizing our approach when treating $\mathbf{x}_0$ as random variable as well. We provide full proofs in the appendix.

Setting: We optimize an energy function $f : \mathbb{R}^D \to \mathbb{R}$ that is assumed to be twice continuously differentiable, $f \in C^2(\mathbb{R}^D, \mathbb{R})$. We assume an initial solution $\mathbf{x}_0 \in \mathbb{R}^D$ is at hand, and seek to improve upon it by optimizing within the sub-level set $\mathcal{A} = \left\{\mathbf{x} \in \mathbb{R}^D \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\right\}$ which we assume to be bounded. Since the sub-level set is closed, Heine-Borel immediately implies compactness of $\mathcal{A}$. In applications, $\mathcal{A}$ is chosen to be a hypercube in $\mathbb{R}^D$. In theory, we need compactness of $\mathcal{A}$. Next we single out distributions p^* for which the optimization (2) can be achieved by minimizing our loss function (3). These are distributions having full interval-support in $f(\mathcal{A})$,

Definition 3.1. A density $p^* : \mathbb{R} \to \mathbb{R}$ has **full interval support in** $f(\mathcal{A})$, if it satisfies the three conditions i) $p^*(y) > 0$ for all $y \in \text{supp}(p^*)$, ii) $\text{supp}(p^*) = [a, b]$ and iii) $\text{supp}(p^*) \subset f(\mathcal{A})$.

Note that many known distributions satisfy i) and ii) but fail iii). Satisfying all three needs a genuinely bounded distribution like uniform, beta, or a truncated version of a familiar one. However, for applications, it turns out it is sufficient that the main part of the mass of p^* lies within $f(\mathcal{A})$. Having full interval support in $f(\mathcal{A})$ means that $\text{supp}(p^*)$ is compact and hence P^{-1} is bounded and an element of the Hilbert space $L^2([0, 1])$. Since our curves $G : [0, 1] \to \mathcal{A}$ map into the compact sub-level set, $f \circ G$ is bounded and hence also a L^2 function. This confirms that the compact interval support of p^* is not a mere technicality, it ensures that $\mathcal{L}(G) = \|f \circ G - P^{-1}\|_2^2$ is well-defined and finite since both functions must lie in $L^2([0, 1])$. It is the compact support condition on p^* that makes the comparison to the Wasserstein 2-metric possible. Having formally defined the constraints on p^*, we are ready to prove that minimizing our loss is "consistent" with our aim of optimization, equivalence 6. In other words: If one minimizes the loss to zero, we have found a generator that both matches the distribution and behaves monotonically. Minimizing our loss produces gradient-like curves G that satisfy $f(G(U)) \sim p^*$. While this follows directly from inverse transform sampling, we choose to prove it via Theorem 3.2 that relates $\mathcal{L}(G)$ to the Wasserstein metric. We denote by $p_G \sim f(G(U))$ the push-forward of U via $f \circ G$. By definition the Wasserstein metric is

$$W_2^2(p_G; p^*) = \int_0^1 \left(P_G^{-1}(u) - P^{-1}(u)\right)^2 du$$

where P_G^{-1} is the quantile function of p_G. Our loss function $\mathcal{L}(G)$ differs from the Wasserstein metric by the (non-negative) **defect term**

$$\Delta_G = 2\left\langle P_G^{-1}, P^{-1}\right\rangle - 2\left\langle f \circ G, P^{-1}\right\rangle \geqslant 0 \tag{8}$$

where $\langle \cdot, \cdot \rangle$ denotes the L^2 inner product on function spaces. Our loss function decomposes into the Wasserstein 2−metric between the induced energy distribution p_G and the target distribution p^* plus a non-negative defect term Δ_G. This is the content of Theorem 3.2.

Theorem 3.2 (**Consistency via Wasserstein metric**). *Let $f \in C^2(\mathbb{R}^D, \mathbb{R})$, $\mathbf{x}_0 \in \mathbb{R}^D$ and $\mathcal{A} = \{x \in \mathbb{R}^D \mid f(x) \leq f(\mathbf{x}_0)\}$ bounded. Let moreover p^* be absolutely continuous and have full interval support in $f(\mathcal{A})$. For the induced energy distribution $p_G \sim f(G(U))$ and the defect term Δ_G the following holds*

i) $\mathcal{L}(G) = W_2^2(p_G, p^*) + \Delta_G$

ii) $W_2(p_G, p^*) \leq \sqrt{\mathcal{L}(G)}$

iii) $\Delta_G = 0 \iff f \circ G$ *is non-decreasing a.e.*0

Proof. Expanding the loss yields $\mathcal{L}(G) = \int_0^1 (f \circ G - P^{-1})^2 \, du = \|f \circ G\|^2 - 2\langle f \circ G, P^{-1}\rangle + \|P^{-1}\|^2$. Doing the same for the Wasserstein-2 distance yields the same expansion with the only difference that $f \circ G$ is sorted and thus P_G^{-1}. Comparing the two expansion yields i) and iii). That the defect term is non-negative is equivalent to the inequality ii) and follows from the Fréchet–Hoeffding inequality. □

Remark 3.3. $\mathcal{L}(G)$ **vs.** W_2^2**:** (i) When training G on W_2^2 instead of $\mathcal{L}(G)$, the objective matches the distribution of generated energies but not the pointwise coupling between P^{-1} and $f \circ G$: in one dimension W_2^2 compares P^{-1} to the *sorted* values of $f \circ G$, discarding the assignment of quantile levels to states. The two are not interchangeable: $f(G(u)) = P^{-1}(1-u)$ has $W_2 = 0$ but $\mathcal{L} > 0$. They do admit the same induced state-space laws π_G, since any G with $W_2 = 0$ can be composed with a measure-preserving map σ of $[0,1]$ so that $\mathcal{L}(G \circ \sigma) = 0$ with $\pi_{G\circ\sigma} = \pi_G$. But such a σ is a rearrangement, not a reparameterization: it is not monotone, so $G \circ \sigma$ is in general discontinuous and its kinetic energy infinite. Monotonizing G after training with W_2 therefore returns the measure and discards the curve; and on $\{W_2 = 0\}$ the kinetic energy is not constant across generators inducing the same π_G, so the selection criterion of Theorem 3.4 is not well posed there. Imposing the ordering through $\mathcal{L}$ is what makes the latent u the quantile level rather than an arbitrary label; we exploit this in §4.2 to interpolate smoothly between an original image and its adversarial perturbation.
(ii) *What is allowd.* Any fixed, strictly increasing bijection $\xi : [0,1] \to [0,1]$ may serve as the network's input coordinate. This not a rearrangement. The curve is still indexed by the quantile level u, so $\mathcal{L}(G)$, π_G and the kinetic energy, all computed in u, are unchanged. In the experiment §4, we us this to give the network an approximately arc-length coordinate.

This theorem provides the mathematical "green light" to replace complex statistical sampling with a deterministic optimization problem. We can now train a generator G to produce outputs with an exact target distribution p^* simply by minimizing $\mathcal{L}$. However, Theorem 3.2 does not address the existence or the uniqueness of generators. A priori training a neural net for minimizing $\mathcal{L}(G)$, could lead to no meaningful generator. This is the content of the next theorem. Existence is guaranteed by excluding critical points. Uniqueness comes from additionally minimizing the kinetic energy functional

$$\mathcal{E}(G) = \tfrac{1}{2} \int_0^1 \|G'\|^2 du. \tag{9}$$

The target energy distribution p^* carves out the region in the state space $\mathbb{R}^D$ where optimization takes place. Our optimization focuses the preimage of the distributions support $\operatorname{supp}(p^*) = [a, b]$ under f being the manifold

$$M_{p^*} := \{x \in \mathbb{R}^D \mid a \leq f(x) \leq b\} = f^{-1}([a, b]). \tag{10}$$

The support inteval can be expressed by $[a, b] = [P^{-1}(0), P^{-1}(1)]$ using the quantile function. It is on the manifold M_{p^*}, where low energy candidates are sought and nowhere else. Thus p^* acts as a prior, dictating not only the distribution of energies we aim for, but the very region M_{p^*} of the

design space within which good solutions ought to be found. In order for a gradient-like generator $G : [0, 1] \to \mathbb{R}^D$ to satisfy $\mathcal{L}(G) = 0$, the initial data $G(1) = \mathbf{x}_0$ and $G'(1) = \mathbf{v}_0$ must be compatible with the target profile $f(G(u)) = P^{-1}(u)$. We say that a curve G has *compatible initial conditions* if

$$f(\mathbf{x}_0) = P^{-1}(1) \quad \text{and} \quad \langle \nabla f(x_0),\, \mathbf{v}_0 \rangle = (P^{-1})'(1). \tag{11}$$

The first compatibility equation ensures that $\mathbf{x}_0$ has the right energy-level $P^{-1}(1)$, while the second matches the derivatives of $f \circ G$ and P^{-1} at the start $u = 1$. To establish existence of gradient-like generators G we assume f to be critical point free on M_{p^*}. Uniqueness then follows by additionally requiring G to be a critical point of the kinetic energy functional (9). This will be formalized in the next theorem.

Theorem 3.4 (**Existence and Uniqueness**). *Let $f \in C^3(\mathbb{R}^D, \mathbb{R})$ and $\mathbf{x}_0 \in \mathbb{R}^D$, and let $\mathcal{A} = \{x \in \mathbb{R}^D \mid f(x) \le f(\mathbf{x}_0)\}$ be bounded. Let $p^* \in C^1$ with full interval support in $f(\mathcal{A})$, and suppose f has no critical points on $M_{p^*} := f^{-1}(supp(p^*))$. Assume that $\mathbf{x}_0, \mathbf{v}_0 \in \mathbb{R}^D$ are compatible initial conditions (11). Then the following holds.*

- **(i) Existence:** *There exists a curve $G : [0, 1] \to \mathcal{A}$ satisfying $\mathcal{L}(G) = 0$ with $G(1) = \mathbf{x}_0$.*
- **(ii) Uniqueness:** *There exist an unique curve $G : [0, 1] \to \mathcal{A}$ satisfying $\mathcal{L}(G) = 0$ with $G(1) = \mathbf{x}_0$, $G'(1) = \mathbf{v}_0$ and additionally solving*

$$G''(u) = \lambda(u)\, \nabla f(G(u)) \tag{12}$$

for all $u \in [0, 1]$. This unique curve is a critical point of the kinetic energy functional (9) restricted to $\mathcal{L}(G) = 0$.

Proof. We formulate the Euler-Lagrange equation of the energy functional subject to $\mathcal{L}(G) = 0$. Existence follows from the fact that we assume no critical points in M_{p^*}. Uniqueness from Picard–Lindelöf. □

This unique curve is a critical point $d\mathcal{E}(G^*) = 0$ of the kinetic energy functional $\mathcal{E} : \mathcal{S} \to \mathbb{R}$ defined in (9), restricted to the constraint manifold

$$\mathcal{S} = \big\{G : [0, 1] \to \mathcal{A} \mid \mathcal{L}(G) = 0, G(1) = \mathbf{x}_0, G'(1) = \mathbf{v}_0\big\}.$$

Critical curves of $\mathcal{E}$ are characterized by (12) and called *geodesics*. A geodesic is uniquely determined by fixing the base point $\mathbf{x}_0$ and direction $\mathbf{v}_0$. Yet a geodesic, however uniquely determined, is not automatically the curve we are after, since being a critical point of $\mathcal{E} : \mathcal{S} \to \mathbb{R}$ does not guarantee minimal energy. Singling out the energy-minimizing curve among these geodesics calls for a genuinely *geometric uniqueness criterion*: If the gradient points in the principal direction of the level sets along the whole curve, then the geodesic locally minimizes energy.

Definition 3.5. We say the gradient ∇f **points in the principal direction of the level sets** along the curve G if $\nabla f(G(u))$ is an eigenvector of the Hessian $d^2 f(G(u))$ for every $u \in [0, 1]$

This admits a precise geometric interpretation in terms of the second fundamental form of the level-set foliation. It turns out that these geodesics are exactly the gradient trajectories of (7). The gradient pointing in the principal direction along a curve G is equivalent to $\nabla f(G(u))$ is an Eigenvector of the Hessian $d^2 f(G(u))$ for all $u \in [0, 1]$.

Proposition 3.6 (**Geometric Uniqueness**). *Under the same assumptions as in Theorem 3.4, suppose further that ∇f points along the principal direction of the level sets along the pure gradient flow line*

$$G'(u) = \frac{1}{p^*(P^{-1}(u))} \cdot \frac{\nabla f(G(u))}{\|\nabla f(G(u))\|^2}$$

with $G_0(1) = \mathbf{x}_0$. Then the following holds.

- i) *The curve satisfies $G'' = \lambda(u) \nabla f(G(u))$*
- ii) *The curve G is a local energy minimum, for sufficiently small $supp(p^*)$.*

Proof. i) Directly follows from differentiating G' and the eigenvector definition. ii) We calculate the index form and show that it is positive using Cauchy-Schwarz, Poincaré inequality and standard calculus. □

It is not clear whether there exists geodesics $G^* \in \mathcal{S}$ having lower kinetic energy than the gradient lines $G_0 \in \mathcal{S}$. A classical result in Riemannian geometry states that G^* is a local minimum of $\mathcal{E}$ on $\mathcal{S}$ if and only if no conjugate points occur along G^*.

3.1 Populating the state-space density - $\mathbf{x}_0$ as random variable

If $\mathbf{x}_0$ is sampled from p_0 and $f(G(u; x_0)) = P^{-1}(u)$ for p_0−a.e. $\mathbf{x}_0$, it follows directly from the tower property that the push-forward law of the random variable $G(U; \mathbf{X}_0)$ is p^*:

$$\mathbb{P}\big(f(G(U; x_0)) \in B\big) = \mathbb{E}_{\mathbf{x}_0 \sim p_0}\Big[\mathbb{P}\big(f(G(U; \mathbf{x}_0)) \in B \mid \mathbf{x}_0\big)\Big] = \int_B p^*.$$

The anchoring law p_0 determines the mass on the level set $f^{-1}(P^{-1}(1))$, a $D-1$ dimensional manifold, and that mass spreads in state-space as prescribed by p^*. This motivates the following state-space generalization of our method

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}_0 \sim p_0}\left[\int_0^1 |f(G(u; \mathbf{x}_0)) - P^{-1}(u)|^2 du + \lambda_0 \|G(1; \mathbf{x}_0) - \mathbf{x}_0\|_2^2\right] \tag{13}$$

as indeed for $\lambda_0 = 0$ and $p_0 = \delta_{\mathbf{x}_0}$ recovers loss (3). The anchoring condition $G(1; \mathbf{x}_0) = \mathbf{x}_0$ can also be explicitly encoded into G and thus $\lambda_0 = 0$, see §4. In the proof-of-concept experiments presented in §4, f is provided and we sample from the level set $f^{-1}(P^{-1}(1))$ through a rejection sampling. This might be challenging in general, especially for high-dimensional D. The restriction disappears if the prescribed law is allowed to depend on the anchor.

Conditioning the law on the anchor. Replace p^* by a family $\{p^*(\cdot \mid \mathbf{x}_0)\}$ with $P^{-1}(1 \mid \mathbf{x}_0) = f(\mathbf{x}_0)$ for p_0-a.e. $\mathbf{x}_0$. Loss (13) is then read with $P^{-1}(u \mid \mathbf{x}_0)$ in place of $P^{-1}(u)$, the anchor convention $G(1; \mathbf{x}_0) = \mathbf{x}_0$ stays exact, and p_0 is free. The pair (p_0, G) therefore still solves the stochastic inverse problem (1), but for the mixture $\bar{p}^* = \mathbb{E}_{\mathbf{x}_0 \sim p_0}\big[p^*(\cdot \mid \mathbf{x}_0)\big]$, see appendix §7.2 for details.

4 Experiments

We compare with Sourcerer and of data-consistent inversion (DCI), two alternative solutions of problem (1) with different selection criteria. On Hartmann-6 and Manywell-5, where the goal is a diverse set of low-energy solutions, we add GFlowNets and classical gradient descent; see §2 for details on these methods. We close by generating adversarial examples on CIFAR-10 at resolution 32×32, that is state-space dimension $D = 3072$. For details and further experiments see §7.

4.1 Optimization through sampling on Hartmann-6 and Manywell-5

In $D = 6$ dimensions, Hartmann-6 (Balandat et al., 2020) has exactly two basins of non-negligible volume: the global basin A (global minimum $f_{\min} = -3.322$, 68% of the cube) and a second basin B (minimum −3.203, 32%). A symmetric version of Manywell-5 in $D = 5$ dimensions has 32 identical wells of energy $f_{\min} = 0$ behind barriers of height 1. Both can be written on the cube $\mathcal{A} = [-1, 1]^D$ - so one explicit anchor $G(u; \mathbf{x}_0) = \tanh(\operatorname{atanh} \mathbf{x}_0 + (1 - \xi(u)) h_\theta(\xi(u), \mathbf{x}_0))$ serves both, where ξ denotes the approximately arc-length coordinate from Remark 3.3(ii). The law is $p^* = f_{\min} + \mathrm{Exp}(1/0.3)$ truncated such that the upper end $P^{-1}(1) = f(\mathbf{x}_0)$, which is −1.64 and 2.05. The level set $f^{-1}(P^{-1}(1))$ is sampled at a tolerance of 0.05 in energy.

Protocol and metrics. We call the samples with $f \le \tau$ the *low-energy set*, where $\tau = P^{-1}(q)$ for a chosen quantile q. On Manywell-5 every well sits at $f_{\min}$, so any q will lead to a diverse set and we take $q = 0.1$. On Hartmann-6, we take $q = 0.4$ such that samples from basin B are part of the set. For gradient-like generators, the low-energy set can be sampled directly with $u \sim \mathcal{U}[0, q]$: the latent *is* the quantile, so every draw is accepted whenever $\Delta = 0$ and one forward pass gives one low-energy sample. The other methods have no such coordinate and reject from N/q draws. W_2 is the distance between the generated energies and p^* with $N = 2048$ samples; note that even a perfect sampler has a distance of 0.012 / 0.014 (Hartmann-6 / ManyWell-5). The defect Δ of Theorem 3.2, $\mathcal{L} = W_2^2 + \Delta$, measures how far the coupling between u and the energy is from monotone and is undefined without a latent coordinate. On the low-energy set, *best E* is the mean of its 100 lowest energies, *spread* the share of samples in basin B or the number of wells holding at least 1% of them, and *yield* the share of draws that fell below τ.

Table 1: HARTMANN-6 (global minimum −3.322; uniform samples: 32% in basin B) and MANYWELL-5 (global minimum 0; uniform samples: 32 wells), with $p^* = f_{\min} + \mathrm{Exp}(1/0.3)$ truncated at $f(\mathbf{x}_0)$. The cut is $q = 0.4$ on HARTMANN-6, the smallest round value for which τ clears basin B, and $q = 0.1$ on MANYWELL-5.

method	$W_2 \downarrow$	$\sqrt{\Delta} \downarrow$	best E	spread	yield	$W_2 \downarrow$	$\sqrt{\Delta} \downarrow$	best E	spread	yield
ours: p_0 on the level set $f^{-1}(P^{-1}(1))$										
$p_0 = \delta_{\mathbf{x}_0}$ (one point of it)	0.043 ±0.001	0.000 ±0.000	−3.203 ±0.000	100 ±0%	100 ±0%	0.012 ±0.002	0.000 ±0.000	0.001 ±0.000	1.0 ±0.0	100 ±0%
$p_0 = \mathcal{U}(f^{-1}(P^{-1}(1)))$ (all of it)	0.025 ±0.004	0.040 ±0.003	−3.310 ±0.001	35 ±1%	89 ±2%	0.030 ±0.006	0.110 ±0.007	0.007 ±0.001	31.6 ±0.5	37 ±4%
same law, other selection rules										
Sourcerer, uniform reference ($\lambda = 0.1$)	0.057 ±0.034	–	−3.311 ±0.004	0 ±0%	38 ±3%	0.035 ±0.003	–	0.008 ±0.003	18.3 ±3.5	5 ±2%
DCI, prior $\mathcal{U}$(box)	0.128 ±0.014	–	−3.229 ±0.010	23 ±25% (4 distinct)	9 ±4%	0.020 ±0.011	–	0.015 ±0.008	4.3 ±3.5 (4 distinct)	5 ±4%
no prescribed law										
GFlowNet, reward e^{-10E}	0.110 ±0.004	–	−3.290 ±0.001	0 ±0%	13 ±1%	0.064 ±0.011	–	0.016 ±0.001	29.7 ±1.5	1 ±0%
multi-start gradient descent	0.351 ±0.001	–	−3.322 ±0.000	32 ±1%	100 ±0%	0.417 ±0.000	–	0.000 ±0.000	32.0 ±0.0	100 ±0%

Results. The anchor law p_0 says which part of the level set is populated, p^* how far down the curves go. A point mass at $\mathbf{x}_0$ (first row) populates one point of it: the generator is gradient-like, $\Delta = 0$, every draw of the low-energy set is accepted, and the samples stay in the basin, or well, of $\mathbf{x}_0$ - this is the mode that improves a given design. Its depth is then capped by its own basin: $\mathbf{x}_0$ lies in B, and −3.203 is within 10^{-4} of B's minimum. That cap is also why its W_2 is 0.043: the law is floored at the global optimum, which lies in A, and a curve anchored in B cannot deliver the bottom of a law it was promised. Sampling the whole level set (second row) populates all of it, and the curves carry that spread downwards: 35% of the low-energy samples lie in basin B, against the level set's own 39%, and 31.6 of the 32 wells are covered.

Given the same law, SOURCERER and DCI also match it, at 0.057 and 0.128 on HARTMANN-6 and 0.035 and 0.020 on MANYWELL-5, against our 0.025 and 0.030. DCI re-weights a fixed pool, and the deeper the law asks the less of that pool can answer, so its HARTMANN-6 distance is worse than the law-free GFlowNet's 0.110. However, that GFlowNet number is also closer than it looks - it is the best of four reward temperatures, and W_2 moves by a factor of 27 across them (3.00 at $\beta = 1$ down to 0.11 at $\beta = 10$), so $\beta = 10$ matching this particular p^* is an accident of the density of states rather than control over the law. What differs is which part of the admissible set each method occupies. SOURCERER spreads by its entropy weight λ_H, and reaches −3.311 - a tie with our −3.310 - but the spread does not survive the restriction to the low-energy sets: *no* low-energy sample in basin B, and only 18.3 wells are visisted. §7.1.1 sweeps λ_H over $\{0.02, 0.1, 0.35, 0.7\}$ and the basin-B column is 0% at every value at which the law is sufficiently matched. DCI's low-energy set on HARTMANN-6 holds 5, 2 and 4 distinct designs over three seeds, so its 23% basin-B share rests on a handful of points, and the 4.3 "wells" it covers on MANYWELL-5 are just its 4.3 distinct points. Multi-start descent reaches the optimum and reproduces the cube's basin-volume prior as expected; the GFlowNet covers 29.7 wells at a yield of 1%. Covering identical wells is easy for any sampler that starts everywhere; choosing the well, the start and the energy level is what the our method adds.

Remark 4.1. In HARTMANN-6, all curves end up in the basin they started in. For MANYWELL-5, this is true for 80% of the samples - significantly higher than chance level for both (53% and 3%). That suggests our curves are gradient-like without explicitly enforcing it.

4.2 GENERATING ADVERSARIAL EXAMPLES IN $D = 3072$ DIMENSIONS

Given an image $\mathbf{x}_0$ with label y and a classifier C with $C(\mathbf{x}_0) = y$, an adversarial example is a perturbation δ that flips the prediction, $C(\mathbf{x}_0 + \delta) \neq y$, while staying imperceptible, $\|\delta\|_p < \varepsilon$ — typically $p = \infty$ with $\varepsilon = 8/255$, or $p = 2$ with $\varepsilon = 0.5$ (Goodfellow et al., 2015; Han et al., 2023). One usually minimises a cost $\mathcal{K}(\mathbf{x}_0, \delta, C)$ over that ball, for instance the cross-entropy between the outputs of $C(\mathbf{x}_0)$ and $C(\mathbf{x}_0 + \delta)$: the larger it is, the further apart the two classes. We take its negative as the energy, that is the log-probability of the true class, $f(x) = \log p_{\mathrm{clf}}(y \mid x)$, so that the clean image is the top of the law, and prescribe $p^* = f(\mathbf{x}_0) - \mathrm{Exp}(1)$. $\mathcal{A}$ is the L_∞ ball of radius $\varepsilon = 8/255$ around a CIFAR-10 image (Krizhevsky & Hinton, 2009) at 32×32, that is $D = 3072$. Figure 2 shows one trajectory against its realised and prescribed law: the latent interpolates smoothly from the clean image at $u = 1$ to a successful attack at $u = 0$.

SOURCERER realises the same law — $W_2 = 0.162$ against our 0.101 — so the two differ not in whether they solve the problem but in what they spend on it. Measured as the mean perturbation per pixel in units of the budget ε, ours spends 0.10 and SOURCERER 0.87; adding an incentive for low kinetic energy halves ours again to 0.055, at no cost in matching ($W_2 = 0.109 \pm 0.029$ against

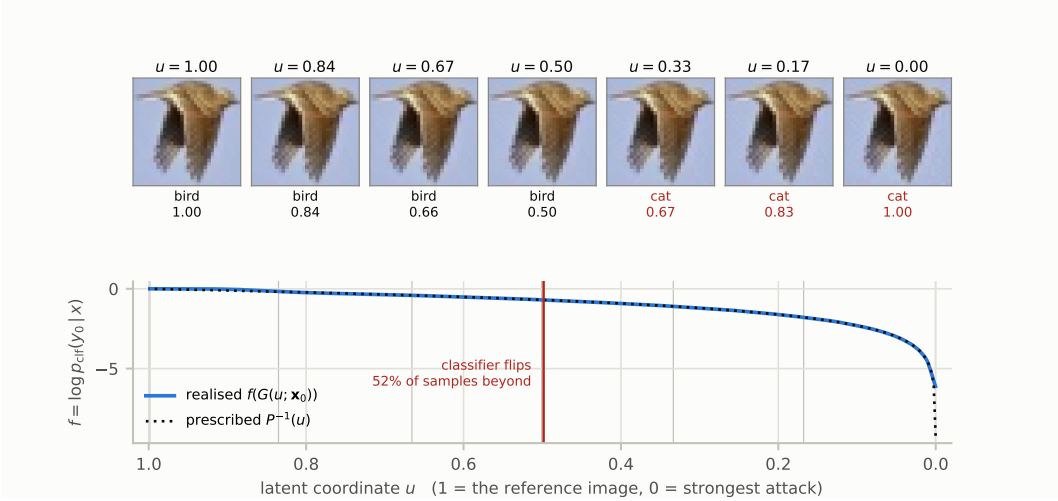


Figure 2: Generated Adversarial examples for $p = \infty$. The latent variable u allows for a smooth interpolation between the original $\mathbf{x}_0$ (a bird, $u = 1$) and an adversarial example ($u = 0$), with the energies following $p^* = f(x_0) - \mathrm{Exp}(1)$. Under each panel: the predicted class and its probability. The classifier flips near $p_{\mathrm{clf}} = 0.5$, so the law itself predicts a 50% attack-success rate; the realised rate is 52%.

0.101 ± 0.031), so the gap reaches a factor of 15. A uniform draw from the threat ball already spends 0.5: SOURCERER's attacks sit further from the image than random points of the ball. Its distance is set by the entropy weight and the geometry of the ball; ours is set by the anchor. DCI re-weights samples it has already drawn, and in $D = 3072$ none of them is a strong attack — the best of its $2 \cdot 10^5$ draws reaches a cross-entropy of 1.6, and its realised mean stops at 0.32 where the law asks for 1.

5 Discussion: Limitations and Open Problems

Theorem 3.2 establishes a new consistency loss for the stochastic inverse problem $f(G(U)) \sim p^*$ with scalar f: the loss upper-bounds the squared Wasserstein-2 distance between the law of $f(G(U))$ and p^*, with equality precisely when $f \circ G$ is non-decreasing. The constraint $\mathcal{L}(G) = 0$ determines the component of G' along ∇f pointwise while leaving the component tangent to the level sets free; Thmeorem 3.4 and Proposition 3.6 resolve that freedom by a geometric criterion for a local minimizer of the kinetic energy.

Inverse transform sampling requires f to be scalar. Multivariate quantile matching or sliced Wasserstein distances are natural extensions allowing for *Vector-valued observations.*

Problem (1) is studied largely independently across different fields. A unifying picture bridging gaps between those areas of research is missing. Stochastic inverse problems are ill-posed in general. Whether every solution π is realizable by our construction — that is, whether for each such π there exist an anchor law p_0 and a generator G with $\mathcal{L}(G) = 0$ such that $G(U; \mathbf{X}_0) \sim \pi$ for $\mathbf{X}_0 \sim p_0$ — is an open question. An affirmative answer would provide a variational characterization of the solutions to problem (1). A further challenge concerns the anchor convention $G(1; \mathbf{x}_0) = \mathbf{x}_0$: at zero loss it forces p_0 to be supported on the level set $f^{-1}(P^{-1}(1))$, which is itself typically hard to sample from. In appendix 7.2, we allow p^* to be anchor dependent allowing for general p_0 but other approaches based on choosing p^* and p_0 from first principles might be possible as well. In both cases, the support of p^* is fixed in advance, with its lower end at the global minimum of f; discovering the law dynamically instead, so that no knowledge of the optimum is required, is left for future work.

Finally, the mathematical framework of Theorem 3.2 and Theorem 3.4 generalizes naturally to the Riemannian setting by replacing $\mathbb{R}^D$ with a Riemannian manifold M with Riemannian metric $\mathfrak{g}$. The admissible set $\mathcal{A}$ could be replaced with an invariant submanifold $\mathcal{N} \subset M$, provided the Euclidean gradient is substituted with the Riemannian gradient $\nabla_{\mathfrak{g}} f$.

Use of large language models

Large language models were used substantially in the preparation of this work, and the authors take full responsibility for all of its content. Their use falls in four places. First, code: much of the experimental harness — the runners, the metric and table generators, and the re-implementation of the Sourcerer and data-consistent-inversion baselines — was written with LLM assistance and then read and tested by the authors. Second, analysis: LLM agents were used to run sweeps, score results and draft summaries of them. Third, writing: the experimental sections and parts of the appendix were drafted with LLM assistance. Fourth, review: an LLM agent was used to check the numbers reported in this paper against the code and the stored results, and to search the draft for internal inconsistencies.

The problem formulation and the two selection criteria — gradient-likeness and minimal kinetic energy — are the authors', as are Theorem 3.4, Proposition 3.6 and the proofs. LLMs were used while developing the analysis rather than to generate results: they helped simplify several of the proofs, and one substantive step came out of that exchange — recognising that the inverse-transform loss decomposes into the squared Wasserstein-2 distance plus a non-negative rearrangement defect, which is the content of Theorem 3.2. The authors checked every statement and proof and take responsibility for them.

References

Tara Akhound-Sadegh, Jarrid Rector-Brooks, Avishek Joey Bose, Sarthak Mittal, Pablo Lemos, Cheng-Hao Liu, Marcin Sendera, Siamak Ravanbakhsh, Gauthier Gidel, Yoshua Bengio, Nikolay Malkin, and Alexander Tong. Iterated denoising energy matching for sampling from Boltzmann densities. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, 2024.

Anders Andreassen, Patrick T. Komiske, Eric M. Metodiev, Benjamin Nachman, and Jesse Thaler. OmniFold: A method to simultaneously unfold all observables. *Physical Review Letters*, 124(18): 182001, 2020. doi: 10.1103/PhysRevLett.124.182001.

Maximilian Balandat, Brian Karrer, Daniel R. Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. BoTorch: A framework for efficient Monte-Carlo Bayesian optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. arXiv:1910.06403.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2106.04399.

Bernd A. Berg and Thomas Neuhaus. Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Physical Review Letters*, 68(1):9–12, 1992. doi: 10.1103/PhysRevLett.68.9.

Thomas B. Berrett, Richard J. Samworth, and Ming Yuan. Efficient multivariate entropy estimation via k-nearest neighbour distances. *The Annals of Statistics*, 47(1):288–318, 2019. doi: 10.1214/18-AOS1688.

Hans-Georg Beyer and Bernhard Sendhoff. Robust optimization – a comprehensive survey. *Computer Methods in Applied Mechanics and Engineering*, 196(33):3190–3218, 2007. ISSN 0045-7825. doi: 10.1016/j.cma.2007.03.003. URL `https://www.sciencedirect.com/science/article/pii/S0045782507001259`.

J. Breidt, T. Butler, and D. Estep. A measure-theoretic computational method for inverse sensitivity problems I: Method and analysis. *SIAM Journal on Numerical Analysis*, 49(5):1836–1859, 2011. doi: 10.1137/100785946.

Oliver J. Britton, Alfonso Bueno-Orovio, Karel Van Ammel, Hua Rong Lu, Rob Towart, David J. Gallacher, and Blanca Rodriguez. Experimentally calibrated population of models predicts and explains intersubject variability in cardiac cellular electrophysiology. *Proceedings of the National Academy of Sciences*, 110(23):E2098–E2105, 2013. doi: 10.1073/pnas.1304382110.

Troy Butler, John Jakeman, and Tim Wildey. Combining push-forward measures and Bayes' rule to construct consistent solutions to stochastic inverse problems. *SIAM Journal on Scientific Computing*, 40(2):A984–A1011, 2018. doi: 10.1137/16M1087229.

Yaofo Chen. PyTorch CIFAR models. `https://github.com/chenyaofo/pytorch-cifar-models`. Accessed: 2026-09-26.

Luc Devroye. Chapter 4 Nonuniform Random Variate Generation. In Shane G. Henderson and Barry L. Nelson (eds.), *Handbooks in Operations Research and Management Science*, volume 13 of *Simulation*, pp. 83–121. Elsevier, January 2006. doi: 10.1016/S0927-0507(06)13004-2. URL `https://www.sciencedirect.com/science/article/pii/S0927050706130042`.

L. C. W. Dixon and G. P. Szegö. The global optimization problem: an introduction. In *Towards Global Optimisation 2*, pp. 1–15. North-Holland, 1978.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P. Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. URL `http://arxiv.org/abs/1412.6572`.

Sicong Han, Chenhao Lin, Chao Shen, Qian Wang, and Xiaohong Guan. Interpreting adversarial examples in deep learning: A review. *ACM Comput. Surv.*, 55(14):328:1–328:38, 2023. ISSN 0360-0300. doi: 10.1145/3594869. URL `https://dl.acm.org/doi/10.1145/3594869`.

Per Christian Hansen. *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*. Society for Industrial and Applied Mathematics, 1998. ISBN 978-0-89871-403-6 978-0-89871-969-7. doi: 10.1137/1.9780898719697. URL `http://epubs.siam.org/doi/book/10.1137/1.9780898719697`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. ResNet-20 is the CIFAR variant of Section 4.2.

Jari P. Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems*, volume 160 of *Applied Mathematical Sciences*. Springer, 2005. ISBN 978-0-387-22073-4 978-0-387-27132-3. doi: 10.1007/b138659. URL `https://link.springer.com/10.1007/b138659`.

L. F. Kozachenko and N. N. Leonenko. Sample estimate of the entropy of a random vector. *Problems of Information Transmission*, 23(2):95–101, 1987.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images.(2009), 2009.

Salem Lahlou, Tristan Deleu, Pablo Lemos, Dinghuai Zhang, Alexandra Volokhova, Alex Hernández-García, Lena Nehale Ezzine, Yoshua Bengio, and Nikolay Malkin. A theory of continuous generative flow networks. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 18269–18300. PMLR. URL `https://proceedings.mlr.press/v202/lahlou23a.html`.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Nikolay Malkin, Moksh Jain, Emmanuel Bengio, Chen Sun, and Yoshua Bengio. Trajectory balance: Improved credit assignment in GFlowNets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2201.13259.

Ilya Mandel, Will M. Farr, and Jonathan R. Gair. Extracting distribution parameters from multiple uncertain observations with selection biases. *Monthly Notices of the Royal Astronomical Society*, 486(1):1086–1093, 2019. doi: 10.1093/mnras/stz896.

John Milnor. *Lectures on the h-cobordism theorem*. Princeton Mathematical Notes. Princeton University Press, 1965. Notes by L. Siebenmann and J. Sondow.

Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

Stephen Smale. On gradient dynamical systems. *Annals of Mathematics*, 74(1):199–206, 1961.

Omar Valsson and Michele Parrinello. Variational approach to enhanced sampling and free energy calculations. *Physical review letters*, 113(9):090601, 2014a.

Omar Valsson and Michele Parrinello. Variational approach to enhanced sampling and free energy calculations. *Physical Review Letters*, 113(9):090601, 2014b. doi: 10.1103/PhysRevLett.113.090601.

Francisco Vargas, Will Grathwohl, and Arnaud Doucet. Denoising diffusion samplers. In *International Conference on Learning Representations (ICLR)*, 2023.

Julius Vetter, Guy Moss, Cornelius Schröder, Richard Gao, and Jakob H. Macke. Sourcerer: Sample-based maximum entropy source distribution estimation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 88772–88806, 2024. doi: 10.52202/079017-2817.

Cédric Villani. *Topics in Optimal Transportation*. American Mathematical Soc., 2021. ISBN 978-1-4704-6726-5. Google-Books-ID: NElDEAAAQBAJ.

Fugao Wang and D. P. Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Physical Review Letters*, 86(10):2050–2053, 2001. doi: 10.1103/PhysRevLett.86.2050.

Guanghong Wei, Wenhui Xi, Ruth Nussinov, and Buyong Ma. Protein ensembles: How does nature harness thermodynamic fluctuations for life? the diverse functional roles of conformational ensembles in the cell. *Chemical Reviews*, 116(11):6516–6551, 2016. doi: 10.1021/acs.chemrev.5b00562.

Qinsheng Zhang and Yongxin Chen. Path integral sampler: A stochastic control approach for sampling. In *International Conference on Learning Representations (ICLR)*, 2022.

Shen-Yi Zhao, Yin-Peng Xie, and Wu-Jun Li. On the convergence and improvement of stochastic normalized gradient descent. *Science China Information Sciences*, 64(3):132103, 2021. ISSN 1869-1919. doi: 10.1007/s11432-020-3023-7. URL `https://doi.org/10.1007/s11432-020-3023-7`.

6 Appendix: Proofs

6.1 Proof of Theorem 3.2

Proof. Both $h = f \circ G$ and $q = P^{-1}$ lie in $L^2[0,1]$, so every term below is finite. We use three standard facts.

1. 1-D optimal transport. For laws on $\mathbb{R}$ with finite second moment, $W_2^2(\mu;\nu) = \int_0^1 \left(F_\mu^{-1} - F_\nu^{-1}\right)^2$ — the L^2 distance of the quantile functions (Villani (2021) Thm. 2.18).
2. Rearrangement = quantile. For $U \sim \mathcal{U}[0,1]$, the nondecreasing rearrangement $h^\uparrow$ is the quantile function of the law $h_\sharp p_U$ of $h(U)$. Moreover h and $h^\uparrow$ are equimeasurable (same law), so $\|h^\uparrow\|^2 = \|h\|^2$.
3. Fréchet–Hoeffding. For $U \sim \mathcal{U}[0,1]$, the comonotone coupling maximizes correlation: $\langle h, q\rangle \le \langle h^\uparrow, q^\uparrow\rangle$. Since q is already nondecreasing, $q^\uparrow = q$, giving $\langle h, q\rangle \le \langle h^\uparrow, q\rangle$, with equality iff h is comonotone with q — under a strictly positive p^*-density (so q strictly increasing), iff $h = h^\uparrow$ a.e.

The loss is a squared L^2 distance; expand it: $\mathcal{L}(G) = \int_0^1 (h-q)^2\,du = \|h\|^2 - 2\langle h, q\rangle + \|q\|^2$. By (1)+(2), the Wasserstein term is the same expansion with h replaced by its quantile function $h^\uparrow$: $W_2^2\big(h_\sharp p_U; p^\star\big) = \int_0^1 (h^\uparrow - q)^2\,du = \|h^\uparrow\|^2 - 2\langle h^\uparrow, q\rangle + \|q\|^2 = \|h\|^2 - 2\langle h^\uparrow, q\rangle + \|q\|^2$, the last step using $\|h^\uparrow\|^2 = \|h\|^2$. The two expressions share $\|h\|^2$ and $\|q\|^2$ — sorting h changes neither — so subtracting leaves only the cross-term: $\mathcal{L}(G) - W_2^2\big(h_\sharp p_U; p^*\big) = 2\big(\langle h^\uparrow, q\rangle - \langle h, q\rangle\big) \ge 0$, nonnegative by (3). Rearranging gives the identity, and the defect $2(\langle h^\uparrow, q\rangle - \langle h, q\rangle)$ vanishes iff $f \circ G$ is a.e. nondecreasing. □

6.2 Proof of Theorem 3.4

Proof. Existence is given by following the gradient up to $G(1) = \mathbf{x}_0$. Uniqueness is achieved by trying to minimize the kinetic energy and failing. Instead we end up with the Euler-Lagrange equation of the energy functional, which gives a sufficient selection criterion.
Existence (i): We pick a curve $\gamma : [a,b] \to \mathcal{A}$ anchored at $\gamma(b) = \mathbf{x}_0$ solving the ODE

$$\frac{d\gamma(y)}{dy} = \frac{\nabla f(\gamma(y))}{\|\nabla f(\gamma(y))\|^2} \quad \Longrightarrow \quad f(\gamma(y)) = y.$$

Since γ stays within the compact, critical-point-free set M_{p^*} the vector field remains bounded there, so the solution can be extended from $(b-\varepsilon, b]$ to the whole interval $[a,b]$ without blowing up or escaping. Reparametrizing the curve via $G(u) = \gamma(P^{-1}(u))$ gives a gradient-like generator satisfying $\mathcal{L}(G) = 0$.
Uniqueness (ii): The three conditions ensuring uniqueness are given by the constraint a) $\mathcal{L}(G) = 0$, parallel acceleration b) $G''(u) = \lambda(u)\nabla f(G(u))$ under initial conditions c) $G(1) = \mathbf{x}_0, G'(1) = \mathbf{v}_0$. We outline the uniqueness of a G satisfying a),b) and c): Differentiating the constrain a) $f(G(u)) = P^{-1}(u)$ twice and inserting b) determines the function $\lambda(u)$ uniquely. Inserting $\lambda(u)$ into b) we get a second order ODE

$$G''(u) = \Phi(u, G, G')$$

with initial condition c) $G(1) = \mathbf{x}_0, G'(1) = \mathbf{v}_0$. Picard–Lindelöf states, that uniqueness of G is given, whenever Φ is locally Lipschitz in (G, G'). The Lipschitz property is satified, by the compactness, continuity and critical point freeness on M_{p^*}. It is left to show that the unique curve G^* satisfying the second order ODE under c) also satisfies a): We define $\varphi(u) = f(G^*(u)) - P^{-1}(u)$ and take the second derivative, with $\lambda(u)$ inserted to get $\varphi''(u) = 0$ for all $u \in [0,1]$. Compatibility then gives $\varphi(1) = \varphi'(1) = 0$. Hence $\varphi = 0$ and $\mathcal{L}(G^*) = 0$. This proves uniqueness. □

6.3 Proof of Proposition 3.6

Proof. We first prove that a geodesic collapses to a gradient flow-line, whenever the gradient points in principal direction of the level-sets. Second we prove that any nearby curve in $\mathcal{S}$ has bigger energy.

We prove i): The derivative of a gradient equation $G_0'(u) = h(u)\,\nabla f(G_0(u))$ yields

$$\begin{aligned} G_0''(u) &= h'(u)\,\nabla f(G_0(u)) + h(u)\big[\nabla f(G_0(u))\big]' && \text{| Product rule} \\ &= h'(u)\,\nabla f(G_0(u)) + h(u)^2\, d^2 f(G_0(u)) \cdot \nabla f(G_0(u)) && \text{| Chain rule} \\ &= \big(h'(u) + h(u)^2\mu(u)\big)\nabla f(G_0(u))) && \text{| Eigenvector} \\ &= \lambda(u)\nabla f(G_0(u)) && \text{| Definition} \end{aligned}$$

We prove ii): **Second variation.** For $\eta \in T_{G_0}\mathcal{S}$ meaning $\eta(u) \perp \nabla f(G(u))$, $\eta(1) = 0$ the index form is

$$I(\eta,\eta) = \int_0^1 \Big(\|\eta'(u)\|^2 - \lambda(u)\langle d^2 f(G_0(u))\eta(u), \eta(u)\rangle \Big)\, du.$$

Positivity. Since $d^2 f(G(u))$ is symmetric with $\nabla f(G(u))$ as an eigenvector, its eigenspaces are orthogonal – the tangential subspace containing $\eta(u)$ is invariant, and the curvature term $\langle d^2 f(G)\eta, \eta\rangle$ picks up only the level set's own shape operator, bounded by $\mu := \max_{M_{p^*}} \|d^2 f\|$. With $\mu' = \max_{[0,1]} |\lambda(u)|$ (compactness of M_{p^*}) and the Poincaré inequality $\int_0^1 \|\eta\|^2\, du \le \frac{1}{2}\|\eta'\|_{L^2}^2$ (from $\eta(1) = 0$) we then get

$$\begin{aligned} I(\eta,\eta) &\ge \int_0^1 \|\eta'(u)\|^2 - \mu\mu' \int_0^1 \|\eta(u)\|^2 du && \text{| Cauchy-Schwarz} \\ &\ge \|\eta'\|_{L^2}^2 - \tfrac{\mu\mu'}{2}\|\eta'\|_{L_2}^2 && \text{| Poincaré Inequality} \\ &= (1 - \tfrac{\mu\mu'}{2})\, \|\eta'\|_{L^2}^2 && \text{| Calculate} \end{aligned}$$

By assuming the support $\mathrm{supp}(p^*) = [a, b]$ to be small enough, continuity on compact sets M_{p^*} achieves $\mu\mu' < 2$ and hence the index-form is positive

$$I(\eta,\eta) > \|\eta'\|_{L_2}^2 > 0$$

for every $\eta \not\equiv 0$. This proves the proposition. □

7 Appendix: Experimental details and further experiments

7.1 Hartmann-6 and Manywell-5

The two energies. Hartmann-6 is the standard six-dimensional test function (Dixon & Szegö, 1978)

$$H(\tilde{x}) \;=\; -\sum_{i=1}^{4} \alpha_i \,\exp\Big(-\sum_{j=1}^{6} M_{ij}\,(\tilde{x}_j - N_{ij})^2\Big), \qquad \tilde{x} \in [0,1]^6,$$

with $\alpha = (1.0,\ 1.2,\ 3.0,\ 3.2)$ and M, N the published constants; we use the `botorch` implementation (Balandat et al., 2020) and write it on the design cube $\mathcal{A} = [-1, 1]^6$ through $\tilde{x} = (x+1)/2$. Its global minimum is -3.32237. Projected gradient descent from $2\cdot 10^4$ uniform starts finds exactly two basins of non-negligible volume — A (minimum -3.32237, 67.9% of the cube) and B (minimum -3.20316, 31.8%) — with 0.3% of starts elsewhere. A sample is labelled by where descent from it converges, which is the standard definition and unambiguous here.

Manywell-5 is a product of one-dimensional double wells,

$$E(x) \;=\; \sum_{i=1}^{5} \big(z_i^2 - 1\big)^2, \qquad z = 1.5\,x, \quad x \in \mathcal{A} = [-1, 1]^5,$$

with $2^5 = 32$ minima of energy 0 at $z_i = \pm 1$, separated by a barrier of height 1 per coordinate at $z_i = 0$. Gradient descent never crosses $z_i = 0$, so the basin of a point is the sign pattern of z: *wells covered* is exact and needs no descent. Uniform draws on the cube have mean energy 2.56 and cover all 32 wells equally.

Metrics All metrics are computed on the low-energy set, the samples with $f \le \tau = P^{-1}(q)$, with one τ per study shared by every method.

W_2 and the defect. W_2 is the one-dimensional Wasserstein-2 distance between the empirical law of the $n = 2048$ realised energies and p^*, computed as the root mean square difference between the sorted energies and P^{-1} at the plotting positions $(i + \frac{1}{2})/N$. It is the same estimator for every row. It has a finite-sample floor: a *perfect* sampler of $N = 2048$ draws scores 0.012 on Hartmann-6 and 0.014 on Manywell-5 (mean over 400 replicates), so a row at those values matches the law to the resolution of the sample and rows below them are not distinguishable.

The defect Δ of Theorem 3.2 is not a distance between distributions and exists only where the run has a latent coordinate. It is computed from the run's own pairs (u_i, e_i) with $e_i = f(G(u_i))$ and $t_i = P^{-1}(u_i)$,

$$\mathcal{L} \;=\; \tfrac{1}{n}\sum_i (e_i - t_i)^2, \qquad W^2_{2,\text{paired}} \;=\; \tfrac{1}{n}\sum_i \big(e_{(i)} - t_{(i)}\big)^2, \qquad \Delta \;=\; \mathcal{L} - W^2_{2,\text{paired}},$$

the second sum over both vectors sorted. $\sqrt{\Delta}$ measures how far the assignment of quantile levels to states is from monotone; it is what distinguishes a run whose histogram of energies is right from one whose $u \mapsto f(G(u))$ is right, and it is the quantity that predicts the yield.

Spread. Every spread figure is computed on the *distinct* points of the low-energy set - rows de-duplicated after rounding to six decimals - not on the raw draws. For a generator with no repeats the two coincide; for a method that resamples a fixed pool with replacement, counting raw draws would let one point that happened to be drawn a thousand times dominate, so de-duplication is the generous reading, and the "n distinct" annotations say when the distinction bites. On Hartmann-6 the figure is then the share of those distinct points whose projected descent converges to basin B ($2 \cdot 10^4$ uniform starts give A 67.9%, B 31.8%, other 0.3%). On Manywell-5 it is the number of sign patterns holding at least 1% of them — a threshold, so a well reached by a single point does not count; uniform coverage would put 3.1% in each.

yield, distinct points. *Yield* is the share of drawn samples that fell below τ; it is 1 by construction for our generator that matches the law, since it draws $u \sim \mathcal{U}[0, q]$, and about q for a method that relies on rejection sampling. *Distinct* counts the distinct rows after rounding to six decimals; it is quoted only where it changes how the spread must be read, i.e. for a method that resamples a fixed pool with replacement.

The generator. The anchor is encoded exactly, not penalised:

$$G(u; \mathbf{x}_0) \;=\; \tanh\big(\operatorname{atanh}\mathbf{x}_0 + (1 - \xi)\, h_\theta(\xi, \mathbf{x}_0)\big),$$

so $G(1; \mathbf{x}_0) = \mathbf{x}_0$ holds by construction and the residual penalty $\lambda_0 = 0$ of loss (13). h_θ is a conditional MLP of hidden widths $(128, 256, 128, 64)$ with the anchor concatenated to the input; the latent is one-dimensional — the quantile alone, re-timed as below — and tanh keeps the output inside the design cube.

Optimisation. Adam, 5000 steps, batch 1024, learning rate 10^{-3} on a cosine schedule, weight decay 0, no gradient clipping, one anchor drawn per sample rather than per step, five seeds.

The input coordinate. The quantile is a poor coordinate for a long-tailed law such as a truncated exponential. We change coordinate to

$$\xi \;=\; \left(\frac{e - f_{\min}}{f(\mathbf{x}_0) - f_{\min}}\right)^{1/2}, \qquad e = P^{-1}(u).$$

The exponent $\frac{1}{2}$ is the constant-speed coordinate near a non-degenerate interior minimum, where arc length grows like $\sqrt{e - e_{\min}}$; it is not exact arc length, and the right exponent is a property of the landscape. For each anchor, ξ is a strictly increasing bijection of $[0, 1]$ with $\xi(1) = 1$, so the anchor constraint is unaffected. This input coordinate improves W_2 from 0.136 to 0.025 on Hartmann-6 and from 0.127 to 0.03 on Manywell-5.

Evaluation. $N = 2048$ draws per run. The W_2 and $\sqrt{\Delta}$ columns use $u \sim \mathcal{U}[0, 1]$; the low-energy columns use $u \sim \mathcal{U}[0, q]$, which is the controlled sampling an anchor-free method can only obtain by rejection.

Baselines. SOURCERER (Vetter et al., 2024), Algorithm 1, entropy weight $\lambda = 0.1$, 5000 steps, batch 1024, learning rate 10^{-3}, sampler width 100, Kozachenko–Leonenko entropy with $k = 5$, three seeds. DCI (Butler et al., 2018) re-weights a pool of 10^6 uniform draws against a 400-bin density of states, three seeds. The continuous GFlowNet (Lahlou et al.) is trajectory balance on the continuous hypercube with reward e^{-10E}, 10 000 steps, mixture-of-Beta forward kernel with 5 components, minimum increment 0.1, MLP policies 256×3, forward batch 100, learning rate 10^{-4}, z-multiplier 100, random-action probability 0.1, three seeds. Multi-start gradient descent is projected Adam, 200 steps, lr 0.05, from uniform starts, five seeds. The anchor-free methods are evaluated on N/q draws so that their low-energy set has comparable size, and every row of a table is filtered with the same τ.

7.1.1 SOURCERER'S ENTROPY WEIGHT, SWEPT

Table 1 reports SOURCERER at $\lambda = 0.1$, the default of its benchmark tasks, and its basin-B and well columns carry the diversity comparison. Because λ is the only knob that trades spread against matching, we sweep it over $\{0.02, 0.1, 0.35, 0.7\}$ with the law, the threshold, the estimator and the seed count held at Table 1's values, so a cell differs from its row in exactly the entropy weight.

Table 2: SOURCERER (Algorithm 1) with the entropy weight swept, against our level-set row. Three seeds, mean ± sample sd. "deepest in B" is the lowest energy the run reaches inside basin B, whose own minimum is -3.20316 and whose samples therefore qualify for the low-energy set only in $[-3.203, -3.169]$; "–" means the run places nothing in B below -3.05. Rows above $\lambda = 0.1$ do not match the law, so their low-energy columns are empty.

SOURCERER (Algorithm 1) under Table 1's law, threshold and estimator, with the entropy weight λ swept. Three seeds, mean ± sample sd. $\lambda = 0.1$ is the row of Table 1; our level-set row is repeated for reference.

	HARTMANN-6					MANYWELL-5			
λ	$W_2 \downarrow$	best E	basin B	*all* samples in B	deepest in B	$W_2 \downarrow$	best E	wells	yield
ours, $p_0 = \mathcal{U}$(level set)	0.025 ±0.004	−3.310 ±0.001	35 ±1%	39 ±0%	−3.202 ±0.001	0.030 ±0.006	0.007 ±0.001	31.6 ±0.5	37 ±4%
$\lambda = 0.02$	0.043 ±0.003	−3.316 ±0.002	0 ±0%	25 ±1%	–	0.031 ±0.013	0.010 ±0.005	8.7 ±4.0	6 ±4%
$\lambda = 0.1$	0.057 ±0.034	−3.311 ±0.004	0 ±0%	47 ±14%	−3.135 ±0.002	0.035 ±0.003	0.008 ±0.003	18.3 ±3.5	5 ±2%
$\lambda = 0.35$	2.920 ±0.008	–	–	31 ±1%	–	1.484 ±0.013	–	–	0 ±0%
$\lambda = 0.7$	2.937 ±0.006	–	–	33 ±1%	–	4.448 ±0.109	–	–	0 ±0%

Two readings. First, the basin-B column is 0% at *every* λ at which SOURCERER matches the law: the entry in Table 1 is not an unlucky setting. Above $\lambda = 0.1$ the method abandons the law — W_2 rises to 2.9 and 1.5 to 4.4, and no draw clears τ — so the 31 to 33% it then shows in B is the cube's volume prior rather than coverage. Second, λ trades matching against spread and both ends are dominated: its best matcher, $\lambda = 0.02$, is also its least diverse (25% of samples in B, 8.7 wells), while $\lambda = 0.1$ spreads most and matches worst. Our row is better than, or tied with, the best λ on both axes at once.

7.2 THE CONDITIONAL EXTENSION: p^* CONDITIONED ON THE ANCHOR

The main text keeps one law for every anchor, which forces p_0 onto the level set $f^{-1}(P^{-1}(1))$. That is restrictive: the level set has to be sampled, and in a real design problem the practitioner rarely has a reference set, only a reference *design*. The extension that removes the restriction is to let the law depend on the anchor.

Construction. Prescribe a family $\{p^*(\cdot \mid \mathbf{x}_0)\}$ rather than a single law, each member satisfying Definition 3.1 and let

$$P^{-1}(1 \mid \mathbf{x}_0) \;=\; f(\mathbf{x}_0) \qquad \text{for } p_0\text{-a.e. } \mathbf{x}_0. \tag{14}$$

The loss of §3 becomes

$$\mathcal{L}(G; p_0) \;=\; \mathbb{E}_{\mathbf{x}_0 \sim p_0} \int_0^1 \left| f(G(u; \mathbf{x}_0)) - P^{-1}(u \mid \mathbf{x}_0) \right|^2 du. \tag{15}$$

Now p_0 is free — any law on the design space — and the anchor convention $G(1; \mathbf{x}_0) = \mathbf{x}_0$ stays exact, because the law moves with the anchor. Two families are natural:

(T) *truncation*: $p^*(\cdot \mid \mathbf{x}_0)$ is a fixed law p^* restricted to $[a, f(\mathbf{x}_0)]$. The energy scale is absolute — the floor a means the same thing for every anchor — and only the upper end moves. This is what every anchor law off the level set uses in this paper.

(R) *relative*: $p^*(\cdot \mid \mathbf{x}_0) = f(\mathbf{x}_0) - r$ for one improvement law $r \geq 0$. Nothing absolute is assumed about f; the request is "improve this design by an amount distributed as r". This is the form a practitioner states a design brief in.

Theorem 3.2 applies to each anchor separately, so at $\mathcal{L} = 0$ the curve $G(\cdot; \mathbf{x}_0)$ is a gradient-like generator for $p^*(\cdot \mid \mathbf{x}_0)$ for p_0-a.e. $\mathbf{x}_0$. What the *marginal* of the generated energies matches is then the mixture

$$\bar{p}^*(y) \;=\; \mathbb{E}_{\mathbf{x}_0 \sim p_0}\big[\, p^*(y \mid \mathbf{x}_0)\,\big],$$

and Theorem 3.2's bound survives at that level.

Proposition 7.1 (Conditional consistency)**.** *Let p_0 be a law on $\mathbb{R}^D$ and $\{p^*(\cdot \mid \mathbf{x}_0)\}$ a family of laws each with full interval support in $f(\mathcal{A})$. Write $\bar{p}_G$ for the law of $f(G(U; X_0))$ with $U \sim \mathcal{U}[0,1]$ independent of $X_0 \sim p_0$. Then*

$$W_2^2\big(\bar{p}_G,\ \bar{p}^*\big) \;\leq\; \mathcal{L}(G; p_0).$$

Proof. The inequality follows directly from Theorem 3.2(i) applied for each $\mathbf{x}_0$ and integrated in p_0, the defects being non-negative. □

The prescribed object is the family, and the marginal is a *consequence* of it and of p_0, not a specification. So the pair (p_0, G) solves $f_\# \pi = \bar{p}^*$, not $f_\# \pi = p^*$: to hit a prescribed marginal one would have to choose the family so that its mixture is that marginal, which needs $f_\# p_0$ — the density of states our construction was designed to avoid estimating. Putting p_0 on a level set is exactly the degenerate case in which the family is constant, the mixture is a single law, and no such estimate is needed.

The relative family (R) is the case where this costs nothing, because it prescribes a different observable. At $\mathcal{L} = 0$ the *improvement* $I = f(\mathbf{x}_0) - f(G(U; \mathbf{x}_0))$ satisfies $I \sim q$ for every anchor, hence $I \sim r$ marginally: the pair (p_0, G) solves the stochastic inverse problem exactly, for the anchor-relative observable $(\mathbf{x}_0, x) \mapsto f(\mathbf{x}_0) - f(x)$.

7.3 Adversarial examples: details

The design set is the L_∞ ball of radius $\varepsilon = 8/255$ around CIFAR-10 image #123 at 32×32, intersected with $[0,1]^{3072}$. The generator is a conditional MLP with hidden widths $(256, 512, 512, 256)$ and latent dimension 1, with the anchor encoded exactly by the convex combination $G(u; \mathbf{x}_0) = u\,\mathbf{x}_0 + (1-u)\,b(h_\theta(u, \mathbf{x}_0))$, b the map onto the box; u is the quantile here, with no re-timing. Adam, 1500 steps, learning rate 10^{-3}, batch 64, no weight decay and no gradient clipping. Evaluation draws 512 samples per run, 3 seeds.

The classifier and the reference image. C is the pretrained CIFAR-10 ResNet-20 (He et al., 2016) of Chen, 92.6% top-1 on the CIFAR-10 test set, evaluated in inference mode on inputs normalised by the standard CIFAR-10 channel statistics. It is an undefended model: no adversarial training, no certified defence, so this study measures control over the law rather than attack strength against a hardened target. The reference image is CIFAR-10 test image #123, which the model calls a *bird* with probability 0.99987, so $f(\mathbf{x}_0) = \log p_{\mathrm{clf}}(y \mid \mathbf{x}_0) = -1.3 \cdot 10^{-4}$ and the reference class is both the label and the prediction.

The distance to the clean image. Both our generator and Sourcerer realise the prescribed law — $W_2 = 0.101 \pm 0.031$ against 0.162 ± 0.005 — so what separates them is not whether they solve the problem but how. Measured as the mean per-pixel perturbation in units of ε, ours spends 0.101 and Sourcerer 0.870. A uniform draw from the threat ball spends 0.5 in expectation, so Sourcerer's attacks sit *further* from the image than random points of the ball: its distance is set by the entropy weight and the geometry of the ball, not by the task. Adding the kinetic selection term at $\lambda_{\mathrm{kin}} = 3{\cdot}10^{-4}$ halves ours again, to 0.055, at $W_2 = 0.109 \pm 0.029$.

7.4 Appendix: A Simple Inverse Problem

An inverse problem consists of input variables describing a (physical) state and a system or model that maps these variables into observations. The goal is to infer either the variables (or the system) given observations and the system (or variables), respectively (Hansen, 1998; Kaipio & Somersalo, 2005).

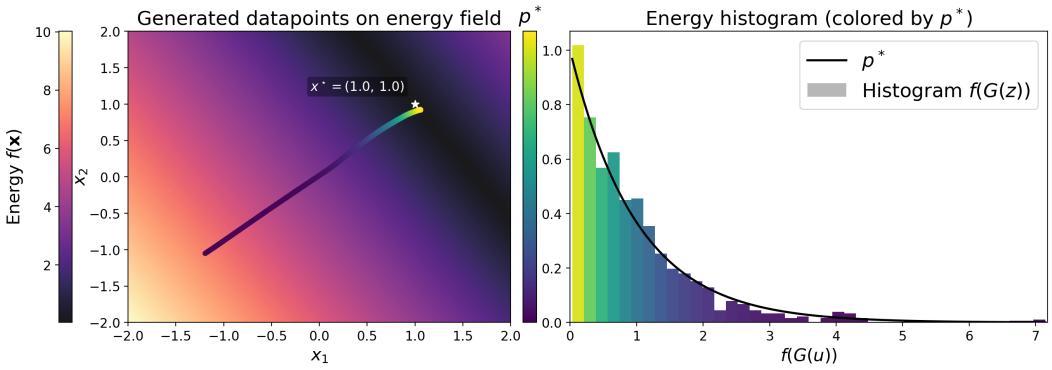


Figure 3: Our method applied on the ill-posed linear inverse problem. The generated solutions concentrate close to the true solution $\mathbf{x}^* = (1, 1)^T$ and samples further away have greater residuals. The exponential distribution p^* serves as an energy prior.

Classical optimization is a special case of such an inverse problem. Consider the linear equation $A\mathbf{x}^* = \mathbf{b}$. Here, A represents the system or model, $\mathbf{b}$ the observation, and $\mathbf{x}^*$ is the true input variable. In real-world problems, measurements are typically noisy and the matrix A ill-conditioned, meaning that small perturbations $\mathbf{b}_\varepsilon = \mathbf{b} + \varepsilon$ can lead to large perturbations $\mathbf{x}^*_\varepsilon$ of the solution $\mathbf{x}^*$. An example of such ill-conditioned problem can be found in Hansen (1998) where

$$A = \begin{pmatrix} 0.16 & 0.1 \\ 0.17 & 0.11 \\ 2.02 & 1.29 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} 0.26 \\ 0.28 \\ 3.31 \end{pmatrix} \quad \text{and thus} \quad \mathbf{x}^* = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{16}$$

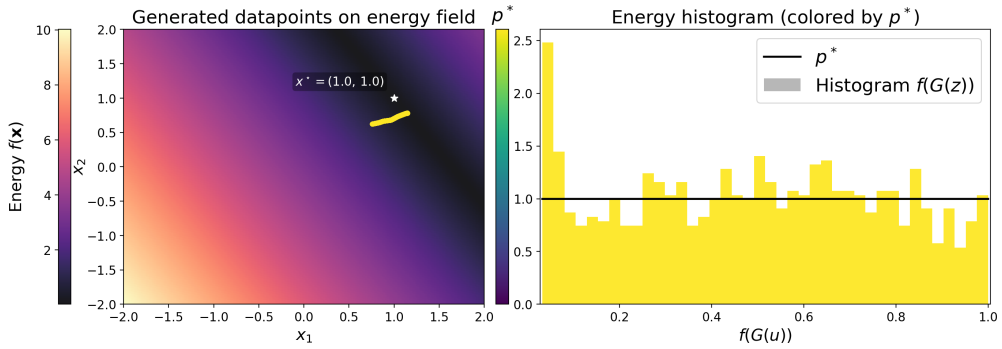
which can be easily verified. With $f(\mathbf{x}) = \|A\mathbf{x} - \mathbf{b}_\varepsilon\|_2^2$, the noisy solution to classical optimization becomes the solution of the least square

$$\mathbf{x}^*_\varepsilon = \arg\min_{\mathbf{x} \in \mathbb{R}^2} \|A\mathbf{x} - \mathbf{b}_\varepsilon\|_2^2 \tag{8}$$

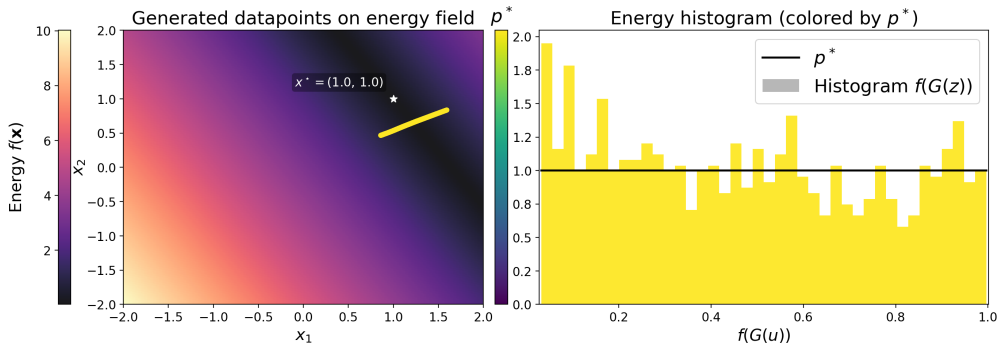
where $\mathbf{b}_\varepsilon = \mathbf{b} + (0.01, -0.03, 0.02)^T$ is a rather small perturbation of $\mathbf{b}$. For this particular perturbation ε, the least square solution is given by $\mathbf{x}^*_\varepsilon = (7.01, -8.4)^T$. Even though $\mathbf{x}^*_\varepsilon$ has residuals in the order of the noise magnitude, $\|A\mathbf{x}^*_\varepsilon - \mathbf{b}_\varepsilon\|_2^2 = 4.7 \cdot 10^{-4}$, it is far away from the unperturbed solution $\mathbf{x}^*$. Typically, to reduce the sensitivity to noise and find good approximate solutions, additional valuable information is injected. For example, in Bayesian inference, a Gaussian prior on $\mathbf{x}$ penalizes solutions with high magnitude.

Our new optimization approach (3) handles the inverse problem nicely. In Figure 3, we show generated samples $G(U), U \sim \mathcal{U}([0, 1])$ (left) with our method using an exponential target distribution p^* (right). The exponential energy distribution regularizes the ill-posed problem: the generated samples concentrate around the unperturbed solution $\mathbf{x}^*$ and the further away samples are from $\mathbf{x}^*$, the greater the residuals as desired (histogram colors correspond to samples on the left). Therefore, p^* serves as a regularizer similar to a prior on $\mathbf{x}$ in Bayesian inference. However, in strong contrast to Bayesian methods, our prior is on the energies f and not on $\mathbf{x}$.

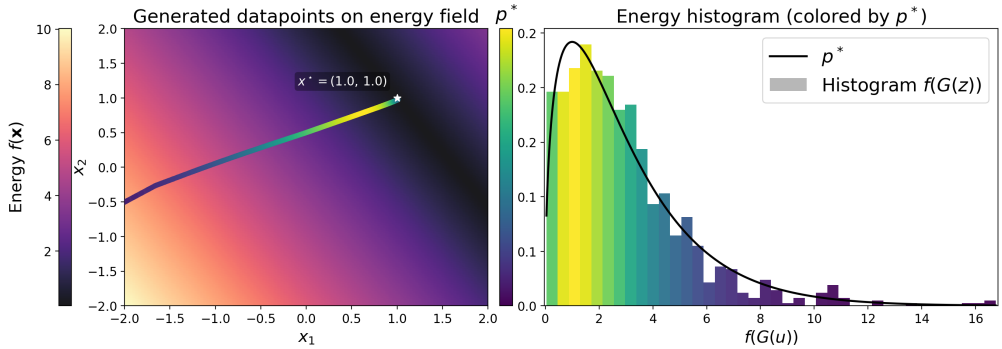
We have chosen the exponential distribution because we know from equation (16) that a good solution is close to $f(\mathbf{x}) = 0$. In figure 4, we show that other choices of p^* yield good solutions as well, as long as the mass concentrates around 0. This illustrates the trade-off of our method: at the one hand, p^* provides design freedom, on the other hand, bad prior information can result in undesired samples.



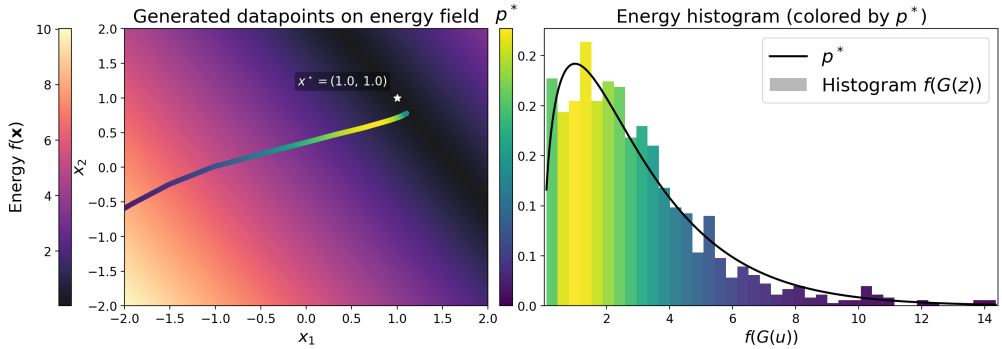
(a) Inverse problem with target distribution $p^* = \mathcal{U}(0, 1)$ and Inverse Transform Sampling loss.



(b) Inverse problem with target distribution $p^* = \mathcal{U}(0, 1)$ and Wasserstein loss.



(c) Inverse problem with target distribution $p^* = \chi^2(3)$ Inverse Transform Sampling loss.



(d) Inverse problem with target distribution $p^* = \chi^2(3)$ Wasserstein loss.

Figure 4: Additional inverse-problem experiments for different target distributions. Each panel shows the learned sample trajectory on the energy field together with the induced energy histogram $f(G_\theta(u))$ using the Inverse Transform Loss and the Wasserstein loss.