A LIFT FOR INPUT-CONVEX NEURAL NET TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Input-convex neural nets parametrize the convex potentials of density models and transport maps, and their convexity requires the inter-layer weights to be non-negative. Projected gradient descent enforces this by projecting after each step, and due to mini-batch noise the boundary is re-crossed indefinitely, which leads to an active set the projection never identifies. The differentiable alternative, direct softplus, optimizes a free latent weight through a softplus positivity map whose derivative attenuates the gradient exponentially where the weight is negative—the shoulder—so a coordinate that reaches it stays for an exponentially long time. To keep this unconstrained parametrization without its slow escape, we propose the lift, which replaces the free latent weight by a learnable slack plus an unconstrained network—the body—that takes a permutation-invariant summary of the training batch as input. The latent weight thus varies with the batch before the positivity map, and couples to the gradient formed on it. We show that this coupling enters the variance of the update to the latent weight at first order in the fluctuation, and that the slack, the batch dependence and the shared batch are each needed for it to act. Where the coupling aligns positively with the loss curvature, that variance is larger under the lift than under direct softplus, and a coordinate leaves the shoulder sooner. We compare the lift with the two existing methods on several applications. Where a constrained weight of direct softplus reaches the shoulder and does not leave, the lift fits the target more closely and reaches the same reconstruction about three times sooner. Where almost none reaches it, the methods agree.

1 INTRODUCTION

Input-convex neural nets (ICNNs) (Amos et al., 2017) are feed-forward networks whose output is convex in the input. They parametrize the negative log-density of a log-concave energy-based model (Lin & Ba, 2023; Vesseron et al., 2025), the potential of a convex potential flow (Huang et al., 2021; Kan et al., 2022), the transport maps that sample the posterior of a Bayesian inverse problem (Taghvaei & Hosseini, 2022; Wang et al., 2025), the Brenier potential of neural optimal transport (Amos, 2023; Bunne et al., 2022; Korotin et al., 2021; Makkuva et al., 2020), and the convex regularizers of learned image reconstruction (Cui et al., 2024; Ehrhardt et al., 2026; Mukherjee et al., 2024). These uses share one parametrization, which requires convex non-decreasing activations and non-negative inter-layer weights $\boldsymbol{\theta} \geq \mathbf{0}$ (Amos et al., 2017).

Two methods enforce this constraint in practice. Projected gradient descent (PGD) (Amos et al., 2017) returns any negative weight to zero by the projection $\max(\boldsymbol{\theta}, \mathbf{0})$ after each step. While a coordinate held at zero receives the full gradient and leaves as soon as that update points inward, mini-batch noise sends it back. This noise does not vanish as the mean gradient does, so the boundary is re-crossed indefinitely and the projection never identifies the coordinates at zero (Duchi & Ruan, 2021). The differentiable alternative, direct softplus, optimizes a free latent weight through the softplus positivity map, whose derivative is exponentially small where that weight is negative. On the shoulder, where this derivative is small, the time to leave is exponential in the barrier height over the attenuated variance (Kramers, 1940). The two methods therefore sit at opposite extremes in how long a coordinate stays near zero once training has taken it there.

We propose the lift, which shortens the time a coordinate spends on the shoulder and keeps the differentiable unconstrained parametrization of direct softplus, changing only what feeds the positivity map. Borrowing the extension of PDE-constrained inverse problems (Symes et al., 2020; van Leeuwen & Herrmann, 2013), we replace the free latent weight by a learnable slack plus the output

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.

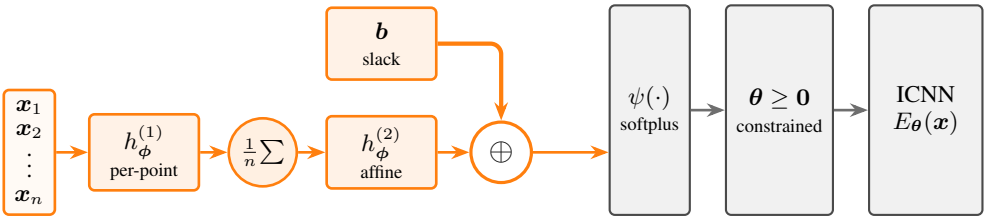


Figure 1: The lift of (2): a permutation-invariant body reads the batch, its output plus a slack passes through the softplus positivity map, and the result is the constrained weight of the ICNN.

of an unconstrained hypernetwork (Ha et al., 2017)—the body—applied to a permutation-invariant summary (Zaheer et al., 2017) of the training batch (Figure 1). Since the batch is resampled at each iteration, the latent weight fluctuates with it, and that fluctuation couples to the gradient on the same batch without passing through the derivative that attenuates it. We show theoretically that the slack, the batch dependence, and the shared batch are each necessary for this coupling (Theorem 2 and the substitutions that follow it). When the coupling is positive and the remainder bounded, the latent weight explores the shoulder faster than under direct softplus, which lowers the exponent of the escape time from the shoulder (Theorem 1 and Corollary 1).

We evaluate the lift against direct softplus on energy-based models, convex potential flows, a transport map for a high-dimensional posterior, and a convex regularizer for image reconstruction. We also run PGD on the energy-based models and on the transport map and the image regularizer, since the published work behind them uses it. The lift is closer to every one-dimensional target than direct softplus on every paired run, and below PGD on every seed of three of the four. When a constrained weight of direct softplus reaches the shoulder and stays, the lift fits the target more closely on every energy-based run that returns a likelihood, and reaches the same reconstruction about three times sooner on the image regularizer. When almost no weight does, the methods agree. Initialization schemes for ICNNs set the starting values of the constrained weights (Hoedt & Klambauer, 2023; Hundrieser et al., 2026), not how long a weight stays near zero. We run the transport map from three initializations, and on the image regularizer direct softplus does not train at all from the one Hoedt & Klambauer (2023) prescribe. What decides whether the methods differ is not the application but how long a weight stays on the shoulder once it has reached it.

The paper is organized as follows. We state the lift in Section 2, derive the escape time it changes in Section 3, report the results in Section 4, and place it among the alternatives in Section 5.

2 The Shoulder and the Lift

The constraint $\boldsymbol{\theta} \geq \mathbf{0}$ on the inter-layer weights of an ICNN (Amos et al., 2017) is the sole obstacle to training one with standard gradient methods. We write $\boldsymbol{\theta} \in \mathbb{R}^d$ for these weights and leave the network $E_{\boldsymbol{\theta}}$'s unconstrained parameters implicit. Our running example is a log-concave energy-based model (EBM) $p(\boldsymbol{x}) \propto \exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))$ trained by a stochastic gradient method on a loss $\mathcal{L}$ over mini-batches $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$.

2.1 The Shoulder of the Positivity Map

The differentiable method applies a monotone positivity map $\boldsymbol{\theta} = \psi(\tilde{\boldsymbol{\theta}})$, with $\psi : \mathbb{R} \to [0, \infty)$ acting coordinate-wise, and optimizes the free latent weight $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^d$ in its place. Monotonicity matters: a non-monotone map with the same infimum, such as the square (Berthier, 2023), has a derivative that vanishes exactly at zero, so a weight that reaches zero receives no gradient at all. The standard choice is the softplus, $\psi(u) = \log(1 + e^u)$, whose derivative is the logistic sigmoid $\psi'(u) = (1 + e^{-u})^{-1} \in (0, 1)$, and we call this method direct softplus. By the chain rule the gradient that reaches the latent weight is the one with respect to the constrained weight, multiplied elementwise by that derivative,

$$\boldsymbol{g} \;=\; \nabla_{\tilde{\boldsymbol{\theta}}} \mathcal{L} \;=\; \psi'(\tilde{\boldsymbol{\theta}}) \odot \nabla_{\boldsymbol{\theta}} \mathcal{L}, \tag{1}$$

and $\psi'(u)$ decays as e^u for negative u. We call the region in which $\psi'(\tilde{\boldsymbol{\theta}}_j) \leq \sigma_{\mathrm{sh}}$ for a threshold $\sigma_{\mathrm{sh}} \ll 1$ the *shoulder*, and the fraction of the d constrained coordinates on it the *shoulder occupancy*,

both at $\sigma_{\mathrm{sh}} = 0.05$ throughout, which puts the shoulder at latent weights below -2.94. A coordinate there has its mean update and its mini-batch noise scaled by that derivative, so at most a fraction σ_{sh}^2 of the gradient variance reaches the latent iterate. The time to cross a barrier is exponential in the ratio of the barrier height to that variance, up to the step size (Hänggi et al., 1990; Kramers, 1940), which Theorem 1 states for this setting. Nothing in the parametrization keeps a constrained weight off the shoulder, either at initialization or over the course of training.

The projected alternative keeps the raw weights and follows each gradient step by the coordinate-wise projection $\max(\cdot, \mathbf{0})$. Since PGD does not differentiate through this projection, each raw coordinate receives $\nabla_{\boldsymbol{\theta}}\mathcal{L}$ undiminished and leaves the boundary as soon as its update points inward. A visit to the boundary—i.e., a coordinate's *dwell*—is therefore the memoryless one of the projected recursion in Proposition 1. We compare PGD with the lift empirically.

2.2 The lift

The lift replaces the free latent weight by the sum of a learnable vector $\boldsymbol{b} \in \mathbb{R}^d$—the *slack*—and a permutation-invariant map $h_{\boldsymbol{\phi}}$ of the mini-batch, the *body*,

$$\boldsymbol{\theta} = \psi(\tilde{\boldsymbol{\theta}}), \qquad \tilde{\boldsymbol{\theta}} = \boldsymbol{b} + h_{\boldsymbol{\phi}}(\boldsymbol{X}). \tag{2}$$

On the energy-based models and the image regularizer the body emits the unconstrained parameters as well, which need no positivity map. Invariance to the ordering of the batch follows from pooling the per-point contributions (Zaheer et al., 2017). Training minimizes the loss over the lifted parameters,

$$\min_{\boldsymbol{\phi},\,\boldsymbol{b}} \; \mathcal{L}\big(\psi(\boldsymbol{b} + h_{\boldsymbol{\phi}}(\boldsymbol{X}))\big), \tag{3}$$

and fixing $h_{\boldsymbol{\phi}} \equiv \mathbf{0}$ recovers direct softplus, so that the slack is the free latent weight defined in Section 2.1 (Figure 1). The body reads the batch on which the gradient of $\mathcal{L}$ is formed at the same iteration. The body acts during training alone: at evaluation the latent weight is computed once and the ICNN is a fixed network with direct softplus's architecture and parameter count.

3 Diffusion and Escape

The results of this section are proved in Appendix A and machine-checked in Lean 4 against mathlib.

3.1 The fluctuation and its coupling

At fixed $\boldsymbol{\phi}$ and $\boldsymbol{b}$ the latent weight of (2) varies with the batch by

$$\delta\tilde{\boldsymbol{\theta}} \;=\; h_{\boldsymbol{\phi}}(\boldsymbol{X}) - \mathbb{E}_{\boldsymbol{X}}\big[h_{\boldsymbol{\phi}}(\boldsymbol{X})\big], \tag{4}$$

the *fluctuation* the lift adds to the training dynamics. While direct softplus trains that weight as a parameter no batch changes, the lift redraws it at each step. The constrained weights of PGD are stored parameters as well, and its coupling vanishes for the same reason (Theorem 4).

3.2 Hypotheses

Fix a reference iterate on the band, the mean latent weight $\bar{\tilde{\boldsymbol{\theta}}} = \boldsymbol{b} + \mathbb{E}_{\boldsymbol{X}}[h_{\boldsymbol{\phi}}(\boldsymbol{X})]$, and a unit direction $\boldsymbol{e}_b$ of the slack subspace, the bias channel. Both methods are read there, under one batch law and one pullback landscape $\tilde{\mathcal{L}}$ along $\boldsymbol{e}_b$, on which the reference action α of Theorem 1(iii) is shared. With $\boldsymbol{g}_0(\boldsymbol{X}) = \nabla_{\tilde{\boldsymbol{\theta}}}\mathcal{L}(\tilde{\boldsymbol{\theta}}; \boldsymbol{X})$ the gradient of a batch at that iterate,

$$\boldsymbol{\xi} = \boldsymbol{g}_0 - \mathbb{E}_{\boldsymbol{X}}[\boldsymbol{g}_0], \quad \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} = \mathbb{E}_{\boldsymbol{X}}[\nabla^2_{\tilde{\boldsymbol{\theta}}}\mathcal{L}(\bar{\tilde{\boldsymbol{\theta}}}; \boldsymbol{X})], \quad \boldsymbol{V} = \mathrm{Cov}_{\boldsymbol{X}}[\delta\tilde{\boldsymbol{\theta}}], \quad \boldsymbol{\Sigma}_\xi = \mathbb{E}_{\boldsymbol{X}}[\delta\tilde{\boldsymbol{\theta}}\,\boldsymbol{\xi}^\top],$$

the gradient of a batch at its own iterate is written $\boldsymbol{g} = \boldsymbol{g}_0 + \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\,\delta\tilde{\boldsymbol{\theta}} + \boldsymbol{r}_2$, which defines the remainder $\boldsymbol{r}_2$, and

$$\sigma^2_{\mathrm{eff}} = \boldsymbol{e}_b^\top \mathrm{Cov}_{\boldsymbol{X}}[\boldsymbol{g}]\,\boldsymbol{e}_b,$$
$$\varrho(\boldsymbol{e}) = 2\,\boldsymbol{e}^\top \mathrm{Cov}_{\boldsymbol{X}}[\boldsymbol{g}_0, \boldsymbol{r}_2]\,\boldsymbol{e} + 2\,\boldsymbol{e}^\top \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} \mathrm{Cov}_{\boldsymbol{X}}[\delta\tilde{\boldsymbol{\theta}}, \boldsymbol{r}_2]\,\boldsymbol{e} + \boldsymbol{e}^\top \mathrm{Cov}_{\boldsymbol{X}}[\boldsymbol{r}_2, \boldsymbol{r}_2]\,\boldsymbol{e}.$$

Throughout, $\boldsymbol{g}_0$, $\delta\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{r}_2$ are square integrable, the mean curvature $\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}$ is symmetric, and the minibatch noise along $\boldsymbol{e}_b$ is non-degenerate, $\sigma^2_{\mathrm{obj}} = \boldsymbol{e}_b^\top \mathrm{Var}_{\boldsymbol{X}}[\nabla_{\boldsymbol{\theta}}\mathcal{L}]\,\boldsymbol{e}_b > 0$.

Assumption A1 (Itô surrogate with i.i.d. batches)**.** *The lifted parameters $\boldsymbol{\varphi} = (\boldsymbol{b}, \boldsymbol{\phi})$ of* (3) *are trained by stochastic gradient descent with step size $\eta > 0$ on mini-batches $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ that are independent and identically distributed, and the recursion is read as the Itô process $d\boldsymbol{\varphi} = -\nabla_{\boldsymbol{\varphi}} \tilde{\mathcal{L}}(\boldsymbol{\varphi})\, dt + \sqrt{\eta}\, \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\varphi})\, d\boldsymbol{B}_t$, with $\tilde{\mathcal{L}}(\boldsymbol{\varphi}) = \mathbb{E}_{\boldsymbol{X}}[\mathcal{L}(\psi(\boldsymbol{b} + h_{\boldsymbol{\phi}}(\boldsymbol{X})))]$ the pullback landscape and $\boldsymbol{\Sigma}(\boldsymbol{\varphi})$ the covariance of the per-step update, whose entries are Lipschitz in $\boldsymbol{\varphi}$ on a region containing the escape interval, on which the drift $\nabla_{\boldsymbol{\varphi}} \tilde{\mathcal{L}}$ is continuous.*

We take a continuous-time process as a surrogate for the recursion, following the stochastic modified equations of Li et al. (2017), and read both methods under one surrogate and one batch law, so that the two exponents inherit the same error. Parts (i) and (ii) hold without it.

Assumption A2 (Band and barrier)**.** *The positivity map ψ is continuously differentiable and non-decreasing, and on the operative band $\mathcal{B} \subset \mathbb{R}$ its derivative lies within a tolerance ε of one constant, $|\psi'(w) - \sigma_s| \le \varepsilon$ for every $w \in \mathcal{B}$, with $\sigma_s = \psi'(\bar{w}_s) > 0$ at a point $\bar{w}_s$ of $\mathcal{B}$ and $0 \le \varepsilon < \sigma_s$. Along $\boldsymbol{e}_b$ the pullback landscape $\tilde{\mathcal{L}}$ is continuous on an escape interval $[a, b]$ and differentiable in its interior, and there are $a < w_b < w_s < b$ such that w_s maximizes and w_b minimizes $\tilde{\mathcal{L}}$ on $[a, b]$, with $\tilde{\mathcal{L}}(w_b) < \tilde{\mathcal{L}}(w_s)$ and $\tilde{\mathcal{L}}$ non-decreasing on $[w_b, w_s]$.*

For softplus the first clause holds automatically (Appendix A.1), so its content is that the operative iterates occupy such a band, one on which σ_s falls below the shoulder threshold of Section 2.1.

Write

$$c \;=\; 2\, \boldsymbol{e}_b^{\top} \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} \boldsymbol{\Sigma}_{\xi}\, \boldsymbol{e}_b \;+\; \varrho(\boldsymbol{e}_b)$$

for the *net coupling* at that iterate along $\boldsymbol{e}_b$: the frozen coupling of the latent-weight fluctuation with the gradient noise of its own batch, net of the remainder channel $\varrho = \varrho(\boldsymbol{e}_b)$. The method does not fix its sign, and the ordering that Corollaries 1 and 2 carry rests on $c > 0$, a hypothesis the experiments do not measure.

3.3 Diffusion excess and escape time

Theorem 1 adds to the exponential order of the classical escape law (Kramers, 1940) an exact accounting of the diffusion coefficient, under a reading of σ^2_{eff} as constant across the barrier band, whose bracketed effective actions we do not show to separate.

Theorem 1 (The coupling channel and the time on the shoulder)**.** *Fix the reference iterate $\bar{\tilde{\boldsymbol{\theta}}}$, the band $\mathcal{B}$ and the unit direction $\boldsymbol{e}_b$ of the slack subspace of Section 3.2, the bias channel. Parts* (i) *and* (ii) *use the regularity stated there and the band clause of Assumption A2, and nothing else; part* (iii) *adds Assumption A1 and the barrier clause of Assumption A2. Then the following hold.*

(i) Channels. *With no term discarded,*

$$\sigma^2_{\mathrm{eff}} \;=\; \underbrace{\sigma_s^2 \sigma^2_{\mathrm{obj}}}_{\text{objective contribution}} \;+\; \underbrace{2\, \boldsymbol{e}_b^{\top} \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} \boldsymbol{\Sigma}_{\xi}\, \boldsymbol{e}_b}_{\text{coupling term}} \;+\; \underbrace{\boldsymbol{e}_b^{\top} \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} \boldsymbol{V} \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\, \boldsymbol{e}_b}_{\text{jitter term}} \;+\; \underbrace{\varrho}_{\text{remainder channel}}, \tag{5}$$

the direct softplus value at that iterate being the objective contribution alone, within the bracket of Assumption A2. The jitter term is $\mathrm{Var}_{\boldsymbol{X}}[\langle \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} \boldsymbol{e}_b, \delta\tilde{\boldsymbol{\theta}} \rangle] \ge 0$, so $\sigma^2_{\mathrm{eff}} - \sigma_s^2 \sigma^2_{\mathrm{obj}} \ge c$ for any coupling.

(ii) Orders. *Scale the fluctuation and the remainder at a fixed batch law, $\delta\tilde{\boldsymbol{\theta}} = t\, \boldsymbol{u}$ and $\boldsymbol{r}_2 = t\, \boldsymbol{w}_1 + t^2 \boldsymbol{w}_2$. Then $\sigma^2_{\mathrm{eff}}(t) - \sigma_s^2 \sigma^2_{\mathrm{obj}}$ is a polynomial in t with no constant term, whose linear coefficient pairs $\boldsymbol{g}_0$ with $\boldsymbol{u}$ through $\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}$ and with $\boldsymbol{w}_1$ directly, and whose quadratic coefficient contains the jitter term $\boldsymbol{e}_b^{\top} \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} \mathrm{Cov}_{\boldsymbol{X}}[\boldsymbol{u}] \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} \boldsymbol{e}_b$; both in full in Lemma 1*(vi)*. A latent weight independent of the batch has $\boldsymbol{\Sigma}_{\xi} = \boldsymbol{0}$ and, for a finite batch law, no first-order remainder pairing either, so its excess begins at second order. A positive linear coefficient gives $\sigma^2_{\mathrm{eff}} > \sigma_s^2 \sigma^2_{\mathrm{obj}}$ at every sufficiently small $t > 0$, and a negative one reverses it.*

(iii) Escape and dwell, the classical law at the two diffusions. *Let $\alpha = \tilde{\mathcal{L}}(w_s) - \tilde{\mathcal{L}}(w_b) > 0$ be the reference action along $\boldsymbol{e}_b$, read σ^2_{eff} as constant across the barrier band, at the direct value $\sigma_s^2 \sigma^2_{\mathrm{obj}}$ for direct softplus and at* (5) *for the lift, and for the diffusion $dw = -\tilde{\mathcal{L}}'(w)\, dt + \sqrt{\eta}\, \sigma_{\mathrm{eff}}\, dB_t$ from a start $x < w_s$ on the escape interval $[a, b]$ of Assumption A2, reflecting at a and absorbing at b, let τ be the first passage to b, $\mathsf{T}(x) = \mathbb{E}_x[\tau]$ its mean, and $\mathsf{D}(x) = \mathbb{E}_x[\int_0^{\tau} \mathbf{1}_{[a, w_s]}(w_t)\, dt]$ the*

mean time spent below the barrier top before it, both under Dynkin's formula for its generator and continuous up to the closed interval. Then $\eta\sigma_{\text{eff}}^2 \log \mathsf{T}(x) \to 2\alpha$ *and* $\eta\sigma_{\text{eff}}^2 \log \mathsf{D}(x) \to 2\alpha$*, while* $\eta\sigma_{\text{eff}}^2 \log\big(\mathsf{T}(x)/\mathsf{D}(x)\big) \to 0$*, as* $\eta\sigma_{\text{eff}}^2 \to 0^+$.

The exponent $2\alpha/(\eta\sigma_{\text{eff}}^2)$ is strictly decreasing in σ_{eff}^2, so the method with the larger effective diffusion has the smaller exponent and, for every sufficiently small step size, the shorter mean first passage and the shorter mean time on the shoulder (Corollaries 1 and 2). Under $c > 0$ part (i) puts the lift on that side; the lift enters part (iii) through σ_{eff}^2 alone.

Remark 1 (Reach of Theorem 1)**.** A latent weight independent of the batch has an empty coupling channel, and its excess over the direct value begins only at second order (Remark 2).

Direct softplus's exponent in part (iii) is of order σ_s^{-2} within the shoulder under no further condition, and the lift's is too wherever the remainder channel is itself of order σ_s^2 with a constant below the leading coefficient of (5). Under $c > 0$ the lift therefore lowers their constant and not their order, by the factor $\sigma_{\text{eff}}^2/(\sigma_s^2\sigma_{\text{obj}}^2)$ of part (i) (Remark 3).

3.4 Joint necessity

Theorem 2 states which parts of the method the frozen coupling and its estimate need. With $\tilde{\boldsymbol{\theta}}_k = \boldsymbol{b} + h_{\boldsymbol{\phi}}(\boldsymbol{X}_k)$ the latent weight of the k-th batch, $\boldsymbol{g}_k$ the gradient of (1) at $\tilde{\boldsymbol{\theta}}_k$ on that batch, and $\bar{\tilde{\boldsymbol{\theta}}}$, $\bar{\boldsymbol{g}}$ their sample means,

$$\widehat{\boldsymbol{\Sigma}}^{(t)} \;=\; \frac{1}{T-1}\sum_{k=1}^{T} \delta\tilde{\boldsymbol{\theta}}_k\, \delta\boldsymbol{g}_k^{\top}, \qquad \delta\tilde{\boldsymbol{\theta}}_k = \tilde{\boldsymbol{\theta}}_k - \bar{\tilde{\boldsymbol{\theta}}}, \quad \delta\boldsymbol{g}_k = \boldsymbol{g}_k - \bar{\boldsymbol{g}}. \tag{6}$$

Theorem 2 (Joint necessity)**.** *Fix an iterate, so that* $\boldsymbol{b}$ *and* $\boldsymbol{\phi}$ *are held, and let* $\widehat{\boldsymbol{\Sigma}}^{(t)}$ *be the estimator of* (6) *formed on* T *batches sampled at that iterate, with either divisor* $T-1$ *or* T*. If the latent weight* $\tilde{\boldsymbol{\theta}}$ *and the gradient* $\boldsymbol{g}_0$ *at the reference iterate have finite fourth moments and are conditionally independent given the parameters—as they are when the body reads a batch drawn independently of the one that forms the gradient—then* $\boldsymbol{\Sigma}_\xi = \mathbf{0}$ *for almost every value of the parameters, and under Assumption A1 the estimate* $\widehat{\boldsymbol{\Sigma}}_\xi^{(t)}$ *formed with* $\boldsymbol{g}_0(\boldsymbol{X}_k)$ *in place of* $\boldsymbol{g}_k$ *has mean zero for every* T*, has an entrywise mean-square error of order* $1/T$*, and converges to zero almost surely as* $T \to \infty$*. A nonzero frozen coupling therefore requires that the latent weight and the gradient at the reference iterate not be conditionally independent.*

The other two pieces zero the estimate by substitution. With $\boldsymbol{J}_{\boldsymbol{b}} = \partial\tilde{\boldsymbol{\theta}}/\partial\boldsymbol{b}$ the slack Jacobian, a latent weight with no learnable additive slack has $\boldsymbol{J}_{\boldsymbol{b}} \equiv \mathbf{0}$ and $\boldsymbol{J}_{\boldsymbol{b}}^{\top}\widehat{\boldsymbol{\Sigma}}^{(t)} = \mathbf{0}$, and a body constant in the conditioning batch has $\delta\tilde{\boldsymbol{\theta}}_k \equiv \mathbf{0}$, so every summand of (6) vanishes; both hold identically, at every iterate and every probe count. Removing any one of the three pieces therefore sets the bias-channel coupling to zero, by Theorem 2 and by the two substitutions above.

4 Experiments

Two questions run through the experiments: whether training enters the shoulder defined in Section 2.1, and what the three methods do where it does. We therefore report the occupancy—the fraction of constrained weights on the shoulder—for every problem. We compare the methods on the training objective they minimize, and read the total-variation distance to the target, the held-out negative log-likelihood, or the reconstruction PSNR at the best held-out read, except where we note otherwise. We pair the methods run for run—i.e., they share the seed and the budget. Training details are in Appendix B and the held-out read along training in Appendix C.

4.1 Log-concave energy-based models

One-dimensional log-concave targets. We train our running example (Section 2) by the forward Kullback–Leibler divergence on the Gumbel, Laplace, Gamma, and Beta targets in one dimension, at thirty paired seeds. Here direct softplus begins with a third of its constrained weights on the shoulder

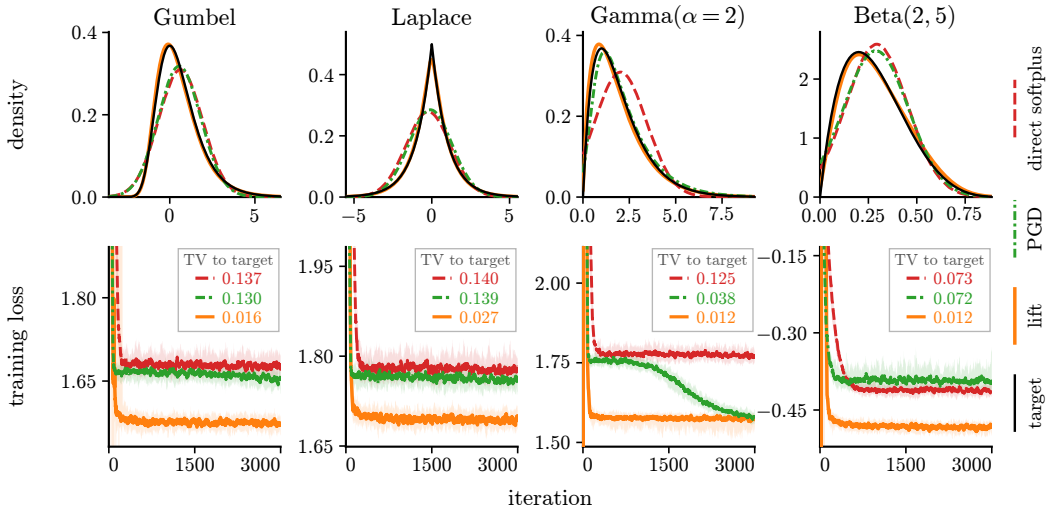


Figure 2: Median-seed densities **(top)** and the training objective **(bottom)**, the median bold inside the seed band. The framed table lists the total-variation distance to the target at the best held-out loss.

and ends with more on three of the four targets. The lift sits closer to the target in total-variation distance than direct softplus on all one hundred twenty paired runs, and its worst seed on each target is below direct softplus's best (Figure 2; statistics in Appendix D). Against PGD the lift is lower on every seed than PGD is on any seed of Gumbel, Laplace, and Beta, while on Gamma it leads at the median and the two seed bands overlap. The lift also reaches the lowest median training loss on all four targets, and on each of them the separation opens within the first few hundred iterations. Direct softplus never closes it, and PGD closes it only on Gamma.

Tabular datasets. POWER, HEPMASS (Baldi et al., 2016), and MiniBooNE (Roe et al., 2005) are the tabular benchmarks of Papamakarios et al. (2017) in 6, 21, and 43 dimensions, run with that paper's splits and preprocessing. On all three the lift is below both baselines on every paired seed that returns a likelihood, on the training objective and on the held-out loss alike, by margins of nats rather than hundredths of a nat (Figure 3). At these widths direct softplus starts with half to two thirds of its constrained weights on the shoulder, and training raises that share. We report each method at the configuration that gave it its best result (Table 3). While the lift needs a ramp on the step in 21 and 43 dimensions, the same ramp costs direct softplus a factor of several hundred on HEPMASS. Direct softplus returns a non-physical likelihood on one HEPMASS seed.

Dwell on the shoulder. Figure 4 measures the dwell on these problems—the quantity Theorem 1(iii) governs for the lift and direct softplus—and the three methods separate there by orders of magnitude. Under direct softplus a visit essentially never ends. On every target but Beta the occupancy series is non-decreasing on all thirty seeds, and a participating coordinate records essentially one visit, which is what an absorbing set gives. PGD reaches the boundary as exact zeros rather than as the shoulder, and nearly half of its visits end within a single snapshot interval. Its visits keep ending at about the same rate across the run, which is the memoryless dwell of the projected recursion in Proposition 1. The lift lies between the two, and its visits end fastest just after entry and at a lower rate thereafter. It also ends with more weights on the shoulder than direct softplus does, so what separates the three is the dwell and not the occupancy. Beta inverts every count. There direct softplus leaves the shoulder about as often as the lift does, its occupancy falls over training, and it is the target direct softplus fits best. A squared map stays several times behind the lift here, for the reason in Section 2.1, and we did not pursue it.

4.2 Convex potential flows

A convex potential flow (Huang et al., 2021) transports the data through the gradient of a strongly convex ICNN potential and evaluates its log-density exactly, with no estimated normalizer. We compare the lift with direct softplus here.

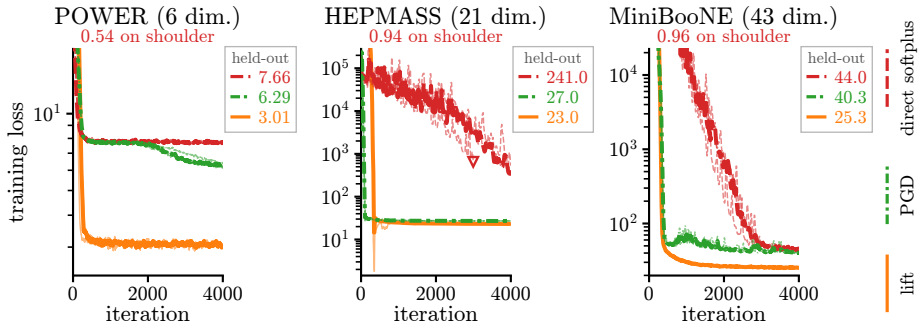


Figure 3: Training objective on the tabular datasets, seeds thin and medians bold. Red is direct softplus's shoulder occupancy at the end, and the framed table lists the best held-out loss.

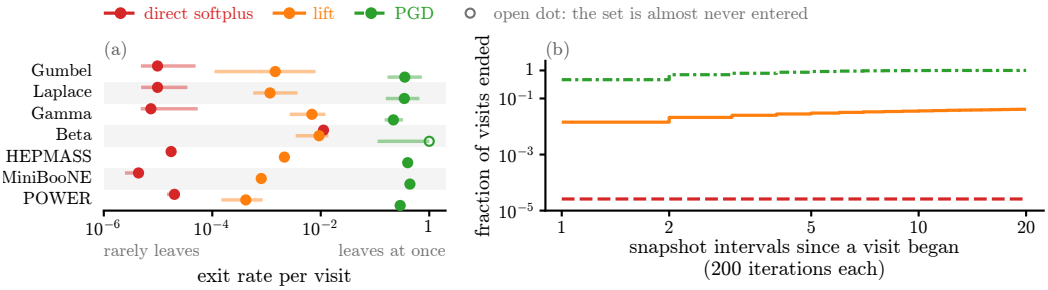


Figure 4: (a) Exit rate per coordinate-visit, from the shoulder for the softplus methods and from the active set for PGD. (b) Fraction of those visits ended by k snapshot intervals, on HEPMASS.

The 8-Gaussians, 2-spirals, and one-moon targets are trained at CP-Flow's published recipe on five paired seeds, with the lift replacing the constrained weights of each block and every number read at the final iterate. Under this recipe direct softplus barely enters the shoulder defined in Section 2.1, so the antecedent of Theorem 1 fails and we expect no separation. None appears (Figure 5). On the training objective the lift ends below direct softplus on every seed of 8-Gaussians and one-moon and level with it on 2-spirals, and the held-out differences run a few hundredths of a nat either way.

4.3 The Darcy transport map

A partially input-convex map (Wang et al., 2025), an ICNN in the unknown, represents the posterior of the Darcy flow inverse problem (Baldassari et al., 2023): 100 Karhunen–Loève coefficients of the log-permeability given 33 noisy readings of the hydraulic head, trained on 50,000 pairs with an exact likelihood over ten seeds (Appendix B). This problem is the paper's control. No constrained weight of direct softplus reaches the shoulder here—occupancy is zero at every evaluation on all ten seeds, and the lowest latent weight stays two units above the threshold defined in Section 2.1—so Theorem 1 has nothing to predict here. Nothing separates the three. Their held-out likelihoods tie within the seed spread and their training curves lie on top of each other (Figure 6).

The tie does not depend on where the run starts. We repeat the comparison from the map's published initialization, the PyTorch default, and a matched start, at five seeds in each of the nine cells. The nine cells agree to a fifth of a nat. Direct softplus sits at zero occupancy in all fifteen of its runs, and the lift stays off the shoulder from every start. The start does change PGD, which from its own published initialization holds more than a quarter of its constrained coordinates at zero on the first step and still above one percent of them on the last. While that set of held coordinates moves by two orders of magnitude between the three starts, the held-out likelihood does not move with it. We render its loss landscape in Appendix E.

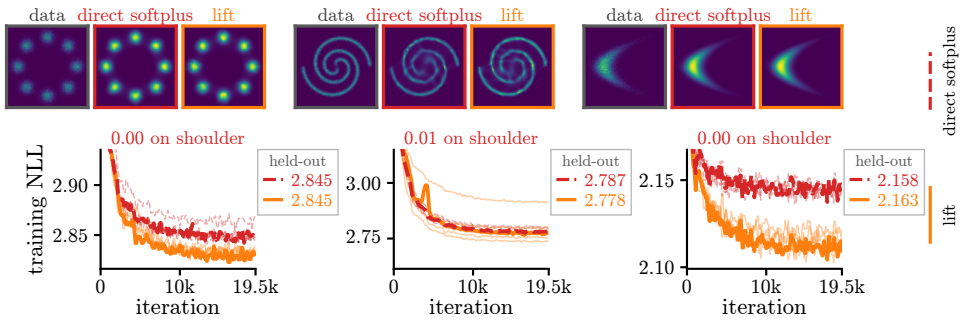


Figure 5: Densities of the data and of the two methods at their median seeds, and the training objective of five paired seeds with the held-out loss framed.

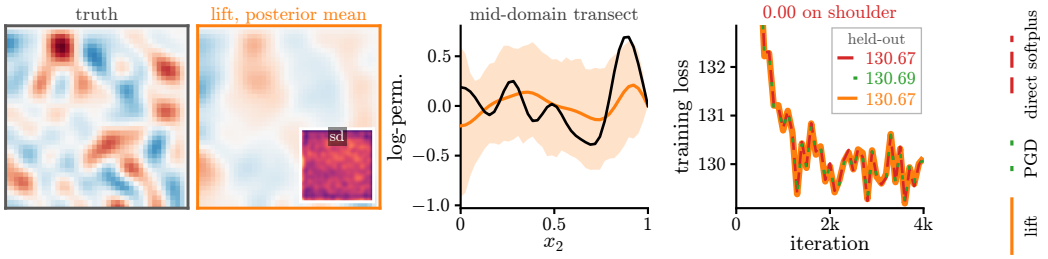


Figure 6: One held-out Darcy observation: the truth, the lift's posterior mean, the mid-domain transect, and the training objective of the three methods with the held-out loss framed.

4.4 An image regularizer

Beyond density models and transport maps, an ICNN serves as the learned regularizer of an inverse problem, where it scores an image and is trained adversarially (Lunz et al., 2018). We reproduce the convex regularizer of Ehrhardt et al. (2026) on grayscale FFHQ (Karras et al., 2019) inpainting at 256×256 as published and change only how positivity is enforced. The regularizer is convex on the image itself, and each is scored by the PSNR of its primal–dual reconstruction on the 20 test images (Figure 8; details in Appendix B). The baselines keep the published recipe's own initialization and the lift an emitter matched to it, so all three begin at the attenuated scale that recipe gives the baselines. Direct softplus then ends with all of its constrained weights on the shoulder, deeper than on any other problem here. We report each method at the iterate of its highest validation PSNR.

The lift reaches that iterate about three times sooner than direct softplus does at the median, and sooner on every one of twenty-three paired seeds, at a test PSNR that matches direct softplus and stands above PGD. Under direct softplus that PSNR arrives only when the budget ends, and as early as the lift's iterate only at four times the published step. The median latent weight climbs at about the step size per iteration for direct softplus and five times faster for the lift (Figure 7). From the initialization Hoedt & Klambauer (2023) prescribe instead, direct softplus does not train this regularizer at all within the published budget, while PGD trains on every seed. The lift trains on every seed of every cell we ran.

5 Related work

While PGD (Amos et al., 2017) identifies the coordinates at zero under exact gradients (Bertsekas, 1999; Sun et al., 2019), the projected stochastic iterate does not (Duchi & Ruan, 2021; Lee & Wright, 2012). A positivity map—i.e., softplus (Huang et al., 2021; Hundrieser et al., 2026) or exponential (Hoedt & Klambauer, 2023; Sill, 1997)—is applied to a free latent weight, and gradient descent on that weight corresponds, in continuous time, to mirror descent on $\boldsymbol{\theta}$ (Amid & Warmuth, 2020). Hoedt & Klambauer (2023) report a worse training loss for an exponential map than for the projection and trace it to the dynamics of the map, which is where our results locate it too. Monotone transport maps and flows (Ahn et al., 2022; Baptista et al., 2024) obtain monotonicity from the architecture

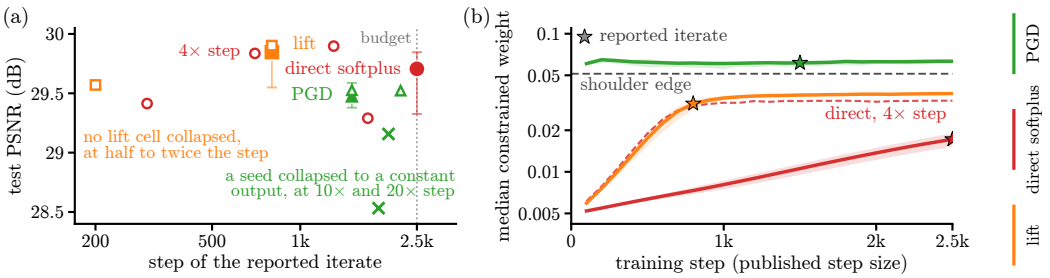


Figure 7: (a) Test PSNR at the reported iterate against the step it reached, at the published step and, as open markers, at half to twenty times that step. (b) The median constrained weight along training.

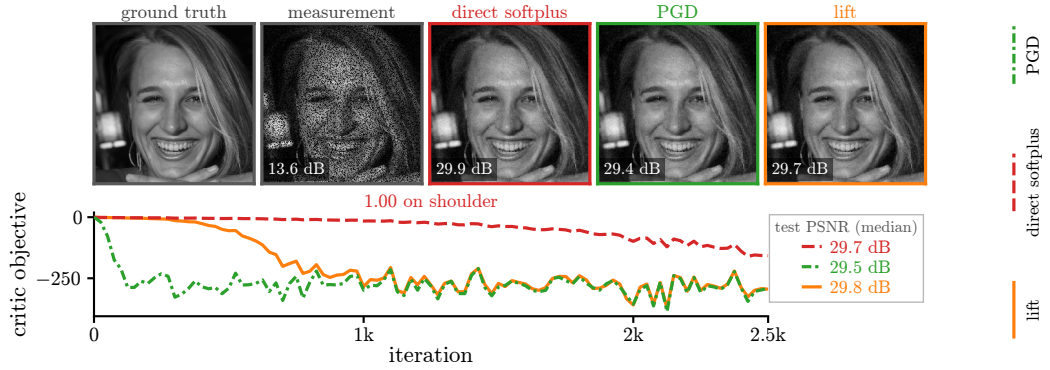


Figure 8: A held-out test image, its inpainting measurement, and the reconstruction under each regularizer at its reported iterate, with the critic objective and the test PSNR framed.

instead, and Hundrieser et al. (2026) widen each layer with maxout branches under an initialization of their own. Each of these maps and architectures acts on the starting point, the architecture, or the map itself, and the gradient on the shoulder remains scaled by its derivative. The lift keeps the map and borrows the extended formulations of PDE-constrained inverse problems (Aghazade et al., 2026; Siahkoohi et al., 2026; Symes et al., 2020; van Leeuwen & Herrmann, 2013), which add an auxiliary unknown and return to the reduced problem by a penalty or a multiplier. Here the batch produces that unknown and there is no penalty, through a hypernetwork (Ha et al., 2017; Mayer et al., 2024; Thatipelli & Siahkoohi, 2026) on a permutation-invariant summary (Zaheer et al., 2017) placed inside the constraint. Annealed and injected noise (Mobahi & Fisher III, 2015; Neelakantan et al., 2016) is scheduled or independent of the batch, batch normalization (Ioffe & Szegedy, 2015) standardizes activations downstream of the weights, and the diffusion view of stochastic gradient descent (Li et al., 2017) describes the noise a method already has. The fluctuation (4) of the lift instead enters the latent weight before the map and is coupled to the gradient through the shared batch.

6 Conclusion

We have presented the lift, which moves the constrained weight of an ICNN around the attenuation of its positivity map by emitting the latent weight from the same batch that forms the gradient. Where this coupling aligns positively with the loss curvature, the constant in the exponent of the escape time from the shoulder falls, and removing the slack, the batch dependence, or the shared batch sets the estimate of that coupling to zero. On the energy-based models, where direct softplus reaches the shoulder and does not leave, the lift fits closer to the target on every paired run that returns a likelihood, and on the image regularizer it reaches the same reconstruction about three times sooner. On the convex potential flows and the transport map, where almost no weight reaches the shoulder, Theorem 1 has nothing to predict and the methods we run agree. The same lift applies to any smooth monotone positivity map whose infimum is zero. Future work involves extending the parameter space of deep network optimization problems by the same mechanism, and asking whether the coupled fluctuation it introduces carries the iterate toward wider minima.

AI USE STATEMENT

Generative AI tools were used to aid and polish the writing of this paper, and to assist in proving its mathematical claims. All AI-assisted work was reviewed by the author, who takes responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The anonymized archive submitted with this paper holds the code that produces every experiment reported here, its configurations and its tests, together with the machine-checked development of the formal results. The architecture, the optimizer and the budget of each experiment are in Appendix B, and the proofs with the assumptions they use are in Appendices A and A.1.

REFERENCES

Kamal Aghazade, Ali Siahkoohi, and Ali Gholami. Scalable Bayesian full waveform inversion via dual augmented Lagrangian SVGD. In *SEG Technical Program Expanded Abstracts*, 2026.

Byeongkeun Ahn, Chiyoon Kim, Youngjoon Hong, and Hyunwoo J. Kim. Invertible monotone operators for normalizing flows. In *Advances in Neural Information Processing Systems*, 2022.

Ehsan Amid and Manfred K. Warmuth. Reparameterizing mirror descent as gradient descent. In *Advances in Neural Information Processing Systems*, 2020.

Brandon Amos. On amortizing convex conjugates for optimal transport. In *International Conference on Learning Representations*, 2023.

Brandon Amos, Lei Xu, and J. Zico Kolter. Input convex neural networks. In *International Conference on Machine Learning*, 2017.

Søren Asmussen. *Applied Probability and Queues*. Springer, 2003.

Lorenzo Baldassari, Ali Siahkoohi, Josselin Garnier, Knut Solna, and Maarten V. de Hoop. Conditional score-based diffusion models for Bayesian inference in infinite dimensions. In *Advances in Neural Information Processing Systems*, 2023.

Pierre Baldi, Kyle Cranmer, Taylor Faucett, Peter Sadowski, and Daniel Whiteson. Parameterized neural networks for high-energy physics. *European Physical Journal C*, 76(5):235, 2016.

Ricardo Baptista, Youssef Marzouk, and Olivier Zahm. On the representation and learning of monotone triangular transport maps. *Foundations of Computational Mathematics*, 24:2063–2108, 2024.

Raphaël Berthier. Incremental learning in diagonal linear networks. *Journal of Machine Learning Research*, 24(171):1–26, 2023.

Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

Charlotte Bunne, Andreas Krause, and Marco Cuturi. Supervised training of conditional Monge maps. In *Advances in Neural Information Processing Systems*, 2022.

Zhuo-Xu Cui, Qingyong Zhu, Jing Cheng, Bo Zhang, and Dong Liang. Deep unfolding as iterative regularization for imaging inverse problems. *Inverse Problems*, 40(2):025011, 2024.

John C. Duchi and Feng Ruan. Asymptotic optimality in stochastic optimization. *The Annals of Statistics*, 49(1):21–48, 2021.

Matthias J. Ehrhardt, Subhadip Mukherjee, and Hok Shing Wong. A primal-dual algorithm for image reconstruction with input-convex neural network regularizers. *SIAM Journal on Imaging Sciences*, 19(2):782–806, 2026.

Mark I. Freidlin and Alexander D. Wentzell. *Random Perturbations of Dynamical Systems*. Springer, 2012.

David Ha, Andrew M. Dai, and Quoc V. Le. HyperNetworks. In *International Conference on Learning Representations*, 2017.

Peter Hänggi, Peter Talkner, and Michal Borkovec. Reaction-rate theory: fifty years after Kramers. *Reviews of Modern Physics*, 62(2):251–341, 1990.

Pieter-Jan Hoedt and Günter Klambauer. Principled weight initialisation for input-convex neural networks. In *Advances in Neural Information Processing Systems*, 2023.

Chin-Wei Huang, Ricky T. Q. Chen, Christos Tsirigotis, and Aaron Courville. Convex potential flows: universal probability distributions with optimal transport and convex optimization. In *International Conference on Learning Representations*, 2021.

Shayan Hundrieser, Insung Kong, and Johannes Schmidt-Hieber. Hyper input convex neural networks for shape constrained learning and optimal transport. *arXiv preprint arXiv:2604.26942*, 2026.

Sergey Ioffe and Christian Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pp. 448–456, 2015.

Kelvin Kan, François-Xavier Aubet, Tim Januschowski, Youngsuk Park, Konstantinos Benidis, Lars Ruthotto, and Jan Gasthaus. Multivariate quantile function forecaster. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Alexander Korotin, Lingxiao Li, Aude Genevay, Justin Solomon, Alexander Filippov, and Evgeny Burnaev. Do neural optimal transport solvers work? A continuous Wasserstein-2 benchmark. In *Advances in Neural Information Processing Systems*, 2021.

Hendrik A. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304, 1940.

Jakob Kruse, Gianluca Detommaso, Ullrich Köthe, and Robert Scheichl. HINT: hierarchical invertible neural transport for density estimation and Bayesian inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8191–8199, 2021.

Sangkyun Lee and Stephen J. Wright. Manifold identification in dual averaging for regularized stochastic online learning. *Journal of Machine Learning Research*, 13(55):1705–1744, 2012.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, 2018.

Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, 2017.

Alexander Lin and Demba Ba. How to train your FALCON: learning log-concave densities with energy-based neural networks. In *Symposium on Advances in Approximate Bayesian Inference*, 2023.

Sebastian Lunz, Ozan Öktem, and Carola-Bibiane Schönlieb. Adversarial regularizers in inverse problems. In *Advances in Neural Information Processing Systems*, 2018.

Ashok Vardhan Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, 2020.

Paul M. Mayer, Lorenzo Luzi, Ali Siahkoohi, Don H. Johnson, and Richard G. Baraniuk. Improving fairness and mitigating MADness in generative models. Technical report, Rice University, 2024.

Hossein Mobahi and John W. Fisher III. A theoretical analysis of optimization by Gaussian continuation. In *AAAI Conference on Artificial Intelligence*, 2015.

Subhadip Mukherjee, Sören Dittmer, Zakhar Shumaylov, Sebastian Lunz, Ozan Öktem, and Carola-Bibiane Schönlieb. Data-driven convex regularizers for inverse problems. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024.

Arvind Neelakantan, Luke Vilnis, Quoc V. Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach, and James Martens. Adding gradient noise improves learning for very deep networks. In *International Conference on Learning Representations, Workshop Track*, 2016.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, 2017.

Byron P. Roe, Hai-Jun Yang, Ji Zhu, Yong Liu, Ion Stancu, and Gordon McGregor. Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nuclear Instruments and Methods in Physics Research Section A*, 543(2–3):577–584, 2005.

Ali Siahkoohi, Kamal Aghazade, and Ali Gholami. Dual-space posterior sampling for Bayesian inference in constrained inverse problems. arXiv preprint arXiv:2603.00393, 2026.

Joseph Sill. Monotonic networks. In *Advances in Neural Information Processing Systems*, 1997.

Frank Spitzer. A combinatorial lemma and its application to probability theory. *Transactions of the American Mathematical Society*, 82(2):323–339, 1956.

Yifan Sun, Halyun Jeong, Julie Nutini, and Mark Schmidt. Are we there yet? Manifold identification of gradient-related proximal methods. In *International Conference on Artificial Intelligence and Statistics*, 2019.

William W. Symes, Huiyi Chen, and Susan E. Minkoff. Full-waveform inversion by source extension: why it works. In *SEG Technical Program Expanded Abstracts*, pp. 765–769, 2020.

Amirhossein Taghvaei and Bamdad Hosseini. An optimal transport formulation of Bayes' law for nonlinear filtering algorithms. In *IEEE Conference on Decision and Control*, 2022.

Anirudh Thatipelli and Ali Siahkoohi. Hypernetwork-based approach for grid-independent functional data clustering. arXiv preprint arXiv:2602.22823, 2026.

Tristan van Leeuwen and Felix J. Herrmann. Mitigating local minima in full-waveform inversion by expanding the search space. *Geophysical Journal International*, 195(1):661–667, 2013.

Nina Vesseron, Louis Béthune, and Marco Cuturi. Sample and map from a single convex potential: generation using conjugate moment measures. In *Advances in Neural Information Processing Systems*, 2025.

Zheyu Oliver Wang, Ricardo Baptista, Youssef Marzouk, Lars Ruthotto, and Deepanshu Verma. Efficient neural network approaches for conditional optimal transport with applications in Bayesian inference. *SIAM Journal on Scientific Computing*, 47(4):C979–C1005, 2025.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, 2017.

A PROOFS

This appendix records the attribution and the discharge of the hypotheses of Section 3, proves Theorem 2, states and proves Lemma 1 and Corollary 1, and closes the escape argument with the proof of Theorem 1, which collects Lemma 1(ii)–(iii),(vi) and Corollaries 1 and 2. Three inputs are modeling steps, and we name each at the point where it enters, with the mathematics that follows once it is in place. The first is the Itô surrogate for the discrete recursion, the register of (A1), which follows the stochastic modified equations of Li et al. (2017); the surrogate is the primitive of every diffusion statement below, and nothing is proved about the passage from the recursion to it. That approximation is weak and holds over horizons of order $1/\eta$; the times of Corollary 1 are exponentially longer. The second is Dynkin's formula, which turns the mean exit time of the bias-channel diffusion into the boundary-value problem of Appendix A.4; that problem has exactly one continuous solution, and everything downstream of it is proved. The third is the same formula at the band indicator, which turns the mean occupation time into the boundary-value problem of Corollary 2, one modeling input applied to a second functional. A fourth step is an approximation—i.e., we read the effective diffusion as constant across the barrier band—and Appendix A.3 brackets the error of this reading at the level of the effective action.

The analysis describes training as it runs by (3), with one batch and one latent weight per step, and adds nothing to it. Its expectations are over the batch law and are never formed, and the one estimate it uses is a diagnostic computed outside training. The notation is that of Section 3: at an iterate with $\boldsymbol{b}$ and $\boldsymbol{\phi}$ held, $\delta\tilde{\boldsymbol{\theta}}$ is the fluctuation of (4), $\bar{\tilde{\boldsymbol{\theta}}} = \boldsymbol{b} + \mathbb{E}_{\boldsymbol{X}}[h_{\boldsymbol{\phi}}(\boldsymbol{X})]$ the reference iterate, $\boldsymbol{g}_0(\boldsymbol{X}) = \nabla_{\tilde{\boldsymbol{\theta}}}\mathcal{L}(\bar{\tilde{\boldsymbol{\theta}}}; \boldsymbol{X})$ the gradient of a batch at that iterate, $\boldsymbol{\xi} = \boldsymbol{g}_0 - \mathbb{E}_{\boldsymbol{X}}[\boldsymbol{g}_0]$ the minibatch noise there, $\boldsymbol{\Sigma}_\xi = \mathbb{E}_{\boldsymbol{X}}[\delta\tilde{\boldsymbol{\theta}}\,\boldsymbol{\xi}^\top]$ the frozen coupling, $\boldsymbol{H}_{\bar{\tilde{\boldsymbol{\theta}}}} = \mathbb{E}_{\boldsymbol{X}}[\nabla^2_{\tilde{\boldsymbol{\theta}}}\mathcal{L}(\bar{\tilde{\boldsymbol{\theta}}}; \boldsymbol{X})]$ the mean loss curvature, $\boldsymbol{V} = \mathrm{Cov}_{\boldsymbol{X}}[\delta\tilde{\boldsymbol{\theta}}]$ the iterate-fluctuation covariance, and $\boldsymbol{r}_2$ the remainder of the chain of Section 3.2. Two conventions fix the rest. The positivity-map slope of (A2) enters the per-step update as a scalar prefactor multiplying every term, and the identity (5) is that decomposition with the prefactor set to one, because the latent-weight gradient $\boldsymbol{g}$ of (1) already includes ψ' and the direct-softplus value $\sigma_s^2\sigma_{\mathrm{obj}}^2$ is the variance of that gradient. Where the prefactor is kept explicit, as it is in Appendix A.6, the excess terms are the paper's divided by σ_s^2, and we say so at the one place where the conversion matters. The estimator of (6) is a T-probe statistic with Bessel's divisor; on the same data its trailing-window form with divisor $1/T$ differs from it by the factor $T/(T-1)$, so every statement about one transports to the other, and the proof of Theorem 2 is written on the probe form.

A.1 THE HYPOTHESES

Each result names what it consumes, from Assumptions A1–A2 of Section 3.2. Table 1 records the attribution, and the two clauses of (A2) are the only hypotheses with content beyond regularity.

Table 1: Every hypothesis of this appendix and the results that consume it.

	statement	consumed by
(A1)	Itô surrogate; i.i.d. batches; $\boldsymbol{\Sigma}$ Lipschitz	Lemma 1(iv)–(v), Corollaries 1 and 2
	the i.i.d. clause alone	Theorem 2, sample form
regularity	$\boldsymbol{g}_0, \delta\tilde{\boldsymbol{\theta}}, \boldsymbol{r}_2$ square integrable; $\boldsymbol{H}_{\bar{\tilde{\boldsymbol{\theta}}}}^\top = \boldsymbol{H}_{\bar{\tilde{\boldsymbol{\theta}}}}$; $\sigma_{\mathrm{obj}}^2 > 0$	Lemma 1(i)–(iii),(vi)
	$\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{g}_0$ of finite fourth moment	Theorem 2
(A2a)	$\lvert\psi'(w) - \sigma_s\rvert \le \varepsilon$ on $\mathcal{B}$, $0 \le \varepsilon < \sigma_s$	Lemma 1
(A2b)	$\alpha = \tilde{\mathcal{L}}(w_s) - \tilde{\mathcal{L}}(w_b) > 0$, extremal on $[a, b]$	Corollary 1
$c > 0$	at the one iterate and the one direction $\boldsymbol{e}_b$	the ordering of Corollaries 1 and 2
Dynkin	the mean exit time solves the boundary-value problem	Corollary 1
	the occupation time likewise	Corollary 2

(A2a) brackets the direct-softplus diffusion,

$$(\sigma_s - \varepsilon)^2\sigma_{\mathrm{obj}}^2 \;\le\; \psi'(w)^2\sigma_{\mathrm{obj}}^2 \;\le\; (\sigma_s + \varepsilon)^2\sigma_{\mathrm{obj}}^2, \qquad w \in \mathcal{B},$$

and its tolerance is the substance of the clause: at $\varepsilon = 0$ the band collapses to a point for any ψ with injective derivative. For softplus the clause is a theorem. From $\psi'' \le \psi'$,

$$e^{-\delta}\psi'(w) \;\le\; \psi'(u) \;\le\; e^{\delta}\psi'(w), \qquad |u - w| \le \delta,$$

so every interval of half-width $\log(1+\rho)$ about w satisfies (A2a) with $\sigma_s = \psi'(w)$ and $\varepsilon = \rho\sigma_s$, at any depth of the shoulder. Its content is therefore that the operative iterates occupy such a band, which is a statement about training and not about ψ.

The slack Jacobian $\boldsymbol{J_b} = \boldsymbol{I}$ is an identity of (2), and it makes the bias-channel estimate of Theorem 2 the estimator itself. The chain $\boldsymbol{g} = \boldsymbol{g}_0 + \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\delta\tilde{\boldsymbol{\theta}} + \boldsymbol{r}_2$ is an identity once $\boldsymbol{r}_2$ is the discrepancy, with $\boldsymbol{\xi}$ of order one in the jitter.

A.2 Joint necessity

Theorem 2 is stated in Section 3, and the two substitutions printed after it there are proved first below.

Proof. Write λ for the divisor, $\lambda = (T-1)^{-1}$ or $\lambda = T^{-1}$, so that $\widehat{\boldsymbol{\Sigma}}^{(t)} = \lambda \sum_{k=1}^{T} \delta\tilde{\boldsymbol{\theta}}_k \, \delta\boldsymbol{g}_k^\top$.

No learnable slack. With $\boldsymbol{J_b} = \boldsymbol{0}$, summand by summand,

$$\boldsymbol{J_b}^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \;=\; \lambda \sum_{k=1}^{T} \boldsymbol{J_b}^\top \, \delta\tilde{\boldsymbol{\theta}}_k \, \delta\boldsymbol{g}_k^\top \;=\; \boldsymbol{0},$$

at every iterate, for every $T \geq 0$, and for both divisors.

A body constant in the batch. A body constant in the batch gives $\tilde{\boldsymbol{\theta}}_k = \boldsymbol{b} + h_{\boldsymbol{\phi}} = \bar{\tilde{\boldsymbol{\theta}}}$ for every k, so for $T \geq 1$

$$\delta\tilde{\boldsymbol{\theta}}_k = \tilde{\boldsymbol{\theta}}_k - \bar{\tilde{\boldsymbol{\theta}}} = \boldsymbol{0}, \qquad \widehat{\boldsymbol{\Sigma}}^{(t)} = \lambda \sum_{k=1}^{T} \boldsymbol{0} \, \delta\boldsymbol{g}_k^\top = \boldsymbol{0}.$$

The contrapositive of the two substitutions is the sentence printed after the theorem,

$$\boldsymbol{J_b}^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \neq \boldsymbol{0} \implies \boldsymbol{J_b} \neq \boldsymbol{0} \;\text{ and }\; \delta\tilde{\boldsymbol{\theta}}_k \neq \boldsymbol{0} \;\text{ for some } k,$$

and neither substitution uses an assumption or a property of ψ.

The population zero. Let $\mathcal{G} = \sigma(\boldsymbol{\phi}, \boldsymbol{b})$. Conditional independence of $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{g}_0$ given $\mathcal{G}$ passes to $\delta\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}} - \mathbb{E}[\tilde{\boldsymbol{\theta}} \mid \mathcal{G}]$ and $\boldsymbol{\xi} = \boldsymbol{g}_0 - \mathbb{E}[\boldsymbol{g}_0 \mid \mathcal{G}]$, each a $\mathcal{G}$-measurable shift of one factor. For measurable, integrable, conditionally independent vectors the conditional expectation of the outer product factorizes through the disintegration kernel, so

$$\boldsymbol{\Sigma}_\xi = \mathbb{E}\big[\delta\tilde{\boldsymbol{\theta}}\,\boldsymbol{\xi}^\top \,\big|\, \mathcal{G}\big] = \mathbb{E}\big[\delta\tilde{\boldsymbol{\theta}} \,\big|\, \mathcal{G}\big]\, \mathbb{E}\big[\boldsymbol{\xi} \,\big|\, \mathcal{G}\big]^\top = \boldsymbol{0} \qquad \text{a.s.}, \tag{7}$$

both factors vanishing by the centering, and $\operatorname{tr}\boldsymbol{\Sigma}_\xi = 0$ with it. At fixed parameters (7) is the unconditional statement for independent integrable centered vectors; independence is what produces the zero, dependent centered pairs with $\boldsymbol{\Sigma}_\xi \neq \boldsymbol{0}$ being available.

The sample version. Fix the parameters and take every expectation under the conditional law. Sample-mean centering annihilates the constant shifts $\mathbb{E}[\tilde{\boldsymbol{\theta}} \mid \mathcal{G}]$ and $\mathbb{E}[\boldsymbol{g}_0 \mid \mathcal{G}]$, so for $T \geq 1$ the estimate formed from the raw pairs $(\tilde{\boldsymbol{\theta}}_k, \boldsymbol{g}_0(\boldsymbol{X}_k))$ is the one formed from $(\delta\tilde{\boldsymbol{\theta}}_k, \boldsymbol{\xi}_k)$, and the sample-cross-covariance identity, exact finite algebra, gives

$$\widehat{\boldsymbol{\Sigma}}_\xi^{(t)} = \lambda \sum_{k=1}^{T} (\delta\tilde{\boldsymbol{\theta}}_k - \overline{\delta\tilde{\boldsymbol{\theta}}})(\boldsymbol{\xi}_k - \bar{\boldsymbol{\xi}})^\top = \lambda \sum_{k=1}^{T} \delta\tilde{\boldsymbol{\theta}}_k \, \boldsymbol{\xi}_k^\top - \lambda T \, \overline{\delta\tilde{\boldsymbol{\theta}}}\, \bar{\boldsymbol{\xi}}^\top, \tag{8}$$

with $\lambda T = T/(T-1)$ at Bessel's divisor and $\lambda T = 1$ at divisor T. Under (A1) the probes are independent and identically distributed, so (7) and the pairwise independence give

$$\mathbb{E}[\delta\tilde{\boldsymbol{\theta}}_k \, \boldsymbol{\xi}_k^\top] = \boldsymbol{0}, \qquad \mathbb{E}[\delta\tilde{\boldsymbol{\theta}}_k \, \boldsymbol{\xi}_l^\top] = \boldsymbol{0} \quad (k \neq l), \tag{9}$$

and the expectation of (8), term by term, is

$$\mathbb{E}\big[\widehat{\boldsymbol{\Sigma}}_\xi^{(t)}\big] = \lambda \sum_{k=1}^{T} \mathbb{E}[\delta\tilde{\boldsymbol{\theta}}_k \, \boldsymbol{\xi}_k^\top] - \frac{\lambda}{T} \sum_{k,l=1}^{T} \mathbb{E}[\delta\tilde{\boldsymbol{\theta}}_k \, \boldsymbol{\xi}_l^\top] = \boldsymbol{0}$$

for every $T \geq 1$ at divisor T and every $T \geq 2$ at Bessel's. On a nonzero target $\mathbf{\Sigma}$ the same two computations give $\mathbf{\Sigma}$ at Bessel's divisor and $(T-1)\mathbf{\Sigma}/T$ at divisor T, for $T \geq 2$, which is where the divisors part company.

The summands $\delta\tilde{\boldsymbol{\theta}}_k \boldsymbol{\xi}_k^\top$ are pairwise independent, identically distributed, integrable, and centered by (9), as are $\delta\tilde{\boldsymbol{\theta}}_k$ and $\boldsymbol{\xi}_k$, so Etemadi's strong law gives

$$\frac{1}{T}\sum_{k=1}^{T} \delta\tilde{\boldsymbol{\theta}}_k\, \boldsymbol{\xi}_k^\top \to \mathbf{0}, \qquad \overline{\delta\tilde{\boldsymbol{\theta}}} \to \mathbf{0}, \qquad \bar{\boldsymbol{\xi}} \to \mathbf{0} \qquad \text{a.s. as } T \to \infty,$$

entrywise and hence in the matrix topology, and $\widehat{\mathbf{\Sigma}}_\xi^{(t)} \to \mathbf{0}$ a.s. by (8), the factor λT tending to one.

For the rate, the variance of an average of T pairwise independent identically distributed centered summands is the summand variance over T. With $\sigma_{ij}^2 = \mathrm{Var}[(\delta\tilde{\boldsymbol{\theta}}\,\boldsymbol{\xi}^\top)_{ij}]$, $\sigma_u^2 = \mathrm{Var}[(\delta\tilde{\boldsymbol{\theta}})_i]$, and $\sigma_v^2 = \mathrm{Var}[(\boldsymbol{\xi})_j]$,

$$\mathbb{E}\Big[\Big(\lambda \sum_{k=1}^{T} (\delta\tilde{\boldsymbol{\theta}}_k\, \boldsymbol{\xi}_k^\top)_{ij}\Big)^2\Big] = \lambda^2 T\, \sigma_{ij}^2, \tag{10}$$

a root mean square of $\sqrt{T}\,\sigma_{ij}/(T-1)$ at Bessel's divisor and $\sigma_{ij}/\sqrt{T}$ at divisor T. The joint law of the $2T$ vectors is the product over probes of the per-probe laws, each of which factorizes by the conditional independence, so the collections $\{\delta\tilde{\boldsymbol{\theta}}_k\}_k$ and $\{\boldsymbol{\xi}_k\}_k$ are independent, hence so are the two sample means, and

$$\mathbb{E}\big[\big((\overline{\delta\tilde{\boldsymbol{\theta}}})_i\,(\bar{\boldsymbol{\xi}})_j\big)^2\big] = \mathbb{E}\big[(\overline{\delta\tilde{\boldsymbol{\theta}}})_i^2\big]\,\mathbb{E}\big[(\bar{\boldsymbol{\xi}})_j^2\big] = \frac{\sigma_u^2\sigma_v^2}{T^2}. \tag{11}$$

This is the one step of the proof that uses the i.i.d. clause of (A1) beyond its pairwise form. By $(x-y)^2 \leq 2x^2 + 2y^2$ applied to (8) at divisor T, with (10) and (11),

$$\mathbb{E}\big[(\widehat{\mathbf{\Sigma}}_\xi^{(t)})_{ij}^2\big] \leq \frac{2\sigma_{ij}^2}{T} + \frac{2\sigma_u^2\sigma_v^2}{T^2},$$

and Chebyshev's inequality makes the entries $O_P(T^{-1/2})$; the variances σ_{ij}^2, σ_u^2, σ_v^2 are finite by the fourth-moment clause of the theorem; an almost-sure rate is neither claimed nor proved. The sample version does not vanish at finite T: for independent standard Gaussian factors in one dimension the population-centered sum at divisor T has $\mathbb{E}[(T^{-1}\sum_k \delta\tilde{\theta}_k\xi_k)^2] = 1/T$, so no nonempty probe set makes it almost surely zero. Every display above is made summand by summand or carries the factor λT, so it holds on the trailing window of (6) as well. □

A.3 Diffusion and barrier

Lemma 1 (Diffusion and barrier; parts (ii)–(iii) are Theorem 1(i) and part (vi) is Theorem 1(ii))**.** *Let Assumptions A1–A2 hold, the second clause of Assumption A2 not being used and the Itô surrogate of Assumption A1 entering parts* (iv)–(v) *alone. Fix the reference iterate $\bar{\tilde{\boldsymbol{\theta}}}$ on the band $\mathcal{B}$ and a unit direction $\boldsymbol{e}_b$ of the slack subspace, and write $\mathbf{\Sigma}_{\mathrm{slack}} = \mathbb{E}_{\boldsymbol{X}}[\delta\tilde{\boldsymbol{\theta}}\,\delta\boldsymbol{g}^\top]$ for the population cross-covariance of* (6)*, whose trace is σ_{Jac}^2, $\sigma_{\mathrm{obj}}^2 = \boldsymbol{e}_b^\top \mathrm{Var}_{\boldsymbol{X}}[\nabla_{\boldsymbol{\theta}}\mathcal{L}]\,\boldsymbol{e}_b$ for the objective variance in the constrained coordinate along $\boldsymbol{e}_b$, and $\sigma_{\mathrm{eff}}^2 = \boldsymbol{e}_b^\top \mathrm{Cov}_{\boldsymbol{X}}[\boldsymbol{g}]\,\boldsymbol{e}_b$ for the projected diffusion of the lifted update. Then the following hold.* (i) $\mathbf{\Sigma}_{\mathrm{slack}} = \mathbf{\Sigma}_\xi + \boldsymbol{V}\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} + \mathrm{Cov}_{\boldsymbol{X}}[\delta\tilde{\boldsymbol{\theta}}, \boldsymbol{r}_2]$. (ii) *With no term discarded,*

$$\sigma_{\mathrm{eff}}^2 = \underbrace{\sigma_s^2\sigma_{\mathrm{obj}}^2}_{\text{objective contribution}} + \underbrace{2\,\boldsymbol{e}_b^\top \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\mathbf{\Sigma}_\xi\,\boldsymbol{e}_b}_{\text{coupling term}} + \underbrace{\boldsymbol{e}_b^\top \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\boldsymbol{V}\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\,\boldsymbol{e}_b}_{\text{jitter term}} + \underbrace{\varrho}_{\text{remainder channel}}, \tag{5}$$

with $\varrho = \varrho(\boldsymbol{e}_b)$ the remainder channel of Section 3.2. (iii) *The direct softplus value at the same iterate is the objective contribution alone, within the bracket of Assumption A2. The jitter term is $\mathrm{Var}_{\boldsymbol{X}}[\langle \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\boldsymbol{e}_b, \delta\tilde{\boldsymbol{\theta}}\rangle] \geq 0$, so $\sigma_{\mathrm{eff}}^2 - \sigma_s^2\sigma_{\mathrm{obj}}^2 \geq c$ with no hypothesis on the coupling.* (iv) *Under the surrogate of Assumption A1 the bias-channel diffusion $dw = -\tilde{\mathcal{L}}'(w)\,dt + \sqrt{\eta}\,\sigma_{\mathrm{eff}}(w)\,dB_t$ has the stationary density $\pi \propto \sigma_{\mathrm{eff}}^{-2}\exp(-(2/\eta)\,U_{\mathrm{eff}})$, $U_{\mathrm{eff}}(w) = \int^w \tilde{\mathcal{L}}'(u)/\sigma_{\mathrm{eff}}^2(u)\,du$, and every zero-flux*

stationary density is of this form. (v) *A barrier of action* $\alpha = \tilde{\mathcal{L}}(w_s) - \tilde{\mathcal{L}}(w_b)$ *climbed monotonically across a band on which* $c_{\min} \le \sigma^2_{\text{eff}} \le c_{\max}$, $c_{\min} > 0$, *has effective action* $U_{\text{eff}}(w_s) - U_{\text{eff}}(w_b) \in [\alpha/c_{\max}, \alpha/c_{\min}]$, *and exactly* $\alpha/\sigma^2_{\text{eff}}$ *on a constant band.* (vi) *Scale the fluctuation and the remainder at a fixed batch law,* $\delta\tilde{\boldsymbol{\theta}} = t\,\boldsymbol{u}$ *and* $\boldsymbol{r}_2 = t\,\boldsymbol{w}_1 + t^2\boldsymbol{w}_2$. *Then* $\sigma^2_{\text{eff}}(t) - \sigma_s^2\sigma^2_{\text{obj}}$ *is a polynomial in* t *with no constant term, of linear coefficient*

$$2\,\boldsymbol{e}_b^\top \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} \mathrm{Cov}_{\boldsymbol{X}}[\boldsymbol{u}, \boldsymbol{g}_0]\,\boldsymbol{e}_b \;+\; 2\,\boldsymbol{e}_b^\top \mathrm{Cov}_{\boldsymbol{X}}[\boldsymbol{g}_0, \boldsymbol{w}_1]\,\boldsymbol{e}_b,$$

and with the jitter term $\boldsymbol{e}_b^\top \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} \mathrm{Cov}_{\boldsymbol{X}}[\boldsymbol{u}] \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} \boldsymbol{e}_b$ *in its quadratic coefficient. A latent weight independent of the batch has* $\boldsymbol{\Sigma}_\xi = \boldsymbol{0}$ *and, for a finite batch law, no first-order remainder pairing either, so its excess begins at second order. A positive linear coefficient gives* $\sigma^2_{\text{eff}} > \sigma_s^2\sigma^2_{\text{obj}}$ *at every sufficiently small* $t > 0$, *and a negative one reverses it.*

Proof. Step 1: the three-term split. The chain of Section 3.2, $\boldsymbol{g} = \boldsymbol{g}_0 + \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\delta\tilde{\boldsymbol{\theta}} + \boldsymbol{r}_2$, defines $\boldsymbol{r}_2$; by the regularity of Section 3.2 $\boldsymbol{g}_0$, $\delta\tilde{\boldsymbol{\theta}}$, and $\boldsymbol{r}_2$ are square integrable and $\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}$ is symmetric. Bilinearity of the covariance and $\mathrm{Cov}_{\boldsymbol{X}}[\delta\tilde{\boldsymbol{\theta}}, \boldsymbol{g}_0] = \mathbb{E}_{\boldsymbol{X}}[\delta\tilde{\boldsymbol{\theta}}\,\boldsymbol{\xi}^\top]$ give

$$\boldsymbol{\Sigma}_{\text{slack}} = \mathrm{Cov}_{\boldsymbol{X}}[\delta\tilde{\boldsymbol{\theta}}, \boldsymbol{g}] = \boldsymbol{\Sigma}_\xi + \boldsymbol{V}\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}^\top + \mathrm{Cov}_{\boldsymbol{X}}[\delta\tilde{\boldsymbol{\theta}}, \boldsymbol{r}_2] = \boldsymbol{\Sigma}_\xi + \boldsymbol{V}\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} + \mathrm{Cov}_{\boldsymbol{X}}[\delta\tilde{\boldsymbol{\theta}}, \boldsymbol{r}_2], \quad (12)$$

which is (i); its trace splits σ^2_{jac}, with $\mathrm{tr}(\boldsymbol{V}\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}) \ge 0$ for a positive semidefinite $\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}$, as for any product of two positive semidefinite matrices.

Step 2: the update covariance, exactly. Both methods are read at the common reference iterate of Section 3.2; the direct one has $\delta\tilde{\boldsymbol{\theta}} \equiv \boldsymbol{0}$ and update $\boldsymbol{g}_0$. By bilinearity, for the square integrable summands of the chain,

$$\begin{aligned}\mathrm{Cov}_{\boldsymbol{X}}[\boldsymbol{g}] = \mathrm{Cov}_{\boldsymbol{X}}[\boldsymbol{g}_0] + \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\boldsymbol{V}\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} + \mathrm{Cov}_{\boldsymbol{X}}[\boldsymbol{r}_2] + \big(\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\boldsymbol{\Sigma}_\xi + \boldsymbol{\Sigma}_\xi^\top \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\big) \\ + \big(\mathrm{Cov}_{\boldsymbol{X}}[\boldsymbol{g}_0, \boldsymbol{r}_2] + \mathrm{Cov}_{\boldsymbol{X}}[\boldsymbol{r}_2, \boldsymbol{g}_0]\big) + \big(\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\mathrm{Cov}_{\boldsymbol{X}}[\delta\tilde{\boldsymbol{\theta}}, \boldsymbol{r}_2] + \mathrm{Cov}_{\boldsymbol{X}}[\boldsymbol{r}_2, \delta\tilde{\boldsymbol{\theta}}]\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\big),\end{aligned} \quad (13)$$

the symmetry of $\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}$ turning $\mathrm{Cov}_{\boldsymbol{X}}[\boldsymbol{g}_0, \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\delta\tilde{\boldsymbol{\theta}}] = \boldsymbol{\Sigma}_\xi^\top \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}^\top$ into $\boldsymbol{\Sigma}_\xi^\top \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}$. Contracting with $\boldsymbol{e}_b$, by $\boldsymbol{e}_b^\top(\boldsymbol{A} + \boldsymbol{A}^\top)\boldsymbol{e}_b = 2\,\boldsymbol{e}_b^\top \boldsymbol{A}\boldsymbol{e}_b$,

$$\sigma^2_{\text{eff}} = \boldsymbol{e}_b^\top \mathrm{Cov}_{\boldsymbol{X}}[\boldsymbol{g}_0]\,\boldsymbol{e}_b + 2\,\boldsymbol{e}_b^\top \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\boldsymbol{\Sigma}_\xi \boldsymbol{e}_b + \boldsymbol{e}_b^\top \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\boldsymbol{V}\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\boldsymbol{e}_b + \varrho(\boldsymbol{e}_b), \quad (14)$$

the six $\boldsymbol{r}_2$ matrices of (13) projecting to the three scalars of $\varrho(\boldsymbol{e}_b)$. With the positivity-map slope s explicit every term carries s^2; the display is that identity at $s = 1$, $\boldsymbol{g}$ already including ψ'. At $\delta\tilde{\boldsymbol{\theta}} \equiv \boldsymbol{0}$ the coupling, jitter, and remainder terms are covariances with a constant and vanish, whatever $\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}$. For $\boldsymbol{e}_b$ a coordinate direction of the slack, as the bias channel is, (1) gives $\boldsymbol{e}_b^\top \boldsymbol{g}_0 = \psi'(\bar{w})\,\boldsymbol{e}_b^\top \nabla_{\boldsymbol{\theta}}\mathcal{L}$ at the coordinate $\bar{w} = \boldsymbol{e}_b^\top \bar{\tilde{\boldsymbol{\theta}}}$ of the reference iterate, so under (A2) the direct value is

$$\boldsymbol{e}_b^\top \mathrm{Cov}_{\boldsymbol{X}}[\boldsymbol{g}_0]\,\boldsymbol{e}_b = \psi'(\bar{w})^2\,\sigma^2_{\text{obj}} \in \big[(\sigma_s - \varepsilon)^2, (\sigma_s + \varepsilon)^2\big]\sigma^2_{\text{obj}}, \quad (15)$$

written $\sigma_s^2\sigma^2_{\text{obj}}$ within the bracket of (A2). This proves (ii) and the first clause of (iii). Substituting (12) into the coupling term,

$$\sigma^2_{\text{eff}} = \sigma_s^2\sigma^2_{\text{obj}} + 2\,\boldsymbol{e}_b^\top \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\boldsymbol{\Sigma}_{\text{slack}}\boldsymbol{e}_b - \boldsymbol{e}_b^\top \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\boldsymbol{V}\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\boldsymbol{e}_b + 2\,\boldsymbol{e}_b^\top \mathrm{Cov}_{\boldsymbol{X}}[\boldsymbol{g}_0, \boldsymbol{r}_2]\boldsymbol{e}_b + \boldsymbol{e}_b^\top \mathrm{Cov}_{\boldsymbol{X}}[\boldsymbol{r}_2]\boldsymbol{e}_b,$$

the self-coupling $\boldsymbol{V}\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}$ inside $\boldsymbol{\Sigma}_{\text{slack}}$ counting the jitter term twice and the middle covariance of ϱ once; (5) is exact on $\boldsymbol{\Sigma}_\xi$ alone.

Step 3: the sign of the excess, and the orders. For symmetric $\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}$ and the covariance $\boldsymbol{V} \succeq 0$,

$$\boldsymbol{e}_b^\top \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\boldsymbol{V}\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\boldsymbol{e}_b = \mathrm{Var}_{\boldsymbol{X}}\big[\langle \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\boldsymbol{e}_b, \delta\tilde{\boldsymbol{\theta}}\rangle\big] \ge 0, \quad (16)$$

with no assumption on the coupling, and the net coupling of Section 3.2, read at $\boldsymbol{e}_b$ with $\|\boldsymbol{e}_b\| = 1$, is

$$c = 2\,\boldsymbol{e}_b^\top \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\boldsymbol{\Sigma}_\xi \boldsymbol{e}_b + \varrho(\boldsymbol{e}_b). \quad (17)$$

Adding (16) and (17) to (14) with (15),

$$\sigma^2_{\text{eff}} \;\ge\; \sigma_s^2\sigma^2_{\text{obj}} + c,$$

which is (iii), and strictly so wherever $c > 0$; a one-dimensional Gaussian instance with $\boldsymbol{r}_2 \equiv \boldsymbol{0}$ has $c \ge 4$ with projected diffusions 1 and 9. We assume the sign of the coupling term wherever a

statement consumes it, for the reasons recorded under the net coupling in Appendix A.1. Scaling the fluctuation as $t\boldsymbol{u}$ and the remainder as $t\boldsymbol{w}_1 + t^2\boldsymbol{w}_2$ at a fixed batch law turns (14) into the exact polynomial

$$\sigma^2_{\text{eff}}(t) - \sigma_s^2\sigma^2_{\text{obj}} = t\big(2\,\boldsymbol{e}_b^\top \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\mathrm{Cov}_{\boldsymbol{X}}[\boldsymbol{u}, \boldsymbol{g}_0]\boldsymbol{e}_b + 2\,\boldsymbol{e}_b^\top \mathrm{Cov}_{\boldsymbol{X}}[\boldsymbol{g}_0, \boldsymbol{w}_1]\boldsymbol{e}_b\big) + t^2\big(\boldsymbol{e}_b^\top \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\mathrm{Cov}_{\boldsymbol{X}}[\boldsymbol{u}]\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\boldsymbol{e}_b + \cdots\big) + t^3(\cdots) + t^4(\cdots),$$

each covariance being homogeneous of degree one in each argument, so that no constant term survives and the jitter term, quadratic in the fluctuation, cannot appear below degree two. A polynomial $a_1t + a_2t^2 + a_3t^3 + a_4t^4$ with $a_1 \neq 0$ has the sign of a_1 at every sufficiently small $t > 0$, so a positive linear coefficient gives the strict ordering there while a negative one reverses it: at small scale the alignment decides and the jitter term, in the quadratic coefficient, decides nothing. A latent weight independent of the batch that forms the gradient has $\boldsymbol{\Sigma}_\xi = \boldsymbol{0}$, and for a finite batch law the pairing $\mathrm{Cov}_{\boldsymbol{X}}[\boldsymbol{g}_0, \boldsymbol{w}_1]$ of the frozen gradient with the curvature-fluctuation part of the remainder vanishes as well, so both first-order terms are absent and the excess begins at second order. This is (vi).

Step 4: the stationary density. Projecting the Itô process of (A1) onto $\boldsymbol{e}_b$, $w = \boldsymbol{e}_b^\top\tilde{\boldsymbol{\theta}}$, gives the scalar equation of (iv), with Fokker–Planck diffusion coefficient $D = (\eta/2)\sigma^2_{\text{eff}}$. Its drift $-\tilde{\mathcal{L}}'(w)$ is the projected mean update, each mean update being the gradient of its own pullback landscape under the domination hypotheses of Appendix A.1. The passage to the stationary Fokker–Planck equation is the register of (A1); what follows is proved. Continuity of $\tilde{\mathcal{L}}'$ and continuity and positivity of σ^2_{eff} supply, by the fundamental theorem of calculus, an antiderivative U_{eff} with

$$U'_{\text{eff}} = \frac{\tilde{\mathcal{L}}'}{\sigma^2_{\text{eff}}}, \qquad \pi = \frac{C}{\sigma^2_{\text{eff}}}\exp\Big(-\frac{2}{\eta}U_{\text{eff}}\Big), \qquad D\pi = \frac{C\eta}{2}\exp\Big(-\frac{2}{\eta}U_{\text{eff}}\Big),$$

the prefactor σ^{-2}_{eff} cancelling against D exactly, so that for $\eta \neq 0$

$$(D\pi)' = -C\,U'_{\text{eff}}\exp\Big(-\frac{2}{\eta}U_{\text{eff}}\Big) = -\tilde{\mathcal{L}}'\,\pi \tag{18}$$

at every point: π has zero probability current, with no expansion in η. Every zero-flux density ρ has this form, (18) for ρ reading $(D\rho\, e^{(2/\eta)U_{\text{eff}}})' = 0$, so

$$\rho = \frac{C_\rho}{\sigma^2_{\text{eff}}}\exp\Big(-\frac{2}{\eta}U_{\text{eff}}\Big), \qquad C_\rho = \sigma^2_{\text{eff}}(0)\,\rho(0)\,e^{(2/\eta)U_{\text{eff}}(0)},$$

which licenses the definite article. The zero-flux condition follows from the equation itself. A non-negative solution of the full stationary equation has a constant current J, $(D\rho)' = -\tilde{\mathcal{L}}'\rho - J$, and

$$D\rho \;=\; \Big(K - J\int_0^w e^{(2/\eta)U_{\text{eff}}(u)}\,du\Big)\,e^{-(2/\eta)U_{\text{eff}}(w)},$$

where for $\eta > 0$ and U_{eff} bounded below the integral tends to $\pm\infty$ as $w \to \pm\infty$, so $J \neq 0$ drives $D\rho$ negative at one end of the line; hence $J = 0$. The lower bound is what a confining loss supplies and cannot be dropped, a potential unbounded below admitting a transient solution of nonzero current. On a constant band $\sigma^2_{\text{eff}} \equiv c$, $U_{\text{eff}} = \tilde{\mathcal{L}}/c$ and $\pi \propto \exp(-2\tilde{\mathcal{L}}/(\eta c))$: the Boltzmann measure is the constant-diffusion case, not a borrowing.

Step 5: the barrier. Let $\alpha = \tilde{\mathcal{L}}(w_s) - \tilde{\mathcal{L}}(w_b) > 0$. On a constant band $\sigma^2_{\text{eff}} \equiv c$ the fundamental theorem of calculus for $U'_{\text{eff}} = \tilde{\mathcal{L}}'/c$ gives

$$U_{\text{eff}}(w_s) - U_{\text{eff}}(w_b) = \frac{1}{c}\int_{w_b}^{w_s}\tilde{\mathcal{L}}'(u)\,du = \frac{\alpha}{c},$$

exactly and with no monotonicity asked. On a band with $c_{\min} \leq \sigma^2_{\text{eff}} \leq c_{\max}$ climbed monotonically, $\tilde{\mathcal{L}}' \geq 0$ on $[w_b, w_s]$ by (A2),

$$\frac{\tilde{\mathcal{L}}'}{c_{\max}} \leq \frac{\tilde{\mathcal{L}}'}{\sigma^2_{\text{eff}}} \leq \frac{\tilde{\mathcal{L}}'}{c_{\min}} \quad \text{pointwise, hence} \quad \frac{\alpha}{c_{\max}} \leq U_{\text{eff}}(w_s) - U_{\text{eff}}(w_b) \leq \frac{\alpha}{c_{\min}},$$

the interval integrability of $\tilde{\mathcal{L}}'$ and $\tilde{\mathcal{L}}'/\sigma^2_{\text{eff}}$ following from their continuity on the band and the antiderivative from continuity and positivity on an open interval containing it. The bounds pinch to

α/c on a constant band and are strict on a non-constant one. Monotone climbing is an added restriction over $\alpha > 0$: it forbids an intervening critical point, the single-barrier profile of (A2) supplies it where Corollary 1 needs it, and without it the pointwise inequalities reverse where $\tilde{\mathcal{L}}' < 0$. Multiplying by $2/\eta$ gives the exponent $2\alpha/(\eta c)$ of Corollary 1 and its bracket $[2\alpha/(\eta c_{\max}),\ 2\alpha/(\eta c_{\min})]$. □

The proposition is a statement about the curvature of $\log \pi$ at one point and not about the width of the measure, except on the constant band, where the ratio statement is about the density itself. It rules out a reading in which the lift adds convexity, since the same $\tilde{\mathcal{L}}''(w^*)$ stands on both sides and only the divisor changes, and it supports the reading in which the direct method, whose divisor is the small value $\sigma_s^2\sigma_{\text{obj}}^2$ on the shoulder, samples a measure concentrated at the point where it was trapped, while the lift divides the same curvature by a strictly larger number.

Remark 2 (The matched-magnitude cell)**.** A perturbation of the latent weight sampled independently of the batch that forms the gradient reproduces $\boldsymbol{V}$ and with it the jitter term, while that independence sends the frozen coupling to zero and, in the finite-sample model, the first-order remainder pairing with it; the coupling channel of a centered latent weight independent of the batch is empty. Its excess over the direct value therefore begins at second order in the perturbation scale, while at a positive linear coefficient the lift's begins at first order, by (vi) and Step 3. The ordering of the lemma survives in that cell in its non-strict form, σ_{eff}^2 at least the direct value, provided the remainder channel is dominated by the jitter term it perturbs; no strict ordering is claimed there. A control that realizes this cell finds it inert, so the lemma's ordering is necessary for the observed effect and demonstrably not sufficient, which is why the two parameter-fluctuation terms are kept as separately named objects throughout.

A.4 Escape time

Corollary 1 (Escape time; Theorem 1(iii) in full)**.** *Let Assumptions A1–A2 hold. Let $\alpha = \tilde{\mathcal{L}}(w_s) - \tilde{\mathcal{L}}(w_b) > 0$ be the reference action along $\boldsymbol{e}_b$, read σ_{eff}^2 as constant across the barrier band, through the bracket of Lemma 1*(v)*, at the direct value $\sigma_s^2\sigma_{\text{obj}}^2$ for direct softplus and at* (5) *for the lift, and let $\mathsf{T}(x)$ be the mean first-passage time of $dw = -\tilde{\mathcal{L}}'(w)\,dt + \sqrt{\eta}\,\sigma_{\text{eff}}\,dB_t$ from a start $x < w_s$ on the escape interval $[a, b]$ of Assumption A2, reflecting at a and absorbing at b, under Dynkin's formula for its generator and continuous up to the closed interval. Then, for every $\delta > 0$, there is $L(\delta) > 0$, depending on δ, on $\tilde{\mathcal{L}}$, on the escape interval, and on the start x, and not on $\eta\sigma_{\text{eff}}^2$, with $\big(2L(\delta)/(\eta\sigma_{\text{eff}}^2)\big)\, e^{2(\alpha-\delta)/(\eta\sigma_{\text{eff}}^2)} \le \mathsf{T}(x) \le \big(2(b-x)(b-a)/(\eta\sigma_{\text{eff}}^2)\big)\, e^{2\alpha/(\eta\sigma_{\text{eff}}^2)}$, written $\mathsf{T}(x) \asymp \exp(2\alpha/(\eta\,\sigma_{\text{eff}}^2))$, and $\eta\sigma_{\text{eff}}^2 \log \mathsf{T}(x) \to 2\alpha$ as $\eta\sigma_{\text{eff}}^2 \to 0^+$. The exponent $2\alpha/(\eta\sigma_{\text{eff}}^2)$ is strictly decreasing in σ_{eff}^2, so whenever $c > 0$ the lift's exponent is strictly the smaller of the two, by Lemma 1*(iii)*, and, under the same condition, for every sufficiently small step size η the lift's mean first-passage time is itself the smaller. Both orderings hold at the level of the exponent.*

Proof. Step 1: the bias-channel reduction. Projected onto $\boldsymbol{e}_b$, the Itô process of (A1) has state $w = \boldsymbol{e}_b^\top\tilde{\boldsymbol{\theta}}$, drift $-\tilde{\mathcal{L}}'(w)$ with $\tilde{\mathcal{L}}$ read along $\boldsymbol{e}_b$, and diffusion coefficient $\eta\sigma_{\text{eff}}^2$: by Step 2 of Appendix A.3, $\eta\sigma_s^2\sigma_{\text{obj}}^2$ for the direct method, whose $\delta\tilde{\boldsymbol{\theta}} \equiv \mathbf{0}$, and η times (5) for the lift. The reduction is a lower bound on the multi-dimensional action and not an identity: for a path $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0 + q\,\boldsymbol{e}_b$ confined to the line, with q differentiable, drift $\boldsymbol{\mu}(\boldsymbol{\gamma}) = m(q)\,\boldsymbol{e}_b$, and diffusion matrix $\boldsymbol{\Sigma}(\boldsymbol{\gamma}) \succ 0$, the two Freidlin–Wentzell (Freidlin & Wentzell, 2012) integrands being interval integrable on $[0, \mathsf{t}]$,

$$\int_0^{\mathsf{t}} \frac{(\dot{q} - m(q))^2}{\boldsymbol{e}_b^\top\boldsymbol{\Sigma}(\boldsymbol{\gamma})\boldsymbol{e}_b}\,dt \;\le\; \int_0^{\mathsf{t}} (\dot{\boldsymbol{\gamma}} - \boldsymbol{\mu}(\boldsymbol{\gamma}))^\top\boldsymbol{\Sigma}(\boldsymbol{\gamma})^{-1}(\dot{\boldsymbol{\gamma}} - \boldsymbol{\mu}(\boldsymbol{\gamma}))\,dt, \tag{19}$$

by $(\boldsymbol{e}_b^\top\boldsymbol{\Sigma}\boldsymbol{e}_b)^{-1} \le \boldsymbol{e}_b^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_b$ for a unit $\boldsymbol{e}_b$ (Cauchy–Schwarz), with equality exactly when $\boldsymbol{e}_b$ is an eigendirection of $\boldsymbol{\Sigma}$. The scalar understates the escape cost off an eigendirection, and the comparison below is stated on the scalar.

Step 2: the boundary-value problem and the bracket. Write $\sigma^2 = \eta\sigma_{\text{eff}}^2$, held at its band value, the constant reading that Lemma 1(v) brackets. Dynkin's formula for the generator $\frac{1}{2}\sigma^2\partial^2 - \tilde{\mathcal{L}}'\partial$ is the hypothesis that the mean exit time solves

$$\tfrac{1}{2}\sigma^2\,\mathsf{T}''(x) - \tilde{\mathcal{L}}'(x)\,\mathsf{T}'(x) = -1 \ \text{ on } (a,b), \qquad \mathsf{T}(b) = 0, \qquad \mathsf{T}'(a^+) = 0, \tag{20}$$

the modeling step of the corollary; everything downstream of it is proved. Under continuity of $\tilde{\mathcal{L}}$ on $[a,b]$ and differentiability inside, two differentiations through the fundamental theorem of calculus show that

$$\mathsf{T}(x) = \frac{2}{\sigma^2}\int_x^b e^{2\tilde{\mathcal{L}}(y)/\sigma^2}\int_a^y e^{-2\tilde{\mathcal{L}}(z)/\sigma^2}\,dz\,dy \tag{21}$$

solves (20), and it is the only solution continuous on $[a,b]$: for two solutions T,V, $\big((\mathsf{T}'-\mathsf{V}')e^{-2\tilde{\mathcal{L}}/\sigma^2}\big)' = 0$ on (a,b) with limit 0 at a^+, so $\mathsf{T}' = \mathsf{V}'$ and $\mathsf{T}-\mathsf{V} \equiv \mathsf{T}(b)-\mathsf{V}(b) = 0$. Every mean exit time given through Dynkin's formula is therefore (21). Fix $\delta > 0$. Continuity of $\tilde{\mathcal{L}}$ at w_s and w_b and the ordering $a < w_b < w_s < b$, $a \le x < w_s$, give windows $[s_1,s_2] \subset [x,b]$ about w_s with $\tilde{\mathcal{L}} \ge \tilde{\mathcal{L}}(w_s) - \delta/2$ and $[t_1,t_2] \subset [a,s_1]$ about w_b with $\tilde{\mathcal{L}} \le \tilde{\mathcal{L}}(w_b) + \delta/2$; the integrands of (21) being positive,

$$\mathsf{T}(x) \ge \frac{2}{\sigma^2}\int_{s_1}^{s_2} e^{2\tilde{\mathcal{L}}(y)/\sigma^2}\int_{t_1}^{t_2} e^{-2\tilde{\mathcal{L}}(z)/\sigma^2}\,dz\,dy \ge \frac{2L(\delta)}{\sigma^2}\,e^{2(\alpha-\delta)/\sigma^2}, \tag{22}$$

where $L(\delta) = (s_2-s_1)(t_2-t_1) > 0$ depends on δ, $\tilde{\mathcal{L}}$, the escape interval, and x, not on σ^2, and degenerates as $\delta \to 0$. The escape-interval clause of (A2) makes $\tilde{\mathcal{L}}(w_s)$ and $\tilde{\mathcal{L}}(w_b)$ the extreme values of $\tilde{\mathcal{L}}$ on $[a,b]$, so

$$\mathsf{T}(x) \le \frac{2}{\sigma^2}\int_x^b e^{2\tilde{\mathcal{L}}(w_s)/\sigma^2}\int_a^y e^{-2\tilde{\mathcal{L}}(w_b)/\sigma^2}\,dz\,dy \le \frac{2(b-x)(b-a)}{\sigma^2}\,e^{2\alpha/\sigma^2}. \tag{23}$$

The pair (22)–(23) is what $\mathsf{T}(x) \asymp \exp(2\alpha/(\eta\sigma^2_{\text{eff}}))$ abbreviates; at fixed diffusion it does not close, $L(\delta) \to 0$, and at the level of exponents it does:

$$2(\alpha-\delta) + \sigma^2\log\frac{2L(\delta)}{\sigma^2} \;\le\; \sigma^2\log\mathsf{T}(x) \;\le\; 2\alpha + \sigma^2\log\frac{2(b-x)(b-a)}{\sigma^2},$$

with $\sigma^2\log(K/\sigma^2) \to 0$ as $\sigma^2 \to 0^+$ and δ arbitrary, so $\sigma^2\log\mathsf{T}(x) \to 2\alpha$, for (21) and hence for any mean exit time given through Dynkin's formula at every diffusion. The sharp Kramers prefactor is not addressed: a function $Ke^{2\alpha/\sigma^2}$ satisfies both bounds on a window of small diffusions, so the prefactor is settled by nothing here. Where the diffusion varies across the band, the exponent lies in $[2\alpha/(\eta c_{\max}),\, 2\alpha/(\eta c_{\min})]$ by Lemma 1(v) and equals $2\alpha/(\eta\sigma^2_{\text{eff}})$ where the band is constant.

Step 3: the comparison. For $\alpha,\eta > 0$ the map $x \mapsto 2\alpha/(\eta x)$ is strictly decreasing on $(0,\infty)$. With $\sigma^2_{\text{dir}} = \sigma_s^2\sigma^2_{\text{obj}}$ and σ^2_{lift} for (5), Lemma 1(iii) at $c > 0$ gives

$$0 < \sigma^2_{\text{dir}} < \sigma^2_{\text{dir}} + c \le \sigma^2_{\text{lift}} \quad\Longrightarrow\quad \frac{2\alpha}{\eta\sigma^2_{\text{lift}}} < \frac{2\alpha}{\eta\sigma^2_{\text{dir}}}, \tag{24}$$

$\sigma^2_{\text{dir}} > 0$ by $\sigma^2_{\text{obj}} > 0$ in Section 3.2 and $\sigma_s > \varepsilon$ in (A2); where the pointwise ordering holds on the band no bracket on either diffusion is needed. For the times themselves choose $\delta < \alpha(1-\sigma^2_{\text{dir}}/\sigma^2_{\text{lift}})$, so that $2(\alpha-\delta)/\sigma^2_{\text{dir}} > 2\alpha/\sigma^2_{\text{lift}}$; then (23) at $\eta\sigma^2_{\text{lift}}$ and (22) at $\eta\sigma^2_{\text{dir}}$ give

$$\mathsf{T}_{\text{lift}}(x) \le \frac{2(b-x)(b-a)}{\eta\sigma^2_{\text{lift}}}\,e^{2\alpha/(\eta\sigma^2_{\text{lift}})} < \frac{2L(\delta)}{\eta\sigma^2_{\text{dir}}}\,e^{2(\alpha-\delta)/(\eta\sigma^2_{\text{dir}})} \le \mathsf{T}_{\text{dir}}(x)$$

for every sufficiently small η, with no gap hypothesis and none on prefactors; under Dynkin's formula for both processes at every η the same holds for the mean first-passage times, by the uniqueness above. The exponent limit at each diffusion gives $\eta\log\big(\mathsf{T}_{\text{dir}}(x)/\mathsf{T}_{\text{lift}}(x)\big) \to 2\alpha(1/\sigma^2_{\text{dir}} - 1/\sigma^2_{\text{lift}})$, and at a fixed η the ratio is pinned by

$$\frac{L(\delta)\,\sigma^2_{\text{lift}}}{(b-x)(b-a)\,\sigma^2_{\text{dir}}}\,e^{\frac{2(\alpha-\delta)}{\eta\sigma^2_{\text{dir}}} - \frac{2\alpha}{\eta\sigma^2_{\text{lift}}}} \;\le\; \frac{\mathsf{T}_{\text{dir}}(x)}{\mathsf{T}_{\text{lift}}(x)} \;\le\; \frac{(b-x)(b-a)\,\sigma^2_{\text{lift}}}{L(\delta)\,\sigma^2_{\text{dir}}}\,e^{\frac{2\alpha}{\eta\sigma^2_{\text{dir}}} - \frac{2(\alpha-\delta)}{\eta\sigma^2_{\text{lift}}}},$$

whose width is the missing prefactor. The action α belongs to the one pullback landscape held fixed across the comparison, as the common reference iterate requires, so the lift enters through σ^2_{eff} alone. □

Remark 3 (The power of the positivity-map slope)**.** On the shoulder the coupling changes the constant in the exponent and not its order. Direct softplus's diffusion there is $\sigma_s^2\sigma_{\text{obj}}^2$, so its exponent is of order σ_s^{-2}, and a diffusion $\sigma_s^2\sigma_{\text{obj}}^2 + \sigma_s q$ with q bounded away from zero on the band, the form a term bearing a single factor of ψ' would supply, would give an exponent of order σ_s^{-1} and a divergent ratio of the two. The coupling term does not take that form. By the chain rule $\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} = \operatorname{diag}(\psi')\boldsymbol{H}_\theta\operatorname{diag}(\psi') + \operatorname{diag}(\psi'' \odot \boldsymbol{g}_\theta)$ and $\boldsymbol{\Sigma}_\xi = \boldsymbol{\Xi}_\theta\operatorname{diag}(\psi')$, with $\boldsymbol{H}_\theta$ and $\boldsymbol{g}_\theta$ the mean Hessian and the mean gradient in constrained units and $\boldsymbol{\Xi}_\theta = \mathbb{E}_{\boldsymbol{X}}[\delta\tilde{\boldsymbol{\theta}}\,\boldsymbol{\xi}_\theta^\top]$ the coupling in those units. At a reference iterate with $\psi' = \sigma_s$ in every coordinate,

$$2\,\boldsymbol{e}_b^\top \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\boldsymbol{\Sigma}_\xi\boldsymbol{e}_b \;=\; 2\sigma_s^3\,\boldsymbol{e}_b^\top \boldsymbol{H}_\theta\boldsymbol{\Xi}_\theta\boldsymbol{e}_b \;+\; 2\sigma_s^2(1-\sigma_s)\,\boldsymbol{e}_b^\top\operatorname{diag}(\boldsymbol{g}_\theta)\boldsymbol{\Xi}_\theta\boldsymbol{e}_b,$$

which is $O(\sigma_s^2)$ and never of order σ_s. The same bound holds at the relative tolerance $\rho = \varepsilon/\sigma_s$ of (A2), since softplus has $\psi'' \le \psi'$. The jitter term and the remainder channel carry the same power, the latter because $\boldsymbol{r}_2 = \psi' \odot \tilde{\boldsymbol{r}}$. Both exponents are therefore of order σ_s^{-2} and their ratio is bounded, at the factor $\sigma_{\text{eff}}^2/(\sigma_s^2\sigma_{\text{obj}}^2)$ of Lemma 1.

Corollary 2 (Time on the shoulder; the dwell half of Theorem 1(iii))**.** *Under the hypotheses of Corollary 1, and under Dynkin's formula for its generator at the band indicator, with the occupation time continuous up to the closed interval, let τ be the first passage of w to b and let $\mathsf{D}(x) = \mathbb{E}_x\big[\int_0^\tau \mathbf{1}_{[a,w_s]}(w_t)\,dt\big]$ be the mean time the diffusion spends below the barrier top before escape. Then $\eta\sigma_{\text{eff}}^2\log\mathsf{D}(x) \to 2\alpha$ and $\eta\sigma_{\text{eff}}^2\log\big(\mathsf{T}(x)/\mathsf{D}(x)\big) \to 0$ as $\eta\sigma_{\text{eff}}^2 \to 0^+$. This time therefore carries the exponent of Corollary 1 and is strictly the shorter under the lift whenever $c > 0$, for every sufficiently small step size, while the fraction $\mathsf{D}(x)/\mathsf{T}(x)$ of the time before escape spent below the barrier top is subexponential in $\eta\sigma_{\text{eff}}^2$.*

Proof. Write $\sigma^2 = \eta\sigma_{\text{eff}}^2$. Dynkin's formula for the generator $\frac{1}{2}\sigma^2\partial^2 - \tilde{\mathcal{L}}'\partial$, the same modeling step that (20) is, states that D solves $\frac{1}{2}\sigma^2\mathsf{D}'' - \tilde{\mathcal{L}}'\mathsf{D}' = -\mathbf{1}_{[a,w_s]}$ at every point of (a,b) other than w_s, with $\mathsf{D}(b) = 0$ and $\mathsf{D}'(a^+) = 0$. At w_s the indicator jumps and D' has a corner, so the second derivative does not exist there and the equation is stated off that point; D' is continuous across it, which is what the uniqueness argument consumes. The two differentiations and the uniqueness argument of Appendix A.4 otherwise apply verbatim to a bounded measurable right-hand side, so

$$\mathsf{D}(x) = \frac{2}{\sigma^2}\int_x^b e^{2\tilde{\mathcal{L}}(y)/\sigma^2}\int_a^y \mathbf{1}_{[a,w_s]}(z)\,e^{-2\tilde{\mathcal{L}}(z)/\sigma^2}\,dz\,dy, \tag{25}$$

which is (21) with the inner integrand restricted. The integrands being non-negative, $\mathsf{D}(x) \le \mathsf{T}(x)$, so (23) bounds $\mathsf{D}(x)$ above. For the lower bound take the windows of (22): there $[t_1,t_2] \subset [a,s_1]$ with $s_1 \le w_s$, so $\mathbf{1}_{[a,w_s]} \equiv 1$ on $[t_1,t_2]$ and the same computation gives $\mathsf{D}(x) \ge (2L(\delta)/\sigma^2)\,e^{2(\alpha-\delta)/\sigma^2}$ for every $\delta > 0$. That the inner window lies below the barrier top is the ordering $a < w_b < w_s$ of (A2), and it is the only step of Appendix A.4 that the indicator touches. The two bounds are those of Corollary 1, so $\sigma^2\log\mathsf{D}(x) \to 2\alpha$ by its exponent argument, and $\sigma^2\log(\mathsf{T}(x)/\mathsf{D}(x)) = \sigma^2\log\mathsf{T}(x) - \sigma^2\log\mathsf{D}(x) \to 0$. The comparison is Step 3 of Appendix A.4 with (25) in place of (21), unchanged. □

A.5 Proof of Theorem 1

Proof. The reference iterate, $\boldsymbol{e}_b$, σ_{obj}^2, and σ_{eff}^2 are fixed as in Lemma 1. Parts (i) and (ii) are Lemma 1(ii)–(iii) and (vi), under that lemma's hypotheses less its parts (iv)–(v), which leaves the regularity of Section 3.2 and the first clause of (A2). Part (iii) is Corollary 1 together with Corollary 2, under their hypotheses, its reading of σ_{eff}^2 as constant across the barrier band bracketed by Lemma 1(v). The sign condition $c > 0$ enters no part of the theorem, only the ordering those corollaries carry. □

A.6 Projection

This subsection states the results behind the comparison with the projection. Projection outputs raw parameters into the loss, so its update includes no factor of ψ', and its constrained weights are stored variables that no batch conditions. We read its objective channel into the diffusion undiminished, $\boldsymbol{\Sigma}_{\text{PGD}} = \boldsymbol{\Sigma}_{\text{obj}}$. This identification is an input to every statement below, and it leaves out what

the return step does to the update on the active set. With the positivity-map slope kept explicit, the three diffusion matrices at a common reference iterate are $\boldsymbol{\Sigma}_{\mathrm{PGD}} = \boldsymbol{\Sigma}_{\mathrm{obj}}$, $\boldsymbol{\Sigma}_{\mathrm{dir}} = \sigma_s^2 \boldsymbol{\Sigma}_{\mathrm{obj}}$, and $\boldsymbol{\Sigma}_{\mathrm{lift}} = \sigma_s^2(\boldsymbol{\Sigma}_{\mathrm{obj}} + \boldsymbol{C} + \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} \boldsymbol{V} \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}})$ with $\boldsymbol{C} = \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} \boldsymbol{\Sigma}_{\xi} + \boldsymbol{\Sigma}_{\xi}^{\top} \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}$ the coupling matrix, and each is read along a unit escape direction $\boldsymbol{v}$ as the quadratic form $\boldsymbol{v}^{\top} \boldsymbol{\Sigma} \boldsymbol{v}$. This is the one place where the conversion of Appendix A's opening applies: the excess terms of (5) are σ_s^2 times $\boldsymbol{v}^{\top}(\boldsymbol{C} + \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} \boldsymbol{V} \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}})\boldsymbol{v}$, and the comparison is stated in the paper's convention. Theorem 3 is stated at a vanishing remainder channel; the threshold reads the excess net of the remainder channel that the net coupling c folds into itself, which is the same arithmetic with the remainder added to the excess. The common reference iterate of Section 3.2 is not itself licensed for a scheme with no latent coordinate, and the comparison assumes it in addition.

No account resting on the size of the exploration alone therefore settles the lift against projection. The direct method is the coldest of the three where the coupling channel is non-negative along $\boldsymbol{v}$, since $0 < \sigma_s < 1$ puts it below projection and $\boldsymbol{v}^{\top} \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} \boldsymbol{V} \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} \boldsymbol{v} \geq 0$ puts it at or below the lift. The lift is above or below projection along $\boldsymbol{v}$ according to the threshold of Theorem 3, in either direction.

Theorem 3 (Directional escape comparison)**.** *Let $\alpha > 0$, $\eta > 0$, and $0 < \sigma_s$, and let $\boldsymbol{v}$ be a direction on which both directional diffusions are positive, $\boldsymbol{v}^{\top} \boldsymbol{\Sigma}_{\mathrm{PGD}} \boldsymbol{v} > 0$ and $\boldsymbol{v}^{\top} \boldsymbol{\Sigma}_{\mathrm{lift}} \boldsymbol{v} > 0$. Then the lift's Arrhenius exponent $2\alpha/(\eta\, \boldsymbol{v}^{\top} \boldsymbol{\Sigma}_{\mathrm{lift}} \boldsymbol{v})$ is strictly below projection's $2\alpha/(\eta\, \boldsymbol{v}^{\top} \boldsymbol{\Sigma}_{\mathrm{PGD}} \boldsymbol{v})$ if and only if*

$$(\sigma_s^{-2} - 1)\, \boldsymbol{v}^{\top} \boldsymbol{\Sigma}_{\mathrm{obj}} \boldsymbol{v} \;<\; \boldsymbol{v}^{\top} \big(\boldsymbol{C} + \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} \boldsymbol{V} \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\big) \boldsymbol{v},$$

equivalently, in the paper's convention, if and only if the coupling and jitter terms of (5) *along $\boldsymbol{v}$ exceed $\sigma_s^{-2} - 1$ times the attenuated objective contribution $\sigma_s^2\, \boldsymbol{v}^{\top} \boldsymbol{\Sigma}_{\mathrm{obj}} \boldsymbol{v}$. If in addition the barrier clause of* (A2) *holds along $\boldsymbol{v}$ with Dynkin's formula for both mean exit times, and the inequality holds, then the lift's mean first-passage time is strictly below projection's for every sufficiently small projection diffusion, the lifted diffusion exceeding it by the surplus that the inequality fixes.*

The threshold is out of reach in the regime the paper is about. Written in absolute terms the bar is $(1 - \sigma_s^2)\, \boldsymbol{v}^{\top} \boldsymbol{\Sigma}_{\mathrm{obj}} \boldsymbol{v}$, which rises to $\boldsymbol{v}^{\top} \boldsymbol{\Sigma}_{\mathrm{obj}} \boldsymbol{v}$ as the positivity map attenuates, while the coupling and jitter terms of (5) are of order σ_s^2 by the power count of Appendix A.3 and fall to zero with it. The inequality therefore fails for every sufficiently small σ_s, and the directional test can be met only where the positivity-map derivative is of order one, away from the shoulder. What the threshold decides is consequently whether projection is overtaken outside the attenuated region, and not whether the mechanism of Section 3 favors the lift within it.

Proof. For $\alpha, \eta > 0$ the map $x \mapsto 2\alpha/(\eta x)$ is strictly decreasing on $(0, \infty)$, so with both directional diffusions positive

$$\begin{aligned} \frac{2\alpha}{\eta\, \boldsymbol{v}^{\top} \boldsymbol{\Sigma}_{\mathrm{lift}} \boldsymbol{v}} < \frac{2\alpha}{\eta\, \boldsymbol{v}^{\top} \boldsymbol{\Sigma}_{\mathrm{PGD}} \boldsymbol{v}} \;&\Longleftrightarrow\; \boldsymbol{v}^{\top} \boldsymbol{\Sigma}_{\mathrm{obj}} \boldsymbol{v} < \sigma_s^2\, \boldsymbol{v}^{\top} \big(\boldsymbol{\Sigma}_{\mathrm{obj}} + \boldsymbol{C} + \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} \boldsymbol{V} \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\big) \boldsymbol{v} \\ &\Longleftrightarrow\; (\sigma_s^{-2} - 1)\, \boldsymbol{v}^{\top} \boldsymbol{\Sigma}_{\mathrm{obj}} \boldsymbol{v} < \boldsymbol{v}^{\top} \big(\boldsymbol{C} + \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} \boldsymbol{V} \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\big) \boldsymbol{v}, \end{aligned}$$

the last step dividing by $\sigma_s^2 > 0$. Dividing by $\boldsymbol{v}^{\top} \boldsymbol{\Sigma}_{\mathrm{obj}} \boldsymbol{v} > 0$ states the same condition on the ratio of the excess to the objective channel against $\sigma_s^{-2} - 1$, and multiplying both sides by σ_s^2 gives the paper's convention. For the exit times let $\sigma_{\mathrm{PGD}}^2 = \eta\, \boldsymbol{v}^{\top} \boldsymbol{\Sigma}_{\mathrm{PGD}} \boldsymbol{v}$ and $\Delta = \eta\, \boldsymbol{v}^{\top} (\boldsymbol{\Sigma}_{\mathrm{lift}} - \boldsymbol{\Sigma}_{\mathrm{PGD}}) \boldsymbol{v} > 0$, the surplus the inequality fixes. Under the barrier clause of (A2) along $\boldsymbol{v}$ and Dynkin's formula for both, the bracket (22)–(23) at the diffusions σ_{PGD}^2 and $\sigma_{\mathrm{PGD}}^2 + \Delta$, with a fixed $\delta \in (0, \alpha)$, gives

$$\mathsf{T}_{\mathrm{lift}}(x) \leq \frac{2(b-x)(b-a)}{\sigma_{\mathrm{PGD}}^2 + \Delta}\, e^{2\alpha/(\sigma_{\mathrm{PGD}}^2 + \Delta)} < \frac{2L(\delta)}{\sigma_{\mathrm{PGD}}^2}\, e^{2(\alpha-\delta)/\sigma_{\mathrm{PGD}}^2} \leq \mathsf{T}_{\mathrm{PGD}}(x)$$

for every sufficiently small σ_{PGD}^2, the middle exponent diverging while the first tends to $2\alpha/\Delta$; at a prescribed pair of diffusions the ordering holds under the explicit gap inequality between the two middle terms. □

The reduction of the multi-dimensional action to the scalar along $\boldsymbol{v}$ is that of Step 1 of Appendix A.4, and the hypothesis that the improving move is a crossing along one fixed direction is inherited from it. The theorem is a biconditional and it is true whatever a measurement of its right-hand side returns; it decides what the paper may claim, not whether the comparison holds. The threshold is a real discriminator: a positively aligned configuration clears it at $\sigma_s = 3/5$ and fails it at $\sigma_s = 1/3$, so positive alignment alone is not sufficient.

Theorem 4 (Projection has no coupling term)**.** *Let the latent weight be independent of the conditioning batch, as it is for projection and for direct softplus. Then for every probe count $T \geq 1$ the cross-covariance estimator of* (6) *vanishes, the frozen coupling $\boldsymbol{\Sigma}_\xi$ vanishes with the fluctuation and the coupling matrix $\boldsymbol{C}$ with it, the lift's diffusion matrix evaluated at slope one, zero coupling, and zero jitter term equals $\boldsymbol{\Sigma}_{\mathrm{PGD}}$, the ratio of the excess to the objective channel along every direction is zero, and the threshold of Theorem 3 fails at slope one and at every slope below one.*

Proof. A latent weight that does not vary with the batch has $\delta\tilde{\boldsymbol{\theta}}_k = \tilde{\boldsymbol{\theta}}_k - \bar{\tilde{\boldsymbol{\theta}}} = \mathbf{0}$ on every probe and $\delta\tilde{\boldsymbol{\theta}} \equiv \mathbf{0}$ in population, so for every $T \geq 1$

$$\widehat{\boldsymbol{\Sigma}}^{(t)} = \mathbf{0} \ \text{ summand by summand in (6)}, \qquad \boldsymbol{\Sigma}_\xi = \mathbb{E}_{\boldsymbol{X}}[\delta\tilde{\boldsymbol{\theta}}\,\boldsymbol{\xi}^\top] = \mathbf{0},$$
$$\boldsymbol{C} = \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\boldsymbol{\Sigma}_\xi + \boldsymbol{\Sigma}_\xi^\top \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} = \mathbf{0}.$$

Substituting into the definitions at $\sigma_s = 1$, $\boldsymbol{C} = \mathbf{0}$, and $\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\boldsymbol{V}\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}} = \mathbf{0}$,

$$\boldsymbol{\Sigma}_{\mathrm{lift}} = \boldsymbol{\Sigma}_{\mathrm{obj}} = \boldsymbol{\Sigma}_{\mathrm{PGD}}, \qquad \frac{\boldsymbol{v}^\top(\boldsymbol{C} + \boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\boldsymbol{V}\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}})\boldsymbol{v}}{\boldsymbol{v}^\top\boldsymbol{\Sigma}_{\mathrm{obj}}\boldsymbol{v}} = 0 \quad \text{for every } \boldsymbol{v},$$

and the threshold of Theorem 3, $\sigma_s^{-2} - 1 < 0$ in ratio form, fails at $\sigma_s = 1$, where it reads $0 < 0$, and at every $\sigma_s < 1$, where $\sigma_s^{-2} - 1 > 0$. □

The theorem does not close the converse of Theorem 2: a structural zero does not by itself say anything about the loss. It says that the alignment route of Theorem 3 is unavailable to projection, whose exploration is whatever direction the minibatch gradient noise supplies and which no amount of temperature redirects, whereas the lift's coupling term is built from the same batch that produces the gradient fluctuation. That leaves the threshold of Theorem 3 as a prediction with an explicit test. Contract the curvature-weighted frozen coupling with the escape direction, $\sigma_s^2\,\boldsymbol{v}^\top\boldsymbol{C}\boldsymbol{v}$, which is the coupling term $2\,\boldsymbol{v}^\top\boldsymbol{H}_{\tilde{\boldsymbol{\theta}}}\boldsymbol{\Sigma}_\xi\boldsymbol{v}$ of (5) along $\boldsymbol{v}$, and compare it against $\sigma_s^{-2} - 1$ times the attenuated objective contribution along $\boldsymbol{v}$. Clearing that bar settles the threshold, since the jitter term the contraction omits is non-negative, while failing it leaves the question open, since an instance with zero coupling and a positive jitter term clears the threshold. The trace of (6) is the wrong monitor for this test, being blind to direction. The test is directional and concerns the frozen coupling alone, we have not run it, and we assert nothing about whether the threshold holds at our operating point.

Proposition 1 (One projected coordinate at the boundary; the Lindley recursion)**.** *Let $\gamma, \sigma, \eta > 0$, let $\zeta_0, \zeta_1, \ldots$ be independent standard normal with distribution function Φ, and let $W_{k+1} = \max\big(0,\, W_k - \eta(\gamma + \sigma\zeta_k)\big)$, $W_0 = 0$, with $\rho = \gamma/\sigma$. A visit to the boundary ends at the first step whose update points into the cone, so it lasts a geometric number of steps of mean $1/\Phi(-\rho)$. At every step the clamped fraction $\mathbb{P}(W_k = 0)$ is free of η and σ, a function of ρ alone, non-decreasing in ρ, and tends to 1 as $\rho \to \infty$. Writing $S_j = -\eta\sigma\sum_{i<j}(\rho + \zeta_i)$, we have $W_k = \max_{0\leq j\leq k} S_j$ in distribution; $M = \sup_{j\geq 0} S_j$ is finite almost surely with finite mean $\mathbb{E}[M] = \eta\sigma\, m(\rho)$, m depending on ρ alone; and $W_k \to M$ in distribution. The law of M is the recursion's unique stationary law, under which $\mathbb{P}(W_\infty = 0) = \exp\big(-\sum_{k\geq 1} k^{-1}\Phi(-\rho\sqrt{k})\big)$ (Spitzer, 1956), decreasing to 0 as $\rho \downarrow 0$.*

Proof. Write $X_k = -\eta(\gamma + \sigma\zeta_k) = -\eta\sigma(\rho + \zeta_k)$, so that $W_{k+1} = (W_k + X_k)^+$ with $X_1, X_2, \ldots$ independent and normal of mean $-\eta\gamma < 0$ and variance $\eta^2\sigma^2$, and $S_j = X_1 + \cdots + X_j$.

The dwell. From $W_k = 0$,

$$W_{k+1} > 0 \iff X_k > 0 \iff \zeta_k < -\rho, \qquad \mathbb{P}(\zeta_k < -\rho) = \Phi(-\rho) \in (0, 1). \tag{26}$$

The ζ_k are independent, so the clamped steps before the release are geometric of that success probability, and the visit, those steps together with the one that ends it, has mean $1/\Phi(-\rho)$, at every step size and at every depth.

The stationary law. Iterating the recursion from $W_0 = 0$ gives $W_k = \max_{0\leq j\leq k} S_j$ in distribution, non-decreasing in k, and $S_j/j \to -\eta\gamma < 0$ almost surely, so $M = \sup_{j\geq 0} S_j$ is finite almost surely and $W_k \to M$ in distribution from every starting point (Asmussen, 2003). Since S_j is normal of mean $-j\eta\gamma$ and variance $j\eta^2\sigma^2$, $\mathbb{P}(S_j > 0) = \Phi(-\rho\sqrt{j})$, free of η, and the ladder-epoch identity $\mathbb{P}(M = 0) = \exp\big(-\sum_{j\geq 1} j^{-1}\mathbb{P}(S_j > 0)\big)$ (Spitzer, 1956) gives the stated form. One clause here is

not machine-checked, in two named parts: this identity and the limit at $\rho \downarrow 0$ it supplies; with it goes the stationarity and uniqueness of the limit law, which the formalization does not state. The identity in distribution $W_k = \max_{0 \le j \le k} S_j$, the convergence $W_k \to M$, the finiteness of M and of its mean, the $\eta\sigma$ scaling, the geometric visit length, and, at every step of the recursion, the clamped fraction's dependence on ρ alone, its monotonicity and its limit at $\rho \to \infty$, are. Each term decreases in ρ, so the clamped fraction increases in ρ; it tends to 1 as $\rho \to \infty$ by dominated convergence, and to 0 as $\rho \downarrow 0$ because $\sum_{j \le \rho^{-2}} j^{-1} \Phi(-\rho\sqrt{j}) \ge \Phi(-1) \sum_{j \le \rho^{-2}} j^{-1}$ diverges.

The offset. The recursion is positively homogeneous in $\eta\sigma$: $M = \eta\sigma \sup_{j \ge 0} \big(- \sum_{i \le j} (\rho + \zeta_i) \big)$, whose supremum depends on ρ alone and has a finite mean, the walk having negative drift and finite variance (Asmussen, 2003). Hence $\mathbb{E}[W_\infty] = \eta\sigma\, m(\rho)$: the stationary iterate sits above the constrained optimum at a distance proportional to the step size. □

What projection pays at the boundary is the re-crossing and not the stay, which is all the proposition licenses.

B Experimental details

Every run trains with Adam (Kingma & Ba, 2015) and reports the iterate of its lowest held-out loss, except the two-dimensional flows, which report their final iterate, and the image regularizer, which reports the iterate of its highest validation PSNR. The tabular benchmarks take the splits and preprocessing of Papamakarios et al. (2017), the two-dimensional flows the published recipe of Huang et al. (2021), and the image regularizer that of Ehrhardt et al. (2026). The energy-based models need the log-normalizer of their target, which is estimated by importance sampling with the trained sampler, a HINT flow (Kruse et al., 2021), as proposal, and by quadrature in one dimension. On the image regularizer the lift scales its permutation-invariant summary to a fixed norm, without which it collapses to a constant output at the published step.

Table 2: Configuration of each experiment: the data dimension, the ICNN width and depth, the body widths, the batch size, the step size, the iteration count, and the seeds each comparison reports.

experiment	D	ICNN	body	n	η	iterations	seeds
one-dimensional targets	1	128×3	64, 64, 96	256	10^{-3}	3,000	30
POWER	6	256×3	96, 96, 128	512	10^{-3}	4,000	3
HEPMASS	21	512×5	64, 64, 96	256	10^{-3}	4,000	3
MiniBooNE	43	512×5	64, 64, 96	256	10^{-3}	4,000	3
two-dimensional flows	2	$5 \times (32 \times 3)$	64, 64, 96	128	5×10^{-3}	19,550	5
Darcy transport map	100	128×4	128, 128, 192	256	10^{-3}	4,000	10
image regularizer	256^2	32 filters	32, 64, 128	8	5×10^{-4}	2,500	23

Table 3: The configuration each method is reported at on the tabular datasets, where a ramp raises the step on the energy, and on MiniBooNE on the sampler as well.

	POWER	HEPMASS	MiniBooNE
lift	none	ramp	ramp and sampler
PGD	ramp	none	ramp and sampler
direct softplus	none	none	none

C The held-out read along training

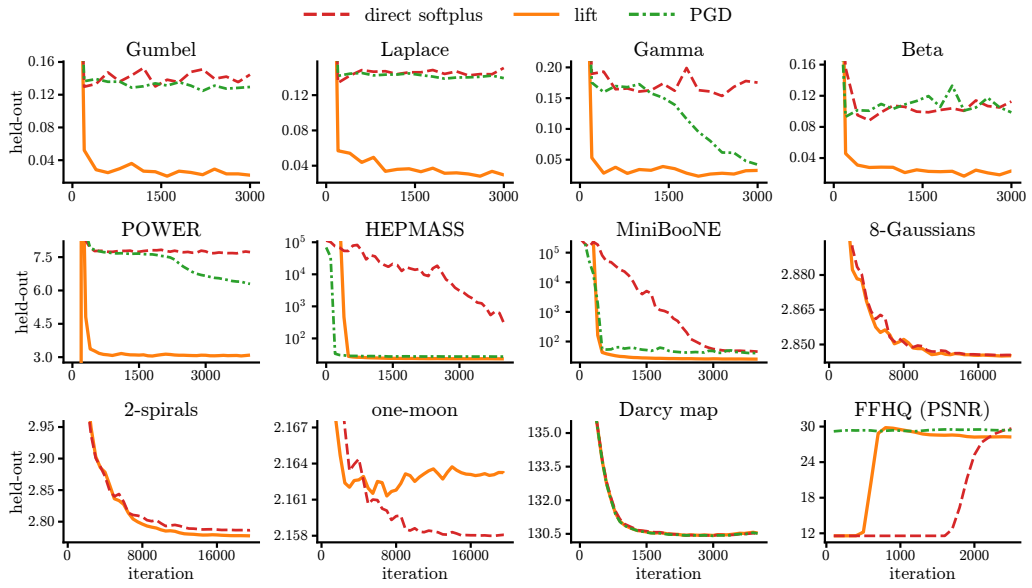


Figure 9: The held-out read along training on each problem, the pointwise median over seeds of each method. Higher is better on the image regularizer and lower on the rest.

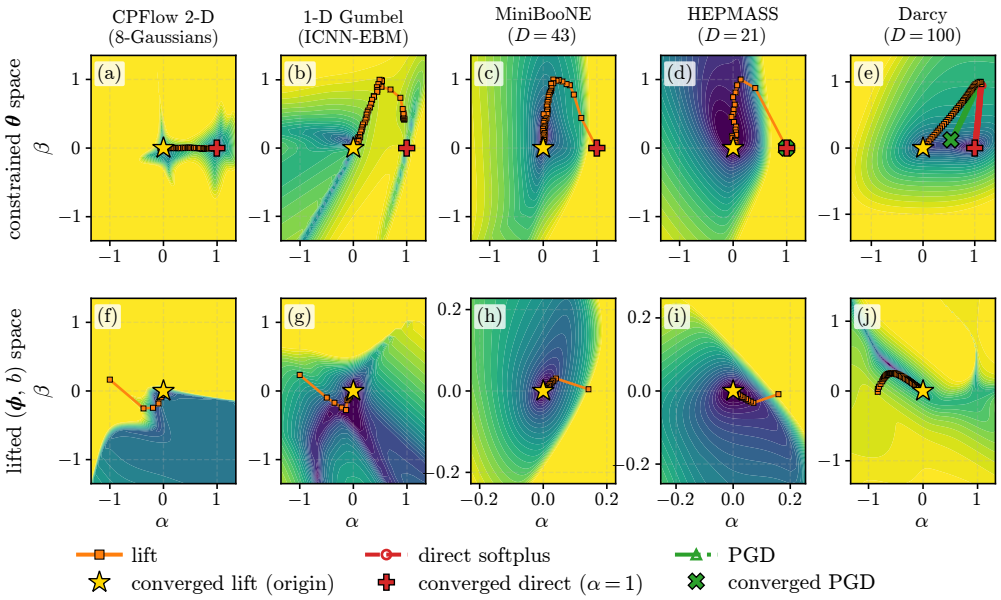
D ONE-DIMENSIONAL GALLERY

Table 4: Total-variation distance to the target at the best held-out loss, the median over seeds with the tenth and ninetieth percentiles. The lift and direct softplus bands are disjoint on every target.

target	lift	direct softplus	PGD
Gumbel	0.016 [0.007, 0.029]	0.137 [0.130, 0.142]	0.130 [0.118, 0.144]
Laplace	0.027 [0.015, 0.044]	0.140 [0.138, 0.142]	0.139 [0.136, 0.145]
Gamma	0.012 [0.006, 0.061]	0.125 [0.113, 0.138]	0.038 [0.024, 0.055]
Beta	0.012 [0.006, 0.025]	0.073 [0.062, 0.087]	0.072 [0.055, 0.090]

E LOSS-LANDSCAPE GEOMETRY

Figure 10 renders the loss along two-dimensional slices through the final iterates of single runs of a convex potential flow, three energy-based models, and the transport map, in the manner of Li et al. (2018). In the constrained coordinates $\boldsymbol{\theta}$ the first direction runs from the final iterate of the lift, at the origin, to that of direct softplus. In the lifted coordinates $(\boldsymbol{\phi}, \boldsymbol{b})$ both directions are principal directions of the lift's trajectory. On the two larger energy-based models the constrained slice places direct softplus's final iterate above the lift's. While the constrained slice separates the two, the lifted slice is a single valley that the lift's trajectory follows. On Darcy, where the three methods tie (Section 4.3), the three final iterates coincide when each is scored under its own map.



Axes are the slice coordinates (α, β) in per-problem frame units. Each panel is shifted to its own minimum and colored on its own $\log_{10}$ scale, so shade reads within a panel and never across panels; per-panel colorbars are omitted.

Figure 10: Loss along two-dimensional slices through the final iterates, in the constrained coordinates with the lift at the origin and direct softplus at the right (top) and in the lifted coordinates (bottom).