TRAINING AS A FLOW OF DISTRIBUTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Analyses of neural-network training usually examine one aspect at a time, the dynamics of the optimizer, the energy of the loss landscape, or the geometry of the underlying space (representational, functional, or parameter), each in isolation. To unify all in a single apparatus, we use the Wasserstein gradient flow, the steepest descent of a free energy in Wasserstein geometry, which coincides with the Fokker–Planck equation and is realized in discrete time by the Jordan–Kinderlehrer–Otto (JKO) proximal scheme. This scheme unifies the energy, geometry, and learning dynamics in one object, making their interaction measurable. The measurements that this framework relies on and employs are high-dimensional and lack a ground truth to validate them, and it might deviate from the correct measurement without any indication of failure. We therefore validate each measurement on small, analytically tractable models, in which the answer is known in closed form. Employing generative classifiers versus discriminative models, and focusing on measurements introduced and known in the related literature, such as calibration, we compare these models and validate these results: (i) a predictor's calibration is controlled by the transport distance of its parameter distribution to the posterior, tracked along the flow down to a finite-sample floor; (ii) the geometry of the landscape sets the rate at which calibration is reached, governed by the slowest mode, (iii) the crossover of discriminate over generative classifiers occurs as a result of misspecification and in the case of the correct assumption, the crossover does not happen.

1 INTRODUCTION

Modern studies of training artificial systems tend to look at one thing at a time (Goodfellow et al., 2016; Bottou et al., 2018). The loss landscape describes the *energy* a network is trying to minimize (Choromanska et al., 2015; Li et al., 2018); the optimizer describes the *dynamics* by which the distribution of the parameters moves (Bottou et al., 2018; Kingma & Ba, 2014); a chosen metric, kernel, or preconditioner describes the *geometry* in which *downhill* is defined (Amari, 1998; Jacot et al., 2018; Kingma & Ba, 2014). These are usually studied with separate tools, and their interactions, where the interactions of energy, geometry, and dynamics have not been addressed.

A change of viewpoint binds the three together. Instead of following one parameter vector down a loss, follow the whole *distribution* of parameters down a *free energy*, and measure distances between distributions with optimal transport (Villani et al., 2009; Gabriel & Marco, 2019). The evolution that does this, the steepest descent of the free energy $F(\rho) = \int V\, d\rho + T \int \rho \log \rho$ in the Wasserstein-2 geometry, is realized by the Fokker–Planck equation and, one sample at a time, by noisy gradient descent (Langevin dynamics)(Risken, 1989; Welling & Teh, 2011). The variational solution of the Fokker–Planck equation, considering its equality to the Wasserstein gradient flow of the energy functional $\nabla_{w_\rho} F(\rho)$, has been developed as the JKO proximal Jordan et al. (1998). Under specific conditions that guarantee the appropriate growth of the potential function, there is a unique stationary distribution $\rho_s(x)$ for the Fokker–Planck equation, which is known as the *Gibbs distribution* (Otto & Villani, 2000; Risken, 1989). This is the distribution which minimizes the free energy functional over all probability densities on $\mathbb{R}^n$. The stationary distribution in the Bayesian framework is exactly the Bayesian posterior $\pi \propto e^{-V/T}$, (for details see 3), and the main quantity that makes a bridge between the theory and our framework and provides a closed form of computations later (details in 3). In the flow of distributions description, the energy F, the geometry W_2, and one step of the dynamics are written together, and their interaction is the ultimate object we aim to *measure*. Figure 1 demonstrates the general framework.

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.

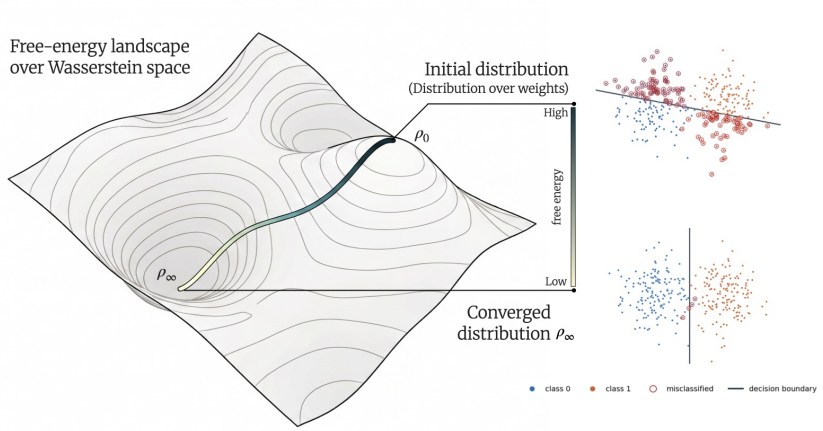

Figure 1: Training viewed as a distribution over the network's weights flowing down a free energy. *Left*: the free energy landscape over Wasserstein space; each point is a distribution ρ over weights, and its height is $F(\rho) = \mathbb{E}_\rho[V] - T\,H(\rho)$. The Wasserstein gradient flow carries ρ_0 downhill to the stationary posterior $\rho_\infty \propto e^{-V/T}$, dissipating free energy. *Right*: each distribution of weights induces a classifier, so a surface point yields a decision on the data. On the same fixed two-class data, ρ_0 gives a misplaced, uncertain boundary with high error, while ρ_∞ gives the correct, sharp boundary with low error.

This formulation yields several precise and testable predictions. First, *calibration*, the agreement between a model's predicted confidence and its empirical accuracy, is governed by proximity to the posterior (Guo et al., 2017; Villani et al., 2009): the expected calibration error is bounded by a constant multiple of the Wasserstein distance between the parameter distribution and the posterior, so that a smaller distance implies a smaller calibration error. Second, the *geometry* of the potential determines the **rate** of convergence: this rate is fixed by the spectrum of the Hessian of V, and in particular by its smallest eigenvalue, which coincides with the log-Sobolev constant that governs the mixing of the flow (Bakry & Émery, 2006; Otto & Villani, 2000; Vempala & Wibisono, 2019). Third, the distributional (generative) formulation is more sample-efficient than a discriminative one when the model is correctly specified; under *misspecification* the discriminative model overtakes it at large sample sizes, the classical crossover, and this crossover is attributable to misspecification rather than to the distributional form of the model (Ng & Jordan, 2001).

The obstacles and limitations do not come from theory, rather it is mostly related to the available instruments. First, none of the quantities on which this viewpoint depends are observed directly: the optimal-transport (Wasserstein) distance between distributions, the potential energy $\mathbb{E}_\rho[V]$, the entropy $H[\rho]$, the free energy $F = \mathbb{E}_\rho[V] - T\,H[\rho]$, and the convergence rate are not observed directly and each must be recovered by an *estimator*. The non-trivial ones are an entropic-transport (Sinkhorn) (Cuturi, 2013; Gabriel & Marco, 2019) solver for the Wasserstein distance and a nearest-neighbor estimator for the entropy (and hence for the free energy), while the convergence rate is read off the trajectory as the exponential decay rate of $W_2(\rho_t, \pi)$, the slope of its log-decay over time. In high dimensions, these estimators can degrade, often without any indication of failure. Second, the predictions above are hypotheses that must be verified. In the setting of ultimate interest, large, over-parameterized networks, no ground truth is available to distinguish a genuine effect from estimator error: a decreasing curve may reflect the predicted phenomenon or the failure of the instrument that produced it, and the two cannot be separated after the fact.

This study is the first stage of a programme that reads training through a unified lens and measures the interaction of its three components: the *energy* (the free energy, together with its potential and entropy terms), the *geometry* (the Wasserstein metric), and the *dynamic* (the flow and its rate of convergence). The program runs in three stages. First, *validation*: establish every measurement on analytically tractable models where the answer is known in closed form (this paper). Second, *confirmation*: apply the validated measurements along the training trajectory of real, over-parameterized

networks, where the phenomena of interest, feature learning, memorization, and the loss of generalization, actually arise, beginning with regimes that admit theoretical guarantees such as the mean-field limit Mei et al. (2018; 2019). Third, *scaling*: the framework is extended to modern large network architectures and to practical tasks such as generative modelling, where substantial computation is justified, and is devoted only to those measurements the earlier, less costly stages have already verified as reliable.

Because the setting is controlled, the results are *foundational, not final*. They show three things: the framework is consistent and computable (energy, geometry, and dynamics are three views of one object); the measurement tools reproduce known answers; and at least one predicted link from the geometry to a downstream outcome, calibration, which is confirmed here empirically. They are *not* statements about deep networks, nor evidence that the flow is a better training algorithm: here the posterior is available in closed form, so the flow is just a way of sampling it. The validated tools, the resolution laws, the comparison methodology, and the test for an optimizer's geometry all transfer to the next stage.

This paper is that beginning. We work in a controlled regime of analytically tractable models whose posterior, and therefore every functional of it, is available in closed form. There we *validate* each instrument against its exact value, and we exhibit the links from geometry to observable behaviour that the larger program is ultimately about. Our contributions are:

- A measurement pipeline whose every estimator is gated against a closed-form reference, including a transport distance estimator validated on non-Gaussian targets and a trajectory-level agreement between the exact proximal scheme and its explicit sampler.
- A demonstration that a predictor's **calibration is controlled by the transport distance** of its parameter distribution to the posterior, tracked along the flow, together with the ablations showing this is a property of the posterior the flow reaches, not of the flow's mechanics.
- Evidence that the **geometry of the landscape sets the rate** at which calibration is attained, governed by the slowest mode of the flow.
- A controlled study of **when a distributional (generative) model outperforms a discriminative one**, identifying model misspecification, not "being generative", as the cause of the classical large-sample crossover.

2 Related work

Gradient flows and their discretization. The variational (minimizing-movement) scheme that realizes Fokker–Planck as steepest descent of a free energy in Wasserstein space is due to Jordan, Kinderlehrer and Otto (Jordan et al., 1998), with the underlying geometry developed by (Otto & Villani, 2000; Ambrosio et al., 2005). Its particle realization is stochastic-gradient Langevin dynamics Welling & Teh (2011), one instance of the broader stochastic-gradient Markov Chain Monte Carlo (MCMC) family (Ma et al., 2015). The sampling floor we plot every convergence curve against, the residual Wasserstein error of an M particle empirical measure, is characterized by the finite-sample rates of (Fournier & Guillin, 2015; Weed & Bach, 2019). Recent machine-learning work solves the proximal step at scale with input-convex networks and related parameterizations (Mokrov et al., 2021; Fan et al., 2021; Xu et al., 2023); our aim is different, we use the scheme as a *measurement lens* and verification target, not as a generative solver.

Distributional views of training. Mean-field analyses show the empirical distribution of a wide network's units evolves as a Wasserstein gradient flow (Chizat & Bach, 2018; Mei et al., 2018; Rotskoff & Vanden-Eijnden, 2022), the theoretical guarantee for treating an ensemble's evolution as a geometric flow. Convergence of the underlying dynamics is governed by log-Sobolev / curvature conditions (Bakry & Émery, 2006; Otto & Villani, 2000; Vempala & Wibisono, 2019); in the Gaussian case, this rate constant is exactly the smallest Hessian eigenvalue (Bakry–Émery).

Calibration and the Bayesian predictive. Point-estimate networks are typically overconfident, and post-hoc temperature scaling is the standard fix (Guo et al., 2017); ensembling improves calibration by averaging predictions over parameter uncertainty (Lakshminarayanan et al., 2017). Our link between calibration error and distance to the posterior follows from Kantorovich–Rubinstein duality and Lipschitz continuity of the predictive map (Villani et al., 2009), giving the bound $\mathrm{ECE} \le c\, W_2$

we test directly. We report calibration at the exact-posterior temperature (T=1); the cold-/tempered-posterior literature Wenzel et al. (2020) documents cases where sharpening ($T < 1$) further helps. The controlled comparison of distributional versus point classifiers, and the role of misspecification, follows the classical analysis of Ng and Jordan Ng & Jordan (2001).

3 METHODS AND SETUP

The flow we run settles, after some time, into an *equilibrium* distribution, and that distribution is a Gibbs (Boltzmann) distribution $\rho_\infty \propto e^{-V(\theta)/T}$, the shape a physical system reaches at temperature T under an energy V. The whole scheme is built around this single function, the *potential* (or energy) V: choosing V chooses what the flow converges to. Setting $T = 1$ and matching this equilibrium to the posterior in Bayes' rule, $\rho_\infty \propto e^{-V} = p(\mathcal{D} \mid \theta)\, p(\theta) \propto p(\theta \mid \mathcal{D})$, fixes the potential to be nothing more than the training loss plus the regularizer,

$$V(\theta) = \underbrace{-\log p(\mathcal{D} \mid \theta)}_{\text{loss}} \underbrace{-\log p(\theta)}_{\text{regularizer}}.$$

In the Fokker-Planck formulation (equivalently, the Langevin equation), the flow relaxes any distribution to the stationary Gibbs measure $e^{-V/T}$; in our Bayesian framework, that measure is the posterior (at $T = 1$), so running the flow is exactly Bayesian inference, and move mass toward low V means "move toward the posterior". The flow only ever needs the gradient ∇V, which is ordinary backprop plus weight decay, and the evidence $p(\mathcal{D})$ enters as a constant with zero gradient, so it drops out.

When the likelihood and the prior are both Gaussian, the loss and the regularizer are quadratic in θ, completing the square gives $V(\theta) = \frac{1}{2}(\theta - \mu^\star)^\top A\,(\theta - \mu^\star) + \text{const}$, where $\mu^\star$ is the best-fit parameter (the bottom of the bowl, considering the bowl as the space in which, the distribution flows) and $A = \nabla^2 V$ is the Hessian. The posterior is then *exactly* Gaussian, $\pi = \mathcal{N}(\mu^\star,\, TA^{-1})$. The matrix A is the *shape* of the bowl, how steeply V rises in each direction, and it plays the role of the geometry throughout: its most-stretched to least-stretched ratio (its condition number, measured by $\frac{\lambda_{\max}}{\lambda_{\min}}$) says how much direction matters; it sets the width of the posterior (TA^{-1}, wide where the bowl is shallow); and its smallest eigenvalue sets how fast the flow converges. The whole scheme is built around this function, the *potential* (or energy) $V(\theta) = -\log p(\mathcal{D} \mid \theta) - \log p(\theta)$, which is the training loss plus the regularizer. Equivalently, since the flow's stationary law is the Gibbs measure $\pi \propto e^{-V/T}$, specifying V is the same as specifying the target the flow converges to.

Two engines for the same flow. There are two engines for the same flow: the implicit and explicit engines. Both engines discretize the same flow to the posterior $\pi = \mathcal{N}(\mu^\star, TA^{-1})$, that its mean follows the $dm/dt = -A(m - \mu^\star)$ and its covariance relaxes toward TA^{-1}.They differ only in how they step through time. **The implicit engine** (Gaussian–JKO) evolves the distribution exactly, by updating its (m, Σ) in closed form. For the mean, it takes one backward-Euler step of $dm/dt = -A(m - \mu^\star)$, giving $m_{k+1} = (I + \tau A)^{-1}(m_k + \tau b)$, where $b = A\mu^\star$ is the data-driven linear term of the potential (equivalently $\mu^\star = A^{-1}b$). Because the step is implicit, this linear solution moves the mean toward $\mu^\star$ without ever overshooting, so it is stable at any step size τ. This engine uses the global quadratic structure (A, b) directly, carries no sampling noise, and serves as the exact reference, but it is limited to the Gaussian (quadratic) case. **The explicit engine** (ensemble SGLD) instead approximates the distribution by a cloud of M particles and moves each one forward with Euler–Maruyama Langevin dynamics: $\theta \leftarrow \theta - \eta \nabla V(\theta) +$ noise, where $\nabla V(\theta) = A(\theta - \mu^\star)$, the empirical distribution of the cloud approximates the current distribution ρ of the parameters. Being explicit, the force is evaluated at the current point; it is stable only for a small step $\eta \propto 1/\lambda_{\max(A)}$, and it carries finite M sampling noise. In return, it touches the potential only through pointwise gradients ∇V, so it generalizes to non-Gaussian targets where no closed form exists.

Models. The study uses two small models built on a common interface, each provides its Hessian, gradient, and posterior mean and covariance, so one pipeline can run on both. *Bayesian linear regression* fits a linear map to data; its parameters are the weights, its potential is squared error plus an ℓ_2 penalty, and its Hessian $A = X^\top X/\sigma^2 + I/\sigma_p^2$ is a full, rotated matrix, a coupled geometry, which is why we use it to recover an optimizer's implicit metric. The *Gaussian class-conditional classifier* treats each class as a Gaussian "blob" and learns only *where each blob's centre sits*; its parameters are the stacked class means (so there are (number of classes)×(feature dimension) of

them). It classifies a new point by asking which centre is nearest, weighted by the class prior (Bayes' rule, i.e. a softmax of distances). This outputs true class probabilities rather than a label, and it makes a real classification decision, which its calibration is well-defined and can be measured. Because both likelihood and prior are Gaussian, the posterior over the centres is again exactly Gaussian; and simply using *unequal class sizes* makes its geometry, characterized by A, stretched while keeping that posterior exact.

Experimental settings. All studies use temperature $T = 1$ (the exact-posterior temperature); the class-conditional classifier uses likelihood scale $\sigma = 1$ and prior scale $\sigma_p = 3$. *(i) Calibration* (Fig. 2): $C = 3$ classes of unequal sizes $(15, 30, 45)$ in feature dimension $d = 8$, i.e. $C \times d = 24$ parameters, evolved by the exact Gaussian–JKO flow (step $\tau = 0.02$, 90 steps) and read out with $M = 600$ particles over 20 seeds; the finite-ensemble floor is estimated from $M = 600$ exact posterior draws. *(ii) Geometry sets the rate* (Fig.3): $C = 3$ classes at three imbalance levels $(20, 20, 20)$, $(8, 20, 60)$, $(5, 15, 120)$ in $d = 6$ (18 parameters), evolved by ensemble SGLD with $M = 800$ particles to a common effective-time window $t_{\max} = 1$. *(iii) Sample efficiency* (Fig.4): $C = 2$ classes in $d = 10$ (20 parameters), training sizes $n \in \{10, 20, 40, 80, 160, 320\}$ each averaged over 20 replicates, with the generative predictor formed from $M = 300$ SGLD particles (400 steps) and evaluated on a test set of 500 points per class.

4 Results

4.1 Calibration is controlled by transport distance, (Fig. 2, left).

The envelope of Fig. 2 rests on a duality that connects a *geometric* quantity we can measure, the transport distance between the parameter cloud ρ and the posterior π, to the network *dynamics* that we aim to study, the calibration of the predictions, which live in a different space. Kantorovich–Rubinstein duality states that the Wasserstein-1 distance equals the largest gap any 1-Lipschitz test function can open between the two measures,

$$W_1(\rho, \pi) = \sup_{\mathrm{Lip}(f) \le 1} \left| \int f \, d\rho - \int f \, d\pi \right|$$

Kantorovich & Rubinshtein (1958); Villani et al. (2009); Gabriel & Marco (2019). In words: if two distributions are close in transport distance, then the average of *any* function that does not change too fast is close between them, so transport distance controls every smooth summary of the two measures at once. This is exactly what we need, because our predictor is such an average of a well-behaved function of the parameters, the map from the parameters to the predictor is Lipschitz. Since the calibration error is a Lipschitz functional of the predicted probabilities, the calibration error inherits the bound

$$\mathrm{ECE}(\rho) \;\le\; c\, W_2(\rho, \pi),$$

See the details of computation in A.1.

Although the posterior π is known in closed form in this controlled setting, the Wasserstein distance we *measure* from a finite ensemble is not zero: representing π by M particles and comparing their empirical summary to π leaves a residual error that comes purely from finite sampling and the randomness of the draw. Concretely, we summarize each ensemble by its empirical mean and covariance and evaluate the closed-form Gaussian (Bures) W_2 between the fitted Gaussian $\mathcal{N}(\hat{m}, \hat{\Sigma})$ and π; the floor is the expected value of this quantity when the M particles, (here $M = 600$), are drawn *exactly* from π. It is a resolution limit: the measured W_2 of a converging flow cannot fall below it, because at best the flow reaches π and is then an M sample picture of π. The floor is dominated by the covariance term and scales as $r_{\mathrm{eff}}/\sqrt{M}$, equivalently $d/\sqrt{M}$ when the covariance is full-rank and isotropic, and $\propto 1/\sqrt{M}$ at fixed dimension, where r_{eff} is the effective rank of the ensemble. This rate follows from the concentration of the empirical mean and covariance Koltchinskii & Lounici (2017); Wainwright (2019) carried through the closed-form Gaussian W_2 Givens & Shortt (1984); Bhatia et al. (2019). We therefore compute the floor empirically, by drawing exact posterior samples.

The left diagram in Figure 2, from upper-right to lower-left, following the color from dark (early training) to bright (late training), shows that the flow pushes the ensemble closer to the true posterior, the transport distance W_2 on the horizontal axis shrinks, and the calibration error on the vertical

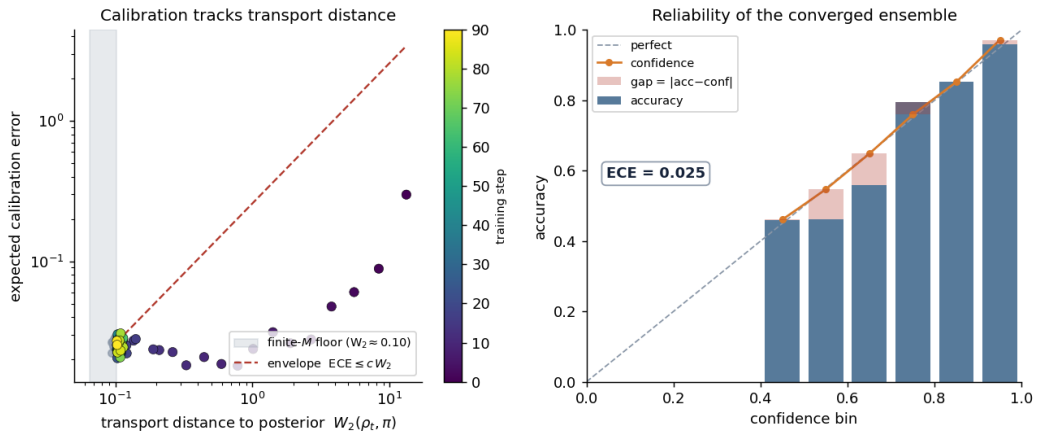

Figure 2: Calibration is controlled by transport distance to the posterior. *Left:* along the flow (colour = step), expected calibration error stays under an ECE $\leq cW_2$ envelope and descends with the transport distance until both saturate at the finite-ensemble floor (shaded). *Right:* reliability of the converged ensemble; bars are accuracy, the orange line is claimed confidence, the diagonal is perfect calibration; the shaded slivers are the per-bin gaps $|\overline{\text{acc}} - \overline{\text{conf}}|$ whose occupancy-weighted average is the reported ECE.

axis is in alignment with it. Every point stays under the red dashed line, which is the guarantee $\text{ECE}(\rho) \leq c\,W_2(\rho, \pi)$, calibration error can never exceed a constant times the distance to the posterior. Both quantities keep dropping until they hit the grey band on the left, the finite-ensemble floor. Calibration is controlled by closeness to the posterior; get closer to the posterior and the probabilities become more trustworthy, down to a floor set purely by how many particles you use.

The curve shows two specific things: the *convergence rate* is how fast the flow ρ_t approaches the posterior π as training proceeds (which connects this result to the result in 3); it sets the descent. The *sampling floor* is the smallest distance we can even measure with a finite ensemble of M particles; it is non-zero even for an ensemble drawn exactly from π. For the details of the background theorems, see A.1.

As mentioned before, this is an apparatus which studies the steepest descent of the energy functional for probability distributions with respect to the W_2 metric. It describes how the distribution in the Wasserstein space should evolve so that its energy functional, $F = \int V d\rho + \int \rho \log \rho$, decreases the most. Fig 5 in A.2, shows how free energy F and its components, V and $\mathbb{E}_\rho$ evolve, particularly in a way that the flow dissipates F and converges to $F_{\min}$ and potential energy to the posterior. This is a complementary result for the left diagram in Fig 2, showing that by converging to the posterior, both free energy and W_2 distance to the posterior decline along the flow, reaching their lowest value.

4.2 Reliability of the converged ensemble, (Fig. 2, right).

The *confidence* of a prediction is the probability the model places on the class it predicts, $\text{conf}(x) = \max_c \hat{p}(c \mid x)$; its *accuracy* is whether that prediction is correct. A classifier is *calibrated* when, among all inputs it predicts with confidence p, a fraction p are in fact correct Guo et al. (2017). The diagram tests this directly and asks whether for the fully converged ensemble, the stated confidence matches reality. Predictions are grouped into equal-width confidence bins, and for each bin we plot the mean claimed confidence (orange line) against the empirical accuracy (blue bars); the diagonal (dashed) marks perfect calibration, where accuracy equals confidence. Bars sitting below the confidence line indicate overconfidence (the model claims more than it delivers), above it underconfidence Niculescu-Mizil & Caruana (2005). The shaded sliver in each bin is that bin's gap $|\overline{\text{acc}}(B_m) - \overline{\text{conf}}(B_m)|$, and the expected calibration error printed in the panel is their occupancy-weighted average, $\text{ECE} = \sum_m \frac{|B_m|}{n} |\overline{\text{acc}}(B_m) - \overline{\text{conf}}(B_m)|$ Naeini et al. (2015); Guo et al. (2017). The shaded area is therefore a direct picture of the ECE, the more (occupancy-weighted) shading, the worse the calibration. Details of the binning and computation can be found in A.1. The per-bin

gaps between claimed confidence and real accuracy are small; their occupancy-weighted average is $\mathrm{ECE} \approx 0.025$.

A single calibration-error number can conceal structure; a model may be overconfident at high confidence and underconfident at low, yet still report a modest average. The reliability diagram exposes the full confidence range at once, and by shading the per-bin gaps, it shows both the ECE *and where it comes from*: here, the converged ensemble is calibrated nearly uniformly, with only slight overconfidence in the mid-confidence bins. It illustrates that the posterior the flow arrives at produces honest probabilities.

To have a systematic comparison between generative and discriminative classifiers (Ng & Jordan, 2001), respecting the calibration metric (Guo et al., 2017), we devised a test to compare the flow-based and logistic models. In this experiment, we compare Bayesian model averaging (the JKO flow), the exact posterior solution (exact draw), and *MAP point*, which reports the posterior's mode as the most-probable parameter. *Logistic MLE* and *Logistic +temperature*, which add the temperature parameter as a regularizer, employed as discriminative models, demonstrate higher expected calibration error (ECE). For more details, see Figure 6 and A.3

4.3 The geometry sets the rate, (Figure 3)

Calibration inherits whatever rate the flow converges at, so the question of *how fast* a model becomes calibrated is a question about the geometry. Holding the flow and the target family fixed and varying only the anisotropy of the landscape, the transport distance descends at a rate set by the slowest-relaxing direction, the smallest eigenvalue of the Hessian (Figure 3). The measured late-time decay rate is proportional to $\lambda_{\min}(A)$ across the sweep. This is one of the simplest statements available in the controlled regime that is about the interaction of geometry and dynamics: it changes when the geometry changes, and it is the tractable version of the rate at which the distribution converges to the posterior, which the next stages of the project aim to track during training. Both engines simulate the flow in discrete steps, and both are time discretizations of the same continuous dynamics: the explicit (Langevin/SGLD) engine takes an Euler–Maruyama step, the implicit (JKO) engine an implicit-Euler step. In each, the step size, η for the explicit engine, τ for the implicit one, is the amount of continuous (physical) time advanced per step, so after k steps the flow has run for a physical time $t = k\eta$ (or $k\tau$). We plot against the *effective time* $t = k\,\eta$, the continuous-time variable of the underlying dynamics after k discrete steps, rather than clock time or the step index. This axis is natural for two reasons. First, the flow relaxes as $e^{-\lambda t}$, so the decay rate reads directly off the slope. Second, different geometries require different stable step sizes $\eta = c/\lambda_{\max}(A)$, where c is a small fixed constant ($c \approx 0.05 - 0.08$); measuring at equal t therefore places all flows on one physical clock, whereas comparing at equal step count would not. Each geometry is accordingly run to a common horizon $t_{\max}$. Over the common time window, the transport distance to the posterior decays more slowly as the geometry becomes more anisotropic. The round bowl ($\mathrm{cond} = 1$, $\lambda_{\min} = 20$) reaches the finite-M floor first, the most anisotropic ($\mathrm{cond} = 24$, $\lambda_{\min} = 5$) last. The *measured late-time W_2 rate* is the exponential decay rate fitted over the late, pre-floor portion of each curve, $\mathrm{rate} = -d\log W_2/dt$, which for this Ornstein–Uhlenbeck relaxation converges to the slowest mode $\lambda_{\min}(A)$ Pavliotis (2014); Bakry et al. (2014). The right panel confirms the relationship: the measured rate is proportional to $\lambda_{\min}$ (linear fit, slope 1.23). This is evidence for the statement "geometry controls dynamics" on a small scale: holding the flow fixed and varying only the geometry, the rate at which the distribution, and hence its calibration, converges is set by the smallest eigenvalue of the Hessian.

4.4 Sample efficiency and the generative/discriminative crossover (Fig 4).

A model that commits to how the data are generated (the distributional / "generative" classifier) is more sample-efficient than one that only learns a decision boundary (the discriminative classifier) when its generative assumptions are correct, and it is overtaken as data accumulate only when those assumptions are wrong, a phenomenon which is known as *misspecification*. This is one of the canonical results of Ng & Jordan (2001), which illustrates that the generative learner reaches its (higher) asymptotic error faster, the discriminative learner reaches its (lower) asymptotic error more slowly; under misspecification, the curves cross. Here we devise a test to observe and measure the crossover in our setting, for correctly specified and misspecified cases. Each panel plots test error (lower

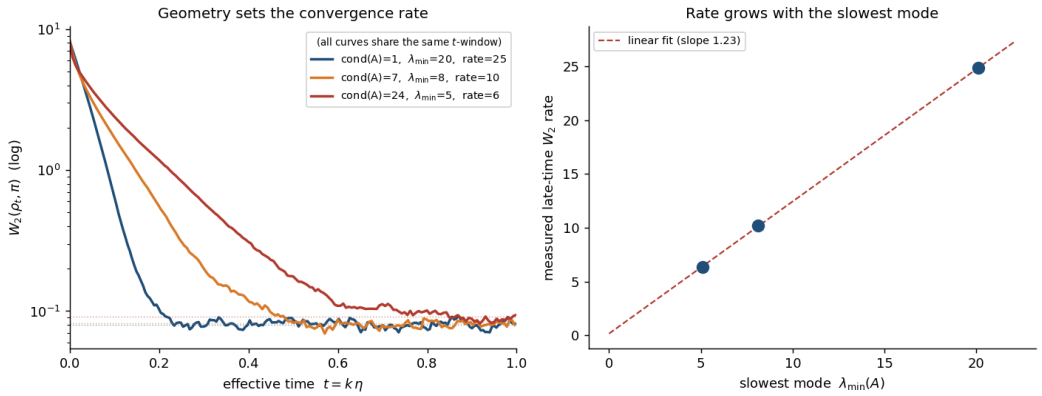

Figure 3: The geometry controls the convergence, and hence calibration, rate. *Left:* transport distance to the posterior for three geometries of increasing anisotropy; the same flow reaches the posterior more slowly as the landscape stiffens. *Right:* the measured late-time decay rate is proportional to the slowest mode $\lambda_{\min}(A)$ (linear fit, slope 1.20), the rate-limiting eigenvalue of the geometry.

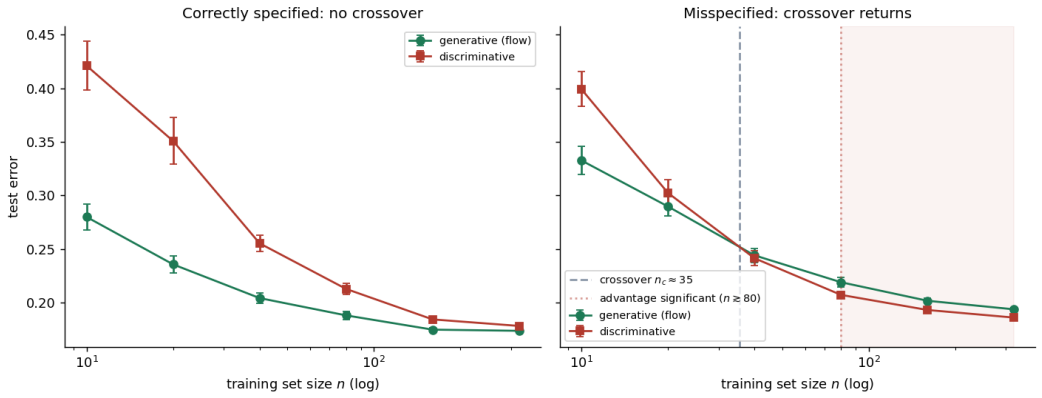

Figure 4: Distributional (green) vs. discriminative (red) classification over 20 replicates. Left: when the model's assumptions hold, and there is no crossover. Right: under misspecification, the discriminative model overtakes at $n \approx 35$, which is the cause of misspecification.

is better), averaged over 20 independent replicates with standard-error bars, against the training-set size $n \in 10, 20, 40, 80, 160, 320$ on a logarithmic axis, comparing the distributional/generative model (green, realized by the flow) with the discriminative logistic model (red) on the same fixed test set, one panel per regime (correctly specified / misspecified), read left to right as training data is added. In the left panel, the data are drawn from the model's own assumptions (Gaussian classes): the generative model is more sample-efficient, below the discriminative one at every n, and the gap only narrows, so there is no crossover. In the right panel, the labels instead follow a logistic rule the generative model does not assume (misspecification): the two mean curves cross at $n_c \approx 35$ (dashed line), and the discriminative model's advantage reaches statistical significance only from $n \geq 80$ (dotted line and shaded band), where the two models' error bars separate at two consecutive sample sizes. We report this because the only thing that differs between the two panels is the correctness of the model's assumptions; the model class, the features, the sampling budget, and the evaluation are all identical, so the experiment isolates misspecification, not just the fact of being distributional/generative, as the true cause of the large-sample crossover. That is a distinction a benchmark run on a single fixed dataset cannot make, and it is the reason this controlled comparison of the calibration and rate results is necessary.

5 DISCUSSION

Our goal is to make the interaction of energy, geometry, and dynamics *measurable* rather than just describable, and the precondition for that is trusting the measurements themselves. We work in a controlled regime where the posterior, and therefore every functional of the flow, has a closed form. This lets us check each measurement against its exact value, and on that verified basis we establish three results connecting the framework's quantities to observable behaviour.

First, calibration is controlled by the transport distance to the posterior. Along the flow, the expected calibration error stays under the envelope $\mathrm{ECE}(\rho) \leq c\, W_2(\rho, \pi)$ and descends with it until both saturate at a finite-ensemble floor scaling as $r_{\mathrm{eff}}/\sqrt{M}$. A comparison, the flow against an exact posterior draw, statistically indistinguishable, locates this behaviour in the posterior the flow *reaches*, not in the mechanics of the proximal scheme. Second, the geometry of the landscape sets the rate at which that calibration is attained: holding the flow and the target family fixed and varying only the anisotropy, the late-time transport distance decays at a rate proportional to the smallest Hessian eigenvalue $\lambda_{\min}(A)$, the finite-dimensional version of the log-Sobolev constant (Bakry & Émery, 2006). Third, in a controlled generative-versus-discriminative comparison, the discriminative model overtakes the generative one only under misspecification (crossover at $n \approx 35$); when the model is correctly specified, no crossover occurs, which isolates misspecification, not the only fact of being distributional, as its cause, in line with Ng and Jordan (Ng & Jordan, 2001).

The scope of these claims is deliberately narrow. Every result lives in an analytically tractable regime with a Gaussian posterior; the closed forms that make validation possible are exactly what a real, over-parameterized network lacks, and we do not claim that the integrity observed here transfers unexamined to that setting. The envelope constant c bundles problem-dependent Lipschitz constants and is read off the data rather than computed, so the bound serves as a validated *relationship*, not a numerical guarantee. The finite-ensemble floor is a resolution limit: with M particles, no effect smaller than $r_{\mathrm{eff}}/\sqrt{M}$ is observable.

Having validated instruments here, the next stage, *confirmation*, carries them to the training trajectories of real, over-parameterized networks, beginning with regimes that admit theoretical guarantees such as the mean-field limit (Mei et al., 2018; 2019), where feature learning, memorization, and the loss of generalization actually occur and no ground truth is available to catch a failing estimator. The final stage, *scaling*, extends the framework to large networks and to practical tasks such as generative modelling, where the substantial computation these measurements demand is justified only once the earlier, cheaper stages have shown the tools read true.

This paper is a first step: on a controlled example with a closed-form answer, we showed that the idea works in principle. As the concrete next step, we will turn the same, now-validated tools on real, over-parameterized networks along their training trajectories, beginning with regimes that carry theoretical guarantees such as the mean-field limit. There we can follow the convergence-rate constant as it evolves during training and ask whether calibration still tracks the distance to the posterior once that posterior is neither Gaussian nor known in advance. Then the framework, without changing the main objects and phenomena, can be extended to large models and to generative tasks such as diffusion. The questions we ultimately want to reach, how the geometry shapes feature learning, when memorization happens and why generalization is lost, are precisely the ones this apparatus is built to measure, and the present controlled study is what makes it valid to pose them at scale.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for polishing and correcting the text. We have not used generative AI tools for method development, experiment design, and implementations, and the data collection or the use of any other assets do not apply to this work. Additionally, we used generative AI tools for the intuitive figure generation based on the human draft of the figure and detailed explanations. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

See the ICLR 2027 AI Policy for Authors for more details. This statement should not be more than 1 page.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

If authors feel that their paper submission raises questions regarding the Code of Ethics, they are encouraged to include a paragraph of Ethics Statement (at the end of the main text before references) to address potential concerns where appropriate. Topics include, but are not limited to, studies that involve human subjects, practices to data set releases, potentially harmful insights, methodologies and applications, potential conflicts of interest and sponsorship, discrimination/bias/fairness concerns, privacy and security issues, legal compliance, and research integrity issues (e.g., IRB, documentation, research ethics). This statement should not be more than 1 page.

REPRODUCIBILITY STATEMENT

The code and experiments can be implemented based on the method section and descriptions of the experiments and what we are aiming to measure.

(This section is **recommended** and does not count toward the page limit.)

It is important that the work published in ICLR is reproducible. Authors are strongly encouraged to include a paragraph-long Reproducibility Statement at the end of the main text (before references) to discuss the efforts that have been made to ensure reproducibility. This paragraph should not itself describe details needed for reproducing the results, but rather reference the parts of the main paper, appendix, and supplemental materials that will help with reproducibility. For example, for novel models or algorithms, a link to an anonymous downloadable source code can be submitted as supplementary materials; for theoretical results, clear explanations of any assumptions and a complete proof of the claims can be included in the appendix; for any datasets used in the experiments, a complete description of the data processing steps can be provided in the supplementary materials. Each of the above are examples of things that can be referenced in the reproducibility statement.

AUTHOR CONTRIBUTIONS

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2005.

Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84: Proceedings*, pp. 177–206. Springer, 2006.

Dominique Bakry, Ivan Gentil, Michel Ledoux, et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the bures–wasserstein distance between positive definite matrices. *Expositiones mathematicae*, 37(2):165–191, 2019.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial intelligence and statistics*, pp. 192–204. PMLR, 2015.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Bradley Efron, Robert J Tibshirani, et al. *An introduction to the bootstrap*. Boca Raton, Florida, 2000.

Jiaojiao Fan, Qinsheng Zhang, Amirhossein Taghvaei, and Yongxin Chen. Variational wasserstein gradient flow. *arXiv preprint arXiv:2112.02424*, 2021.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3):707–738, 2015.

Peyré Gabriel and Cuturi Marco. Computational optimal transport with applications to data sciences. *Foundations and trends® in machine learning*, 11(5-6):355–607, 2019.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

Clark R Givens and Rae Michael Shortt. A class of wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240, 1984.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Leonid Vasilevich Kantorovich and SG Rubinshtein. On a space of totally additive functions. *Vestnik of the St. Petersburg University: Mathematics*, 13(7):52–59, 1958.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, pp. 110–133, 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient mcmc. *Advances in neural information processing systems*, 28, 2015.

David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on learning theory*, pp. 2388–2464. PMLR, 2019.

Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M Solomon, and Evgeny Burnaev. Large-scale wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 34:15243–15256, 2021.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14, 2001.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632, 2005.

Felix Otto and Cédric Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

Grigorios A Pavliotis. Stochastic processes and applications. *Texts in applied mathematics*, 60: 41–43, 2014.

Hannes Risken. Fokker-planck equation. In *The Fokker-Planck equation: methods of solution and applications*, pp. 63–95. Springer, 1989.

Grant Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75 (9):1889–1935, 2022.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688, 2011.

Florian Wenzel, Kevin Roth, Bastiaan S Veeling, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, Sebastian Nowozin, et al. How good is the bayes posterior in deep neural networks really? *arXiv preprint arXiv:2002.02405*, 2020.

Chen Xu, Xiuyuan Cheng, and Yao Xie. Normalizing flow neural networks by jko scheme. *Advances in Neural Information Processing Systems*, 36:47379–47405, 2023.

A TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

A.1 FIG. 2

Two rates not to be conflated Because we summarize each ensemble by its mean and covariance, this floor shrinks quickly, like $r_{\text{eff}}/\sqrt{M}$, where the *effective rank* r_{eff} is the number of directions in which the distribution genuinely spreads (at most the dimension d). This is far faster than the floor one would get by comparing the raw cloud of M points to π without a Gaussian fit, which suffers the curse of dimensionality, $M^{-1/d}$ Fournier & Guillin (2015) (improving to M^{-1/d^*} only when the data concentrate near a low-dimensional set of intrinsic dimension $d^* \leq d$ Weed & Bach (2019)). The distinction matters downstream: the floor sets the ensemble size we need, since any effect we hope to see- a change in the convergence rate, a spike in a residual, a shift in calibration- must exceed the floor at the chosen M to be visible. Where no closed-form target exists, the floor is instead estimated by resampling (e.g. splitting the ensemble in two), and the same caution applies at scale: the Fréchet Inception Distance used to evaluate generators is itself a Gaussian W_2 and carries the same finite-sample floor, which must be accounted for.

The Kantorovich–Rubinstein bridge. The posterior-predictive probability $\bar{p}_\rho(y \mid x) = \int p(y \mid x, \theta)\, \rho(d\theta)$. The map $\theta \mapsto p(y \mid x, \theta)$ is Lipschitz for the Gaussian classifier, the class probabilities are a softmax of the negative squared distances $-\|x - \mu_c\|^2/2\sigma^2$, the softmax is 1-Lipschitz Gao & Pavel (2017), and each squared-distance term has bounded gradient $(x - \mu_c)/\sigma^2$ on the bounded region the particles occupy, so a bounded change in the class means produces a bounded change in the predicted probabilities. Consequently $\left|\bar{p}_\rho(y \mid x) - \bar{p}_\pi(y \mid x)\right| \leq L\, W_1(\rho, \pi)$, and since the calibration error is itself a Lipschitz functional of the predictive while the posterior predictive $\bar{p}_\pi$ is calibrated by construction, the calibration error inherits the bound,

$$\text{ECE}(\rho) \;\leq\; c\, W_1(\rho, \pi) \;\leq\; c\, W_2(\rho, \pi),$$

where we use $W_1 \leq W_2$ Villani et al. (2009) to pass to the quantity the flow actually contracts and the one plotted. The constant c bundles the Lipschitz constants and is problem-dependent, which is why we read its value off the data as an empirical envelope rather than computing it in closed form.

Which bins are shown. We partition confidence into ten equal-width bins on $[0, 1]$, yet only about seven are populated. The reason is structural: with C classes, the predicted (most-probable) class must carry probability at least $1/C$, the largest of C numbers summing to one cannot fall below their average $1/C$, so no prediction can be less than $1/C$ confident. For $C = 3$ the confidence floor is $1/3$, and the three bins below 0.33 are necessarily empty. We additionally omit any bin holding only a handful of points, whose accuracy is pure sampling noise (a single correct point reads as accuracy 1.0) while its ECE weight $|B_m|/n$ is negligible. The displayed range $[1/C,\ 1]$ is thus the entire *achievable*, statistically meaningful confidence range, not a cropped view of it.

A.2 FIG. 2 COMPLEMENTARY FREE ENERGY EVOLUTION.

In this figure, the horizontal axis is time along the flow, and the vertical axis is energy. The black curve is the free energy $F(\rho_t) = \int V d\rho_t + T \int \rho_t \log \rho_t$; it decreses monotonically to the dashed floor $F_{\min} = F(\rho_\infty)$, because the dynamics is the steepest descent of F in the Wasserstein-2 geometry, along which F can only decrease (Jordan et al., 1998; Ambrosio et al., 2005). The two colored curves are its ingredients, and they move in opposite directions: the energy part $\int V d\rho_t$ (green) falls as the mass slides toward the potential minimum and concentrates. In contrast, the entropy part $\int \rho_t \log \rho_t$ (orange) rises as the initial distribution contracts and its differential entropy drops. Importantly, it is their sum, not either term alone, that is monotone, the hallmark of a free energy, describing how lowering potential energy is balanced against losing entropy (Ambrosio et al., 2005). The dashed floor is the vlaue at the stationary Gibbs measure $\rho_\infty \propto e^{-V/T}$ (the Bayesian posterior at $T = 1$), and the residual gap $F(\rho_t) - F_{\min}$ equals $T\text{KL}(\rho_t || \rho_\infty)$, so the curve reaching its floor is precisely the distribution reaching the posterior (Jordan et al., 1998). In one line: the flow provably dissipates free energy to its minimum, trading potential energy against entropy, and converges to the posterior.

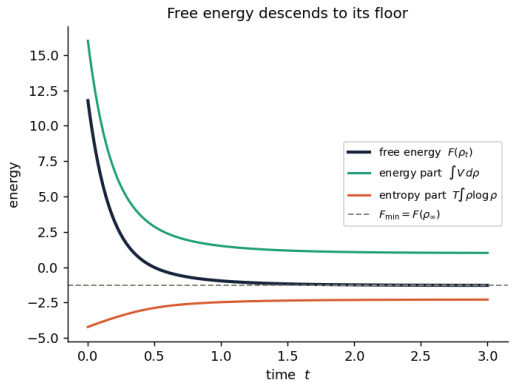

Figure 5: The figure illustrates the descent of the free energy functional $F(\rho_t)$ and the variation of its components, the descent of the potential V, and the ascent of the entropy $\mathbb{E}_\rho$ over time. The free energy F converges to its floor $F_{\min}$ over the Wassertein gradient descent process.

A.3 Controlled comparison of predictors (Fig. 6).

The panel isolates what each ingredient contributes to calibration by changing one factor at a time, so the difference between two bars attributes the change to that factor. *Flow (ours)* is the posterior-averaged predictive, $\bar{p}(c \mid x) = \frac{1}{M}\sum_j p(c \mid x, \theta_j)$: every particle predicts and we average the probability vectors, so parameter uncertainty becomes predictive uncertainty (Bayesian model averaging MacKay (1992), in its ensemble form Lakshminarayanan et al. (2017)). *Exact draw* replaces the flow with i.i.d. samples from the closed-form posterior, isolating the flow's own contribution beyond reaching π. *MAP point* collapses the posterior to its mode and predicts once, isolating the value of averaging over a distribution rather than a single point. *Logistic MLE* is the discriminative point-estimate baseline, the classical counterpart in the generative-versus-discriminative comparison Ng & Jordan (2001). *Logistic+temperature* adds one post-hoc temperature parameter fit on a held-out split, the standard calibration baseline Guo et al. (2017); including it prevents crediting the flow with an advantage a single scalar can buy.

Each bar is the mean ECE over 20 random seeds with a 95% bootstrap confidence interval Efron et al. (2000); the dashed line marks the flow's value and the vertical divider separates posterior-based predictors from discriminative ones. The comparison is a set of pre-registered contrasts. The flow and the exact posterior draw are statistically indistinguishable (flow and exact draw 0.025, paired difference $+0.0005$, CI $[-0.001, +0.002]$), so the flow adds nothing beyond reaching the posterior; the calibration is a property of π, not of the dynamics, consistent with the envelope reading of the left panel. The *MAP point* estimate, which keeps only the most probable parameter (posterior mode), is slightly worse (0.027), showing that averaging helps a little. The gap between the point and the averaged predictors thus isolates the effect of propagating parameter uncertainty into the prediction: a point estimate cannot be certain which θ is right and is mildly overconfident, so its ECE is slightly higher (0.027 vs. 0.025 for the draw). The untuned discriminative baseline (logistic MLE) is clearly worse (0.063), and temperature scaling recovers most of the gap (0.030). Details of the models and computation can be found in A.1. Reporting the tuned baseline rather than excluding it keeps the comparison between honest alternatives; the residual, statistically resolved advantage of the distributional predictor is the classical small-sample effect of Ng & Jordan (2001), not a claim specific to the proximal scheme.

The posterior-based predictors. The three left-hand bars use the *same* posterior over the parameters, $\pi = \mathcal{N}(\mu^\star, TA^{-1})$, and differ only in how much of it they use. *Exact draw* samples M parameter vectors from π and averages their predictions, giving the full posterior predictive $\bar{p}(y \mid x) = \int p(y \mid x, \theta)\, \pi(d\theta)$ (Bayesian model averaging MacKay (1992)). *Flow* reaches the same π by running the JKO flow rather than sampling it directly and is therefore statistically indistinguishable from the exact draw. *MAP point* keeps only the single most-probable parameter, the posterior mode, which for a Gaussian equals the mean $\mu^\star$, and predicts once, discarding the posterior's spread.

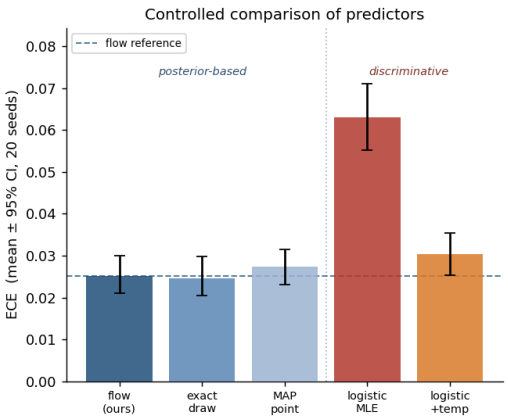

Figure 6: *Right:* controlled comparison over 20 seeds (mean $\pm$ 95% CI); the dashed line marks the flow's calibration and the vertical divider separates posterior-based predictors from discriminative ones. The flow is statistically indistinguishable from drawing the posterior directly (difference $+0.0005$, CI $[-0.001, +0.002]$), confirming the effect is a property of the posterior the flow reaches; the point estimate and the untuned discriminative baseline are worse, and post-hoc temperature scaling recovers most of the gap.

The temperature-scaled baseline. Temperature scaling is a one-parameter post-hoc recalibration applied after the discriminative model is trained: its logits z are divided by a scalar $T > 0$ before the softmax, $p_T(c \mid x) = \mathrm{softmax}(z/T)$, with T fit on a held-out split by minimizing the negative log-likelihood. Because dividing all logits by the same positive constant preserves their ordering, the predicted class, and hence the accuracy, is unchanged; only the confidence is rescaled ($T > 1$ softens an overconfident model, $T < 1$ sharpens it). Here the untuned logistic classifier is overconfident (ECE 0.086), and the fitted temperature $T = 1.84$ brings it to 0.026: a single scalar recovers most of the calibration gap, which is why we report it rather than exclude it, so that the distributional predictor is compared against an honestly-tuned discriminative one.