

Nonreversible gauge fields in Fokker-Planck dynamics: Supersymmetric Hamiltonians and learned finite forces

Masayuki Ohzeki ^{*}

Graduate School of Information Sciences, Tohoku University, Sendai 980-8579, Japan;

Department of Physics, School of Science, Institute of Science Tokyo, Tokyo 152-8551, Japan;

*Research and Education Institute for Semiconductors and Informatics, Kumamoto University, Kumamoto 860-8555, Japan;
and Sigma-i Co., Ltd., Tokyo 108-0075, Japan*



(Received 4 June 2026; accepted 14 July 2026; published 10 August 2026)

We formulate stationary-density-preserving nonreversible perturbations of Fokker-Planck dynamics as gauge fields that deform relaxation spectra while leaving the invariant state fixed. When detailed balance holds, a similarity transformation maps the reversible Fokker-Planck operator to the scalar sector of a Witten-Laplacian-type supersymmetric Hamiltonian; nonreversible gauges then appear as non-Hermitian perturbations that preserve the zero mode but modify the excited spectrum. This operator viewpoint gives a common language for relaxation gaps, circulating probability currents, hypocoercive acceleration, and finite control costs. We represent admissible gauge currents by antisymmetric tensor fields and identify the detailed-balance-violating Ohzeki-Ichiki force as a constant symplectic example whose infinite-strength limit is Hamiltonian dynamics. The continuous-time spectral gap alone does not select a finite gauge strength, so we introduce a finite-time regularized objective and an actor-critic procedure for learning the gauge. An exactly solvable anisotropic Gaussian Ornstein-Uhlenbeck benchmark separates the spectral transition from the finite-time optimum and shows that the learned gauge recovers the Lyapunov-equation optimum. A double-well benchmark then illustrates the same constrained selection in a nonconvex metastable landscape. Stochastic gradient methods enter this framework as physically relevant Fokker-Planck systems: Minibatch noise acts as an effective diffusion tensor, and adaptive methods such as Adam correspond to metric choices with possible nonequilibrium currents.

DOI: [10.1103/tg4c-7cnr](https://doi.org/10.1103/tg4c-7cnr)

I. INTRODUCTION

Relaxation in stochastic dynamics is controlled not only by the stationary distribution but also by the probability currents that carry a density toward that distribution. Detailed balance is a sufficient condition for stationarity, but it is not necessary. Nonreversible currents that preserve the invariant density while violating detailed balance are known to accelerate relaxation in Langevin and Markov-chain dynamics, and have been studied in nonequilibrium statistical physics, stochastic sampling, and variance reduction [1–5]. The generalized Ohzeki-Ichiki dynamics provides a particularly transparent construction: A symplectic nonconservative force preserves an arbitrary Gibbs distribution, interpolates between reversible Langevin dynamics and Hamiltonian dynamics, and reaches the Hamiltonian Monte Carlo (HMC) limit at infinite gauge strength [6]. This interpolation, however, does not by itself determine the finite strength of the nonreversible force. A physical principle is needed once

relaxation speed, force amplitude, control cost, possible entropy production, discretization stability, and possible Metropolis rejection are all considered.

The antisymmetric structures considered here are also related to odd transport phenomena, such as odd viscosity and odd diffusivity, where time-reversal symmetry breaking produces transverse nondissipative response coefficients [7–13]. In active chiral fluids, such odd coefficients can arise from microscopic nonequilibrium driving or self-rotation. In passive systems with magnetic-field-like odd variables, one may formulate generalized reciprocity or detailed-balance relations by reversing the odd parameter together with time. In contrast, the gauge field optimized in this paper is an externally chosen deterministic drift for a fixed target density. Thus, “detailed-balance breaking” below means violation of the ordinary fixed-parameter detailed-balance condition, while stationarity of the target density is preserved. This distinction is important when comparing the present weak fluctuation-dissipation construction with microscopic odd-transport models.

This paper develops an operator-theoretic framework for this finite-gauge problem. We write Fokker-Planck dynamics as a Hamiltonian time evolution and, when detailed balance holds, map the reversible operator by a similarity transformation to the scalar sector of a Witten-Laplacian-type supersymmetric Hamiltonian [14,15]. In this representation, relaxation is governed by low-lying energy gaps. A stationary-density-preserving nonreversible perturbation acts as a gauge

^{*}Contact author: mohzeki@tohoku.ac.jp

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

field: It leaves the zero mode, and hence the invariant density, unchanged while deforming the excited spectrum. The central physical message is that nonreversible gauges provide a controlled way to design the relaxation spectrum of a Fokker-Planck operator without changing its stationary state.

Stochastic gradient methods provide a useful and timely class of Fokker-Planck systems to which this structure applies. Stochastic gradient Langevin dynamics (SGLD) connects stochastic optimization and scalable Bayesian sampling [16], and its finite-step behavior has been analyzed through Fokker-Planck equations and Itô processes [17]. Constant-step Stochastic Gradient Descent (SGD) can be interpreted as approximate Bayesian inference [18], while the ratio between learning rate and batch size controls an effective noise scale [19,20]. In the present framework, minibatch noise becomes an effective diffusion tensor or temperature, adaptive methods such as Adam become metric choices, and nonequilibrium deviations from reversible Langevin dynamics are organized as gauge currents.

The contributions of this paper are as follows. First, we formulate a supersymmetric-Hamiltonian representation of reversible stochastic-gradient Fokker-Planck dynamics. This gives a spectral language in which minibatch noise, injected Langevin temperature, learning-rate decay, and batch-size growth are interpreted as changes of the low-lying Fokker-Planck spectrum and the stationary density.

Second, we represent stationary-density-preserving nonreversible currents by antisymmetric tensor gauge fields. This identifies the Ohzeki-Ichiki force as a concrete constant-gauge construction and clarifies its relation to the HMC limit. The gauge changes the excited spectrum while keeping the ground state fixed, so the problem becomes one of constrained spectral design.

Third, we show that the continuous-time spectral gap alone does not select a finite gauge strength. We therefore introduce a finite-time regularized objective that balances relaxation against gauge-force cost, and we formulate an actor-critic procedure for learning the gauge. An exactly solvable anisotropic Gaussian Ornstein-Uhlenbeck benchmark shows that the learned gauge matches the finite-time optimum computed from Lyapunov equations. A nonconvex double-well benchmark then illustrates the same constrained selection in a metastable landscape.

II. FROM STOCHASTIC GRADIENTS TO FOKKER-PLANCK HAMILTONIANS

A. Stochastic gradient methods

Let $\mathcal{D} = \{z_n\}_{n=1}^N$ be a training dataset, where $z_n \in \mathcal{Z}$ denotes the n th data point and N is the number of samples. Let $\theta \in \mathbb{R}^d$ be the parameter vector to be optimized, and $\ell : \mathbb{R}^d \times \mathcal{Z} \rightarrow \mathbb{R}$ be the loss associated with a single data point. The empirical loss is then defined by

$$L(\theta) = \frac{1}{N} \sum_{n=1}^N \ell(\theta; z_n). \quad (1)$$

Throughout the paper, ∇ denotes the gradient with respect to θ .

For a minibatch $B \subset \{1, \dots, N\}$ of size $|B| = b$, the minibatch gradient is

$$\hat{g}_B(\theta) = \frac{1}{b} \sum_{n \in B} \nabla \ell(\theta; z_n). \quad (2)$$

We decompose it as

$$\hat{g}_B(\theta) = \nabla L(\theta) + \xi_B(\theta), \quad \mathbb{E}_B[\xi_B(\theta)] = 0, \quad (3)$$

where $\xi_B(\theta)$ is the minibatch gradient noise and $\mathbb{E}_B$ denotes expectation over the random choice of the minibatch. Its covariance is denoted by

$$\Sigma_B(\theta) = \mathbb{E}_B[\xi_B(\theta)\xi_B(\theta)^\top]. \quad (4)$$

With learning rate η and preconditioner $G(\theta)$, a local approximation of SGD or Adam-type updates is

$$\theta_{k+1} = \theta_k - \eta G(\theta_k) \hat{g}_B(\theta_k). \quad (5)$$

We here use the diffusion approximation. The diffusion approximation is obtained by matching the first two conditional moments of one discrete SGD step with those of an Itô stochastic differential equation. Starting from

$$\theta_{k+1} = \theta_k - \eta G(\theta_k) \{\nabla L(\theta_k) + \xi_B(\theta_k)\}, \quad (6)$$

the one-step increment is

$$\Delta\theta_k := \theta_{k+1} - \theta_k = -\eta G(\theta_k) \nabla L(\theta_k) - \eta G(\theta_k) \xi_B(\theta_k). \quad (7)$$

Conditioned on the current parameter value $\theta_k = \theta$, its mean and covariance are

$$\mathbb{E}_B[\Delta\theta_k | \theta_k = \theta] = -\eta G(\theta) \nabla L(\theta), \quad (8)$$

$$\text{Cov}_B[\Delta\theta_k | \theta_k = \theta] = \eta^2 G(\theta) \Sigma_B(\theta) G(\theta)^\top. \quad (9)$$

We now introduce the continuous time variable

$$t = k\eta, \quad \Delta t = \eta, \quad (10)$$

and seek an Itô diffusion whose increment over a short time interval Δt has the same leading-order mean and covariance. For an SDE

$$d\theta_t = b(\theta_t) dt + \sigma(\theta_t) dW_t, \quad (11)$$

where W_t is a standard d -dimensional Brownian motion, the increment over Δt satisfies

$$\mathbb{E}[\Delta\theta_t | \theta_t = \theta] = b(\theta) \Delta t + o(\Delta t), \quad (12)$$

$$\text{Cov}[\Delta\theta_t | \theta_t = \theta] = \sigma(\theta) \sigma(\theta)^\top \Delta t + o(\Delta t). \quad (13)$$

Matching these expressions with the discrete SGD moments and using $\Delta t = \eta$, we obtain

$$b(\theta) = -G(\theta) \nabla L(\theta), \quad \sigma(\theta) \sigma(\theta)^\top = \eta G(\theta) \Sigma_B(\theta) G(\theta)^\top. \quad (14)$$

Thus, the diffusion approximation of minibatch SGD is

$$d\theta_t = -G(\theta_t) \nabla L(\theta_t) dt + \sqrt{\eta G(\theta_t) \Sigma_B(\theta_t) G(\theta_t)^\top} dW_t^{\text{mb}}. \quad (15)$$

Here, W_t^{mb} is a standard Brownian motion representing the accumulated minibatch fluctuation, and $\sqrt{A}$ denotes any matrix square root satisfying $\sqrt{A} \sqrt{A}^\top = A$. This approximation

is understood as a small-learning-rate, many-step limit: The deterministic part of the SGD update becomes the drift term, while the accumulated minibatch fluctuations are treated as Brownian noise.

In stochastic gradient Langevin dynamics, an additional artificial thermal noise is added. Using the same empirical loss $L(\theta)$ as the potential, and denoting the target temperature by $T > 0$, we write

$$d\theta_t = -G(\theta_t)\nabla L(\theta_t)dt + \sqrt{\eta G(\theta_t)\Sigma_B(\theta_t)G(\theta_t)^\top}dW_t^{\text{mb}} + \sqrt{2T G(\theta_t)}dW_t^{\text{th}}, \quad (16)$$

where W_t^{mb} and W_t^{th} are independent Brownian motions. The first noise term comes from minibatch sampling, while the second one is the deliberately injected Langevin noise. Equivalently, the two independent noises can be combined into a single diffusion tensor

$$D_{\text{eff}}(\theta) = T G(\theta) + \frac{\eta}{2} G(\theta)\Sigma_B(\theta)G(\theta)^\top. \quad (17)$$

Here and below, ∂_t denotes differentiation with respect to time t , and

$$\partial_i := \frac{\partial}{\partial \theta_i}$$

denotes differentiation with respect to the i th component of θ . We use the Einstein summation convention for repeated coordinate indices.

Thus, the density $\rho(\theta, t)$ of θ_t obeys the Fokker-Planck equation

$$\begin{aligned} \partial_t \rho(\theta, t) = & \partial_i [G(\theta)\nabla L(\theta)]_i \rho(\theta, t) \\ & + \partial_i \partial_j [(D_{\text{eff}}(\theta))_{ij} \rho(\theta, t)]. \end{aligned} \quad (18)$$

Equivalently, this can be written as a continuity equation

$$\partial_t \rho = -\nabla \cdot J, \quad J_i = -\{G\nabla L\}_i \rho - \partial_j [(D_{\text{eff}})_{ij} \rho]. \quad (19)$$

The vector field $J(\theta, t)$ is the probability current associated with the density $\rho(\theta, t)$. When $T = 0$, this equation reduces to the diffusion approximation of minibatch SGD. When the minibatch covariance term is neglected, it reduces to the usual Langevin Fokker-Planck equation with temperature T .

B. Supersymmetric Hamiltonian

We now rewrite the Fokker-Planck equation obtained above as a quantum-mechanical time evolution. Equation (18) can be written as

$$\partial_t \rho = -\mathcal{H}_{\text{FP}} \rho, \quad (20)$$

where the Fokker-Planck Hamiltonian is the differential operator

$$\mathcal{H}_{\text{FP}} = -\partial_i [G\nabla L]_i \cdot - \partial_i \partial_j [(D_{\text{eff}})_{ij} \cdot]. \quad (21)$$

The dot indicates the argument on which the operator acts. At this stage, no similarity transformation has been performed. Therefore, $\mathcal{H}_{\text{FP}}$ is generally non-Hermitian as an operator on $L^2(\mathbb{R}^d, d\theta)$, the space of square-integrable functions with respect to the Lebesgue measure $d\theta$.

If an invariant density $\pi(\theta)$ exists, it satisfies

$$\mathcal{H}_{\text{FP}} \pi = 0. \quad (22)$$

Thus, π is the right zero eigenfunction of $\mathcal{H}_{\text{FP}}$, while probability conservation implies that the constant function 1 is the left zero eigenfunction. The relaxation rate is controlled by the nonzero eigenvalues of $\mathcal{H}_{\text{FP}}$. We define the spectral gap by

$$\Delta = \min_{\lambda \in \text{Spec}(\mathcal{H}_{\text{FP}}) \setminus \{0\}} \text{Re } \lambda, \quad (23)$$

assuming that the zero eigenvalue is isolated and that the real parts of all nonzero eigenvalues are positive. For nonreversible dynamics, this gap should be interpreted with care, because the generator is generally non-normal. Finite-time relaxation can therefore depend on eigenvectors and transient amplification, not only on the eigenvalues. The detailed balance with respect to π then means that the stationary probability current vanishes:

$$J_i^\pi = -\{G\nabla L\}_i \pi - \partial_j [(D_{\text{eff}})_{ij} \pi] = 0. \quad (24)$$

In contrast, preserving the stationary density only requires the weaker stationarity condition

$$\partial_i J_i^\pi = 0. \quad (25)$$

Thus, a nonzero divergence-free stationary current can break detailed balance without changing the invariant density. This stationary-density-preserving current is the gauge degree of freedom that will later be represented by antisymmetric tensors. The Ohzeki-Ichiki force is a concrete construction of such a detailed-balance-breaking current for Langevin dynamics [2,3].

We now specialize the SGLD Fokker-Planck Hamiltonian to the reversible scalar-temperature approximation. In general, $D_{\text{eff}}(\theta)$ is state dependent and anisotropic. To obtain the standard supersymmetric representation without additional geometric correction terms, we assume that G is constant, symmetric, and positive definite, and approximate the minibatch covariance by

$$G\Sigma_B(\theta)G^\top \simeq \frac{\sigma_G^2}{b} G, \quad b = |B|. \quad (26)$$

Here, b is the minibatch size and σ_G^2 is the gradient-noise strength measured in the metric defined by G . Then,

$$D_{\text{eff}} \simeq T_{\text{eff}} G, \quad T_{\text{eff}} = T + \frac{\eta \sigma_G^2}{2b}. \quad (27)$$

For plain minibatch SGD without injected Langevin noise, $T = 0$; for the Euclidean case $G = I$, this reduces to $T_{\text{eff}} \simeq \eta \sigma^2 / (2b)$.

In this limit, the stationary density satisfying detailed balance is

$$\pi(\theta) = \frac{1}{Z} \exp \left[-\frac{L(\theta)}{T_{\text{eff}}} \right], \quad Z = \int_{\mathbb{R}^d} \exp \left[-\frac{L(\theta)}{T_{\text{eff}}} \right] d\theta. \quad (28)$$

The similarity transformation

$$\rho(\theta, t) = \pi(\theta)^{1/2} \psi(\theta, t) \quad (29)$$

leads to

$$\partial_t \psi = -H_{\text{SUSY}} \psi, \quad H_{\text{SUSY}} := \pi^{-1/2} \mathcal{H}_{\text{FP}}^{\text{rev}} \pi^{1/2}. \quad (30)$$

Thus, in the reversible case, the similarity-transformed Fokker-Planck Hamiltonian is the scalar, or bosonic 0-form,

sector of the Witten-Laplacian-type supersymmetric Hamiltonian. The full supersymmetric complex would include differential forms of all degrees; in this paper, we work only with the density sector.

For constant G , its scalar, or 0-form, part is

$$H_{\text{SUSY}} = -T_{\text{eff}} \partial_i (G_{ij} \partial_j) + \frac{1}{4T_{\text{eff}}} \partial_i L G_{ij} \partial_j L - \frac{1}{2} \partial_i [G_{ij} \partial_j L]. \quad (31)$$

Equivalently,

$$H_{\text{SUSY}} = A_i^\dagger G_{ij} A_j, \quad A_i = \sqrt{T_{\text{eff}}} \left(\partial_i + \frac{\partial_i L}{2T_{\text{eff}}} \right), \quad (32)$$

where $A_i^\dagger$ denotes the adjoint of A_i in $L^2(\mathbb{R}^d, d\theta)$. The ground state is

$$\psi_0(\theta) = \pi(\theta)^{1/2} \propto \exp \left[-\frac{L(\theta)}{2T_{\text{eff}}} \right], \quad H_{\text{SUSY}} \psi_0 = 0. \quad (33)$$

This is the 0-form Witten-Laplacian-type Hamiltonian associated with the loss landscape $L(\theta)$ in the reversible SGLD approximation [14,15].

The advantage of this transformation is that relaxation becomes an energy-gap problem. Since H_{SUSY} is Hermitian and non-negative in the reversible case, the first nonzero eigenvalue

$$\Delta = \min_{\lambda \in \text{Spec}(H_{\text{SUSY}}) \setminus \{0\}} \lambda \quad (34)$$

controls the leading relaxation mode:

$$\rho(\theta, t) - \pi(\theta) \simeq a_1 e^{-\Delta t} \pi(\theta)^{1/2} \psi_1(\theta), \quad (35)$$

where ψ_1 is the first excited eigenfunction and a_1 is its projection coefficient determined by the initial density.

In metastable loss landscapes, such as double-well-like landscapes, the low-temperature gap is exponentially small because transitions between basins require barrier crossing. If ΔL_b denotes the loss barrier height between two basins, then the Eyring-Kramers form gives

$$\Delta(T_{\text{eff}}) \simeq A_{\text{EK}} \exp \left[-\frac{\Delta L_b}{T_{\text{eff}}} \right], \quad (36)$$

up to a prefactor A_{EK} determined by the local curvatures near the minimum and the saddle point [21]. Combining Eq. (27) with Eq. (36), minibatch noise increases the effective temperature and can enlarge the first nonzero Fokker-Planck eigenvalue in metastable regimes. This is the precise sense in which minibatch noise opens a relaxation gap. The more accurate statement beyond the scalar approximation is that SGD carries the anisotropic diffusion tensor $D_{\text{eff}}(\theta)$, which can promote barrier crossing differently along different directions. This connects with fluctuation-dissipation and effective-noise analyses of SGD, as well as studies of anisotropic gradient noise and flat-solution selection [22–26].

It is important, however, not to confuse increasing the gap with optimizing the learning objective. Increasing T_{eff} improves mixing across barriers, but the stationary density $\pi(\theta) \propto \exp[-L(\theta)/T_{\text{eff}}]$ becomes less concentrated near low-loss regions. If $L_\star$ is the loss at a local minimum and d_{eff} is the effective quadratic dimension of that basin, equipartition

gives the local estimate

$$\mathbb{E}_\pi[L(\theta) - L_\star] \simeq \frac{d_{\text{eff}}}{2} T_{\text{eff}}. \quad (37)$$

Thus, a useful temperature is finite. To escape an undesirable basin with barrier height ΔL_{esc} within an exploration time τ_{exp} , a Kramers estimate gives

$$T_{\text{eff}} \gtrsim \frac{\Delta L_{\text{esc}}}{\ln(A_{\text{esc}} \tau_{\text{exp}})}. \quad (38)$$

Here, A_{esc} is the Kramers prefactor for escape. On the other hand, to keep the expected excess loss below a tolerance ε , Eq. (37) gives

$$T_{\text{eff}} \lesssim \frac{2\varepsilon}{d_{\text{eff}}}. \quad (39)$$

Similarly, to remain in a favorable basin with barrier height ΔL_{hold} for a timescale τ_{hold} , one needs approximately

$$T_{\text{eff}} \lesssim \frac{\Delta L_{\text{hold}}}{\ln(A_{\text{hold}} \tau_{\text{hold}})}, \quad (40)$$

where A_{hold} is the corresponding Kramers prefactor. Therefore, a useful effective temperature is neither obtained by maximizing the gap alone nor obtained by increasing T_{eff} without bound. The temperature should be large enough to assist exploration and barrier crossing, but small enough to concentrate the stationary density near low-loss regions. This interpretation has a concrete algorithmic consequence. Since

$$T_{\text{eff}} = T + \frac{\eta \sigma_G^2}{2b},$$

learning-rate decay and batch-size growth both reduce the same effective diffusion scale. Thus, they are not independent heuristics in the SGLD/Fokker-Planck view; they are two ways of cooling the same diffusion process. This is the Fokker-Planck analog of simulated-annealing schedules, although real SGD noise is generally anisotropic and state dependent [27,28]. This observation is consistent with the noise-scale analysis of SGD, where the relevant scale is controlled by the ratio of learning rate to batch size [19,20]. It is also consistent with the empirical finding that learning-rate decay can often be replaced by increasing the batch size during training, while preserving similar training and test curves [29]. From the present viewpoint, such a schedule gradually lowers T_{eff} : Early training keeps enough noise to explore and escape sharp or metastable regions, whereas late training reduces the noise so that the dynamics can concentrate near low-loss regions. This should not be read as a claim that smaller batches or larger noise are always better; rather, the useful quantity is the scheduled noise scale. Indeed, carefully controlled experiments show that stochastic-gradient noise can improve generalization in some regimes, but that its benefit depends on the training budget and hyperparameter tuning [30].

The main point of the Supersymmetry (SUSY)-Hamiltonian representation is not merely to rederive known noise-scale heuristics. Rather, it provides a unified spectral language in which several empirical training practices can be compared on the same footing. Minibatch noise, injected Langevin temperature, learning-rate decay, and batch-size growth all change the effective diffusion tensor and hence

the low-lying spectrum of the Fokker-Planck operator. In the reversible scalar-temperature limit, this spectrum is equivalently the spectrum of the supersymmetric Hamiltonian H_{SUSY} . Thus exploration, metastable escape, concentration near low-loss regions, and annealing schedules can be discussed in terms of how they modify the relaxation gap and the stationary density. This unified viewpoint is the first benefit of the SUSY formulation: It turns a collection of optimizer heuristics into spectral statements about a single operator.

III. ADAM-TYPE ADAPTIVE LEARNING RATES AS GAUGES

A. Adaptive learning rates

Adam-type methods [31] introduce an extended state with first and second gradient moments,

$$m_{t+1} = \beta_1 m_t + (1 - \beta_1) \hat{g}_t, \quad v_{t+1} = \beta_2 v_t + (1 - \beta_2) \hat{g}_t^2, \quad (41)$$

and update

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}. \quad (42)$$

Under a continuous-time and adiabatic approximation, this is a Langevin-type equation with diagonal preconditioner

$$G_{ii}(\theta) \simeq (\sqrt{v_i(\theta)} + \epsilon)^{-1}. \quad (43)$$

Parameter space, therefore, no longer looks Euclidean; it carries a position-dependent metric G .

If one wants an adaptive-preconditioned process to be an exact reversible sampler for a fixed Gibbs density, the drift and diffusion must be matched. For a target density $\pi(\theta) \propto \exp[-L(\theta)/T]$, this requires the thermal noise associated with the metric G and the Itô correction $T \nabla \cdot G$. With this correction, the quantum-mechanical operator replaces the ordinary Laplacian by a metric-dependent Laplace operator, and geometric connection terms enter the transformed Hamiltonian. Standard Adam is not this reversible sampler. An Adam-like reversible sampler is possible only in the more specific sense that a diagonal, data-adaptive preconditioner is coupled to the matching noise and correction terms, as in preconditioned or Riemannian Langevin dynamics [32–34]. If the metric is estimated online from stochastic gradients, exact reversibility is recovered only in an enlarged Markov state with a consistent fluctuation-dissipation structure, or after a Metropolis correction. Otherwise, the adaptive metric and auxiliary moment variables generate nonequilibrium currents. Decomposing those currents into gauge fields allows one to analyze Adam's speed as a mixture of curvature preconditioning and circulating nonreversible flow.

Thus, Adam should not be viewed only as a coordinate-wise learning-rate rule. In the continuous-time Fokker-Planck description, it is more naturally interpreted as a choice of a metric, or preconditioner, together with possible nonequilibrium currents. In Euclidean coordinates, we write this general metric-gauge Langevin form as

$$d\theta_t = [-G(\theta_t) \nabla L(\theta_t) + T \nabla \cdot G(\theta_t) + c(\theta_t)] dt + \sqrt{2T} E(\theta_t) dW_t, \quad G(\theta) = E(\theta) E(\theta)^T. \quad (44)$$

Here, $G(\theta)$ is a symmetric positive semidefinite preconditioning matrix, $E(\theta)$ is any matrix square root of $G(\theta)$, and

$$\{\nabla \cdot G(\theta)\}_i := \partial_j G_{ij}(\theta).$$

The vector field $c(\theta)$ represents a nonreversible drift. If it satisfies

$$\partial_i \{\pi(\theta) c_i(\theta)\} = 0, \quad (45)$$

then it preserves the stationary density $\pi(\theta)$, although it generally breaks detailed balance. Equation (44) should be read as a weak fluctuation-dissipation construction. The diffusion matrix is tied to the symmetric metric G through the noise amplitude, while the antisymmetric part enters as the deterministic drift c . This is different from a microscopic odd-transport model in which the noise, mobility, and odd variables may obey a stronger generalized reciprocity relation. For the optimization problem considered here, c is an externally chosen control field whose role is to change probability currents while leaving the target density fixed.

The idea of adding stationary-density-preserving currents is standard in nonequilibrium statistical physics: Detailed balance is sufficient for stationarity, but it is not necessary. A nonzero stationary current can break detailed balance while preserving the same invariant density. In the stochastic-gradient Markov Chain Monte Carlo (MCMC) literature, Ma, Chen, and Fox recast this idea as a general recipe for constructing samplers from a positive semidefinite diffusion matrix and an antisymmetric matrix field [34]. Nonreversible Langevin samplers have also been studied as a systematic way to improve convergence and reduce variance [4,5]. In the stochastic-gradient setting, detailed-balance violation has also been introduced into stochastic gradient Langevin dynamics to accelerate stochastic dynamics [35]. Related overdamped dynamics with Lorentz-force-like terms can exhibit subtle flux and entropy-production properties after eliminating fast variables, emphasizing again that the present gauge should be regarded as a controlled effective drift rather than as a unique microscopic realization [36–38].

In the notation used here, a stationary-density-preserving nonreversible drift can be generated from an antisymmetric tensor field $A(\theta)$ as

$$c_i(\theta) = \frac{1}{\pi(\theta)} \partial_j [A_{ij}(\theta) \pi(\theta)], \quad A_{ij}(\theta) = -A_{ji}(\theta). \quad (46)$$

This form automatically satisfies Eq. (45), since

$$\partial_i \{\pi(\theta) c_i(\theta)\} = \partial_i \partial_j [A_{ij}(\theta) \pi(\theta)] = 0. \quad (47)$$

Thus, $A(\theta)$ is a gauge potential for nonreversible probability flow: It changes the stationary current without changing the stationary density. The Ohzeki-Ichiki force is a concrete physical construction of such a detailed-balance-breaking current for Langevin dynamics.

B. Role in stochastic optimization

The metric-gauge form above also clarifies why modern stochastic optimizers exhibit different relaxation and generalization behavior. Adam, RMSProp, AdaGrad, AdamW, Adafactor, Sophia, and Lion all use adaptive coordinate-wise or curvature-informed scaling to reduce local anisotropy in

the effective landscape [31,39–42]. In the present notation, these methods primarily modify the metric G and, when their auxiliary variables are treated dynamically, can also generate nonequilibrium currents. This explains why adaptive methods often accelerate early training in strongly anisotropic problems, while also making clear that a standard Adam update is not automatically a reversible Langevin sampler.

The same formulation also separates metric adaptation from the temperaturelike effect of SGD noise. Large-batch generalization gaps, sharp minima, noise-scale arguments, and empirical gradient-noise-scale laws all concern the effective diffusion tensor in Eq. (15) [19,20,30,43,44]. Switching or bounding adaptive methods, as in SWATS and AdaBound, can be viewed as changing the balance between early metric acceleration and later SGD-like diffusion [45,46]. Decoupled weight decay in AdamW prevents the adaptive metric from unintentionally deforming the radial confining part of the dynamics [39]. Failures or different solution selection by adaptive methods, including the convergence issues motivating AMSGrad, reflects the fact that a rapidly changing metric need not satisfy a fluctuation-dissipation relation for a fixed target density [47,48].

High-learning-rate phenomena such as the edge of stability are related but not identical to the continuous-time Fokker-Planck limit. The observed stability boundary in full-batch gradient descent and the stochastic sharpness gap in minibatch SGD require discrete-time effects beyond the small-step diffusion approximation used here [49,50]. Similarly, sharpness-aware minimization explicitly modifies the effective landscape by penalizing local sharpness, whereas minibatch SGD does so implicitly through diffusion [51]. These developments support the operator viewpoint without being reduced to it: Metric, temperature, and discrete-time stability are distinct mechanisms.

Our focus in the remainder of the paper is the third mechanism, the stationary-density-preserving gauge freedom. After a metric and diffusion tensor have been specified, an antisymmetric gauge current can accelerate relaxation without changing the target density. This is the part of the dynamics that is optimized below.

IV. GAUGE FIELDS AND STATIONARY CURRENTS

Antisymmetric tensor field

A stationary-density-preserving nonreversible drift can be generated by an antisymmetric tensor field. To avoid confusion with the probability current J , we denote this tensor by $A(\theta)$. Let

$$A_{ij}(\theta) = -A_{ji}(\theta).$$

Then,

$$c_i(\theta) = \frac{1}{\pi(\theta)} \partial_j [A_{ij}(\theta) \pi(\theta)]. \quad (48)$$

Equivalently, the stationary current generated by the gauge alone,

$$K_i^\pi(\theta) = \pi(\theta) c_i(\theta),$$

is written as

$$K_i^\pi(\theta) = \partial_j [A_{ij}(\theta) \pi(\theta)]. \quad (49)$$

Its divergence vanishes identically:

$$\partial_i K_i^\pi = \partial_i \partial_j [A_{ij}(\theta) \pi(\theta)] = 0, \quad (50)$$

because A_{ij} is antisymmetric while $\partial_i \partial_j$ is symmetric. Thus, $A(\theta)$ is a gauge potential for stationary probability flow.

In two dimensions, the antisymmetric tensor is represented by a single scalar stream function $\chi(\theta_1, \theta_2)$. Writing

$$A(\theta) \pi(\theta) = \begin{pmatrix} 0 & \chi(\theta) \\ -\chi(\theta) & 0 \end{pmatrix}, \quad (51)$$

we obtain

$$K^\pi(\theta) = \pi(\theta) c(\theta) = \begin{pmatrix} \partial_2 \chi(\theta) \\ -\partial_1 \chi(\theta) \end{pmatrix}, \quad (52)$$

up to the sign convention used for A_{12} . This is the usual stream-function representation of a divergence-free current.

In an arbitrary dimension, the same construction is the Hodge representation of a divergence-free vector field. If

$$K_i^\pi(\theta) = \pi(\theta) c_i(\theta)$$

satisfies $\partial_i K_i^\pi = 0$, then locally it can be represented by an antisymmetric tensor potential:

$$K_i^\pi(\theta) = \partial_j \chi_{ij}(\theta), \quad \chi_{ij}(\theta) = -\chi_{ji}(\theta). \quad (53)$$

One may further write

$$\chi_{ij} = \partial_i \Omega_j - \partial_j \Omega_i, \quad (54)$$

with a gauge condition such as $\partial_i \Omega_i = 0$. Then, Ω solves a Poisson-type equation whose source is the gauge-induced stationary current K^π . This is the higher-dimensional analog of the two-dimensional stream function.

In the Hamiltonian representation, the nonreversible gauge has a particularly transparent meaning. The reversible Fokker-Planck operator is mapped by a similarity transformation to the Hermitian 0-form SUSY Hamiltonian H_{SUSY} . A stationary-density-preserving gauge field A adds a non-Hermitian perturbation:

$$\tilde{H}_{\text{FP}} = H_{\text{SUSY}} + \tilde{H}_A.$$

Because the gauge preserves the target density, the zero mode is unchanged. Thus, the gauge acts by deforming the excited spectrum while keeping the ground state fixed. In this sense, gauge optimization is a spectral-design problem: One changes the low-lying relaxation modes without changing the stationary distribution. For reversible dynamics, the spectrum is real and relaxation is purely dissipative, whereas a nonreversible gauge can produce complex eigenvalues whose imaginary parts represent circulating probability currents. In particular, Ichiki and Ohzeki showed that detailed-balance violation can generate imaginary components in the relaxation spectrum [1]. Because the gauge-perturbed operator is non-Hermitian and generally non-normal, the real parts of the eigenvalues are not the whole finite-time story. Transient amplification, pseudospectral sensitivity, and possible coalescence of eigenvalues at exceptional points can also affect relaxation. The gap is therefore used below as an asymptotic spectral diagnostic,

while the finite-time objective is evaluated directly from the dynamics. The Ohzeki-Ichiki symplectic force is a concrete example of such a perturbation. They further showed that, in the strong-gauge limit, the Hamiltonian, energy-preserving part dominates and the dynamics approaches the HMC limit [6]. This spectral viewpoint clarifies why a finite gauge may be preferable: One wants to increase the real parts of the slow relaxation modes without paying excessive control cost or collapsing into purely energy-shell motion.

V. VARIATIONAL PRINCIPLE FOR THE OPTIMAL GAUGE

The Ohzeki-Ichiki method shows that adding a nonreversible force while preserving the stationary distribution can accelerate relaxation. However, it does not determine how strong the force should be. In continuous time, an excessively strong force carries control cost and, depending on the physical realization, may also carry entropy production; in discrete time, it may worsen numerical error and Metropolis rejection. The optimal gauge is therefore not the strongest gauge in isolation, but the gauge that gives the best balance between finite-time relaxation and a specified cost. This specification is essential: Without a force-amplitude, thermodynamic, discretization, or rejection constraint, there is generally no parameter-free finite strength to derive.

We formulate this balance as a finite-time transport problem in parameter space. Let $\rho_0(\theta)$ be the initial density of the stochastic parameters and let $\pi(\theta)$ be the target stationary density. For a given antisymmetric gauge field $A(\theta)$, let $\rho_A(\theta, t)$ be the density path generated by the Fokker-Planck dynamics with the corresponding drift $c_A(\theta)$. The relevant question is how efficiently $\rho_A(\cdot, t)$ is transported from ρ_0 toward π over a finite-time horizon $[0, \tau]$.

The Benamou-Brenier formulation gives a useful reference for this transport viewpoint [52]. If arbitrary velocity fields were allowed, the least-action density path between the prescribed end points would be obtained by minimizing

$$\inf_{\rho, v} \int_0^\tau \int_{\mathbb{R}^d} \frac{1}{2} \rho(\theta, t) |v(\theta, t)|^2 d\theta dt, \quad \partial_t \rho + \partial_i(\rho v_i) = 0, \quad (55)$$

with end-point conditions

$$\rho(\theta, 0) = \rho_0(\theta), \quad \rho(\theta, \tau) = \pi(\theta). \quad (56)$$

This problem defines the least-action transport from the initial density to the target density. It is also related to the Wasserstein gradient-flow interpretation of the reversible Fokker-Planck equation [53].

The gauge problem is more constrained. The actual Fokker-Planck dynamics cannot realize an arbitrary velocity field. With a gauge drift c_A , the admissible current velocity is

$$v_{A,i}[\rho] = v_{0,i}[\rho] + c_{A,i}, \quad v_{0,i}[\rho] = -\{G\nabla L\}_i - \frac{1}{\rho} \partial_j [(D_{\text{eff}})_{ij} \rho]. \quad (57)$$

The density path ρ_A is therefore determined by

$$\partial_t \rho_A + \partial_i [\rho_A v_{A,i}[\rho_A]] = 0, \quad \rho_A(\theta, 0) = \rho_0(\theta). \quad (58)$$

One may interpret the gauge design as a proximal projection of an ideal transport flow onto this admissible family. However, computing the Benamou-Brenier velocity is itself a hard optimal-transport problem. We therefore use it only as motivation for the density-transport viewpoint and define the practical gauge objective directly in terms of measurable finite-time relaxation.

Let $\mathcal{R}_\tau[\rho_A, \pi]$ be a non-negative relaxation loss that measures how close the density path ρ_A comes to the target π during the time interval $[0, \tau]$. The regularized gauge objective is

$$\mathcal{J}_\lambda[A] = \mathcal{R}_\tau[\rho_A, \pi] + \frac{\lambda}{2} \int_0^\tau \int_{\mathbb{R}^d} \rho_A(\theta, t) |c_A(\theta)|^2 d\theta dt, \quad (59)$$

where $\lambda > 0$ is the prescribed cost coefficient for the gauge force. The quadratic term in Eq. (59) is a control-power proxy for the externally imposed drift c_A . It should not be identified, in general, with stochastic thermodynamic entropy production. A microscopic thermodynamic formulation would require the mobility, noise, eliminated variables, and time-reversal parities to be specified, and would typically lead to a mobility-weighted current or entropy-production functional rather than the simple Euclidean norm used here [38,54]. For sampling, one may choose

$$\mathcal{R}_\tau[\rho_A, \pi] = D_{\text{KL}}(\rho_A(\cdot, \tau) \parallel \pi)$$

or its time-integrated version. For observable relaxation, one may choose

$$\mathcal{R}_\tau[\rho_A, \pi] = \int_0^\tau |\mathbb{E}_{\rho_A(t)}[f] - \mathbb{E}_\pi[f]| dt.$$

In the double-well experiment as shown below, the target distribution is symmetric and $\mathbb{E}_\pi[x] = 0$. Therefore, we use

$$\mathcal{R}_\tau[\rho_A, \pi] = \int_0^\tau |\mathbb{E}_{\rho_A(t)}[x]| dt. \quad (60)$$

Thus, the learning algorithm does not require solving the Benamou-Brenier problem. It only requires simulating, or otherwise propagating, the Fokker-Planck dynamics under candidate gauge fields and estimating the objective in Eq. (59).

VI. LEARNING CONTINUOUS GAUGE FIELDS BY ACTOR-CRITIC OPTIMIZATION

Objective (59) suggests a learning-based route to gauge design. In general, the optimal antisymmetric tensor field $A^*(\theta)$ is not available in closed form. We therefore parametrize the admissible gauge and optimize its parameters using finite-time information generated by the controlled dynamics.

We use the term actor-critic in analogy with policy-gradient methods in reinforcement learning [55,56]. In the standard actor-critic setting, the actor is a parametrized policy, while the critic estimates the value or gradient information needed to update that policy. In the present continuous-time setting, the actor is not a discrete action policy. It is the parametrized antisymmetric gauge field A_ψ . The critic estimates the finite-time objective $\mathcal{J}_\lambda[\psi]$ and, when needed, its gradient with respect to the actor parameters ψ . This estimate

may be obtained from rollouts, as in the numerical experiment below, or from an adjoint Fokker-Planck equation.

Thus, the actor is an antisymmetric tensor field

$$A_\psi(\theta) = -A_\psi(\theta)^\top,$$

with trainable parameters ψ . It generates the nonreversible drift

$$c_{\psi,i}(\theta) = \frac{1}{\pi(\theta)} \partial_j [A_{\psi,ij}(\theta) \pi(\theta)]. \quad (61)$$

This hard parametrization preserves the target density in continuous time, because

$$\partial_i \{\pi(\theta) c_{\psi,i}(\theta)\} = \partial_i \partial_j [A_{\psi,ij}(\theta) \pi(\theta)] = 0.$$

The normalization constant of π is not needed. Thus, learning changes the stationary current, not the target density.

Let $\rho_\psi(\theta, t)$ be the density path generated by the Fokker-Planck dynamics with gauge drift c_ψ . Then, the practical objective is $\mathcal{J}_\lambda[A_\psi]$, the parametrized version of Eq. (59). For observable relaxation, a convenient terminal choice of $\mathcal{R}_\tau$ is

$$\mathcal{R}_\tau[\rho_\psi, \pi] = \frac{1}{2} \sum_a w_a \left[\int_{\mathbb{R}^d} f_a(\theta) \rho_\psi(\theta, \tau) d\theta - \int_{\mathbb{R}^d} f_a(\theta) \pi(\theta) d\theta \right]^2, \quad w_a \geq 0, \quad (62)$$

where f_a are chosen observables. A time-integrated version may also be used when one wants to penalize slow relaxation throughout the whole interval $[0, \tau]$. For sampling, $\mathcal{R}_\tau$ may instead be chosen as a terminal or time-integrated discrepancy between ρ_ψ and π , such as relative entropy, Wasserstein distance, or total variation distance.

We now describe the adjoint form of the critic. This derivation is not needed for the scalar finite-difference experiment below, but it shows what the continuum actor-critic gradient computes. For a fixed parameter value ψ , let $\rho_\psi(\theta, t)$ solve

$$\partial_t \rho_\psi = \mathcal{L}_\psi^\dagger \rho_\psi, \quad \rho_\psi(\theta, 0) = \rho_0(\theta), \quad (63)$$

where $\mathcal{L}_\psi$ is the backward generator of the dynamics with gauge drift c_ψ . We write

$$\mathcal{L}_\psi f = \mathcal{L}_0 f + c_\psi \cdot \nabla f, \quad (64)$$

where $\mathcal{L}_0$ is the baseline generator without the learned gauge.

Consider the terminal observable loss

$$\Phi(\rho_\psi(\tau)) = \frac{1}{2} \sum_a w_a [\rho_\psi(\tau)(f_a) - \pi(f_a)]^2, \quad (65)$$

with

$$\begin{aligned} \rho_\psi(t)(f_a) &= \int_{\mathbb{R}^d} f_a(\theta) \rho_\psi(\theta, t) d\theta, \\ \pi(f_a) &= \int_{\mathbb{R}^d} f_a(\theta) \pi(\theta) d\theta. \end{aligned} \quad (66)$$

The full finite-time objective is

$$\begin{aligned} \mathcal{J}_\lambda[\psi] &= \Phi(\rho_\psi(\tau)) + \int_0^\tau \int_{\mathbb{R}^d} \rho_\psi(\theta, t) r_\psi(\theta) d\theta dt, \\ r_\psi(\theta) &= \frac{\lambda}{2} |c_\psi(\theta)|^2. \end{aligned} \quad (67)$$

The first term measures finite-time relaxation of the chosen observables, while the running cost r_ψ penalizes the gauge force.

To differentiate $\mathcal{J}_\lambda$, first vary the forward equation. A perturbation $\delta\psi$ induces δc_ψ and $\delta\rho$, with

$$\partial_t \delta\rho = \mathcal{L}_\psi^\dagger \delta\rho + \delta\mathcal{L}_\psi^\dagger \rho_\psi, \quad \delta\rho(\theta, 0) = 0. \quad (68)$$

Since the parameter enters through the drift c_ψ ,

$$\delta\mathcal{L}_\psi f = \delta c_\psi \cdot \nabla f, \quad \delta\mathcal{L}_\psi^\dagger \rho_\psi = -\partial_i [\rho_\psi \delta c_{\psi,i}]. \quad (69)$$

The adjoint critic $a(\theta, t)$ is defined backward in time by

$$-\partial_t a = \mathcal{L}_\psi a + r_\psi, \quad a(\theta, \tau) = \frac{\delta\Phi}{\delta\rho_\tau}(\theta). \quad (70)$$

For the terminal observable loss (65), the terminal condition is

$$a(\theta, \tau) = \sum_a w_a [\rho_\psi(\tau)(f_a) - \pi(f_a)] f_a(\theta). \quad (71)$$

The source term r_ψ appears in Eq. (70) because the objective contains the time-integrated control cost.

Using Eq. (68) and integrating by parts in time and space, with vanishing boundary terms, one obtains

$$\delta\mathcal{J}_\lambda = \int_0^\tau \int_{\mathbb{R}^d} \rho_\psi \delta c_\psi \cdot \nabla a d\theta dt + \int_0^\tau \int_{\mathbb{R}^d} \rho_\psi \delta r_\psi d\theta dt. \quad (72)$$

Since

$$r_\psi = \frac{\lambda}{2} |c_\psi|^2, \quad \delta r_\psi = \lambda c_\psi \cdot \delta c_\psi,$$

the gradient with respect to a parameter component ψ_k is

$$\frac{\partial \mathcal{J}_\lambda}{\partial \psi_k} = \int_0^\tau \int_{\mathbb{R}^d} \rho_\psi(\theta, t) \partial_{\psi_k} c_\psi(\theta) \cdot [\nabla a(\theta, t) + \lambda c_\psi(\theta)] d\theta dt. \quad (73)$$

Thus, the computation proceeds in three steps. First, solve or sample the forward dynamics (63) to obtain ρ_ψ . Second, solve the backward adjoint equation (70). Third, use Eq. (73) to update the actor parameter:

$$\psi_{n+1} = \psi_n - \alpha_n \nabla_{\psi} \mathcal{J}_\lambda[\psi_n]. \quad (74)$$

In this sense, A_ψ is the actor, because it determines the gauge drift c_ψ , while $a(\theta, t)$ is the critic, because it evaluates how changes in the drift affect the finite-time objective.

VII. OHZEKI-ICHIKI DYNAMICS AND THE HMC LIMIT

We now connect the gauge formulation to the Ohzeki-Ichiki construction. Let the state variable be

$$\theta = (x, y) \in \mathbb{R}^{2d},$$

where $x \in \mathbb{R}^d$ is the original variable and $y \in \mathbb{R}^d$ is an auxiliary variable. Consider the separable energy

$$H(x, y) = H_x(x) + H_y(y), \quad (75)$$

and the target density

$$\pi(x, y) = \frac{1}{Z} \exp \left[-\frac{H(x, y)}{T} \right]. \quad (76)$$

The generalized Ohzeki-Ichiki dynamics is

$$dx_i = [-\partial_{x_i} H_x(x) + \gamma \partial_{y_i} H_y(y)] dt + \sqrt{2T} dW_i^x, \quad (77)$$

$$dy_i = [-\partial_{y_i} H_y(y) - \gamma \partial_{x_i} H_x(x)] dt + \sqrt{2T} dW_i^y. \quad (78)$$

Here, $T > 0$ is the temperature, $\gamma \geq 0$ is the strength of the nonreversible force, and W^x, W^y are independent Brownian motions.

Introduce the constant antisymmetric matrix

$$S = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}, \quad S^\top = -S. \quad (79)$$

Then, Eq. (78) can be written compactly as

$$d\theta_t = [-\nabla H(\theta_t) + \gamma S \nabla H(\theta_t)] dt + \sqrt{2T} dW_t. \quad (80)$$

The added drift is

$$c_\gamma(\theta) = \gamma S \nabla H(\theta). \quad (81)$$

It preserves the target density because

$$\partial_i \{\pi(\theta) c_{\gamma,i}(\theta)\} = \gamma \partial_i [\pi(\theta) S_{ij} \partial_j H(\theta)] = 0. \quad (82)$$

Indeed, the term $S_{ij} \partial_i \partial_j H$ vanishes by antisymmetry of S , while $S_{ij} \partial_i H \partial_j H$ also vanishes because it contracts an antisymmetric matrix with a symmetric product. Thus, the Ohzeki-Ichiki force is a constant-antisymmetric-gauge current in the notation of Eq. (48). Equivalently, it is generated by the antisymmetric tensor

$$A_\gamma(\theta) = -\gamma T S, \quad (83)$$

since

$$\frac{1}{\pi} \partial_j [A_{\gamma,ij} \pi] = \gamma S_{ij} \partial_j H \quad (84)$$

up to the sign convention chosen in Eq. (48). This is a symplectic nonreversible flow in arbitrary dimension. For $d = 1$, the state space is two dimensional and the current appears as a rotation in the (x, y) plane.

Ichiki and Ohzeki showed that the large- γ limit connects this dynamics to Hamiltonian dynamics [6]. After rescaling time by the strength of the nonreversible force, the dominant part of Eq. (80) becomes

$$\frac{dx_i}{ds} = \partial_{y_i} H_y(y), \quad \frac{dy_i}{ds} = -\partial_{x_i} H_x(x). \quad (85)$$

For

$$H_y(y) = \sum_i \frac{y_i^2}{2m_i},$$

these are Hamilton's equations with masses m_i . Thus, Hamiltonian Monte Carlo arises as the infinite-strength detailed-balance-breaking limit of the Ohzeki-Ichiki dynamics.

This viewpoint also explains why the HMC limit is not automatically optimal. Hamiltonian flow transports probability rapidly along energy surfaces, but it preserves the energy $H(x, y)$. Transitions between energy surfaces must therefore come from momentum resampling, noise, friction, or Metropolis-Hastings correction [57–60]. A finite nonreversible Langevin gauge combines two effects: symplectic motion along energy surfaces and stochastic diffusion across them. Once finite-time transport performance, control cost,

discretization stability, and possible Metropolis rejection are included, the optimal gauge strength need not be $\gamma = \infty$.

In the terminology of this paper, the Ohzeki-Ichiki family is a one-parameter subfamily of admissible antisymmetric gauges. The learned-gauge problem asks whether a finite gauge field A^* , or in this restricted case a finite strength γ^* , gives a better balance than either reversible Langevin dynamics ($\gamma = 0$) or the Hamiltonian limit ($\gamma = \infty$). When the resulting nonreversible Langevin proposal is used for exact sampling after time discretization, it can be combined with a Metropolis correction. We refer broadly to this finite-gauge Langevin strategy as Stochastic Langevin Gradient Descent (SLGD), emphasizing Langevin dynamics augmented by a symplectic or nonreversible gauge force rather than the infinite-strength HMC limit.

VIII. ANALYTICAL GAUSSIAN BENCHMARK

Before turning to a nonconvex metastable landscape, we consider an exactly solvable convex benchmark. Let $x \in \mathbb{R}^d$ and consider the quadratic potential

$$U(x) = \frac{1}{2} x^\top K x, \quad K = \Sigma^{-1}, \quad (86)$$

where K is symmetric positive definite. The target density is

$$\pi(x) = \frac{1}{Z} \exp \left[-\frac{1}{2T} x^\top K x \right]. \quad (87)$$

At $T = 1$, the stationary covariance is Σ ; for general T , it is $T\Sigma$. The reversible Langevin dynamics is the Ornstein-Uhlenbeck (OU) process

$$dx_t = -Kx_t dt + \sqrt{2T} dW_t. \quad (88)$$

If K has a small eigenvalue, relaxation along the corresponding flat direction is slow.

We now add a constant antisymmetric gauge. Let $S^\top = -S$ and define the stationary-density-preserving drift

$$c_a(x) = aSKx, \quad (89)$$

where $a \in \mathbb{R}$ is a scalar gauge strength. This is generated by the antisymmetric tensor potential $-aTS$ in the convention of Eq. (48). The gauge-modified OU process is

$$dx_t = -M_a x_t dt + \sqrt{2T} dW_t, \quad M_a = (I - aS)K. \quad (90)$$

The mean $\mu_a(t) = \mathbb{E}[x_t]$ obeys

$$\dot{\mu}_a(t) = -M_a \mu_a(t). \quad (91)$$

Thus, the relaxation of the mean is determined exactly by the spectrum of the non-Hermitian drift matrix M_a .

In two dimensions, take

$$K = \begin{pmatrix} k_1 & 0 \\ 0 & k_2 \end{pmatrix}, \quad 0 < k_1 < k_2, \quad S = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \quad (92)$$

The eigenvalues of M_a are

$$\lambda_{\pm}(a) = \frac{k_1 + k_2}{2} \pm \sqrt{\frac{(k_1 - k_2)^2}{4} - a^2 k_1 k_2}. \quad (93)$$

For $a = 0$, the spectral gap is k_1 , the slow flat direction. When

$$|a| \geq a_{\text{crit}} := \frac{|k_1 - k_2|}{2\sqrt{k_1 k_2}} = \frac{1}{2} \left| \frac{\sigma_1}{\sigma_2} - \frac{\sigma_2}{\sigma_1} \right|, \quad (94)$$

where $k_i = \sigma_i^{-2}$, the eigenvalues become a complex conjugate pair and

$$\text{Re } \lambda_{\pm}(a) = \frac{k_1 + k_2}{2}. \quad (95)$$

Thus, the gauge converts independent dissipative relaxation into a chiral spiral and removes the asymptotic slow-mode bottleneck. For $k_1 = 0.1$ and $k_2 = 1$, the asymptotic mean relaxation rate improves from 0.1 to 0.55, a factor of 5.5.

The critical value a_{crit} , however, is not the same as the optimum of a finite-time regularized objective. It is the smallest gauge strength at which the asymptotic spectral rate reaches the plateau. The actual finite optimum depends on the time horizon, the initial condition, the chosen observable, and the force penalty. This distinction is useful because it is exactly the distinction made throughout this paper: Spectral acceleration alone does not select a unique finite gauge.

For the finite-time objective, let

$$Q_a(t) = \mathbb{E}[x_t x_t^T]$$

be the second moment. It satisfies the Lyapunov equation

$$\dot{Q}_a(t) = -M_a Q_a(t) - Q_a(t) M_a^T + 2TI. \quad (96)$$

The exact regularized cost used in the benchmark is

$$\begin{aligned} C_{\lambda}^{(\tau)}(a) &= \int_0^{\tau} \|\mu_a(t)\|^2 dt + \lambda \int_0^{\tau} \text{Tr}[B_a^T B_a Q_a(t)] dt, \\ B_a &= aSK. \end{aligned} \quad (97)$$

The first term measures relaxation of the mean, and the second term is the expected gauge-force cost $\mathbb{E}[|c_a(x_t)|^2]$. Equations (91) and (96) make this benchmark deterministic: No Monte Carlo sampling is needed to compute the critic.

In the numerical benchmark, we set

$$k_1 = 0.1, \quad k_2 = 1, \quad T = 1, \quad \tau = 10, \quad \lambda = 5.0 \times 10^{-3},$$

and initialize the mean along the slow direction,

$$\mu_0 = (1, 0)^T.$$

The spectral transition occurs at

$$a_{\text{crit}} = 1.423.$$

Direct minimization of Eq. (97) gives

$$a_{\lambda}^* = 2.909, \quad C_{\lambda}^{(10)}(a_{\lambda}^*) = 1.401.$$

For comparison,

$$C_{\lambda}^{(10)}(0) = 4.323, \quad C_{\lambda}^{(10)}(a_{\text{crit}}) = 2.089.$$

The finite optimum is larger than a_{crit} because the finite-time mean relaxation still benefits from additional rotation until the force penalty balances the gain. This is not a contradiction with the spectral analysis: a_{crit} marks the onset of the complex-spectrum plateau, whereas a_{λ}^* minimizes a specified finite-time cost.

We then apply the same scalar actor update used later in the double-well experiment. Starting from $a_0 = 0.6$, using a centered finite-difference critic with $\epsilon = 0.05$ and step size

$$\alpha_n = \frac{0.5}{1 + 0.05n},$$

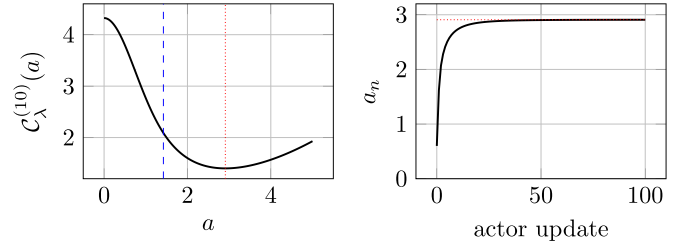


FIG. 1. Finite-gauge learning in the anisotropic Gaussian OU benchmark. In the left panel, the black solid curve is the finite-time objective, the blue dashed vertical line is the spectral threshold a_{crit} , and the red dotted vertical line is the finite-time optimum $a_{\lambda}^* \simeq 2.909$. In the right panel, the black solid curve is the actor trajectory and the red dotted horizontal line is a_{λ}^* .

100 actor updates give

$$a_{\text{learned}} = 2.909,$$

matching the finite-time optimum a_{λ}^* computed from the Lyapunov equations. Figure 1 shows the objective curve and the actor trajectory. Figure 2 shows the mean trajectory: The reversible process remains on the slow axis, while the learned gauge produces a chiral relaxation path toward the origin.

IX. DOUBLE-WELL VALIDATION BY FORWARD SIMULATION

As a toy validation, we use the double-well potential

$$U(x) = \frac{x^4}{4} - \frac{x^2}{2}. \quad (98)$$

We introduce an auxiliary variable y and define

$$H(x, y) = U(x) + \frac{1}{2}y^2, \quad S = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \quad (99)$$

The one-parameter Ohzeki-Ichiki gauge family is

$$c_{\gamma}(x, y) = \gamma S \nabla H(x, y) = \gamma \begin{pmatrix} y \\ -U'(x) \end{pmatrix}. \quad (100)$$

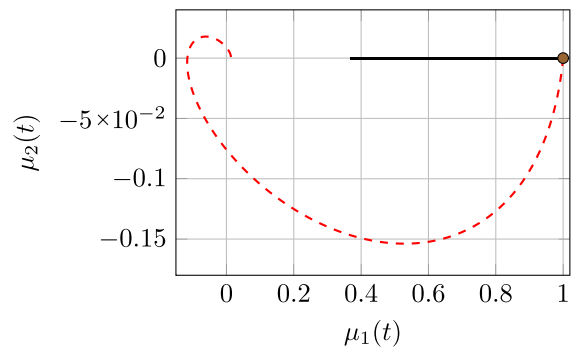


FIG. 2. Mean trajectory in the anisotropic Gaussian OU benchmark from $\mu_0 = (1, 0)^T$. The black solid curve is the reversible trajectory $a = 0$, the red dashed curve is the learned gauge $a = 2.909$, and the marker denotes the initial mean. The reversible trajectory decays only along the slow direction, while the learned nonreversible gauge bends the flow into a chiral path and accelerates relaxation.

Thus, the learned actor is the scalar gauge strength γ , which is a one-dimensional restriction of the general antisymmetric gauge field A_ψ .

For a given γ , we simulate the forward Langevin dynamics

$$d\begin{pmatrix} x_t \\ y_t \end{pmatrix} = [-\nabla H(x_t, y_t) + c_\gamma(x_t, y_t)]dt + \sqrt{2T} dW_t. \quad (101)$$

The target stationary density is symmetric in x , so $\mathbb{E}_\pi[x] = 0$. We therefore measure relaxation from the right well by the finite-time observable error $|\mathbb{E}_\gamma[x_t]|$.

For a time horizon τ , the rollout objective is

$$\mathcal{C}_\lambda^{(\tau)}(\gamma) = \int_0^\tau |\mathbb{E}_\gamma[x_t]| dt + \lambda \int_0^\tau \mathbb{E}_\gamma[|c_\gamma(x_t, y_t)|^2] dt. \quad (102)$$

Equivalently,

$$|c_\gamma(x, y)|^2 = \gamma^2[y^2 + \{U'(x)\}^2]. \quad (103)$$

The first term measures finite-time relaxation, and the second term penalizes the strength of the nonreversible force.

We optimize γ directly from forward simulations. At actor update n , we evaluate

$$\widehat{\mathcal{C}}_\lambda^{(\tau)}(\gamma_n + \epsilon), \quad \widehat{\mathcal{C}}_\lambda^{(\tau)}(\gamma_n - \epsilon)$$

with common random numbers and form the finite-difference estimate

$$\widehat{g}_n = \frac{\widehat{\mathcal{C}}_\lambda^{(\tau)}(\gamma_n + \epsilon) - \widehat{\mathcal{C}}_\lambda^{(\tau)}(\gamma_n - \epsilon)}{2\epsilon}. \quad (104)$$

The scalar actor update is

$$\gamma_{n+1} = \Pi_{[0, \gamma_{\max}]}[\gamma_n - \alpha_n \widehat{g}_n]. \quad (105)$$

Thus, the numerical procedure is simply: simulate the dynamics forward, estimate the expectations in Eq. (102), compute a finite-difference gradient, and update γ .

In the numerical experiment, we set

$$T = 0.45, \quad \tau = 24, \quad \lambda = 3.0 \times 10^{-3}, \\ \epsilon = 0.15, \quad \gamma_{\max} = 6.$$

The actor is initialized at $\gamma_0 = 0.7$ and updated for 160 iterations with

$$\alpha_n = \frac{0.25}{1 + 0.10n}.$$

Training rollouts use time step 2.0×10^{-3} and 800 trajectories. Validation rollouts use time step 1.0×10^{-3} and 2500 trajectories with independent random seeds. All trajectories start from $x(0) = 1$, and $y(0)$ is sampled from the Gaussian equilibrium marginal $N(0, T)$.

With these settings, the learned value is

$$\gamma_{\text{learned}} = 2.006. \quad (106)$$

For validation, the learned gauge gives

$$\int_0^{24} |\mathbb{E}[x_t]| dt = 1.243, \quad \mathcal{C}_\lambda^{(24)} = 1.589.$$

This is slightly below the fixed $\gamma = 2$ baseline, which gives 1.248 and 1.592, respectively. The stronger fixed $\gamma = 3$ baseline has a smaller raw relaxation area, 1.016, but the larger force penalty raises its regularized objective to 1.787. Thus,

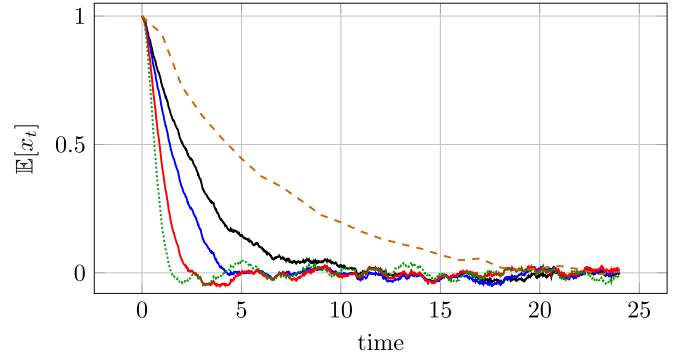


FIG. 3. Relaxation of the mean position in the double-well model. The black, blue, and red solid curves show $\gamma = 0$, $\gamma = 1$, and the learned finite gauge $\gamma = 2.006$, respectively; the green dotted curve shows $\gamma = 3$ and the orange dashed curve shows HMC. The learned finite gauge improves the regularized objective $\mathcal{C}_\lambda$, while the stronger fixed gauge $\gamma = 3$ is faster in raw relaxation but has a larger control cost.

the optimized finite gauge is not the strongest gauge; it is the one that best balances relaxation and control cost under $\mathcal{C}_\lambda$.

Figure 3 shows the relaxation of $\mathbb{E}[x_t]$. The learned nonreversible gauge accelerates reversible Langevin dynamics and the fixed $\gamma = 1$ baseline. The fixed $\gamma = 3$ curve relaxes faster in the raw observable, but its control cost is larger. HMC is included as a reference for the Hamiltonian limit; in this setting, its relaxation per unit physical time is slower because transitions between energy shells depend on momentum resampling. For the HMC reference, we use leapfrog step size 0.025, 40 leapfrog steps per proposal, full momentum refresh between proposals, and Metropolis correction. The corresponding proposal time is 1.0, and the observed acceptance rate is 0.9999. This high acceptance rate indicates a conservative discretization. The HMC curve is therefore included as a reference for the Hamiltonian, infinite-gauge limit, not as a fully optimized HMC or No-U-Turn Sampler (NUTS) benchmark.

X. CONCLUSION AND OUTLOOK

We developed an operator-theoretic framework for nonreversible Fokker-Planck dynamics with stationary-density-preserving gauge fields. In the reversible case, a similarity transformation maps the Fokker-Planck operator to a Witten-Laplacian-type supersymmetric Hamiltonian. More precisely, the density dynamics is mapped to the scalar sector of this supersymmetric Hamiltonian. The invariant density becomes the ground state, while relaxation is controlled by the low-lying excited spectrum. Nonreversible gauges generated by antisymmetric tensor fields preserve this ground state but deform the excited spectrum, producing circulating probability currents and, in general, complex relaxation eigenvalues.

This perspective unifies several structures that are often treated separately. Minibatch stochastic gradients appear as effective diffusion tensors or temperatures; adaptive gradient methods appear as metric choices; and detailed-balance-breaking forces appear as gauge currents. The gauge current considered here is an effective external control preserving

a fixed target density, not a unique microscopic realization of odd transport or entropy production. The Ohzeki-Ichiki force is a concrete symplectic gauge in this language, and its infinite-strength limit connects to Hamiltonian dynamics and HMC. However, the strongest nonreversible force is not automatically optimal. The continuous-time spectral gap may improve or saturate with increasing gauge strength, while finite-time transport, force cost, discretization stability, and Metropolis rejection impose additional constraints.

We therefore formulated finite-gauge selection as a regularized relaxation problem. The anisotropic Gaussian OU benchmark gives an exactly solvable test: The spectral transition occurs at $a_{\text{crit}} = 1.423$, whereas the chosen finite-time objective selects $a_{\lambda}^* = 2.909$, and the actor update recovers this Lyapunov-equation optimum. The double-well benchmark shows the same principle in a nonconvex metastable setting: The learned value $\gamma_{\text{learned}} = 2.006$ slightly improves the regularized objective relative to the fixed $\gamma = 2$ baseline, while the stronger $\gamma = 3$ gauge relaxes faster in the raw observable but pays a larger control cost. Thus, finite nonreversible gauges are constrained optima, not simply approximations to the HMC limit.

Several directions remain open. First, the spectral effect of state-dependent diffusion tensors and adaptive metrics should be characterized more rigorously in high dimension, for example, using hypocoercivity or curvature-based methods [61]. Second, practical optimizers such as Adam

should be analyzed as extended-state Fokker-Planck dynamics, separating metric adaptation from genuine stationary currents. Third, scalable parametrizations of antisymmetric tensor fields A_{ψ} are needed, together with rollout-based or adjoint-based critics for finite-time objectives. Fourth, finite-gauge Langevin methods with optional Metropolis correction should be benchmarked against SGLD, HMC, nonreversible Langevin samplers, learned MCMC proposals, and Adam-type optimizers [62,63].

The central physical message is that nonreversible gauges provide a controlled way to deform the excited spectrum of a Fokker-Planck operator while preserving its stationary state. This makes gauge learning a problem of physically constrained spectral design for stochastic dynamics.

ACKNOWLEDGMENT

The author acknowledges financial support from the Cross-ministerial Strategic Innovation Promotion Program (SIP) of the Cabinet Office, Government of Japan, No. 23836436.

DATA AVAILABILITY

There are no publicly available research data or software supporting this manuscript. Requests for further information or data should be sent to the authors.

- [1] A. Ichiki and M. Ohzeki, Violation of detailed balance accelerates relaxation, *Phys. Rev. E* **88**, 020101(R) (2013).
- [2] M. Ohzeki and A. Ichiki, Langevin dynamics neglecting detailed balance condition, *Phys. Rev. E* **92**, 012105 (2015).
- [3] M. Ohzeki and A. Ichiki, Mathematical understanding of detailed balance condition violation and its application to Langevin dynamics, *J. Phys.: Conf. Ser.* **638**, 012003 (2015).
- [4] L. Rey-Bellet and K. Spiliopoulos, Improving the convergence of reversible samplers, *J. Stat. Phys.* **164**, 472 (2016).
- [5] A. B. Duncan, T. Lelièvre, and G. A. Pavliotis, Variance reduction using nonreversible Langevin samplers, *J. Stat. Phys.* **163**, 457 (2016).
- [6] A. Ichiki and M. Ohzeki, Stochastic dynamics without detailed balance condition connecting simple gradient method and Hamiltonian Monte Carlo, [arXiv:2107.08633](https://arxiv.org/abs/2107.08633).
- [7] J. E. Avron, R. Seiler, and P. G. Zograf, Viscosity of quantum Hall fluids, *Phys. Rev. Lett.* **75**, 697 (1995).
- [8] J. E. Avron, Odd viscosity, *J. Stat. Phys.* **92**, 543 (1998).
- [9] M. Fruchart, C. Scheibner, and V. Vitelli, Odd viscosity and odd elasticity, *Annu. Rev. Condens. Matter Phys.* **14**, 471 (2023).
- [10] D. Banerjee, A. Souslov, A. G. Abanov, and V. Vitelli, Odd viscosity in chiral active fluids, *Nat. Commun.* **8**, 1573 (2017).
- [11] T. Markovich and T. C. Lubensky, Odd viscosity in active matter: Microscopic origin and 3D effects, *Phys. Rev. Lett.* **127**, 048001 (2021).
- [12] C. Hargus, J. M. Epstein, and K. K. Mandadapu, Odd diffusivity of chiral random motion, *Phys. Rev. Lett.* **127**, 178001 (2021).
- [13] E. Kalz, H. D. Vuijk, I. Abdoli, J.-U. Sommer, H. Löwen, and A. Sharma, Collisions enhance self-diffusion in odd-diffusive systems, *Phys. Rev. Lett.* **129**, 090601 (2022).
- [14] H. Risken, *The Fokker-Planck Equation: Methods of Solution and Applications* (Springer, Berlin, 1984).
- [15] E. Witten, Supersymmetry and Morse theory, *J. Differ. Geom.* **17**, 661 (1982).
- [16] M. Welling and Y. W. Teh, Bayesian learning via stochastic gradient Langevin dynamics, in *Proceedings of the 28th International Conference on Machine Learning*, edited by L. Getoor and T. Scheffer (Omnipress, Madison, WI, 2011), pp. 681–688.
- [17] I. Sato and H. Nakagawa, Approximation analysis of stochastic gradient Langevin dynamics by using Fokker-Planck equation and Itô process, in *Proceedings of the 31st International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 32 (PMLR, Beijing, China, 2014), pp. 982–990.
- [18] S. Mandt, M. D. Hoffman, and D. M. Blei, Stochastic gradient descent as approximate Bayesian inference, *J. Mach. Learn. Res.* **18**, 1 (2017).
- [19] S. Jastrzębski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey, Width of minima reached by stochastic gradient descent is influenced by learning rate to batch size ratio, in *Artificial Neural Networks and Machine Learning - ICANN 2018*, Lecture Notes in Computer Science Vol. 11141 (Springer, Cham, 2018), pp. 392–402.
- [20] S. L. Smith and Q. V. Le, A Bayesian perspective on generalization and stochastic gradient descent, in *International Conference on Learning Representations* (OpenReview.net, Vancouver, Canada, 2018).
- [21] J.-F. Bony, D. Le Peutrec, and L. Michel, Eyring-Kramers law for Fokker-Planck type differential operators, *J. Eur. Math. Soc.* **27**, 4347 (2025).

- [22] S. Yaida, Fluctuation–dissipation relations for stochastic gradient descent, in *International Conference on Learning Representations* (OpenReview.net, New Orleans, LA, 2019).
- [23] F. Mignacco and P. Urbani, The effective noise of stochastic gradient descent, *J. Stat. Mech.* (2022) 083405.
- [24] Z. Zhu, J. Wu, B. Yu, L. Wu, and J. Ma, The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects, in *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 97 (PMLR, Long Beach, CA, 2019), pp. 7654–7663.
- [25] X. Dai and Y. Zhu, On large batch training and sharp minima: A Fokker–Planck perspective, *J. Stat. Theory Pract.* **14**, 53 (2020).
- [26] N. Yang, C. Tang, and Y. Tu, Stochastic gradient descent introduces an effective landscape-dependent regularization favoring flat solutions, *Phys. Rev. Lett.* **130**, 237101 (2023).
- [27] S. Geman and D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Trans. Pattern Anal. Mach. Intell.* **6**, 721 (1984).
- [28] B. Hajek, Cooling schedules for optimal annealing, *Math. Oper. Res.* **13**, 311 (1988).
- [29] S. L. Smith, P.-J. Kindermans, C. Ying, and Q. V. Le, Don’t decay the learning rate, increase the batch size, in *International Conference on Learning Representations* (OpenReview.net, Vancouver, Canada, 2018).
- [30] S. L. Smith, E. Elsen, and S. De, On the generalization benefit of noise in stochastic gradient descent, in *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 119 (PMLR, Virtual, 2020), pp. 9058–9067.
- [31] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, in *International Conference on Learning Representations* (OpenReview.net, San Diego, CA, 2015).
- [32] M. Girolami and B. Calderhead, Riemann manifold Langevin and Hamiltonian Monte Carlo methods, *J. Stat. Soc.: Ser. B* **73**, 123 (2011).
- [33] S. Patterson and Y. W. Teh, Stochastic gradient Riemannian Langevin dynamics on the probability simplex, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc. Red Hook, NY, 2013), Vol. 26.
- [34] Y.-A. Ma, T. Chen, and E. B. Fox, A complete recipe for stochastic gradient MCMC, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc. Red Hook, NY, 2015), Vol. 28.
- [35] M. Ohzeki, Stochastic gradient method with accelerated stochastic dynamics, *J. Phys.: Conf. Ser.* **699**, 012019 (2016).
- [36] H.-M. Chun, X. Durang, and J. D. Noh, Emergence of nonwhite noise in Langevin dynamics with magnetic Lorentz force, *Phys. Rev. E* **97**, 032117 (2018).
- [37] H. D. Vuijk, J. M. Brader, and A. Sharma, Anomalous fluxes in overdamped Brownian dynamics with Lorentz force, *J. Stat. Mech.* (2019) 063203.
- [38] H.-M. Chun, L. P. Fischer, and U. Seifert, Effect of a magnetic field on the thermodynamic uncertainty relation, *Phys. Rev. E* **99**, 042128 (2019).
- [39] I. Loshchilov and F. Hutter, Decoupled weight decay regularization, in *International Conference on Learning Representations* (OpenReview.net, New Orleans, LA, 2019).
- [40] N. Shazeer and M. Stern, Adafactor: Adaptive learning rates with sublinear memory cost, in *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 80 (PMLR, Stockholm, Sweden, 2018), pp. 4596–4604.
- [41] H. Liu, Z. Li, D. L. W. Hall, P. Liang, and T. Ma, Sophia: A scalable stochastic second-order optimizer for language model pre-training, in *International Conference on Learning Representations* (OpenReview.net, Vienna, Austria, 2024).
- [42] X. Chen, C. Liang, D. Huang, E. Real, K. Wang, Y. Liu, H. Pham, X. Dong, T. Luong, C.-J. Hsieh, Y. Lu, and Q. V. Le, Symbolic discovery of optimization algorithms, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc. Red Hook, NY, 2023), Vol. 36.
- [43] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, On large-batch training for deep learning: Generalization gap and sharp minima, in *International Conference on Learning Representations* (OpenReview.net, Toulon, France, 2017).
- [44] S. McCandlish, J. Kaplan, D. Amodei, and O. D. Team, An empirical model of large-batch training, [arXiv:1812.06162](https://arxiv.org/abs/1812.06162).
- [45] N. S. Keskar and R. Socher, Improving generalization performance by switching from Adam to SGD, [arXiv:1712.07628](https://arxiv.org/abs/1712.07628).
- [46] L. Luo, Y. Xiong, Y. Liu, and X. Sun, Adaptive gradient methods with dynamic bound of learning rate, in *International Conference on Learning Representations* (OpenReview.net, New Orleans, LA, 2019).
- [47] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht, The marginal value of adaptive gradient methods in machine learning, in *Advances in Neural Information Processing Systems 30* (Curran Associates, Red Hook, NY, 2017), Vol. 30.
- [48] S. J. Reddi, S. Kale, and S. Kumar, On the convergence of Adam and beyond, in *International Conference on Learning Representations* (OpenReview.net, Vancouver, Canada, 2018).
- [49] J. Cohen, S. Kaur, Y. Li, J. Z. Kolter, and A. Talwalkar, Gradient descent on neural networks typically occurs at the edge of stability, in *International Conference on Learning Representations* (OpenReview.net, Virtual, 2021).
- [50] F. Liao, A. Kolomvaki, and A. Kyriklidis, SGD at the edge of stability: The stochastic sharpness gap, [arXiv:2604.21016](https://arxiv.org/abs/2604.21016).
- [51] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, Sharpness-aware minimization for efficiently improving generalization, in *International Conference on Learning Representations* (OpenReview.net, Virtual, 2021).
- [52] J.-D. Benamou and Y. Brenier, A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem, *Numer. Math.* **84**, 375 (2000).
- [53] R. Jordan, D. Kinderlehrer, and F. Otto, The variational formulation of the Fokker–Planck equation, *SIAM J. Math. Anal.* **29**, 1 (1998).
- [54] F. Ghimenti, L. Berthier, G. Szamel, and F. van Wijland, Sampling efficiency of transverse forces in dense liquids, *Phys. Rev. Lett.* **131**, 257101 (2023).
- [55] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, Policy gradient methods for reinforcement learning with function approximation, in *Advances in Neural Information Processing Systems* (2000), Vol. 12.
- [56] V. R. Konda and J. N. Tsitsiklis, Actor-critic algorithms, in *Advances in Neural Information Processing Systems* (2000), Vol. 12.

- [57] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, Equation of state calculations by fast computing machines, *J. Chem. Phys.* **21**, 1087 (1953).
- [58] W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**, 97 (1970).
- [59] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, Hybrid Monte Carlo, *Phys. Lett. B* **195**, 216 (1987).
- [60] M. D. Hoffman and A. Gelman, The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo, *J. Mach. Learn. Res.* **15**, 1593 (2014).
- [61] D. Bakry and M. Émery, Diffusions hypercontractives, in *Séminaire de Probabilités XIX 1983/84*, Lecture Notes in Mathematics (Springer, Berlin, Heidelberg, 1985), Vol. 1123, pp. 177–206.
- [62] J. Song, S. Zhao, and S. Ermon, A-NICE-MC: Adversarial training for MCMC, in *Advances in Neural Information Processing Systems 30* (Curran Associates, Red Hook, NY, 2017), Vol. 30.
- [63] D. Levy, M. D. Hoffman, and J. Sohl-Dickstein, Generalizing Hamiltonian Monte Carlo with neural networks, in *International Conference on Learning Representations* (OpenReview.net, Vancouver, Canada, 2018).