INVERSE DESIGN OF IMPLICIT BIAS BY SYMMETRY, STOCHASTICITY AND HOLOMORPHIC PARAMETRIZATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Can we design a model such that its stochastic training favours a desired class of solutions without enforcing an explicit penalty? Under suitable conditions, the interplay between symmetries of a model's weight parametrization and stochastic training favours particular solutions, inducing an implicit bias. Building on this mechanism, we develop a framework for inverse-designing such biases by constructing novel parametrizations and their associated symmetries. We show how holomorphic functions make this construction and calculation simple and explicit. Specifically, we introduce a new parametrization that biases learned weights toward the binary values $\{-1, +1\}$. Numerical experiments confirm the theoretical predictions. They also show that our parametrization reproduces the same preference induced by an explicitly regularized model without adding a penalty to the training loss.

1 INTRODUCTION

Modern overparametrized learning systems can fit their training data in many ways. The loss function considers all interpolating solutions as equivalent, but the training procedure does not. The model parametrization, initialization, learning rate, mini-batch sampling, and noise during training influence which predictor the optimizer converges to. This selection mechanism, usually called implicit bias or implicit regularization, helps explain why predictors with similar training error can have different structural and generalization properties (Vardi, 2023).

Most studies of implicit bias address the following question: 'Given a model and an optimizer, which solutions does training favour?' Here, we study the reverse question: can we design a parametrization such that stochastic training favours prescribed predictors, thereby inducing a regularization without adding an explicit penalty term to the loss? We refer to this problem as *inverse design of implicit bias*. Thus, our aim is to encode the desired predictor preference in the geometry of the parametrization and let the optimizer reach it.

The key idea is that different parameter values can specify the same predictor. A predictor-preserving symmetry is a continuous transformation of the parameters that leaves the model's predictions unchanged. Parameters can therefore vary along symmetry directions without changing the function represented by the model. When the infinitesimal symmetry generator is the gradient of a scalar function of the parameters, continuous-time gradient flow preserves that function, which we call a *charge* by analogy with conserved quantities in theoretical physics. A finite optimization step can break this conservation law because, although its first-order contribution vanishes by symmetry, higher-order terms do not. We will show that, under suitable conditions, repeated stochastic optimization steps can drive the charge toward a specific value, thus biasing training toward particular predictors among those that fit the data.

Charge conservation, finite-step breaking of conservation laws, and stochastic selection have been studied from different perspectives and with different objectives by Kunin et al. (2021); Tanaka & Kunin (2021). Noise-induced selection has also been investigated by Liu Ziyin et al. (2024); Ziyin et al. (2025). We start from these ideas and results to introduce the key mathematical objects needed to formulate and address the problem of inverse design of implicit bias.

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.

Contributions. Our first and foremost contribution is a constructive approach to the inverse design of implicit biases under stochastic training, using holomorphic functions. The Cauchy–Riemann relations provide a simple way to construct predictor-preserving symmetries and their associated charges. At the critical points of the holomorphic function, the symmetry generator vanishes, so the corresponding parameter values remain unchanged under the symmetry transformation. Under suitable conditions, we can enforce these points to map to targeted predictor values.

Our second contribution connects geometry to stochastic training dynamics. We derive a finite-step update for the charge and establish a sufficient condition on the charge second moment that ensures its convergence to a prescribed level under stochastic training.

Our third contribution is an explicit calculation of the bias on the selected charge level, under the assumption that the predictor evolves on a much faster timescale than the charge.

Finally, our simulations provide a controlled test for the proposed mechanism. We introduce a novel parametrization that biases learned weights toward the binary values $\{-1, +1\}$. The performance of the new model is on a par with that of a tuned, explicitly regularized baseline.

Together, these contributions provide a framework for the inverse design of implicit bias, with the potential to shape which solutions models learn and, consequently, their generalization properties.

2 Related Work

Implicit bias and optimization. Foundational results connect gradient descent to maximum-margin classification on separable data (Soudry et al., 2018), relate solution selection to optimization geometry (Gunasekar et al., 2018), and characterize function-space margins for gradient flow in infinitely wide two-layer networks with homogeneous activations (Chizat & Bach, 2020). More recent work extends these analyses to logistic regression at the edge of stability (Wu et al., 2023) and to broader families of multi-class losses (Ravi et al., 2024). For non-homogeneous networks, Cai et al. (2025) establish convergence under an exponential loss to a margin-maximization problem.

Parametrizations and implicit bias. Mirror descent provides a way to control implicit bias through the choice of mirror potential (Sun et al., 2023). Related work connects continuous mirror descent to gradient flow under suitable reparametrizations (Amid & Warmuth, 2020), including a characterization for commuting parametrizations (Li et al., 2022b). These results provide a complementary route to controlling implicit bias through optimization geometry. Additional related work characterizes conservation laws of gradient flows using the Lie algebra generated by the model's Jacobian fields (Marcotte et al., 2023).

Geometry and stochasticity. The structure of gradient noise during training can shape the solutions selected by stochastic gradient descent (SGD). Zhu et al. (2019) relate escape from sharp minima to the alignment between noise covariance and loss curvature, while Xie et al. (2021) develop a diffusion theory of flat-minimum selection. Wu et al. (2022) connect noise alignment to linear stability and bounds on the sharpness of accessible minima in square-loss models. Additional work describes landscape-dependent noise (Feng & Tu, 2021) and effective regularization favouring flat solutions (Yang et al., 2023). Recent work also examines the relation between noise covariance and the Hessian (Zhang et al., 2026). More broadly, noise covariance shapes implicit bias (HaoChen et al., 2021; Li et al., 2022a). Label noise also induces implicit regularization favouring flatter solutions (Blanc et al., 2020; Damian et al., 2021). For diagonal and quadratic parametrizations (called *Hadamard parametrizations*), stochasticity can promote sparse solutions (Pesme et al., 2021; Pillaud-Vivien et al., 2022), while finite-step effects also alter the bias (Even et al., 2023). Parameter symmetries provide another perspective on stochastic selection through noise-induced equilibration and noise balance (Liu Ziyin et al., 2024; Ziyin et al., 2025). We adopt this latter perspective in our paper. Related work derives an entropy formula from group-orbit volumes in deep linear networks (Menon & Yu, 2025). For Langevin dynamics with projected noise, Menon et al. (2026) identify an implicit regularization effect governed by orbit volume.

Singular learning theory. Singular learning theory studies how model singularities affect Bayesian evidence and generalization (Watanabe, 2009; 2024). Our construction also involves singularities: the joint predictor–charge Jacobian loses rank at designed critical points.

3 METHODOLOGY

We begin with the well-known sparsity-promoting Hadamard parametrization to introduce the main mathematical objects of our framework: the predictor-preserving symmetry, its associated charge, and the joint predictor–charge map. We examine how discrete stochastic training affects the charge and induces an implicit bias. We derive the finite-step charge dynamics and give a sufficient condition on the second charge moment for concentration on the zero-charge level. Under an explicit adiabatic assumption, discussed below, this concentration selects a parameter set. The geometry of this set determines the implicit bias, which we compute using the coarea formula (Federer, 1969). This formula combines a change of variables with integration over level sets, allowing us to express parameter volume in terms of predictor and charge coordinates. The subsequent sections reverse the construction and use holomorphic functions to jointly design this charge and the predictor to favour targeted solutions.

3.1 HADAMARD PARAMETRIZATION

Let Θ denote the parameter space and let $h : \Theta \to \mathcal{W}$ be the predictor map. Denote the predictor by $w = h(\theta)$. We consider the Hadamard parametrization: each predictor is the product $w = u \odot v$, where $w_i = u_i v_i$. Since the parametrization acts coordinatewise, we begin with the scalar predictor $w = h(u, v) = uv$. For $w \neq 0$, the set of parameters that gives the same predictor is $h^{-1}(w) = \{(u, v) \in \mathbb{R}^2 : uv = w\}$, i.e., a hyperbola.

Symmetry and charge. The transformations $Q_\tau(u, v) = (e^\tau u, e^{-\tau} v)$ form a one-parameter Lie group that preserves h. Its infinitesimal generator $\xi(u, v) = \partial_\tau Q_\tau(u, v)|_{\tau=0} = (u, -v)$ is tangent to each element in the set $h^{-1}(w)$. We suppose that the generator is the gradient of a scalar function, the charge χ. This condition will be satisfied by definition in the holomorphic construction of Section 3.2. A simple calculation shows:

$$\chi(u, v) = \frac{u^2 - v^2}{2}, \qquad \xi(u, v) = \nabla\chi(u, v), \qquad [\nabla h \cdot \nabla\chi](u, v) = 0. \tag{1}$$

Thus predictor and charge level sets intersect orthogonally (Figure 1). In particular, for any batch $\mathcal{B}$, gradient flow on the sampled loss $L = \ell_{\mathcal{B}}(h)$ conserves χ, since $\dot\chi = -\ell'_{\mathcal{B}}(h)\nabla\chi \cdot \nabla h = 0$ (Kunin et al., 2021; Tanaka & Kunin, 2021).

In a discrete setting, the charge is no longer conserved but becomes a dynamical quantity. Let $g_{\mathcal{B}} \equiv g_{\mathcal{B}_k}(w) = \partial_w \ell_{\mathcal{B}_k}(w)$ denote the loss gradient at w for an independently sampled batch $\mathcal{B}$ and $\theta = (u, v)$, where k is the step number. A discrete SGD step is written as $\theta_{k+1} = \theta_k - \eta g_{\mathcal{B}} \nabla h(\theta_k)$. A Taylor expansion of the charge around the kth-iterate gives

$$\chi(\theta_{k+1}) - \chi(\theta_k) = \frac{\eta^2 g_{\mathcal{B}}^2}{2} \nabla h^\mathsf{T} \nabla^2 \chi \nabla h + O(\eta^3). \tag{2}$$

The first-order term vanishes by symmetry since $\nabla\chi^\mathsf{T}\nabla h = 0$, and, in the Hadamard case, there is no $O(\eta^3)$ term as χ is quadratic. Hence

$$\chi_{k+1} = (1 - \eta^2 g_{\mathcal{B}}^2)\chi_k. \tag{3}$$

Thus, the charge dynamics appears at order η^2. The exact parameter-level derivation and the resulting second-moment recursion are given in Appendix B.

In order to calculate the implicit bias, we first have to determine where the charge concentrates. To do so, we analyse the dynamics of the charge second moment. Fix the predictor value w and define the jth-moment $\mu_j(w) := \mathbb{E}_{\mathcal{B}}[g_{\mathcal{B}}(w)^j]$ and the conditional charge law $\nu_w(\cdot, k)$. Then the charge second moment is

$$M_{2,w}(k) = \int_{\mathbb{R}} \chi^2 \, \nu_w(\mathrm{d}\chi, k) = \mathbb{E}[\chi_k^2 \mid w]. \tag{4}$$

In the Hadamard case, since the exact identity in equation (3) holds for every sampled batch, equation (4) obeys a closed recursion. Squaring equation (3) and taking the expectation over the sampled batch gives

$$M_{2,w}(k + 1) = \rho(w)\, M_{2,w}(k). \tag{5}$$

where

$$\rho(w) := 1 - 2\eta^2\mu_2(w) + \eta^4\mu_4(w) = \mathbb{E}_{\mathcal{B}}\big[(1-\eta^2 g_{\mathcal{B}}(w)^2)^2\big] \tag{6}$$

Therefore, if $\rho(w) < 1$, then

$$M_{2,w}(k) = \rho(w)^k M_{2,w}(0) \to 0, \qquad \text{when} \quad \eta^2\mu_4(w) < 2\mu_2(w). \tag{7}$$

Consequently, $\nu_w(\cdot, k) \Rightarrow \delta(\chi)$, i.e. the conditional charge law concentrates at $\chi = 0$.
Appendix B derives this recursion and proves convergence of the conditional charge law. It also shows that, when $\mu_4(w) > 0$, concentration around any centre $\chi_\star(w)$ necessarily implies $\chi_\star(w) = 0$ for the Hadamard case. Thus, the Hadamard dynamics selects the balanced level $\chi = 0$, where $u^2 = v^2$. For a general parametrization, equation (4) need not yield a closed recursion for $M_{2,w}(k)$, whose convergence must therefore be established case by case.

Adiabatic description and implicit bias. As discussed above, a discrete SGD step changes the predictor at order $O(\eta)$, whereas the charge changes at order $O(\eta^2)$ (see Appendix B). This motivates an adiabatic assumption: the charge evolves on a much slower timescale than the predictor.

For the Hadamard parametrization, condition (7) implies that the charge concentrates at the zero-charge level. The selected support is

$$\mathcal{S}_w = \{\theta : h(\theta) = w,\ \chi(\theta) = 0\}. \tag{8}$$

Any parameter law at fixed w, if it exists, is therefore supported on (8). Figure 1 illustrates the selection at fixed $w \neq 0$. The predictor level set $h = w$ (blue curves) intersects the selected charge level $\chi = 0$ (solid orange lines) at exactly two points. These points form $\mathcal{S}_w$. The other charge levels (dashed orange curves) illustrate what is excluded dynamically as the charge concentrates at zero level. We use the coarea formula (Federer, 1969) to quantify the parameter volume associated with each predictor value w on the selected charge level $\chi = 0$. This weights each parameter representation by the inverse Jacobian determinant of the joint map $\Psi(\theta) = (h(\theta), \chi(\theta))$. Selecting the charge level is expressed by the distributional measure $\delta(\chi(\theta))\,\mathrm{d}\theta$. For a test function f supported at regular predictor values, the coarea identity gives

$$\int_\Theta f(h(\theta))\delta(\chi(\theta))\,\mathrm{d}\theta = \int_{\mathbb{R}} f(w)\Gamma_0(w)\,\mathrm{d}w. \tag{9}$$

The resulting weight is

$$\Gamma_0(w) = \sum_{\theta\in\mathcal{S}_w} \frac{1}{|\det D\Psi(\theta)|}, \tag{10}$$

for regular $(w, 0)$. As mentioned above, coarea assigns each selected point the inverse-Jacobian weight appearing in equation (10). We interpret the corresponding quantity as a form of *implicit bias*

$$L_{\mathrm{IB}}(w) = -\log\Gamma_0(w). \tag{11}$$

Indeed, for equal loss values, a larger $\Gamma_0(w)$ represents more parameter volume on the selected charge level and hence a geometric preference for w. Full details are given in Appendix C.

Hadamard implicit bias. For the Hadamard parametrization $\chi(u, v) = \frac{1}{2}(u^2 - v^2)$. Fix $w \neq 0$. Charge contraction selects $\chi = 0$, so the selected parameter values satisfy $uv = w$ and $u^2 = v^2$. There are exactly two solutions to those constraints. The joint Jacobian is

$$D\Psi(u, v) = \begin{pmatrix} v & u \\ u & -v \end{pmatrix}, \qquad |\det D\Psi(u, v)| = u^2 + v^2. \tag{12}$$

On the selected level $u^2 + v^2 = 2|w|$. Equation (10) contains therefore the sum of two inverse Jacobians, each equal to $1/(2|w|)$:

$$\Gamma_0(w) = \frac{1}{|w|}. \tag{13}$$

The reasoning can be easily extended to multiple coordinates, leading to

$$\Gamma_0(w) = \prod_{i=1}^{d} |w_i|^{-1}, \qquad L_{\mathrm{IB}}(w) = \sum_{i=1}^{d} \log|w_i|. \tag{14}$$

This example links a predictor-preserving symmetry, a dynamically selected charge level, and implicit bias. We now see how these mathematical objects can be inverse designed using holomorphic functions to target specific Γ_0 .

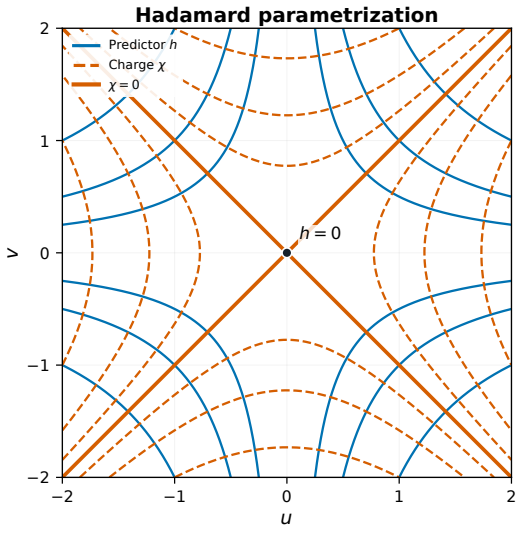

Figure 1: **Geometry of the Hadamard parametrization.** Predictor level sets $uv = \text{const}$ (blue) and charge levels $\chi = (u^2 - v^2)/2 = \text{const}$ (dashed orange) intersect orthogonally away from the origin. The zero-charge set consists of the balanced lines $u = \pm v$ (solid orange). The origin is the unique fixed point of the symmetry generator $\nabla\chi = (u, -v)$.

3.2 Inverse design of the implicit bias by holomorphic functions

Given a finite collection of target predictor values, we seek a parametrization whose induced implicit bias favours them. Holomorphic functions provide such a construction, since their real and imaginary parts have orthogonal gradients, $\nabla\chi^\top \nabla h = 0$. We can therefore identify the former with the charge and the latter with the predictor. Importantly, as shown below, by designing the critical points of the holomorphic function and their associated predictor values, we can shape the bias toward the prescribed targets.

From holomorphic functions to symmetries, charges, and the implicit bias. Let the function

$$F(\zeta) = A(u, v) + \mathrm{i}B(u, v), \qquad \zeta = u + \mathrm{i}v \tag{15}$$

be holomorphic. We define the charge and predictor by $\chi(u, v) = A(u, v)$ and $h(u, v) = B(u, v)$. The Cauchy–Riemann relations, $\partial_u A = \partial_v B$ and $\partial_v A = -\partial_u B$, imply

$$\nabla\chi^\mathsf{T}\nabla h = 0. \tag{16}$$

Consequently, $\nabla\chi$ generates a local predictor-preserving transformation. It vanishes exactly at the critical points $F'(\zeta) = 0$. Crucially, note that

$$|F'(\zeta)|^2 = (\partial_u A)^2 + (\partial_v A)^2 = \|\nabla\chi\|_2^2 = |\det D\Psi| \tag{17}$$

where the Jacobian is that of the joint map $\Psi = (h, \chi)$. Thus, each regular representative on $\chi = 0$ contributes $1/|F'(\zeta)|^2$ to $\Gamma_0(w)$ in equation (10). By placing critical points $F'(\zeta_j) = 0$ on the dynamically selected charge level and their associated predictor values through h, we can design where the induced implicit bias is concentrated. Appendix A gives the full calculation.

Thus, the final step in the inverse design is to construct a holomorphic function with prescribed critical points. For a target polynomial holomorphic function $F = P$, we prescribe its critical points $\zeta_0, \ldots, \zeta_{K-1}$ by inverse design the zeros of its derivative

$$P'(\zeta) = \alpha \prod_{j=0}^{K-1} (\zeta - \zeta_j), \qquad \alpha \neq 0 \tag{18}$$

and then integrating to obtain P.

As an example, we recover the well-known sparsity-promoting Hadamard parametrization within this framework. To favour zero predictors, the minimal choice is $P'(\zeta) = \zeta$, which integrates to $P(\zeta) = \zeta^2/2$. Taking its real and imaginary parts gives $\chi(u, v) = (u^2 - v^2)/2$ and $h(u, v) = uv$, recovering equations (1).

3.3 Inverse design: binary quantization

For the target dictionary $w_\star = \{-1, +1\}$, we seek a polynomial P whose critical points $\zeta_\pm$ lie on the zero-charge level and map to the predictor values $w = \pm 1$. Following equation (15) we define $\chi = \operatorname{Re} P$ and $h = \operatorname{Im} P/2$. These requirements amount to

$$P'(\zeta_\pm) = 0, \qquad P(\zeta_\pm) = \pm 2\mathrm{i}. \tag{19}$$

The factor $1/|P'(\zeta)|^2$ then produces the desired singular bias at the target values, as follows from equations (10) and (17). To construct such a polynomial, we choose the critical points at $\zeta_\pm = \pm\mathrm{i}$ and take the simplest derivative with these zeros,

$$P'(\zeta) = a(\zeta - \mathrm{i})(\zeta + \mathrm{i}). \tag{20}$$

Integrating with zero constant gives $P(\zeta) = a(\zeta^3/3 + \zeta)$, so that $P(\pm\mathrm{i}) = \pm 2a\mathrm{i}/3$. The prescribed critical values therefore fix $a = 3$, yielding

$$P'(\zeta) = 3(\zeta - \mathrm{i})(\zeta + \mathrm{i}), \qquad P(\zeta) = \zeta^3 + 3\zeta. \tag{21}$$

Taking the real and imaginary parts of P yields the wanted parametrization h and charge χ:

$$\chi(u, v) = u(u^2 - 3v^2 + 3), \qquad h(u, v) = \frac{1}{2}v(3u^2 - v^2 + 3). \tag{22}$$

Figure 2 (a) illustrates the corresponding predictor–charge geometry and the critical representatives $(0, \pm 1)$. More generally, imposing simultaneously $\chi = 0$ and $h = w = \pm 1$ is equivalent to $P(\zeta) = 2\mathrm{i}w$. This specializes the set in equation (8) to this parametrization: $\mathcal{S}_w = \{\zeta : P(\zeta) = 2\mathrm{i}w\}$. Using equation (10), where $|\det D\Psi(\chi, h)| = \frac{1}{2}|P'(\zeta)|^2$, we have:

$$\Gamma_0(w) = \sum_{\zeta:\, P(\zeta)=2\mathrm{i}w} \frac{2}{|P'(\zeta)|^2}, \qquad w \notin \{-1, +1\}. \tag{23}$$

To characterize the implicit bias near each target value $w_\star \in \{-1, +1\}$, we expand P around the corresponding critical point $\zeta = \mathrm{i}w_\star$. This yields

$$\Gamma_0(w) \sim \frac{1}{6|w - w_\star|}, \qquad L_{\mathrm{IB}}(w) = \log|w - w_\star| + O(1) \quad \text{as } w \to w_\star. \tag{24}$$

Appendix D gives the full derivation.

Discrete dynamics and branch attraction. The calculations above complete the geometric part of the construction. The parametrization was chosen so that the critical points $(0, \pm 1)$ make the Jacobian vanish, producing the singular behaviour of $\Gamma_0(w)$ as $w \to \pm 1$. Thus the implicit bias identifies the prescribed binary values as the geometrically preferred predictors. It remains to see how the discrete stochastic dynamics actually drives the parameters toward the zero-charge branch containing these critical representatives. Writing $g_{\mathcal{B}}$ for the sampled loss derivative, a simultaneous SGD step gives

$$\begin{aligned} u_{k+1} &= (1 - 3\eta v_k g_{\mathcal{B}})u_k, \\ v_{k+1} &= v_k - \frac{3\eta g_{\mathcal{B}}}{2}(u_k^2 - v_k^2 + 1). \end{aligned} \tag{25}$$

Under the sufficient negative-logarithmic-drift condition derived in Appendix D, the multiplicative update drives $u_k \to 0$ and, for bounded v_k, also gives $\chi_k \to 0$. Hence the dynamics approaches the invariant branch $u = 0$. On this branch, the remaining v-update vanishes at $v = \pm 1$, so the corresponding critical representatives preserve the binary predictor values $h = \pm 1$ under SGD (see Appendix D).

4 Experiments

In the following, we test whether the proposed parametrization reproduces the preference for binary weights induced by explicit regularization, without adding a penalty to the loss. We also track charge evolution to assess the proposed selection mechanism.

4.1 Recovering binary Ising interactions from spin energies

We apply a model with the binary parametrization described in Section 3.3 to the problem of recovering signed pairwise interactions in an Ising model (Ising, 1925). The model is linear in the unknown interactions and provides a controlled setting in which to examine charge evolution, concentration near prescribed predictor values, and reconstruction accuracy. We also measure the ratio of charge to predictor update norms as a diagnostic of the timescale separation underlying the adiabatic assumption. Our aim is to test how the proposed parametrization biases optimization towards binary interactions without an explicit penalty, and to compare its reconstruction performance with that of a tuned, explicitly regularized model. We also include an unregularized vanilla model as a baseline.

Model, data and measurements. We consider a classical Ising model with energy

$$E(s) = \sum_{i<j} J^*_{ij} s_i s_j = \frac{1}{2} s^\mathsf{T} J^* s,$$

where $s \in \{-1, +1\}^D$ and the interaction matrix $J^* \in \mathbb{R}^{D\times D}$ is symmetric with zero diagonal and independent, uniformly sampled couplings $J^*_{ij} \in \{-1, +1\}$ for $i < j$. The factor $1/2$ ensures that each interaction is counted once. Positive and negative couplings are therefore equally probable, without enforcing an exact balance. Our goal is to recover these couplings from a limited number of energy observations. We use $D = 60$, giving $d = D(D-1)/2 = 1770$ independent unknowns. For independently sampled uniform spin configurations s_μ, we observe the noiseless normalized energies

$$E_\mu = \frac{1}{\sqrt{d}} \sum_{i<j} J^*_{ij} s_{\mu i} s_{\mu j} = \frac{s_\mu^\mathsf{T} J^* s_\mu}{2\sqrt{d}}, \qquad \mu = 1, \dots, N. \tag{26}$$

These are measurements at random spin configurations, not samples from an unknown equilibrium distribution. Writing the ground-truth weights as $w^* = (J^*_{ij})_{i<j} \in \mathbb{R}^d$, we construct $X \in \mathbb{R}^{N\times d}$ with one column for each pair $i < j$:

$$X_{\mu,(i,j)} = \frac{s_{\mu i} s_{\mu j}}{\sqrt{d}}.$$

Thus, each row X_μ contains d features and $E_\mu = X_\mu w^*$. We use $N = 1062$ observations for training and evaluate each model on 2048 additional independent spin configurations with the same interaction matrix.

Parametrized, regularized and vanilla training. The parametrized model applies the map h in (22) independently to each coupling, setting $\widehat{J}_{ij} = h(u_{ij}, v_{ij})$ for $i < j$ and enforcing symmetry and a zero diagonal. An explicitly regularized baseline optimizes these couplings directly with a double-well penalty that favours the same binary values. Finally, a vanilla baseline optimizes couplings directly, without any penalty. For minibatch $\mathcal{B}$, the objectives are

$$\begin{aligned} L^{\mathrm{param}}(u, v) &= \frac{1}{2|\mathcal{B}|} \sum_{\mu\in\mathcal{B}} \big(X_\mu h(u, v) - E_\mu - \xi^{\mathrm{param}}_{\mu,k}\big)^2, \\ L^{\mathrm{reg}}(w) &= \frac{1}{2|\mathcal{B}|} \sum_{\mu\in\mathcal{B}} \big(X_\mu w - E_\mu\big)^2 + \frac{\lambda}{d} \sum_{e=1}^{d} (w_e^2 - 1)^2, \\ L^{\mathrm{van}}(w) &= \frac{1}{2|\mathcal{B}|} \sum_{\mu\in\mathcal{B}} \big(X_\mu w - E_\mu\big)^2. \end{aligned} \tag{27}$$

Here, $h(u, v)$ is applied component-wise. For the parametrized model, independently resampled Gaussian noise is added to the energy targets at each training iteration, with $\xi^{\mathrm{param}}_{\mu,k} \sim \mathcal{N}(0, \sigma^2_{\mathrm{param}})$. This training noise sustains stochastic fluctuations near interpolation and is intended to maintain the finite-step stochastic mechanism active underlying the proposed charge-concentration mechanism (see Figure 2, (e)). Neither baseline uses added target noise. We set the batch size to $|\mathcal{B}| = 1$ for all three models and sample training indices uniformly with replacement at each iteration. This choice yields good convergence while maximizing minibatch sampling noise.

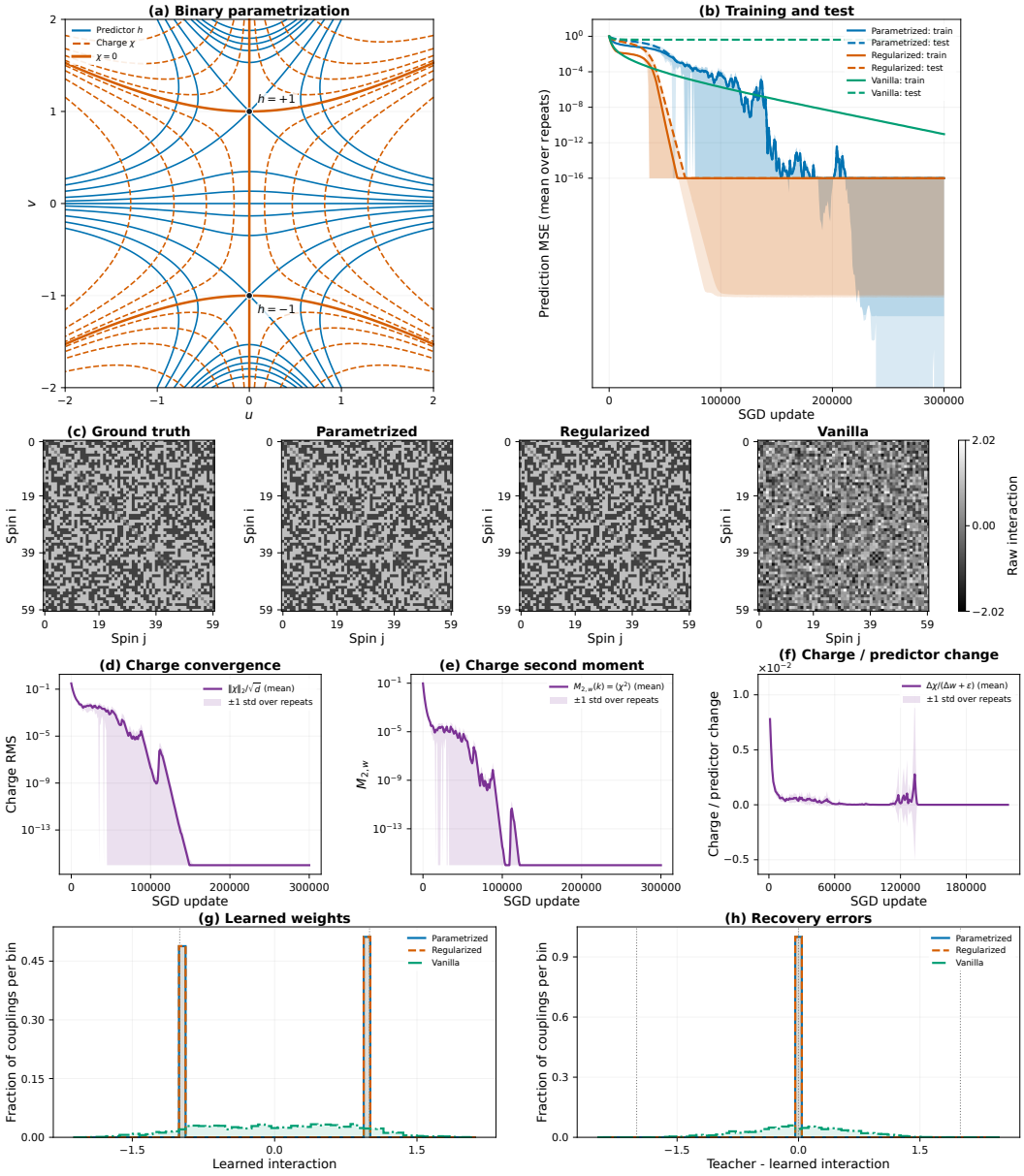

Figure 2: **Binary Ising interaction recovery with three independently tuned methods. Top row: (a)** Predictor level sets (blue), charge level sets (dashed orange), and the zero-charge set (thick orange); the marked critical points yield $h = \pm 1$. **(b)** Prediction MSE for training (solid) and test (dashed) for the parametrized (blue), regularized (orange), and vanilla (green) models. **Second row: (c)** Ground-truth, parametrized, regularized, and vanilla interaction matrices from the first run, using a shared colour scale (all runs gave qualitatively similar results). **Third row: (d)** Charge RMS $\|\boldsymbol{\chi}\|_2/\sqrt{d}$ for the parametrized model. **(e)** Charge second moment $M_{2,w} = \langle \boldsymbol{\chi}^2 \rangle$ for the parametrized model. **(f)** Charge-to-predictor change ratio for the parametrized model. **Bottom row: (g)** Distributions of raw learned couplings for the same run as in (c). **(h)** Distributions of true-minus-learned coupling errors for the same run as in (c).

Each model is trained for 3×10^5 iterations using SGD. For each coupling, we initialize $u_{ij} \sim \mathcal{N}(0.1, 0.02^2)$ independently and set $v_{ij} = 0$, giving $h(u_{ij}, v_{ij}) = 0$. Both baseline models start from the same effective weights.

We select hyperparameters for each method by minimizing the final coupling reconstruction error $D_1 = \sum_{i<j} |\widehat{J}_{ij} - J^*_{ij}|$ on a single initial run. For the regularized model, we search $\lambda \in \{0.01, 0.1, 0.3, 0.6, 1\}$ and $\eta \in \{0.1, 0.9, 1.2\}$. For the parametrized model, we search $\eta \in \{0.1, 0.6, 0.9\}$ and $\sigma_{\text{param}} \in \{0.6, 0.7, 0.8\}$. For vanilla regression, we search $\eta \in \{0.1, 0.6, 0.9, 1.2\}$. The search grids were chosen through preliminary experiments covering hyperparameters regimes with low D_1. Candidates reuse the same initialization and random seeds for sampling and training noise. The selected settings are $(\eta, \lambda) = (1.2, 0.1)$ for the regularized model, $(\eta, \sigma_{\text{param}}) = (0.6, 0.6)$ for the parametrized model, and $\eta = 0.6$ for vanilla regression. We then repeat the training procedure with these fixed selected settings across ten independent runs, varying the random seeds for stochastic training, while preserving the true interaction matrix, training data, and initialization.

Results. Across the ten runs, the parametrized and regularized models achieve coupling reconstruction errors of $D_1 = (3.8 \pm 2.1) \times 10^{-16}$ and $D_1 = (2.4 \pm 0.1) \times 10^{-12}$, respectively (mean $\pm$ one sample standard deviation). Vanilla regression instead achieves $D_1 = 892.96$ with a standard deviation below 0.01. The parametrized and regularized models therefore recover the couplings much more accurately than vanilla regression, and achieve near-exact numerical recovery across all runs.

The learning curves and coupling distributions in Figure 2 (b) and (g), respectively, compare prediction accuracy and concentration near the binary targets of the learned weights. The inverse-designed parametrization produces a preference for binary values without an explicit penalty, whereas the regularized baseline imposes this preference through its double-well penalty.
For the parametrized model, we track the charge RMS, $\|\boldsymbol{\chi}_k\|_2/\sqrt{d}$. We compare charge and predictor updates using $R = \langle \|\Delta\boldsymbol{\chi}\|_2 \rangle / (\langle \|\Delta w\|_2 \rangle + \varepsilon)$, with $\varepsilon = 10^{-12}$, where the brackets denote averages over non-overlapping windows of 1000 SGD updates.

The underlying parameter trajectories are shown in Figure 3 in Appendix D: u approaches zero while individual v-coordinates approach either $+1$ or -1, consistent with the branch-attraction and critical-representative dynamics derived in Appendix D.

Curves and histograms in Figure 2 show means and bands showing one sample standard deviation across the ten independent runs. The models share the same true interaction matrix, data, and effective initialization across all runs, varying only the stochastic training noise. The displayed interaction matrices refer to the first run.

5 Conclusions and Future Perspectives

We introduced a constructive framework for designing implicit biases through holomorphic parametrizations. The Cauchy–Riemann relations connect predictor-preserving symmetries to conserved charges, while the coarea formula characterizes the implicit bias induced by the geometry of the charge level set selected by training. This framework recovers the known sparsity-promoting Hadamard parametrization and provides the construction of a novel parametrization that favours the learning of binary weights. In a controlled numerical experiment, the latter reproduces the effect of an explicit regularizer towards weight binarization without adding a penalty to the loss, thus supporting the proposed mechanism for the inverse design of a chosen implicit bias.

A theoretical direction for future work is to move beyond constructions that parametrize each predictor coordinate separately. Allowing parameters to interact across coordinates could support richer symmetry structures with multiple associated charges. Developing such an extension requires identifying predictor-preserving generators that guarantee closure relations under Lie brackets. Further work could explore larger target dictionaries and more complex models or datasets, potentially going beyond holomorphic parametrizations, and using such theory to guide model design and training. This work lays the theoretical foundations for such developments by showing how implicit bias can be designed to shape which solutions models learn, with potential implications for the control of learned representations and ultimately generalization.

Reproducibility Statement

Section 3 defines the constructions and theoretical assumptions; Appendices A, B, C, and D provide detailed derivations and proofs. Section 4.1 specifies the synthetic data generation, model parametrizations, training objectives, initialization, hyperparameter selection, and evaluation protocol. The experimental code is available in anonymized form at `https://github.com/anonymous-submittion/inverse-design`.

AI Use Statement

We used generative AI tools to assist with the following aspects of this work.

Writing and organization: Revising English grammar, improving clarity and concision, and organizing the presentation of the theoretical framework and experiments.

Literature and references. Locating and checking related work on implicit bias, parameter symmetries, stochastic optimization, and geometric approaches to learning, including checking bibliographic details and the correspondence between cited results and our descriptions.

Experimental implementation. Refining code for the Ising interaction-recovery experiments, including synthetic data generation, model training, evaluation, and visualization.

All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final manuscript, including its claims, proofs, code, and experimental results.

References

Ehsan Amid and Manfred K. Warmuth. Reparameterizing mirror descent as gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8430–8439, 2020.

Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an Ornstein–Uhlenbeck like process. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 483–513. PMLR, 2020.

Yuhang Cai, Kangjie Zhou, Jingfeng Wu, Song Mei, Michael Lindsey, and Peter L. Bartlett. Implicit bias of gradient descent for non-homogeneous deep networks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6426–6504. PMLR, 2025.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on learning theory*, pp. 1305–1338. PMLR, 2020.

Alex Damian, Tengyu Ma, and Jason D. Lee. Label noise SGD provably prefers flat global minimizers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 27449–27461, 2021.

Mathieu Even, Scott Pesme, Suriya Gunasekar, and Nicolas Flammarion. (s) gd over diagonal linear networks: Implicit bias, large stepsizes and edge of stability. *Advances in Neural Information Processing Systems*, 36:29406–29448, 2023.

Herbert Federer. *Geometric Measure Theory*. Springer, Berlin, 1969.

Yu Feng and Yuhai Tu. The inverse variance-flatness relation in stochastic gradient descent is critical for finding flat minima. *Proceedings of the National Academy of Sciences*, 118(9):e2015617118, 2021. doi: 10.1073/pnas.2015617118.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pp. 1832–1841. PMLR, 2018.

Jeff Z. HaoChen, Colin Wei, Jason Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. In *Proceedings of the Thirty-Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 2315–2357, 2021.

Ernst Ising. Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Physik*, 31(1):253–258, February 1925. ISSN 0044-3328. doi: 10.1007/BF02980577.

Daniel Kunin, Javier Sagastuy-Brena, Surya Ganguli, Daniel LK Yamins, and Hidenori Tanaka. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. In *ICLR*, 2021.

Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after SGD reaches zero loss? – a mathematical framework. In *International Conference on Learning Representations*, 2022a.

Zhiyuan Li, Tianhao Wang, Jason D. Lee, and Sanjeev Arora. Implicit bias of gradient descent on reparametrized models: On equivalence to mirror descent. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34626–34640, 2022b.

Liu Ziyin, Mingze Wang, Hongchao Li, and Lei Wu. Parameter symmetry and noise equilibrium of stochastic gradient descent. In *Advances in Neural Information Processing Systems*, 2024.

Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Abide by the law and follow the flow: Conservation laws for gradient flows. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Govind Menon and Tianmin Yu. An entropy formula for the deep linear network. *arXiv preprint arXiv:2509.09088*, 2025.

Govind Menon, Austin Stromme, and Adrien Vacher. On the implicit regularization of langevin dynamics with projected noise. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 5159–5187. PMLR, 2026.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of SGD for diagonal linear networks: A provable benefit of stochasticity. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29218–29230, 2021.

Loucas Pillaud-Vivien, Julien Reygner, and Nicolas Flammarion. Label noise (stochastic) gradient descent implicitly solves the Lasso for quadratic parametrisation. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of the Thirty-Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 2127–2159. PMLR, 2022.

Hrithik Ravi, Clayton Scott, Daniel Soudry, and Yutong Wang. The implicit bias of gradient descent on separable multiclass data. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

Haoyuan Sun, Khashayar Gatmiry, Kwangjun Ahn, and Navid Azizan. A unified approach to controlling implicit regularization via mirror descent. *Journal of Machine Learning Research*, 24 (393):1–58, 2023.

Hidenori Tanaka and Daniel Kunin. Noether's learning dynamics: Role of symmetry breaking in neural networks. *Advances in Neural Information Processing Systems*, 34:25646–25660, 2021.

Gal Vardi. On the implicit bias in deep-learning algorithms. *Communications of the ACM*, 66(6): 86–93, 2023.

Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, 2009. doi: 10.1017/CBO9780511800474.

Sumio Watanabe. Recent advances in algebraic geometry and Bayesian statistics. *Information Geometry*, 7(Suppl 1):187–209, 2024. doi: 10.1007/s41884-022-00083-9.

Jingfeng Wu, Vladimir Braverman, and Jason D. Lee. Implicit bias of gradient descent for logistic regression at the edge of stability. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Lei Wu, Mingze Wang, and Weijie J. Su. The alignment property of SGD noise and how it helps select flat minima: A stability analysis. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *International Conference on Learning Representations*, 2021.

Ning Yang, Chao Tang, and Yuhai Tu. Stochastic gradient descent introduces an effective landscape-dependent regularization favoring flat solutions. *Physical Review Letters*, 130:237101, 2023. doi: 10.1103/PhysRevLett.130.237101.

Yikuan Zhang, Ning Yang, and Yuhai Tu. On the superlinear relationship between SGD noise covariance and loss landscape curvature, 2026. URL https://arxiv.org/abs/2602.05600.

Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen 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*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7654–7663, 2019.

Liu Ziyin, Hongchao Li, and Masahito Ueda. Noise balance and stationary distribution of stochastic gradient descent. *Physical Review E*, 111(6):065303, 2025.

SUPPLEMENTARY MATERIAL

OVERVIEW

The supplementary material follows the logical construction used in the main text. Appendix A shows why harmonic conjugates (the real and imaginary part of a holomorphic function) generate predictor-preserving symmetries and relates their critical points to the implicit bias. Appendix B then studies the Hadamard parametrization, where the discrete charge dynamics closes exactly and proves concentration on the balanced level $\chi = 0$. Appendix C takes this dynamically selected level as input and derives the corresponding predictor-space implicit bias by coarea formula. Finally, Appendix D applies the same geometric construction to the inverse-designed binary parametrization and separately studies whether discrete SGD is attracted toward the part of the zero-charge set containing the designed critical points.

A HARMONIC CONJUGATES GENERATE LOCAL LIE SYMMETRIES

The purpose of this section is to establish two properties used later in the inverse-design construction. First, the real part of a holomorphic function generates a direction along which its imaginary part is unchanged. Second, the same complex derivative controls the Jacobian entering the implicit bias.

Let $F(\zeta) = A(u, v) + \mathrm{i}B(u, v)$ be holomorphic on an open domain $\Omega \subset \mathbb{C} \simeq \mathbb{R}^2$, where $\zeta = u + \mathrm{i}v$. Define $\chi(u, v) = A(u, v)$, $h(u, v) = B(u, v)$, and $X = \nabla\chi$. The Cauchy–Riemann equations give $\partial_u B(u, v) = -\partial_v A(u, v)$ and $\partial_v B(u, v) = \partial_u A(u, v)$. Therefore

$$\begin{aligned} X[h] &= \nabla A \cdot \nabla B \\ &= (\partial_u A)(\partial_u B) + (\partial_v A)(\partial_v B) \\ &= -(\partial_u A)(\partial_v A) + (\partial_v A)(\partial_u A) = 0. \end{aligned} \tag{28}$$

Since X is smooth, it generates a local flow Φ_t. Along this flow,

$$\frac{\mathrm{d}}{\mathrm{d}t} h(\Phi_t(\theta)) = X[h](\Phi_t(\theta)) = 0.$$

Thus $h \circ \Phi_t = h$ wherever the flow is defined. Flow uniqueness gives $\Phi_0 = \mathrm{id}$, $\Phi_t^{-1} = \Phi_{-t}$, and $\Phi_t \circ \Phi_s = \Phi_{t+s}$ wherever these expressions are defined. Hence X generates a local one-parameter Lie symmetry of the predictor and of every loss depending only on it. Thus the same holomorphic function simultaneously provides the predictor h and a charge χ whose gradient moves along predictor level sets. The next calculation shows why critical points of that function are also distinguished geometrically.

The same derivatives give the coarea Jacobian used in the main text. Indeed,

$$F'(\zeta) = \partial_u A + \mathrm{i}\partial_u B = \partial_u A - \mathrm{i}\partial_v A,$$

and, for $\Psi = (h, \chi) = (B, A)$,

$$\begin{aligned} |\det D\Psi| &= |(\partial_u B)(\partial_v A) - (\partial_v B)(\partial_u A)| \\ &= (\partial_u A)^2 + (\partial_v A)^2 = |F'(\zeta)|^2. \end{aligned} \tag{29}$$

At regular points, each representative therefore contributes $1/|F'(\zeta)|^2$ to the coarea sum in equation (10). If the predictor is rescaled by a constant factor, the joint Jacobian is rescaled by the same factor. This is the only modification needed in the binary example, where $h = \mathrm{Im}\, P/2$.

For a coordinatewise construction on disjoint parameter pairs, the generators X_i preserve every predictor coordinate and satisfy $[X_i, X_j] = 0$. Their commuting local flows define a local action of the abelian Lie group $(\mathbb{R}^d, +)$. Since $|F'(\zeta)|^2 = \|\nabla\chi(\zeta)\|_2^2$, the fixed points of each local flow are precisely the corresponding critical points of F.

B HADAMARD CHARGE DYNAMICS

This section establishes the dynamical selection of the Hadamard zero-charge level. We first derive the exact finite-step evolution of the charge. We then show that, after freezing the predictor according to

the adiabatic approximation, its second moment contracts geometrically and the complete conditional charge law converges to δ_0. Finally, we show that no nonzero deterministic charge value can be a second-moment concentration center.

Set at step k, the predictor $w_k = u_k v_k$, the charge $\chi_k = \frac{1}{2}(u_k^2 - v_k^2)$ and the derivative for a minibatch $\mathcal{B}$, $g_{\mathcal{B}} = g_{\mathcal{B}_k}(w_k)$. One SGD step gives, $u_{k+1} = u_k - \eta g_{\mathcal{B}} v_k$ and $v_{k+1} = v_k - \eta g_{\mathcal{B}} u_k$. For the predictor,

$$\begin{aligned} w_{k+1} &= u_{k+1} v_{k+1} \\ &= (u_k - \eta g_{\mathcal{B}} v_k)(v_k - \eta g_{\mathcal{B}} u_k) \\ &= w_k - \eta g_{\mathcal{B}}(u_k^2 + v_k^2) + \eta^2 g_{\mathcal{B}}^2 w_k. \end{aligned} \tag{30}$$

And for the charge,

$$\begin{aligned} 2\chi_{k+1} &= u_{k+1}^2 - v_{k+1}^2 \\ &= (u_k - \eta g_{\mathcal{B}} v_k)^2 - (v_k - \eta g_{\mathcal{B}} u_k)^2 \\ &= u_k^2 - 2\eta g_{\mathcal{B}} u_k v_k + \eta^2 g_{\mathcal{B}}^2 v_k^2 - v_k^2 + 2\eta g_{\mathcal{B}} u_k v_k - \eta^2 g_{\mathcal{B}}^2 u_k^2 \\ &= u_k^2 - v_k^2 - \eta^2 g_{\mathcal{B}}^2 (u_k^2 - v_k^2) \\ &= 2(1 - \eta^2 g_{\mathcal{B}}^2)\chi_k. \end{aligned} \tag{31}$$

The $O(\eta)$ terms cancel exactly, leaving a purely multiplicative $O(\eta^2)$ change of the charge.

B.1 Charge second moment

The exact SGD process couples w_k and χ_k. Under the adiabatic assumption, we fix w, and let $(\mathcal{B}_k)_{k\geq 0}$ be independently sampled minibatches and define, $A_k(w) = 1 - \eta^2 g_{\mathcal{B}}^2$. The conditional charge dynamics is, $\chi_{k+1} = A_k(w)\chi_k$. Let $\nu_w(\cdot, k)$ be the law of χ_k under this conditioned process. Assume that $\nu_w(\cdot, 0)$ has a finite second moment and that, $\mu_2(w) = \mathbb{E}_{\mathcal{B}}[g_{\mathcal{B}}(w)^2] < \infty$ and $\mu_4(w) = \mathbb{E}_{\mathcal{B}}[g_{\mathcal{B}}(w)^4] < \infty$. Define

$$M_{2,w}(k) = \int_{\mathbb{R}} \chi^2 \nu_w(\mathrm{d}\chi, k) \tag{32}$$

Because $\mathcal{B}_k$ is sampled independently of the charge produced by the preceding minibatches,

$$\begin{aligned} M_{2,w}(k+1) &= \int_{\mathbb{R}} \mathbb{E}_{\mathcal{B}}[A_k(w)^2 \chi_k^2] \nu_w(\mathrm{d}\chi, k) \\ &= \mathbb{E}_{\mathcal{B}}[A_k(w)^2] \int_{\mathbb{R}} \chi^2 \nu_w(\mathrm{d}\chi, k) \\ &= \mathbb{E}_{\mathcal{B}}[(1 - \eta^2 g_{\mathcal{B}}(w)^2)^2] M_{2,w}(k) \end{aligned} \tag{33}$$

By expanding the square we recover equation (5). Iterating the recursion gives,

$$M_{2,w}(k) = \rho(w)^k M_{2,w}(0). \tag{34}$$

where $\rho(w) = 1 - 2\eta^2 \mu_2(w) + \eta^4 \mu_4(w)$, therefore, if $0 \leq \rho(w) < 1$, then $M_{2,w}(k) \to 0$.

B.2 Convergence of the charge law

The second-moment estimate has a direct probabilistic meaning: as k grows, the charge mass outside every fixed neighborhood of zero must vanish. For every $\varepsilon > 0$,

$$\nu_w(\{|\chi| \geq \varepsilon\}, k) \leq \frac{M_{2,w}(k)}{\varepsilon^2} = \frac{\rho(w)^k M_{2,w}(0)}{\varepsilon^2} \longrightarrow 0. \tag{35}$$

Thus the charge converges to zero in probability under the frozen-predictor reduced process. Let $f : \mathbb{R} \to \mathbb{R}$ be bounded and continuous. Given $\varepsilon > 0$, continuity of f at zero provides $\delta > 0$ such that $|f(\chi) - f(0)| < \varepsilon$, whenever $|\chi| < \delta$. Then

$$\begin{aligned} \left| \int_{\mathbb{R}} f(\chi)\, \nu_w(d\chi, k) - f(0) \right| &\leq \int_{\{|\chi|<\delta\}} |f(\chi) - f(0)|\, \nu_w(d\chi, k) \\ &\quad + \int_{\{|\chi|\geq\delta\}} |f(\chi) - f(0)|\, \nu_w(d\chi, k) \\ &\leq \varepsilon + 2\|f\|_\infty \nu_w\left(\{|\chi| \geq \delta\}, k\right). \end{aligned} \tag{36}$$

The second term converges to zero. Since ε is arbitrary, $\int_{\mathbb{R}} f(\chi)\,\nu_w(d\chi, k) \longrightarrow f(0)$. Therefore

$$\nu_w(\cdot, k) \Rightarrow \delta_0. \tag{37}$$

For each fixed w, this identifies zero as the selected charge value. The corresponding selected parameter set is $\mathcal{S}_w = \{\theta : h(\theta) = w,\ \chi(\theta) = 0\}$. The support implication for any parameter law, and the implicit bias assigned to this selected set, are established in Appendix C.

B.3 Concentration center

We have shown that concentration at zero occurs under the contraction condition. We now check whether the same dynamics could instead concentrate around some other deterministic charge value. The following calculation shows that it cannot. For a candidate concentration center $\chi_\star(w)$, define its centered second moment by

$$M_{2,w}(k) := \int_{\mathbb{R}} \big(\chi - \chi_\star(w)\big)^2 \nu_w(\mathrm{d}\chi, k). \tag{38}$$

Thus concentration in second moment at $\chi_\star(w)$ means $M_{2,w}(k) \to 0$. At the fixed predictor w, equation (31) gives

$$\chi_{k+1} - \chi_\star(w) = \big(1 - \eta^2 g_{\mathcal{B}}(w)^2\big)\big(\chi_k - \chi_\star(w)\big) - \eta^2 g_{\mathcal{B}}(w)^2 \chi_\star(w). \tag{39}$$

The new batch is independent of the charge at step k. Squaring this identity and taking expectations therefore yields

$$\begin{aligned} M_{2,w}(k+1) = \rho(w) M_{2,w}(k) \\ &- 2\eta^2 \chi_\star(w)\big(\mu_2(w) - \eta^2 \mu_4(w)\big) \int_{\mathbb{R}} \big(\chi - \chi_\star(w)\big) \nu_w(\mathrm{d}\chi, k) \\ &+ \eta^4 \mu_4(w) \chi_\star(w)^2. \end{aligned} \tag{40}$$

By Cauchy–Schwarz,

$$\left| \int_{\mathbb{R}} \big(\chi - \chi_\star(w)\big) \nu_w(\mathrm{d}\chi, k) \right| \le \sqrt{M_{2,w}(k)}. \tag{41}$$

Consequently, if $M_{2,w}(k) \to 0$, the first two terms on the right of equation (40) vanish. Its left side also vanishes, so

$$0 = \eta^4 \mu_4(w) \chi_\star(w)^2. \tag{42}$$

For $\eta > 0$ and $\mu_4(w) > 0$, any deterministic second-moment concentration center must therefore satisfy $\chi_\star(w) = 0$. At this center, equation (40) reduces to equation (5).

C Implicit bias on the selected charge level

The previous section answered a dynamical question: at fixed predictor w, which charge value survives asymptotically? This section answers a different, geometric question: once the selected value is known to be $\chi = 0$, how much parameter-space weight is associated with each predictor w on that level?

Let $\Theta \subseteq \mathbb{R}^2$, and let $h, \chi : \Theta \to \mathbb{R}$ be continuously differentiable. Work at regular values for which the inverse images below are discrete and their sums are well defined.

The level selected by charge concentration. Under the adiabatic assumption, fix w. Suppose the reduced charge law satisfies

$$M_{2,w}(k) \longrightarrow 0.$$

By equation (37), its limiting charge distribution is δ_0. Hence the compatible parameter values at this fixed predictor form

$$\mathcal{S}_w = \{\theta \in \Theta : h(\theta) = w,\ \chi(\theta) = 0\}.$$

If parameter laws supported on $h^{-1}(w)$ have a weakly convergent subsequence, continuity of h and χ implies that its limit is supported on $\mathcal{S}_w$. This conclusion identifies the selected level, without determining a measure on that level.

Coarea at the fixed predictor. Set $\Psi = (h, \chi)$. On a regular inverse branch, change of variables from $\theta = (u, v)$ to $(w, c) = (h(\theta), \chi(\theta))$ gives

$$\mathrm{d}\theta = \frac{\mathrm{d}w\,\mathrm{d}c}{|\det D\Psi(\theta)|}.$$

Summing over inverse branches and evaluating at the selected charge $c = 0$ yields, for a test function f,

$$\int_\Theta f(\theta)\,\delta(h(\theta) - w)\,\delta(\chi(\theta))\,\mathrm{d}\theta = \sum_{\theta\in\mathcal{S}_w} \frac{f(\theta)}{|\det D\Psi(\theta)|}. \tag{43}$$

The delta factors denote a distribution evaluated through this identity. They are not ordinary functions whose pointwise product is being integrated.

Taking $f = 1$, when the inverse-branch sum is finite, gives the total geometric factor

$$\Gamma_0(w) = \sum_{\theta\in\mathcal{S}_w} \frac{1}{|\det D\Psi(\theta)|}. \tag{44}$$

Thus charge concentration tells us to evaluate coarea at $\chi = 0$, while change of variables supplies the inverse-Jacobian factors. Equation (43) specifies the resulting unnormalized geometric weight at the fixed predictor.

The same calculation can be expressed while allowing w to vary as a coordinate in the change-of-variables formula. For a test function f of the predictor,

$$\int_\Theta f(h(\theta))\delta(\chi(\theta))\,\mathrm{d}\theta = \int_{\mathbb{R}} f(w)\Gamma_0(w)\,\mathrm{d}w, \tag{45}$$

whenever both sides are defined on the regular region under consideration. This identity says that the geometric distribution on the selected charge level has predictor weight $\Gamma_0(w)$. Varying w here is a change-of-variables calculation: the charge dynamics above was analyzed separately at each fixed w.

Where $0 < \Gamma_0(w) < \infty$, the associated geometric bias is $L_{\mathrm{IB}}(w) = -\log\Gamma_0(w)$. Neither this definition nor equation (45) asserts that SGD samples predictors according to $\Gamma_0(w)\,\mathrm{d}w$.

C.1 Hadamard implicit bias

We now evaluate the preceding general formula for the Hadamard parametrization. The dynamics has already supplied $\chi = 0$, so only the Jacobian and the representatives in $\mathcal{S}_w$ remain to be computed. For $h(u, v) = uv$ and $\chi(u, v) = \frac{1}{2}(u^2 - v^2)$, the dynamically selected charge is $\chi_\star = 0$. Thus, $\Psi(u, v) = \left(uv, \frac{u^2-v^2}{2}\right)$, hence

$$D\Psi(u, v) = \begin{pmatrix} v & u \\ u & -v \end{pmatrix}, \qquad |\det D\Psi(u, v)| = u^2 + v^2. \tag{46}$$

Fix $w \neq 0$. Solving $uv = w$ and $\chi = 0$ gives exactly

$$(u, v) = \left(\sqrt{|w|}, \mathrm{sgn}(w)\sqrt{|w|}\right) \quad \text{and} \quad (u, v) = \left(-\sqrt{|w|}, -\mathrm{sgn}(w)\sqrt{|w|}\right).$$

At both values, $|\det D\Psi| = 2|w|$. Equation (43) therefore becomes the explicit distributional identity

$$\begin{aligned}\int_{\mathbb{R}^2} \varphi(u, v)\,\delta(uv - w)\,\delta\left(\frac{u^2 - v^2}{2}\right)\,\mathrm{d}u\,\mathrm{d}v \\ = \frac{\varphi\left(\sqrt{|w|}, \mathrm{sgn}(w)\sqrt{|w|}\right) + \varphi\left(-\sqrt{|w|}, -\mathrm{sgn}(w)\sqrt{|w|}\right)}{2|w|}.\end{aligned} \tag{47}$$

In particular, its total coarea factor is

$$\Gamma_0(w) = \frac{1}{|w|}, \qquad L_{\mathrm{IB}}(w) = \log|w|. \tag{48}$$

For d independent coordinates, the joint Jacobian is block diagonal and each coordinate has two balanced representatives. Consequently, for $w_i \neq 0$ throughout,

$$\Gamma_0(w) = 2^d \prod_{i=1}^{d} \frac{1}{2|w_i|} = \prod_{i=1}^{d} |w_i|^{-1}, \qquad L_{\mathrm{IB}}(w) = \sum_{i=1}^{d} \log |w_i|. \tag{49}$$

Thus, the geometric weight factorizes across coordinates, recovering the sparsity-promoting logarithmic bias of the Hadamard parametrization.

D BINARY CONSTRUCTION: GEOMETRIC BIAS AND CHARGE DYNAMICS

This appendix derives the binary construction introduced in Section 3.3. The argument has two parts. First, we inverse-design a holomorphic polynomial whose critical points represent the targeted binary predictors and compute its implicit bias. Second, we study the exact discrete SGD dynamics and determine under which conditions the parameters approach the component of the zero-charge set containing those critical points. These two statements are kept separate: the coarea calculation determines the geometric preference on the selected charge level, whereas the discrete dynamics determines whether training approaches that level.

Throughout, the target dictionary is $w_\star = \{-1, +1\}$.

D.1 INVERSE-DESIGNED POLYNOMIAL AND ITS CRITICAL REPRESENTATIVES

We seek a holomorphic polynomial P for which the charge and predictor are defined by

$$\chi(u, v) = \operatorname{Re} P(\zeta), \qquad h(u, v) = \frac{1}{2} \operatorname{Im} P(\zeta), \qquad \zeta = u + \mathrm{i}v, \tag{50}$$

and whose critical points represent the target values $w_\star = \pm 1$.

For a critical point ζ, the required conditions are

$$P'(\zeta) = 0, \qquad \operatorname{Re} P(\zeta) = 0, \qquad \frac{1}{2} \operatorname{Im} P(\zeta) = w_\star. \tag{51}$$

Equivalently,

$$P'(\zeta) = 0, \qquad P(\zeta) = 2\mathrm{i}w_\star. \tag{52}$$

We choose the two critical locations symmetrically at $\zeta = \mathrm{i}w_\star$ for $w_\star \in \{-1, +1\}$. The polynomial derivative of smallest degree with zeros at $\zeta = \pm\mathrm{i}$ is

$$P'(\zeta) = a(\zeta - \mathrm{i})(\zeta + \mathrm{i}) = a(\zeta^2 + 1), \tag{53}$$

where $a \neq 0$. Integrating and choosing the additive constant equal to zero gives

$$P(\zeta) = a\left(\frac{\zeta^3}{3} + \zeta\right). \tag{54}$$

At the two critical points,

$$P(\mathrm{i}) = \frac{2a}{3}\mathrm{i}, \qquad P(-\mathrm{i}) = -\frac{2a}{3}\mathrm{i}. \tag{55}$$

Condition (52) therefore fixes $a = 3$, and hence

$$P'(\zeta) = 3(\zeta - \mathrm{i})(\zeta + \mathrm{i}), \qquad P(\zeta) = \zeta^3 + 3\zeta. \tag{56}$$

Expanding P at $\zeta = u + \mathrm{i}v$ gives

$$\begin{aligned} P(u + \mathrm{i}v) &= (u + \mathrm{i}v)^3 + 3(u + \mathrm{i}v) \\ &= u^3 - 3uv^2 + 3u + \mathrm{i}\left(3u^2 v - v^3 + 3v\right). \end{aligned} \tag{57}$$

Therefore

$$\chi(u, v) = u(u^2 - 3v^2 + 3), \qquad h(u, v) = \frac{1}{2} v(3u^2 - v^2 + 3). \tag{58}$$

which recover equation (22). Differentiating gives

$$\nabla h = \begin{pmatrix} 3uv \\ \frac{3}{2}(u^2 - v^2 + 1) \end{pmatrix}, \qquad \nabla \chi = \begin{pmatrix} 3(u^2 - v^2 + 1) \\ -6uv \end{pmatrix}. \tag{59}$$

Their inner product vanishes identically:

$$\nabla \chi^{\mathsf{T}} \nabla h = 9uv(u^2 - v^2 + 1) - 9uv(u^2 - v^2 + 1) = 0. \tag{60}$$

Thus $\nabla \chi$ generates a local predictor-preserving direction.

The generator vanishes when $uv = 0$ and $u^2 - v^2 + 1 = 0$. If $u = 0$, the second equation gives $v^2 = 1$. If $v = 0$, it would give $u^2 = -1$, which has no real solution. Hence the only real fixed points are

$$(u, v) = (0, +1), \qquad (u, v) = (0, -1). \tag{61}$$

Equivalently, these are precisely the two critical points $P'(\zeta) = 0$. Their predictor values are

$$h(0, w_\star) = w_\star, \qquad w_\star \in \{-1, +1\}. \tag{62}$$

The zero-charge set itself is larger. From equation (58),

$$\{\chi = 0\} = \{u = 0\} \cup \{u^2 = 3v^2 - 3\}. \tag{63}$$

The designed critical representatives belong to the component $u = 0$. This distinction will become important in the discrete dynamical analysis below.

D.2 Implicit bias of the binary parametrization

We now compute the geometric weight on the zero-charge level. Define the joint map $\Psi(u, v) = \big(h(u, v), \chi(u, v)\big)$. Because the binary predictor is $h = \operatorname{Im} P/2$, rather than $h = \operatorname{Im} P$, the predictor row of the joint Jacobian carries an additional factor $1/2$ compared with the general holomorphic construction of Appendix A.

Using equation (59),

$$D\Psi = \begin{pmatrix} 3uv & \frac{3}{2}(u^2 - v^2 + 1) \\ 3(u^2 - v^2 + 1) & -6uv \end{pmatrix}. \tag{64}$$

Hence

$$|\det D\Psi| = 18u^2v^2 + \frac{9}{2}(u^2 - v^2 + 1)^2. \tag{65}$$

On the other hand,

$$|P'(\zeta)|^2 = 9(u^2 - v^2 + 1)^2 + 36u^2v^2. \tag{66}$$

Comparing with equation (65) gives

$$|\det D\Psi| = \frac{1}{2}|P'(\zeta)|^2. \tag{67}$$

At a fixed predictor value w, imposing simultaneously $\chi = 0$ and $h = w$ is equivalent to $P(\zeta) = \chi(\zeta) + 2\mathrm{i}h(\zeta) = 2\mathrm{i}w$. Thus the selected set is

$$\mathcal{S}_w = \{\zeta \in \mathbb{C} : P(\zeta) = 2\mathrm{i}w\}. \tag{68}$$

For a regular predictor value, the geometric weight is therefore

$$\Gamma_0(w) = \sum_{\zeta:\, P(\zeta) = 2\mathrm{i}w} \frac{1}{|\det D\Psi(\zeta)|} = \sum_{\zeta:\, P(\zeta) = 2\mathrm{i}w} \frac{2}{|P'(\zeta)|^2}. \tag{69}$$

This is a geometric weight on the zero-charge level.

We next derive its behaviour near the targeted binary values. Fix $w_\star$, with corresponding critical point $\zeta = \mathrm{i}w_\star$. At the target value itself, the polynomial factors exactly as

$$P(\zeta) - 2\mathrm{i}w_\star = (\zeta - \mathrm{i}w_\star)^2(\zeta + 2\mathrm{i}w_\star). \tag{70}$$

Thus the designed critical representative $\zeta = \mathrm{i}w_\star$ is a double root, while $\zeta = -2\mathrm{i}w_\star$ is a distinct regular representative of the same predictor.

To determine the behaviour for nearby regular values, write $\zeta = \mathrm{i}w_\star + \xi$. Since $P'(\mathrm{i}w_\star) = 0$, direct expansion gives

$$P(\mathrm{i}w_\star + \xi) = 2\mathrm{i}w_\star + 3\mathrm{i}w_\star\xi^2 + \xi^3. \tag{71}$$

Hence $P(\zeta) = 2\mathrm{i}w$ is equivalent to

$$\xi^2\,(3\mathrm{i}w_\star + \xi) = 2\mathrm{i}(w - w_\star). \tag{72}$$

Two roots approach $\xi = 0$ as $w \to w_\star$. For either of these roots,

$$|\xi|^2 = \frac{2|w - w_\star|}{|3\mathrm{i}w_\star + \xi|} = \frac{2}{3}|w - w_\star|\big(1 + o(1)\big). \tag{73}$$

Moreover,

$$\begin{aligned} P'(\mathrm{i}w_\star + \xi) &= 6\mathrm{i}w_\star\xi + 3\xi^2 \\ &= 3\xi\,(2\mathrm{i}w_\star + \xi)\,. \end{aligned} \tag{74}$$

Therefore, using equation (73),

$$\begin{aligned} |P'(\mathrm{i}w_\star + \xi)|^2 &= 9|\xi|^2|2\mathrm{i}w_\star + \xi|^2 \\ &= 24|w - w_\star|\big(1 + o(1)\big). \end{aligned} \tag{75}$$

Equation (67) then yields

$$|\det D\Psi| = 12|w - w_\star|\big(1 + o(1)\big). \tag{76}$$

Consequently, each of the two representatives approaching the critical point contributes

$$\frac{1}{12|w - w_\star|}\big(1 + o(1)\big) \tag{77}$$

to $\Gamma_0(w)$.

The third representative approaches $\zeta = -2\mathrm{i}w_\star$. Since $P'(-2\mathrm{i}w_\star) = -9$, its contribution remains finite. The two singular contributions therefore dominate, giving

$$\Gamma_0(w) = \frac{1 + o(1)}{6|w - w_\star|}, \qquad w \to w_\star. \tag{78}$$

Accordingly,

$$\begin{aligned} L_{\mathrm{IB}}(w) &= -\log \Gamma_0(w) \\ &= \log|w - w_\star| + \log 6 + o(1). \end{aligned} \tag{79}$$

Thus the geometric weight diverges as the predictor approaches either targeted binary value. The inverse-Jacobian formula is used only for regular values approaching $w_\star$. At $w = w_\star$, the Jacobian vanishes at the critical representative and its inverse is not defined there. Moreover, the singular factor $1/|w - w_\star|$ is not locally integrable, so equation (78) should be interpreted as the local behaviour of the geometric coarea weight, not as a normalized predictor probability density.

D.3 Exact finite-step charge dynamics

We now turn to discrete SGD. For minibatch $\mathcal{B}_k$, define $g_{\mathcal{B}} \equiv g_{\mathcal{B}_k} = \partial_w \ell_{\mathcal{B}_k}\big(h(u_k, v_k)\big)$. Since

$$\nabla h(u_k, v_k) = \begin{pmatrix} 3u_k v_k \\ \frac{3}{2}(u_k^2 - v_k^2 + 1) \end{pmatrix}, \tag{80}$$

To derive the charge change, define the parameter increments

$$\Delta u_k := u_{k+1} - u_k = -3\eta g_{\mathcal{B}} u_k v_k, \qquad \Delta v_k := v_{k+1} - v_k = -\frac{3}{2}\eta g_{\mathcal{B}}(u_k^2 - v_k^2 + 1). \tag{81}$$

Since $\chi(u, v) = u^3 - 3uv^2 + 3u$ is a cubic polynomial, its Taylor expansion around (u_k, v_k) is exact at third order. Writing

$$A_k := u_k^2 - v_k^2 + 1, \tag{82}$$

the required derivatives are

$$\begin{aligned}
\partial_u\chi(u_k,v_k) &= 3A_k, & \partial_v\chi(u_k,v_k) &= -6u_kv_k,\\
\partial_{uu}\chi(u_k,v_k) &= 6u_k, & \partial_{uv}\chi(u_k,v_k) &= -6v_k,\\
\partial_{vv}\chi(u_k,v_k) &= -6u_k, & \partial_{uuu}\chi &= 6,\\
\partial_{uvv}\chi &= -6.
\end{aligned} \tag{83}$$

Therefore the exact expansion is

$$\begin{aligned}
\chi_{k+1}-\chi_k = {}& 3A_k\Delta u_k - 6u_kv_k\Delta v_k\\
&+ 3u_k(\Delta u_k)^2 - 6v_k\Delta u_k\Delta v_k - 3u_k(\Delta v_k)^2\\
&+ (\Delta u_k)^3 - 3\Delta u_k(\Delta v_k)^2.
\end{aligned} \tag{84}$$

The first-order contribution is

$$\begin{aligned}
3A_k\Delta u_k - 6u_kv_k\Delta v_k &= 3A_k\left(-3\eta g_{\mathcal{B}}u_kv_k\right)\\
&\quad - 6u_kv_k\left(-\frac{3}{2}\eta g_{\mathcal{B}}A_k\right)\\
&= -9\eta g_{\mathcal{B}}u_kv_kA_k + 9\eta g_{\mathcal{B}}u_kv_kA_k\\
&= 0.
\end{aligned} \tag{85}$$

Thus the complete $O(\eta)$ contribution cancels exactly, as required by the predictor-preserving symmetry. We next compute the quadratic contribution. From equation (81),

$$\begin{aligned}
(\Delta u_k)^2 &= 9\eta^2 g_{\mathcal{B}}^2u_k^2v_k^2,\\
\Delta u_k\Delta v_k &= \frac{9}{2}\eta^2 g_{\mathcal{B}}^2u_kv_kA_k,\\
(\Delta v_k)^2 &= \frac{9}{4}\eta^2 g_{\mathcal{B}}^2A_k^2.
\end{aligned} \tag{86}$$

Hence

$$\begin{aligned}
&3u_k(\Delta u_k)^2 - 6v_k\Delta u_k\Delta v_k - 3u_k(\Delta v_k)^2\\
&= 27\eta^2 g_{\mathcal{B}}^2u_k^3v_k^2 - 27\eta^2 g_{\mathcal{B}}^2u_kv_k^2A_k - \frac{27}{4}\eta^2 g_{\mathcal{B}}^2u_kA_k^2\\
&= \frac{27}{4}\eta^2 g_{\mathcal{B}}^2u_k\left[4u_k^2v_k^2 - 4v_k^2A_k - A_k^2\right].
\end{aligned} \tag{87}$$

Since $4u_k^2v_k^2 - 4v_k^2A_k = 4v_k^2(v_k^2-1)$, the quadratic term reduces to

$$\frac{27}{4}\eta^2 g_{\mathcal{B}}^2u_k\left[4v_k^2(v_k^2-1) - A_k^2\right]. \tag{88}$$

The cubic contribution is

$$\begin{aligned}
(\Delta u_k)^3 - 3\Delta u_k(\Delta v_k)^2 = {}& -27\eta^3 g_{\mathcal{B}}^3u_k^3v_k^3\\
&+ \frac{81}{4}\eta^3 g_{\mathcal{B}}^3u_kv_kA_k^2\\
= {}& -\frac{27}{4}\eta^3 g_{\mathcal{B}}^3u_kv_k\left[4u_k^2v_k^2 - 3A_k^2\right].
\end{aligned} \tag{89}$$

Combining equations (85), (88), and (89) gives the exact finite-step charge update

$$\begin{aligned}
\chi_{k+1}-\chi_k = {}& \frac{27}{4}\eta^2 g_{\mathcal{B}}^2u_k\left[4v_k^2(v_k^2-1) - (u_k^2-v_k^2+1)^2\right]\\
&- \frac{27}{4}\eta^3 g_{\mathcal{B}}^3u_kv_k\left[4u_k^2v_k^2 - 3(u_k^2-v_k^2+1)^2\right].
\end{aligned} \tag{90}$$

Thus the predictor-preserving symmetry cancels the complete first-order change in the charge. Because χ is cubic, the remaining terms terminate exactly at orders η^2 and η^3.

D.4 A SUFFICIENT CONDITION FOR ATTRACTION TO THE INVARIANT BRANCH

We now derive a sufficient condition under which a trajectory starting with $u_0 \neq 0$ approaches the invariant branch $u = 0$.

From equation (81), $u_{k+1} = (1 - 3\eta v_k g_{\mathcal{B}})u_k$. Assume that $1 - 3\eta v_k g_{\mathcal{B}} \neq 0$, and suppose that, conditionally on the current parameters (u_k, v_k), there exists a constant $\varepsilon > 0$ such that

$$\mathbb{E}_{\mathcal{B}}\left[1 - 3\eta v_k g_{\mathcal{B}} \mid u_k, v_k\right] \leq -\varepsilon \qquad \text{for every } k. \tag{91}$$

Assume additionally that the corresponding conditional second moments are uniformly bounded,

$$\sup_k \mathbb{E}_{\mathcal{B}}\left[(1 - 3\eta v_k g_{\mathcal{B}})^2 \mid u_k, v_k\right] < \infty. \tag{92}$$

Iterating the multiplicative update gives the exact identity

$$u_N = u_0 \prod_{k=0}^{N-1} (1 - 3\eta v_k g_{\mathcal{B}}), \tag{93}$$

and therefore

$$\log \frac{|u_N|}{|u_0|} = \sum_{k=0}^{N-1} \log |1 - 3\eta v_k g_{\mathcal{B}}|. \tag{94}$$

We decompose each logarithmic increment into its conditional mean and its zero-mean fluctuation,

$$\begin{aligned} \log |1 - 3\eta v_k g_{\mathcal{B}}| = {} & \mathbb{E}_{\mathcal{B}}\left[\log |1 - 3\eta v_k g_{\mathcal{B}}| \mid u_k, v_k\right] \\ & + \left(\log |1 - 3\eta v_k g_{\mathcal{B}}| - \mathbb{E}_{\mathcal{B}}\left[\log |1 - 3\eta v_k g_{\mathcal{B}}| \mid u_k, v_k\right]\right). \end{aligned} \tag{95}$$

Because a fresh minibatch is sampled at each step, the terms in parentheses are martingale differences. Under the uniform second-moment bound in equation (92), their empirical average converges to zero:

$$\frac{1}{N} \sum_{k=0}^{N-1} \left(\log |1 - 3\eta v_k g_{\mathcal{B}}| - \mathbb{E}_{\mathcal{B}}\left[\log |1 - 3\eta v_k g_{\mathcal{B}}| \mid u_k, v_k\right]\right) \longrightarrow 0. \tag{96}$$

Hence, dividing equation (94) by N,

$$\begin{aligned} \limsup_{N\to\infty} \frac{1}{N} \log \frac{|u_N|}{|u_0|} &= \limsup_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} \mathbb{E}_{\mathcal{B}}\left[\log |1 - 3\eta v_k g_{\mathcal{B}}| \mid u_k, v_k\right] \\ &\leq -\varepsilon. \end{aligned} \tag{97}$$

Therefore $u_N \longrightarrow 0$. If v_N remains bounded, then $|\chi_N| \to 0$. Thus attraction to the invariant branch $u = 0$ also implies convergence of the charge to zero, provided v_N remains bounded.

D.5 DYNAMICAL SELECTION OF THE CRITICAL REPRESENTATIVES

The previous subsection establishes, under the stated conditions, attraction toward $u = 0$. This is enough to imply $\chi_k \to 0$ when v_k remains bounded, but it is not enough by itself to determine the limiting value of v_k.

On the invariant branch $u = 0$, the predictor and the SGD update reduce to

$$h(0, v) = \frac{1}{2}(-v^3 + 3v), \qquad v_{k+1} = v_k - \frac{3}{2}\eta g_{\mathcal{B}}(1 - v_k^2). \tag{98}$$

The factor $1 - v_k^2$ already distinguishes the two values $v = \pm 1$: at these points the parameter update vanishes for every sampled gradient. Moreover, $h(0, \pm 1) = \pm 1$. Thus the designed binary representatives are precisely the points on the invariant branch where the remaining SGD motion in v stops.

Suppose now that the predictor converges to a targeted value $h(u_k, v_k) \longrightarrow w_\star$, and that v_k remains bounded. Every accumulation point then satisfies

$$u = 0, \qquad h(0, v) = w_\star. \tag{99}$$

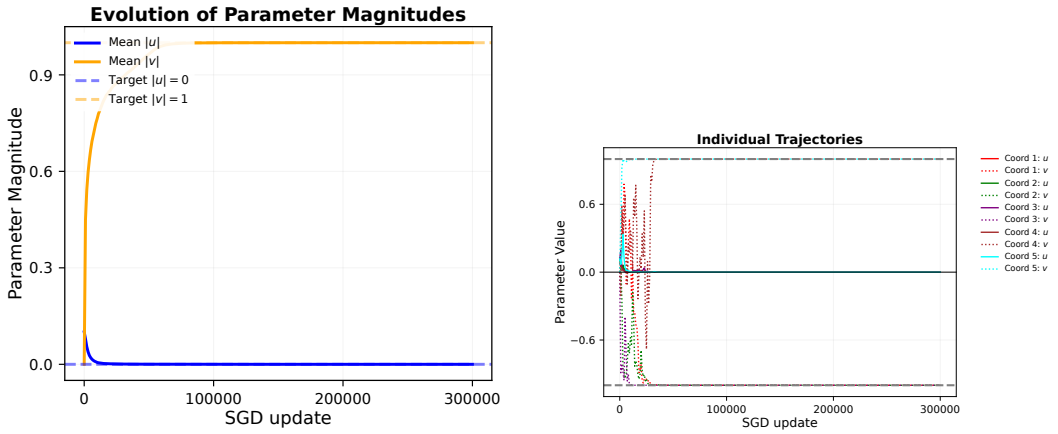

Figure 3: **Parameter dynamics of the parameterized model. Left:** evolution of the mean parameter magnitudes during training. The u-parameters approach zero while the magnitude of the v-parameters approaches one. **Right:** five individual parameter trajectories from the first run. The u-coordinates approach zero, while the corresponding v-coordinates approach either $+1$ or -1. These trajectories are consistent with attraction toward the invariant branch $u = 0$ and stabilization at the critical representatives $(0, \pm 1)$ derived above.

The latter equation factors exactly as

$$\begin{aligned} h(0, v) - w_\star &= \frac{-v^3 + 3v}{2} - w_\star \\ &= -\frac{1}{2}(v - w_\star)^2(v + 2w_\star). \end{aligned} \tag{100}$$

Thus there are exactly two distinct compatible parameter representatives:

$$(u, v) = (0, w_\star), \qquad (u, v) = (0, -2w_\star). \tag{101}$$

These are exactly the two distinct points already identified by the factorization

$$P(\zeta) - 2\mathrm{i}w_\star = (\zeta - \mathrm{i}w_\star)^2(\zeta + 2\mathrm{i}w_\star). \tag{102}$$

The first is the designed critical representative and occurs as a double root. The second is the regular representative that has a finite contribution equal to $2/81$.

At the designed representative, $v = w_\star$ and $1 - v^2 = 0$, equation (98) gives $v_{k+1} = v_k = w_\star$, for every sampled gradient. Equivalently,

$$\nabla h(0, w_\star) = 0. \tag{103}$$

Hence the critical representative is absorbing under SGD.

At the alternative representative $(0, -2w_\star)$, instead, $\partial_v h = -9/2 \neq 0$. Writing $v_{k+1} = -2w_\star + \delta_k$ where $\delta_k = \frac{9}{2}\eta g_{\mathcal{B}}$, the predictor after one step is

$$h_{k+1} - w_\star = -\frac{9}{2}\delta_k + 3w_\star\delta_k^2 - \frac{1}{2}\delta_k^3. \tag{104}$$

Hence a generic nonzero sampled gradient moves the predictor away from $w_\star$. The critical representative is therefore the point at which the binary predictor value is preserved under the stochastic dynamics, whereas the regular representative produces the same predictor only instantaneously.