MIND THE DRIFT: DIAGONAL LINEAR NETWORKS UNDER LARGE LEARNING RATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large learning rates can qualitatively change the trajectory of neural network training, often pushing optimization into regimes far from classical gradient-flow behavior. The Edge of Stability (EoS) offers a valuable lens on the dynamics such learning rates induce. We study corresponding dynamics in diagonal linear networks, where we uncover a competition between two distinct implicit biases that jointly determine the sparsity of the recovered solution in regression settings. Complementary to the Gain, which captures the average discretization error accumulated by Gradient Descent relative to Gradient Flow, we derive a closely associated but overlooked quantity: the Drift. Under large learning rates, it describes an imbalance between different discretization errors and represents a systematic shift in the optimization trajectory. While the Gain grows monotonically in certain regimes, and can bias towards denser, flatter interpolators, the impact of the Drift depends on its alignment with potential solutions, which can either counteract or reinforce the effect of the Gain. Consequently, its behavior drives model selection, particularly during early training epochs. To validate our theoretical insights, we introduce an intervention that actively steers the Gain to recover sharper, sparser solutions. Thus, our analysis reveals that large learning rates do not universally hinder the recovery of sparse solutions. On the contrary, they can be harnessed to control the implicit bias of training.

1 INTRODUCTION

Large learning rates play an important role in the success of deep learning. In overparameterized models, where many solutions fit the training data, this raises a fundamental question: How does the learning rate influence which solution is selected? For small learning rates, gradient flow provides a useful approximation for understanding the implicit regularization governing this selection. At larger learning rates, however, training can deviate substantially from gradient-flow behavior. Progressive sharpening (Damian et al., 2022) can bring training to the Edge of Stability (Cohen et al., 2021), where sharpness fluctuates around the stability threshold and the loss exhibits persistent oscillations. Understanding how these dynamics affect implicit regularization at large learning rates, requires accounting for the discretization effects of gradient descent relative to gradient flow, that persist throughout training.

These persistent discretization effects can be explicitly characterized in diagonal linear networks. Under gradient flow, the initialization determines the implicit regularizer, which approaches an ℓ_1-bias at small initialization scales and an ℓ_2-bias at large scales (Woodworth et al., 2020). For discrete training, Even et al. (2023) characterize changes to the effective initialization scale through the so-called Gain. Under gradient descent, this Gain can accumulate heterogeneously across features, with larger values on support coordinates increasing their relative implicit penalty. Consequently, the resulting bias can favor denser solutions and hinder sparse recovery. Their convergence guarantees, however, require step-size restrictions that do not cover the Edge-of-Stability regime.

At large learning rates, a second effect can overturn the preference suggested by Gain alone and restore sparse recovery. We isolate this effect as the Drift, which captures the contribution of the shifted reference point in the implicit regularizer. While Gain changes the relative penalties on features, Drift introduces a directional preference among interpolating solutions. When suitably aligned, it can offset the penalty on support coordinates and favor a sharper, sparser solution over

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.

a flatter, denser one. Understanding solution selection at large learning rates therefore requires understanding the competition between Gain and Drift. This leads to the central question of this work: How do Gain and Drift interact during training, and when does their competition favor sparse recovery in selecting a final solution?

Concretely, we show that their relative influence changes across phases associated with the Edge of Stability regime. Under progressive sharpening, the Drift can accumulate significantly . In the oscillatory regime, it saturates, while the Gain can still accumulate. At sufficiently large learning rates, the Drift can compete with the Gain by shifting the preference among interpolating solutions, either promoting or hindering sparse recovery. Note that even in a minimal sparse regression problem (Fig. 1) gradient descent with large learning rates can recover sharper, sparser solutions despite the opposing bias of Gain.

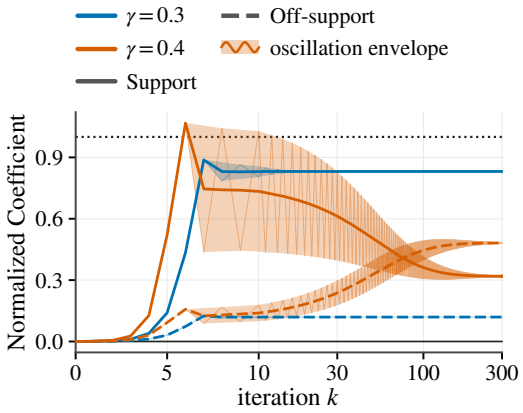


Figure 1: **Illustration.** A diagonal linear network with learning rate $\gamma = 0.3$ and 0.4 to show the evolution of the predictor coefficient. The smaller step size does not reach the Edge of Stability regime, while the larger one does. We recover a sparser solution for the former.

This perspective also suggests a way to control solution selection. We introduce a modification of gradient descent that changes how the Gain is distributed across coordinates, using only the observed gradients. In our sparse-recovery setting, this intervention counteracts the relative penalization of support coordinates and improves recovery where ordinary gradient descent would favor a denser solution. This shows that large learning rates need not uniformly undermine sparse recovery. Their effect depends on the joint evolution and alignment of the discretization corrections.

Contributions:

- We introduce the *Drift*, a post-trajectory quantity that captures the directional effect of discretization on solution selection. We characterize how the Drift and the Gain determine the preference among interpolating solutions across different training regimes.
- We analyze the interaction between the Gain and Drift, deriving a sparse-recovery characterization in Theorem 2.4 and show how their effects change across large-learning-rate training regimes in Lemmas 2.5 and 2.6. We then make this competition explicit in a tractable example in § 2.2.
- To corroborate our insights, we introduce a *Gain control* scheme in §3 that actively redistributes Gain and Drift to improve sparse recovery. Our characterization of its implicit-bias and our convergence proof explain the underlying mechanism and experiments validate the improved support recovery in large-learning-rate and edge-of-stability regimes where standard (S)GD fails.

Related work. Implicit regularization in linear networks connects solution selection to parameterization, initialization, and optimization dynamics (Gunasekar et al., 2017; Vaškevičius et al., 2019; Woodworth et al., 2020). For diagonal networks, initialization controls the transition between ℓ_1- and ℓ_2-type biases under gradient flow (Woodworth et al., 2020). Most closely related, the implicit bias of discrete GD and SGD has been characterized through a trajectory-dependent effective initialization and a shifted Bregman reference point (Even et al., 2023). We build on this characterization by isolating the directional contribution of the shifted reference point as Drift and quantifying how its interaction with Gain determines sparse recovery at large learning rates.

Studies of the Edge of Stability document progressive sharpening and persistent oscillations near the stability threshold (Cohen et al., 2021), explain self-stabilization (Damian et al., 2022), and develop continuous approximations of oscillatory dynamics (Cohen et al., 2024; Regis & Chewi, 2026). Complementary work establishes decreasing gradient-flow endpoint sharpness along GD trajectories in scalar networks (Kreisler et al., 2023) and identifies competing norm and sharpness biases

in diagonal networks (Fojtík et al., 2025). Our analysis offers a complementary account of these dynamics and competing preferences to the evolution of Gain and Drift, showing how their interaction can favor sharper, sparser interpolators. Our gain-control intervention further demonstrates how these discretization effects can be steered to improve sparse recovery. An extended discussion on the related works is presented in Appendix B.

1.1 PRELIMINARIES

We consider the overparameterized regression problem class $X\beta = y$ for data $X \in \mathbb{R}^{n\times d}$ and outputs $y \in \mathbb{R}^n$, where $d > n$. There exist infinitely many solutions $\beta \in \mathbb{R}^d$ to this system, to which we refer as interpolators, and their collection as the interpolation manifold $\mathcal{M} = \{\beta^* \in \mathbb{R}^d : X\beta^* = y\}$. We assume $\mathcal{M}$ to be non-empty.

Diagonal linear networks. Diagonal linear networks provide a tractable way to analyze the training dynamics in terms of the network weights. Although the problem is linear in x, the parameterization of β influences the selected solution. We consider the two-layer network $x \mapsto \langle u, \sigma(\mathrm{diag}(v)x)\rangle = \langle u \odot v, x\rangle$, where $u, v \in \mathbb{R}^d$, u represents the output weights, $\mathrm{diag}(v)$ is the hidden-layer weight matrix, and σ is the Identity activation. We refer to $\beta = u \odot v$ as the predictor. Throughout, we initialize $u_0 = \sqrt{2}\alpha\mathbf{1}$ and $v_0 = 0$, where $\alpha > 0$, so that $\beta_0 = 0$. We use the standard quadratic loss $L(\beta) = \frac{1}{2n}\|X\beta - y\|_2^2$ for training. For simplicity of our analysis, we consider a fixed learning rate γ for all optimization steps.

Sparse Solution. We fix a s-sparse reference interpolator $\beta^\star \in \mathcal{M}$ and let $S := \mathrm{supp}(\beta^\star) = \{j \in [d] : \beta_j^\star \neq 0\}$, with $|S| = s \ll d$. We refer to the coordinates in S as the *support coordinates* and those in $S^c = [d] \setminus S$ as *off-support coordinates*. For any β, the vector β_{S^c} records coefficients outside the target support. We call β denser than $\beta^\star$ when $\|\beta\|_0 = |\{j \in [d] : \beta_j \neq 0\}| > s$.

Gradient Flow. We first consider Gradient Flow (GF), the continuous-time limit of Gradient Descent (GD). Let $g_t := \nabla L(\beta_t) = \frac{1}{n}X^\top(X\beta_t - y)$ be the gradient of the loss. GF evolves the weights as follows:

$$\dot{u}_t = -v_t \odot g_t, \qquad \dot{v}_t = -u_t \odot g_t. \tag{1}$$

Consequently, the predictor evolves as $\dot{\beta}_t = -(u_t^{\odot 2} + v_t^{\odot 2}) \odot g_t$, where $u^{\odot 2}$ denotes the elementwise square of u. Moreover, the "balance" is conserved, $u_t^{\odot 2} - v_t^{\odot 2} = u_0^{\odot 2} - v_0^{\odot 2}$, which induces the hyperbolic entropy mirror geometry governing the implicit bias of GF.

Definition 1.1 (Hyperbolic entropy)**.** For $a \in \mathbb{R}^d_{>0}$, define

$$\psi_a(\beta) = \frac{1}{2}\sum_{j=1}^{d}\left[\beta_j \,\mathrm{arcsinh}\left(\frac{\beta_j}{a_j}\right) - \sqrt{\beta_j^2 + a_j^2} + a_j\right]. \tag{2}$$

Its associated Bregman divergence is $\mathcal{D}_{\psi_a}(\beta, \hat{\beta}) = \psi_a(\beta) - \psi_a(\hat{\beta}) - \langle \nabla\psi_a(\hat{\beta}), \beta - \hat{\beta}\rangle$.

Under vanishing initialization $\alpha \to 0$, ψ in Definition 1.1 resembles the l_1 norm and characterizes the minimization problem in Eq. (8) leading to sparse recovery guarantees (Candes et al., 2005). In the implicit-bias literature, $\alpha \to 0$ corresponds to the sparse "rich" regime, whereas $\alpha \to \infty$ approaches an ℓ_2 geometry associated with the "kernel" regime (Woodworth et al., 2020).

Implicit bias at finite step size. GD with $g_k = \nabla L(\beta_k)$ and $z_k = \gamma g_k$ satisfies

$$u_{k+1} = u_k - z_k \odot v_k, \qquad v_{k+1} = v_k - z_k \odot u_k.$$

Unlike GF, these updates do not preserve the balance underlying the fixed mirror geometry:

$$u_{k+1}^{\odot 2} - v_{k+1}^{\odot 2} = (\mathbf{1} - z_k^{\odot 2}) \odot (u_k^{\odot 2} - v_k^{\odot 2}). \tag{3}$$

Consequently, the implicit bias depends on the training trajectory through an *effective* initialization. This change is analyzed using a post-trajectory quantity named the Gain.

Definition 1.2 (Gain ((Even et al., 2023))). For $z \neq \pm 1$, let $q(z) = -\log|1 - z^2|$. For a GD trajectory avoiding $z_{k,j} = \pm 1$, define

$$Q_k = \sum_{\ell=0}^{k-1} q(z_\ell), \qquad a_k = \alpha^2 \exp(-Q_k). \tag{4}$$

We call Q_k the Gain and a_k the effective initialization or the scale. When Q_k has a finite limit, we write $Q_\infty = \lim_{k\to\infty} Q_k$ and $a_\infty = \alpha^2 \exp(-Q_\infty)$. The Gain vector is defined as $\text{Gain}_\gamma := \ln(\frac{\alpha^2 \mathbf{1}}{a_k}) = \sum_k q(\gamma \nabla_{\mathcal{B}_k} \mathcal{L}(\beta_k)) \in \mathbb{R}^d$.

With this notation in place, we can restate the implicit bias characterization of GD next.

Theorem 1.3 (Implicit bias of GD (Even et al., 2023)). *Suppose GD initialized at $(u_0, v_0) = (\sqrt{2}\alpha \mathbf{1}, \mathbf{0})$ satisfies $\beta_k = u_k \odot v_k \to \beta_\infty \in \mathcal{M}$, with $\sum_{k\geq 0} q_\pm(z_k)$ finite coordinatewise. Then*

$$\beta_\infty = \arg\min_{\beta \in \mathcal{M}} \mathcal{D}_{\psi_{a_\infty}}(\beta, \widetilde{\beta}_0), \qquad \widetilde{\beta}_0 = \tfrac{1}{2}(a_{+,\infty} - a_{-,\infty}), \tag{5}$$

where $q_\pm(z) = \mp 2z - 2\log|1 \mp z|$ and $a_{\pm,\infty} = \alpha^2 \exp(-\sum_{k\geq 0} q_\pm(z_k))$.

The corrections $q_\pm$ capture GD's discretization effects relative to GF, producing a trajectory-dependent scale a_∞ and a shifted Bregman reference point $\widetilde{\beta}_0$. Under conservative step-size bounds, Even et al. (2023) establish convergence to β_∞ and characterize the Gain through a_∞, but do not further analyze the shift in $\widetilde{\beta}_0$.

2 Theoretical analysis of drift-gain decomposition

The implicit-bias formulation in Eq. (5) reveals two trajectory-dependent quantities: the scale $a_{\pm,k}$ and the shifted initialization $\widetilde{\beta}_0$. While the Gain captures the former, the effect of the shifted initialization is less explicit. To isolate this effect, we introduce the *Drift*, a post-trajectory statistic analogous to the Gain that captures the asymmetry of the GD discretization across the coupled layer dynamics. We provide rigorous proofs for all lemmas, propositions and theorems in Appendix C.

Definition 2.1 (Drift increment). For $z \neq \pm 1$, define

$$d(z) := \frac{q_+(z) - q_-(z)}{2} = \log\left|\frac{1+z}{1-z}\right| - 2z, \qquad d(z) \xrightarrow{z\to\pm 1} \pm\infty. \tag{6}$$

Definition 2.2 (Drift). Let $z = \gamma \nabla_{\mathcal{B}} \mathcal{L}$, and $q_\pm(z) = \mp 2z - 2\log|1 \mp z|$, $a_{\pm,\infty} = \alpha^2 \exp\left(-\sum_{k=0}^{\infty} q_\pm(z_k)\right)$. For $k \in \{0, \ldots, K-1\}$ (S)GD steps, we define the accumulated drift as

$$D_K := \sum_{k=0}^{K-1} d(\gamma \nabla_{\mathcal{B}^k} \mathcal{L}(\beta_k)) = \frac{1}{2} \ln \frac{a_{-,K}}{a_{+,K}} \tag{7}$$

where $d(z)$ is the drift increment Eq. (6), and $|\mathcal{B}^k|$ is the batchsize. For a converging sequence, $D_\infty = \lim_{k\to\infty} D_k$.

To make the contribution of the drift explicit, we rewrite the Bregman objective in Theorem 1.3 as a hyperbolic-entropy potential with a linear correction to highlight the origins of the drift.

Proposition 2.3 (Gain–Drift decomposition). *Under the assumptions of and with quantities defined as in Theorem. 1.3,*

$$\beta_\infty = \arg\min_{\beta \in \mathcal{M}} \left\{ \psi_{a_\infty}(\beta) + \frac{1}{2}\langle D_\infty, \beta \rangle \right\}. \tag{8}$$

Moreover, $\tilde{\psi}(\beta, a_\infty, D_\infty) := \psi_{a_\infty}(\beta) + \frac{1}{2}\langle D_\infty, \beta\rangle$ attains its global minimum at $\beta = a_\infty \sinh(-D_\infty) \neq 0$ when $D_\infty \neq 0$.

This reformulation separates two effects: The Gain changes the curvature of the potential, while the Drift shifts its global minimum. Hence, the Drift introduces a directional bias distinct from the Gain. See Fig. 8.

2.1 SPARSE RECOVERY

In the small-initialization regime, Even et al. (2023) show that a heterogeneous accumulation of the Gain across support and off-support coordinates can lead to a loss of sparsity. We extend this picture by deriving a new *sparse recovery theorem* that quantifies the joint influence of the Gain and the Drift on the recovered predictor, making their competition across different training regimes explicit.

Theorem 2.4. *(Sparse Recovery Theorem) Let $X \in \mathbb{R}^{n\times d}$ with $d > n$ and $X\beta^* = y$, where $\beta^* \in \mathbb{R}^d$ has non-empty finite support $S \subsetneq \{1,\dots,d\}$. Let S^c be the set of the off-support coordinates. Suppose that for some finite $c \geq 0$,*

$$\|h_S\|_1 \leq c\|h_{S^c}\|_1 \qquad \text{for every } h \in \ker X. \tag{NSP}$$

Let ψ be the potential function in Eq. (2) and $\hat{\beta}$ the minimizer of the implicit bias in Eq. (8). Suppose that GD converges to $\widehat{\beta}$, Q and D are the final Gain and Drift vectors and a is the effective initialization in Eq. (4). Let $A_{S^c} := \sum_{j\in S^c} a_j = \alpha^2 \sum_{j\in S^c} e^{-Q_j}$ be the accumulated scale of the off-support coordinates and let $R_S := \left\| \operatorname{arsinh}\left(\frac{\beta^\star_S}{a_S}\right) + D_S \right\|_\infty$. Then we have that

$$\|\widehat{\beta} - \beta^\star\|_1 \leq (1+c) A_{S^c} \sinh(cR_S + \|D_{S^c}\|_\infty)\,.$$

Proof Sketch. For $h = \widehat{\beta} - \beta^\star \in \ker X$, NSP gives the bound $\|h\|_1 \leq (1+c)\|h_{S^c}\|_1$. We split the optimality condition into support and off-support terms, bounding the off-support contribution using the Gain and Drift. Finally, Jensen's inequality yields $\|h_{S^c}\|_1 \leq A_{S^c}\sinh(cR_S + \|D_{S^c}\|_\infty)$, giving the stated recovery bound. See App. C.1 for a detailed proof.

The Null-Space Property (NSP) (Cohen et al., 2008), a standard sufficient condition for exact and stable sparse recovery in compressed sensing, links changes on the support to the off-support mass that any competing interpolator must introduce. In particular, when $c < 1$, the sparse target β^* is the unique minimum-ℓ_1 interpolator. This provides a natural way to interpret the theorem in terms of the support S and off-support S^c coordinates. If more Gain accumulates on the support S, the sparse target becomes relatively more expensive. The Drift instead acts through its alignment with feasible directions in $\ker X$. It can either reinforce this effect or counteract it and favor the sparse solution. Theorem 2.4 makes this competition explicit. See App. C.1 for a summary.

We have now established the relationship between the accumulation of Gain and Drift to the sparsity of the recovered solution. However, these quantities do not evolve in the same way during training. During progressive sharpening, gradients can remain sign-coherent for many steps. So both quantities accumulate. Once training enters the oscillatory regime near the Edge of Stability, consecutive gradients approximately reverse the sign. The Drift then largely cancels, while the Gain can continue to accumulate. When support coordinates also carry larger gradients than off-support coordinates, the continued Gain accumulation becomes heterogeneous and can increasingly penalize the support. We formalize these two regimes in Lemmas 2.5 and 2.6.

2.2 DRIFT AT THE EDGE OF STABILITY

Progressive Sharpening Phase. The progressive sharpening phase (Damian et al., 2022) is the early training regime under large learning rates, where the sharpness $\bar{S}_k = \lambda_{\max}(\nabla^2_{(u_k,v_k)}\mathcal{L}(\beta_k))$ increases towards the stability threshold $2/\gamma$ (where $\lambda_{\max}$ is the largest eigenvalue). Let K_{edge} be the first GD iteration for which $\gamma\bar{S} > 2$, i.e. $K_{\text{edge}} := \inf\{k \geq 0 : \gamma\bar{S}_k \geq 2\}$. Thus $\bar{S}_k < \frac{2}{\gamma}$ for all $k < K_{\text{edge}}$. Additionally, we assume that the predictor gradients maintain a consistent sign in each coordinate. Then the following lemma quantifies the resulting accumulation of the Drift.

Lemma 2.5 (Drift accumulation)**.** *Let $k \in \{0,\dots,K_{edge}-1\}$. Let $D_{k+1} - D_k = d(z_k)$ be the drift increment as in Definition 2.1. Suppose that, for some $0 < \rho < 1$ and $\sigma \in \{-1,1\}^d$,*

$$\|z_k\|_\infty \leq \rho, \qquad \sigma_j z_{k,j} \geq 0 \quad \text{for every coordinate } j \text{ and step } 0 \leq k < K. \tag{9}$$

We thus assume bounded gradients and sign-coherence between the gradients for GD with learning rate $\gamma > 0$. Then, for any finite $1 \leq K \leq K_{edge}$ coordinatewise,

$$\frac{2}{3}\sum_{k=0}^{K-1}|z_{k,j}|^3 \leq \sigma_j D_{K,j} = |D_{K,j}| \leq \frac{2}{3(1-\rho^2)}\sum_{k=0}^{K-1}|z_{k,j}|^3. \tag{10}$$

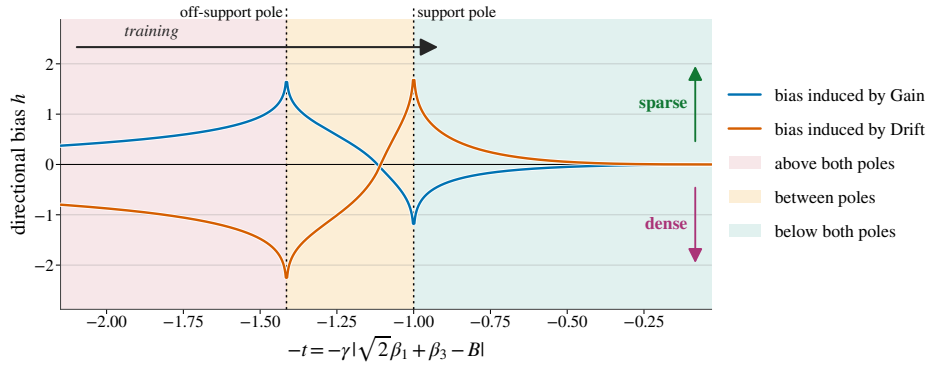


Figure 2: Alternating behavior of the Gain and the Drift. At initialization, the training starts from the left and proceeds to the right as the gradients shrink. The upper half of the plot indicates a sparsity promoting behavior, while the bottom half represents bias towards denser predictors. The two dotted lines represent the support and offsupport poles. See Table 1 for a summary.

Thus for $z_k \neq 0$, each coordinate accumulates drift in a fixed direction.

In this phase, both the Gain and the Drift can accumulate based on the gradient history and the learning rate. The Drift additionally depends on the coherence between the signs of consecutive gradients. Gain increments are quadratic in the scaled gradients (Even et al., 2023) and therefore invariant to their signs. In particular, at large learning rates, both Gain and Drift accumulate rapidly in the early epochs, then slow down as the model fits the data and the gradients shrink.

Oscillatory Phase. Upon reaching the Edge of Stability threshold, GD iterates oscillate, which alters the accumulation of Drift. We consider intervals in which consecutive scaled predictor gradients are approximately opposite ($z_{k+1} \approx -z_k$).

Lemma 2.6 (Drift cancellation at EoS). *Let $d(z)$ be the drift-increment in Eq. (2.1) and suppose $\|z_k\|_\infty \leq \rho < 1$ for all $k_0 \leq k < k_0 + 2m$. For $\ell = 0, \ldots, m-1$, let $e_\ell = z_{k_0+2\ell+1} + z_{k_0+2\ell}$ denote the deviation of the ℓ-th pair of gradient steps from exact sign reversal. Then*

$$\|D_{k_0+2m} - D_{k_0}\|_\infty \leq \frac{2\rho^2}{1-\rho^2} \sum_{\ell=0}^{m-1} \|e_\ell\|_\infty. \tag{11}$$

In particular, if every pair is exactly sign-reversed (i.e. $e_\ell = 0$), then $D_{k_0+2m} = D_{k_0}$.

In this phase, GD enters a period-two regime (i.e. $z_{k+1} \approx -z_k$), in which consecutive gradients nearly reverse sign. Since the drift increment in Eq. (6) is odd, each such pair of steps contributes minimally to the Drift. Hence, it starts to saturate. See Fig. 9.

The accumulated Drift in the progressive sharpening phase (Lemma 2.5) influences the predictor selection in the early epochs. However, above the EoS, Lemma 2.6 shows that the additional Drift arising from the oscillatory phase approximately cancels in contrast to the Gain, which keeps accumulating. In the absence of a well-aligned Drift, this additional Gain is able to regularize the support coordinates and ruin the sparsity of the predictor. Fig. 2 illustrates the competition between the Gain and the Drift with an example case, where we can analytically track the bias induced by them.

Example case of gain and drift interaction. We consider a linear problem $X\beta = y$, where

$$X = \begin{pmatrix} \sqrt{2} & 0 & 1 \\ 0 & \sqrt{2} & 1 \end{pmatrix} \in \mathbb{R}^{2\times 3}, \qquad y = B\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad B \in \mathbb{R}_+.$$

The system admits a sharper, sparser solution $\beta_s^* = B\,(0,0,1)^\top$ and a flatter, denser solution $\beta_d^* = B\,(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0)$. We refer to the third feature column (i.e. β_s^*) as the "support" and the first two feature columns as "off-support". Let $\mathcal{S} = \{\beta(t) : t \in [0,1]\}$ be a line segment , where

$$\beta(t) = \beta_s^* + Bth = B\left(\frac{t}{\sqrt{2}}, \frac{t}{\sqrt{2}}, 1-t\right)^\top \qquad h = \left(\tfrac{1}{\sqrt{2}}, \tfrac{1}{\sqrt{2}}, -1\right) \in \ker X.$$

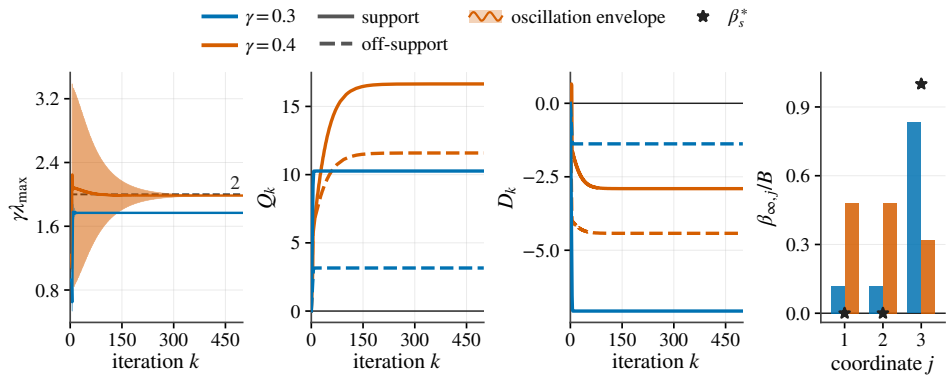


Figure 3: A diagonal linear network example case with control parameter $B = 3.1$ and learning rates $\gamma = 0.3$ and 0.4. (*Left*) We plot the product of sharpness and step size. The smaller step size run converges below the EoS regime while larger learning rate oscillates. (*Middle*) In both cases, the Gain and the Drift accumulate comparatively (Lemma 2.5 and 2.6). (*Right*) We plot the predictor coefficient, recovering a sparser solution for the former run and denser for the latter.

Here, h satisfies Eq. (NSP). At $t = 0$, we recover the sparse solution β_s^*, while $t = 1$ produces the dense counterpart β_d^*. For $0 \le t < 1$, the predictor ℓ_1 norm $\|\beta(t)\|_1 = B\big(1 + (\sqrt{2} - 1)t\big)$ increases while the minimum sharpness $\bar{S}_\infty(t) = B(2 - (2 - \sqrt{2})t)$ decreases. Along the direction h, the coefficients of $\beta(t)$ at the two off-support coordinates increases and trades sparsity for flatness. The gradient of the loss is $\nabla\mathcal{L} = x^\top r_k$, where the residual r_k after k steps of GD is $r_k = \frac{1}{\sqrt{2}}(\beta_{1,k} + \beta_{2,k}) + \beta_{3,k} - B$. Since the Gain and the Drift are singular at $|\gamma\,\nabla\mathcal{L}| = 1$ so that $\gamma|r_k||x| = 1$, they have 2 poles: an off-support pole at $\sqrt{2}$ and a support pole at 1.

Alternating behavior of Gain and Drift. To inspect the competition between the Gain and Drift in early training, we train a diagonal linear network for 500 iterations with setup $(\gamma, B, \alpha) = (0.1, 16, 0.04)$. Table 1 and Fig. 2 summarize the changing directional biases of Gain and Drift across the three residual regimes. Derivations are in App. C.2.1–C.2.3.

Sparse Recovery. Figs. 1 and 3 illustrate how this competition shapes solution selection for $B = 3.1$, comparing GD trajectories with $\gamma = 0.3$ and 0.4 over 500 iterations. For $\gamma = 0.3$, Drift substantially offsets the Gain-induced preference for β_d^*, yielding a predictor closer to the sparse target β_s^*. For $\gamma = 0.4$, the trajectory enters the oscillatory EoS regime, where Gain continues to accumulate while Drift largely saturates (Lemma 2.6). The accumulated Drift provides less compensation, and the resulting predictor assigns more mass to off-support coordinates.

3 Rebalancing the Gain

Building on Proposition 2.3, we introduce an intervention that reverses the Gain imbalance induced by GD. Through the layer-imbalance dynamics in Eq. (3), larger gradient coordinates undergo stronger contraction of the effective initialization scale in Eq. (4) and hence receive a larger implicit penalty (Theorem. 2.4). In our example, this penalizes the support coordinate. Our modification reverses this ordering to promote sparse recovery.

Definition 3.1. (Gain Control (GC) scheme). Let $\beta = u \odot v$, where $u, v \in \mathbb{R}^d$ are the layer weights, and learning rate $\gamma > 0$. Let $z_k = \gamma\nabla\mathcal{L}(\beta_k)$ and $M_k = \max_j(1 + z_{k,j}^2 + z_{k,j}^4)$. We define

$$u_{k+1,j} = \frac{(1 + z_{k,j}^2)u_{k,j} - z_{k,j}v_{k,j}}{\sqrt{M_k}}, \quad \text{and} \quad v_{k+1,j} = \frac{(1 + z_{k,j}^2)v_{k,j} - z_{k,j}u_{k,j}}{\sqrt{M_k}}. \tag{12}$$

The addition of the quadratic term in the iterative scheme in Eq. (12) reinforces the existing weights. The normalization introduces a common contribution to the gain across all coordinates. The quadratic correction offsets this contribution more strongly on coordinates with larger gradients. Together, they concentrate gain on coordinates with smaller gradients—corresponding to the off-support coordinates in our example—and thereby strengthen their associated implicit penalty.

We highlight the net effect on the Gain:

$$\Delta Q_{k,j} = \underbrace{\log M_k}_{\text{shared gain}} - \underbrace{\log(1 + z_{k,j}^2 + z_{k,j}^4)}_{\text{larger gradients} \Rightarrow \text{less gain}}.$$

As discussed in Thm. 2.4, GD induces larger Gain on the support through its larger gradients, thereby increasing the implicit penalty on the support and favoring a denser solution. The update in Definition 3.1 reverses this effect. It concentrates Gain on coordinates with smaller gradients, which in our setting are the off-support coordinates. This strengthens their implicit penalty and promotes sparse recovery without requiring prior knowledge of the support. In contrast to SGD, which tends to homogenize the Gain across coordinates, our scheme instead deliberately redistributes the Gain toward the off-support coordinates. The following lemma makes this modification explicit.

Lemma 3.2. *Let $(u_k, v_k)_{k\geq 0}$ be the iterates generated by the update in Definition 3.1. For every $K \geq 1$ and $i \in \{1, \dots, d\}$, the accumulated Gain and Drift are as follows:*

$$Q_{K,i}^{\mathrm{GC}} = \sum_{k=0}^{K-1} \log\left(\frac{M_k}{1 + z_{k,i}^2 + z_{k,i}^4}\right) \quad \textit{and} \quad D_{K,i}^{\mathrm{GC}} = \sum_{k=0}^{K-1} \left[\log\left(\frac{1 + z_{k,i} + z_{k,i}^2}{1 - z_{k,i} + z_{k,i}^2}\right) - 2z_{k,i}\right].$$

The implicit bias of the resulting dynamics however retains the same structure as GD.

Proposition 3.3. *(Implicit Bias of Gain Control (GC) scheme) Suppose GC converges to an interpolator $\beta_\infty \in \mathcal{M}$ such that $X\beta_\infty = y$ and the Gain and Drift converge to finite Q_∞ and D_∞. Then, for a scale $a_{\infty,i} = \alpha^2 e^{-Q_{\infty,i}}$ in Eq. (4) and potential function $\psi_a(\beta)$ as defined in Eq. (2), we have*

$$\beta_\infty = \arg\min_{\beta \in \mathcal{M}} \left\{\psi_{a_\infty}(\beta) + \frac{1}{2}\langle D_\infty, \beta\rangle\right\}.$$

Thus, GC preserves the implicit-bias formulation of GD in Eq. (8) but with the modified Gain and Drift from Lemma 3.2. The preceding characterization is conditional on convergence. We therefore show next that, for sufficiently small constant step size, the full-batch GC dynamics indeed converge to an interpolating solution and admit finite limiting Gain and Drift.

Theorem 3.4. *(Convergence of the Gain Control scheme) Assume $X \neq 0$, $y \in \mathrm{range}(X)$, and $u_{0,i} > |v_{0,i}|$ for all i. For a fixed b satisfying $Xb = y$, set $a_{0,i} = (u_{0,i}^2 - v_{0,i}^2)/2$, and define $\kappa = \frac{1}{n}\max_{1\leq i\leq d} \|X_{:,i}\|_2^2$, $R = 4\mathcal{D}_{a_0}(b, \beta_0) + 2\|a_0\|_1 + 2\|b\|_1 \left[1 + \log\left(1 + \frac{4\|b\|_1}{\min_i a_{0,i}}\right)\right]$, where $\mathcal{D}_a$ is the Bregman divergence generated by ψ_a. For every constant step size $0 < \gamma \leq (32\kappa R)^{-1}$, the full-batch iterates of the Gain Control scheme in Eq. (12) satisfy $(u_k, v_k) \longrightarrow (u_\infty, v_\infty)$ and $X(u_\infty \odot v_\infty) = y$, while the accumulated Gain and Drift converge to finite limits $Q_\infty^{\mathrm{GC}}, D_\infty^{\mathrm{GC}} \in \mathbb{R}^d$.*

4 Observation in practice

Experiments. While our Gain Control scheme converges provably for smaller learning rates (Thm. 3.4), we wish to test sparse recovery for larger learning rates experimentally. First, we consider noiseless sparse regression with (n=48) observations and (d=768) features, using a normalized Gaussian design matrix X with teacher generated columns and labels ($y = X\beta^{\mathrm{sp}}$) where $|\mathrm{supp}(\beta^{\mathrm{sp}})| = 4$. We train a 2-layer diagonal linear networks initialized with $\alpha = 0.03$ for 10,000 epochs, comparing full-batch GD, SGD with batch size 12, and our Gain Control scheme (GC). We use a learning rate $\gamma = 0.35$ for the first 500 epochs, then decrease it to $\gamma = 0.05$ to ensure convergence. All the runs reach the EoS and oscillate around the stability threshold. Fig. 4 shows that GC recovers the sparse teacher much better than (S)GD. The learning rate switch helps only convergence, but has no effect on the performance of (S)GD, whose Gain and Drift prevents accurate recovery. We provide further experiments in App. A.1.

Exploratory analysis of transformer query-key pairs. As an exploratory test of whether the Gain/Drift competition extends beyond diagonal linear networks, we consider transformers with query-key product $\mathbf{QK}^\top$, whose singular values govern the effective rank of the attention logits, and define a per-coordinate Gain and Drift on its singular values in a frame fixed at initialization, in

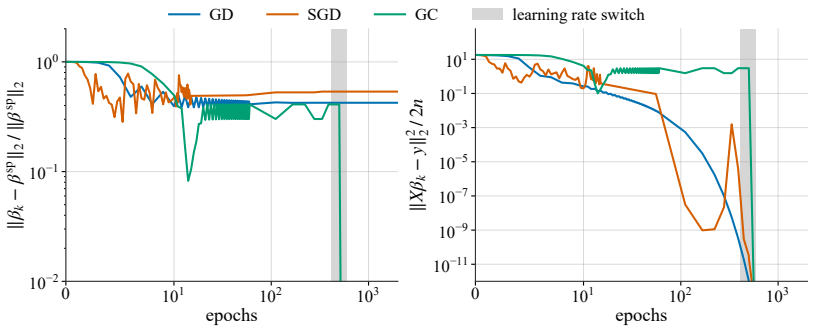


Figure 4: Sparse recovery for GD, SGD and GC on the noiseless sparse regression problem. (*Left*) Distance to the sparse teacher. Due to its implicit bias (Prop. 3.3), GC recovers a predictor substantially closer to the sparse teacher. (*Right*) Loss curve for the three methods. The learning rate switch is to allow GC to converge.

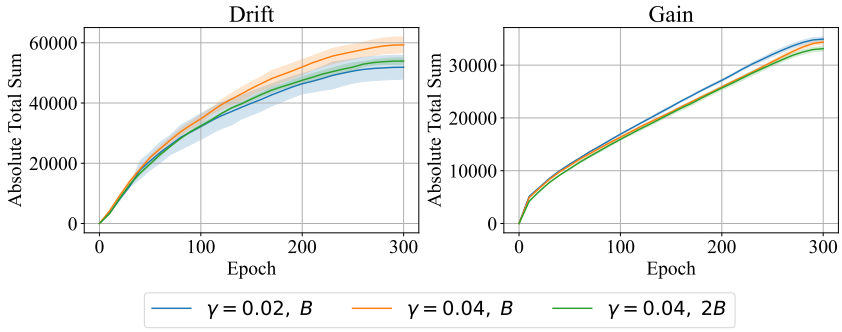


Figure 5: Accumulated Drift and Gain of all Query-Key pairs for a DeiT-Small trained on ImageNet with Muon, across three (batch size, learning rate) configurations. Both quantities grow substantially early in training, but Drift starts plateauing while Gain continues to accumulate throughout training.

analogy with Definitions 1.2 and 2.1.[1] Sparsity of this now corresponds to low rank. We pretrain a DeiT-Small (Touvron et al., 2020) on ImageNet (Deng et al., 2009) for 300 epochs with Muon (Jordan et al., 2024), storing snapshots every 10 epochs and computing per-head Gain and Drift in this fixed frame (Fig. 5). Consistent with the diagonal-network picture, Drift accumulates significantly earlier. It then starts plateauing while Gain keeps growing. Across our three configurations (Table 2), the runs with the largest Drift at epoch 300 also attain the lowest effective rank, hinting that the analyzed competition may shape solution sparsity also in this setting.

5 Conclusion

This work connects training dynamics at large learning rates with the sparsity of solutions in linear regression settings. Central to this connection is a hidden driver, which we term the Drift, that implicitly competes with the Gain to shape the early stages of training, which is consequential for model selection and thus generalization. Using diagonal linear networks as an analytically tractable framework, we characterize the role of the Drift and demonstrate experimentally that these concepts could extend to modern architectures and optimization algorithms. Our findings further highlight how understanding trajectory-dependent quantities such as the Drift and Gain can inform the development of methods that exploit their behavior. As a first step in this direction, we introduce a Gain Control scheme that biases the optimization trajectory toward sparser solutions, even in the EoS regime. Extending this scheme and evaluating its applicability to modern networks and architectures are promising directions for future work.

[1]This ignores rotation of the singular subspaces and the fact that Muon is not coordinatewise GD. So we treat the result as suggestive only, not a validation of the theory.

AI USE STATEMENT

In this work, we used generative AI tools for overall writing, polishing statements and presentation. We have not used generative AI tools for idea generation, identification of research gaps, or critical explanations. All ideas were developed by the authors. Additionally, we used generative AI tools for assist with the presentation of the proof of Theorem 3.4, as well as assisting in running toy experiments and preliminary data analysis. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Proofs of all lemmas, theorems and propositions discussed in the main section have been provided in App. C. Relevant code for the plots have been attached as supplementary material. Details of the experiments are provided in App. A.1 and D.

REFERENCES

Maksym Andriushchenko, Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Sgd with large step sizes learns sparse features. *ArXiv*, abs/2210.05337, 2022. URL https://api.semanticscholar.org/CorpusID:252815402.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/c0c783b5fc0d7d808f1d14a6e9c8280d-Paper.pdf.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on edge of stability in deep learning. *ArXiv*, abs/2205.09745, 2022. URL https://api.semanticscholar.org/CorpusID:248887658.

David Barrett and Benoit Dherin. Implicit gradient regularization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=3q5IqUrkcF.

Emmanuel Candes, Justin Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements, 2005. URL https://arxiv.org/abs/math/0503066.

Hung-Hsu Chou, Johannes Maly, and Holger Rauhut. More is less: Inducing sparsity via overparameterization. *ArXiv*, abs/2112.11027, 2021. URL https://api.semanticscholar.org/CorpusID:245353663.

Hung-Hsu Chou, Holger Rauhut, and Rachel A. Ward. Robust implicit regularization via weight normalization. *ArXiv*, abs/2305.05448, 2023. URL https://api.semanticscholar.org/CorpusID:258564503.

Hung-Hsu Chou, Johannes Maly, and Dominik Stöger. How to induce regularization in linear models: A guide to reparametrizing gradient flow, 2024.

Albert Cohen, Wolfgang Dahmen, and Ronald DeVore. Compressed sensing and best k-term approximation. *Journal of the American Mathematical Society*, 22:211–231, 07 2008. doi: 10.1090/S0894-0347-08-00610-3.

Jeremy Cohen, Alex Damian, Ameet Talwalkar, Zico Kolter, and Jason D. Lee. Understanding optimization in deep learning with central flows. *ArXiv*, abs/2410.24206, 2024. URL https://api.semanticscholar.org/CorpusID:273707747.

Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *ArXiv*, abs/2103.00065, 2021. URL https://api.semanticscholar.org/CorpusID:232076011.

Jeremy M. Cohen, B. Ghorbani, Shankar Krishnan, Naman Agarwal, Sourabh Medapati, Michal Badura, Daniel Suo, David E. Cardoze, Zachary Nado, George E. Dahl, and Justin Gilmer. Adaptive gradient methods at the edge of stability. *ArXiv*, abs/2207.14484, 2022. URL https://api.semanticscholar.org/CorpusID:251196732.

Alexandru Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. *ArXiv*, abs/2209.15594, 2022. URL https://api.semanticscholar.org/CorpusID:252668622.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Mathieu Even, Scott Pesme, Suriya Gunasekar, and Nicolas Flammarion. (s)gd over diagonal linear networks: Implicit bias, large stepsizes and edge of stability. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 29406–29448. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/5da6ce80e97671b70c01a2e703b868b3-Paper-Conference.pdf.

Vít Fojtík, Maria Matveev, Hung-Hsu Chou, Gitta Kutyniok, and Johannes Maly. Conflicting biases at the edge of stability: Norm versus sharpness regularization. *ArXiv*, abs/2505.21423, 2025. URL https://api.semanticscholar.org/CorpusID:278911796.

Leonardo Galli, Curtis Fox, Wiebke Bartolomaeus, Mark Schmidt, and Holger Rauhut. Flatland: The adventures of gradient descent with large step sizes, 2026. URL https://arxiv.org/abs/2606.06722.

Benjamin Gess and Daniel Heydecker. Large spikes in stochastic gradient descent: A large-deviations view. *ArXiv*, abs/2603.10079, 2026. URL https://api.semanticscholar.org/CorpusID:286457095.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/58191d2a914c6dae66371c9dcdc91b41-Paper.pdf.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1832–1841. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/gunasekar18a.html.

Rustem Islamov, Michael Crawshaw, Jeremy Cohen, and Robert M. Gower. Non-euclidean gradient descent operates at the edge of stability. *ArXiv*, abs/2603.05002, 2026. URL https://api.semanticscholar.org/CorpusID:286256457.

Tom Jacobs and Rebekka Burkholz. Mask in the mirror: Implicit sparsification. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=U47ymTS3ut.

Tom Jacobs and Guido Montúfar. Implicit bias of mirror flow in homogeneous neural networks: Sparse and dense feature learning. 2026. URL https://api.semanticscholar.org/CorpusID:288654339.

Tom Jacobs, Chao Zhou, and Rebekka Burkholz. Mirror, mirror of the flow: How does regularization shape implicit bias? In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=MLiR9LS5PW.

Tom Jacobs, Chao Zhou, and Rebekka Burkholz. Never saddle for reparameterized steepest descent as mirror flow. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=YgudIlQ9nC.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Chris Kolb, Tobias Weber, Bernd Bischl, and David Rügamer. Deep weight factorization: Sparse learning through the lens of artificial symmetries. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vNdOHr7mn5.

Itai Kreisler, Mor Shpigel Nacson, Daniel Soudry, and Yair Carmon. Gradient descent monotonically decreases the sharpness of gradient flow solutions in scalar networks and beyond. *ArXiv*, abs/2305.13064, 2023. URL https://api.semanticscholar.org/CorpusID:258832278.

Shuang Liang and Guido Montúfar. Gradient descent with large step sizes: Chaos and fractal convergence region. *ArXiv*, abs/2509.25351, 2025. URL https://api.semanticscholar.org/CorpusID:281681436.

Shuang Liang and Guido Montúfar. Implicit bias of mirror flow for shallow neural networks in univariate regression. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=IF0Q9KY3p2.

Shuang Liang, Tom Jacobs, and Guido Montúfar. Implicit bias of sgd in multivariate relu networks: Effective width collapse. *ArXiv*, abs/2607.03613, 2026. URL https://api.semanticscholar.org/CorpusID:289934596.

Tongtong Liang, Alexander Cloninger, Rahul Parhi, and Yu-Xiang Wang. Generalization below the edge of stability: The role of data geometry. *ArXiv*, abs/2510.18120, 2025. URL https://api.semanticscholar.org/CorpusID:282246553.

Philip M. Long and Peter L. Bartlett. Sharpness-aware minimization and the edge of stability. *Journal of Machine Learning Research*, 25(179):1–20, 2024. URL http://jmlr.org/papers/v25/23-1285.html.

Hancheng Min, Lachlan Ewen MacDonald, and René Vidal. On the convergence, implicit bias, and edge of stability of gradient descent in deep learning: Reviewing recent progress [special issue on the mathematics of deep learning]. *IEEE Signal Processing Magazine*, 43(3):37–51, 2026. doi: 10.1109/MSP.2026.3665010.

Rotem Mulayoff and Sebastian U. Stich. On the stability of nonlinear dynamics in gd and sgd: Beyond quadratic potentials. In *Annual Conference Computational Learning Theory*, 2026. URL https://api.semanticscholar.org/CorpusID:285615894.

Sree Harsha Nelaturu, Advait Gadhikar, and Rebekka Burkholz. TurboPrune: High-Speed Distributed Lottery Ticket Training. URL https://github.com/nelaturuharsha/TurboPrune.

Scott Pesme and Nicolas Flammarion. Saddle-to-saddle dynamics in diagonal linear networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=iuqCXg1Gng.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of SGD for diagonal linear networks: a provable benefit of stochasticity. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=vvi7KqHQiA.

E. R. Regis and Sinho Chewi. Rod flow: A continuous-time model for gradient descent at the edge of stability. *ArXiv*, abs/2602.01480, 2026. URL https://api.semanticscholar.org/CorpusID:285270093.

Anna Shalova, André Schlichting, and Mark A. Peletier. Singular-limit analysis of gradient descent with noise injection. *ArXiv*, abs/2404.12293, 2024. URL https://api.semanticscholar.org/CorpusID:269213972.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 2020. URL https://api.semanticscholar.org/CorpusID:229363322.

Mario Tuci, Caner Korkmaz, Umut cSimcsekli, and Tolga Birdal. Generalization at the edge of stability. *ArXiv*, abs/2604.19740, 2026. URL https://api.semanticscholar.org/CorpusID:287636709.

Tomas Vaškevičius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery, 2019. URL https://arxiv.org/abs/1909.05122.

Yuqing Wang, Zhenghao Xu, Tuo Zhao, and Molei Tao. Good regularity creates large learning rate implicit biases: edge of stability, balancing, and catapult. *Journal of Machine Learning Research*, 26(273):1–68, 2025. URL http://jmlr.org/papers/v26/23-1691.html.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3635–3673. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/woodworth20a.html.

Liu Ziyin and Zihao Wang. spred: Solving l1 penalty with sgd. In *International Conference on Machine Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:259075663.

A EXPERIMENTS

A.1 NOISELESS SPARSE REGRESSION

In this section, we show that GC converges on the noiseless regression problem $y = X\beta$ for moderate learning rates. We use a training setup of $(n, d, \alpha) = (48, 768, 0.03)$. We train a diagonal linear network using GD with learning rate $\gamma = 0.1$ for 10,000 epochs. We use a batch size of 4 for SGD. Fig 6 shows that GC recovers the sparse teacher much better than (S)GD, although converging slower.

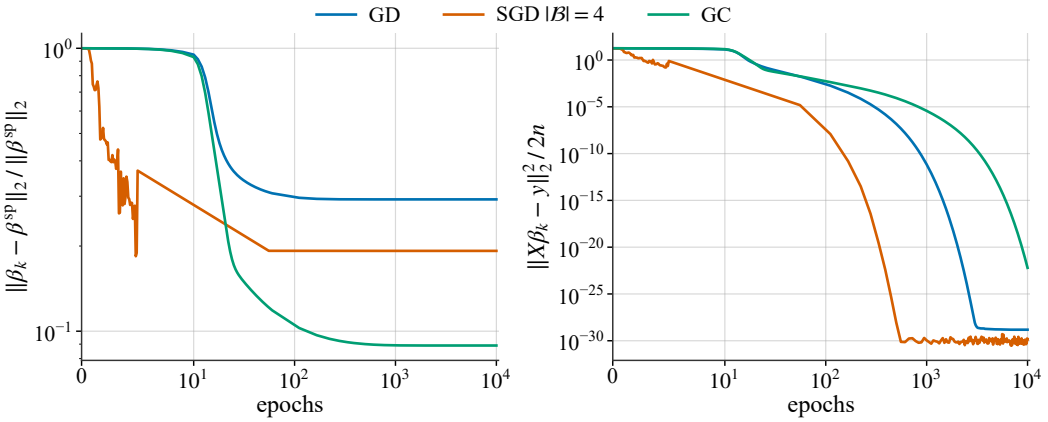


Figure 6: Sparse recovery in noiseless regression with $\gamma = 0.1$. Left: normalized recovery error. Right: normalized training loss. GC achieves substantially lower recovery error although it converges slower compared to (S)GD.

We show in Fig 4 in Sec §4 that GC recovers the sparse teacher for larger learning rates ($\gamma = 0.35$), which is lowered to $\gamma = 0.05$ for convergence. Next, we further experiment with cosine annealing as a learning rate schedule, where

$$\gamma_k = \gamma_{\min} + (\gamma_{\max} - \gamma_{\min}) \cdot \frac{1 + \cos(\pi k / K)}{2}$$

with $(\gamma_{\min},\ \gamma_{\max},\ K) = (0.05,\ 0.35,\ 10000)$. Fig. 7 shows that the recovery advantage of GC also holds under cosine annealing in noiseless settings.

In the same setup as that of the noiseless regression (App. A.1) with $(n, d, \alpha) = (48, 768, 0.03)$ and a 4-sparse teacher β^{sp}, trained with GD at $\gamma = 0.35$ upto 500 iterations then dropped to $\gamma = 0.05$. Fig. 8 shows that ignoring the drift in the potential recovers an incorrect, denser solution and is needed to approximate the GD solution accurately. We also show the accumulated Gain and Drift during the oscillatory phase of EoS to highlight Lemma 2.6 in Fig. 9.

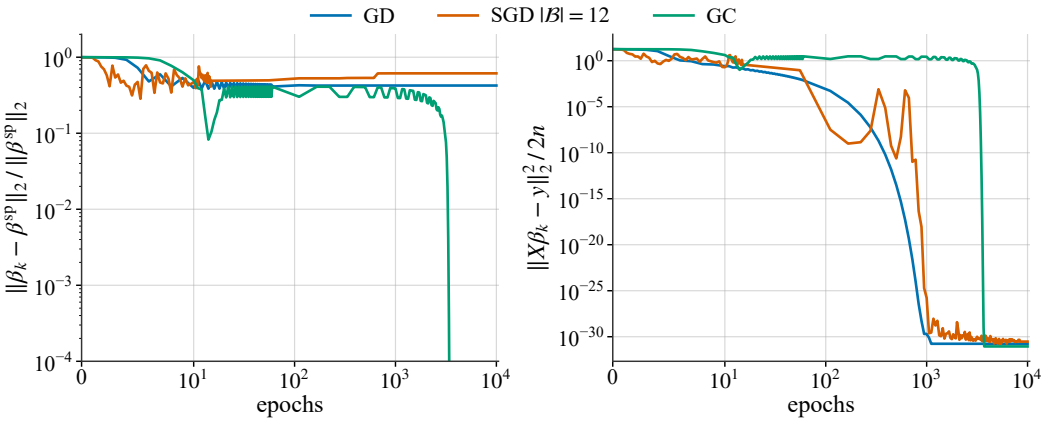


Figure 7: Sparse recovery in noiseless regression with cosine annealing and an initial large initial learning rate, $\gamma = 0.35$. After 500 iterations, the learning rate is reduced to $\gamma = 0.05$ (shaded region). Left: normalized distance to the sparse teacher. Right: normalized training loss; GC recovers the sparse teacher substantially better than (S)GD.

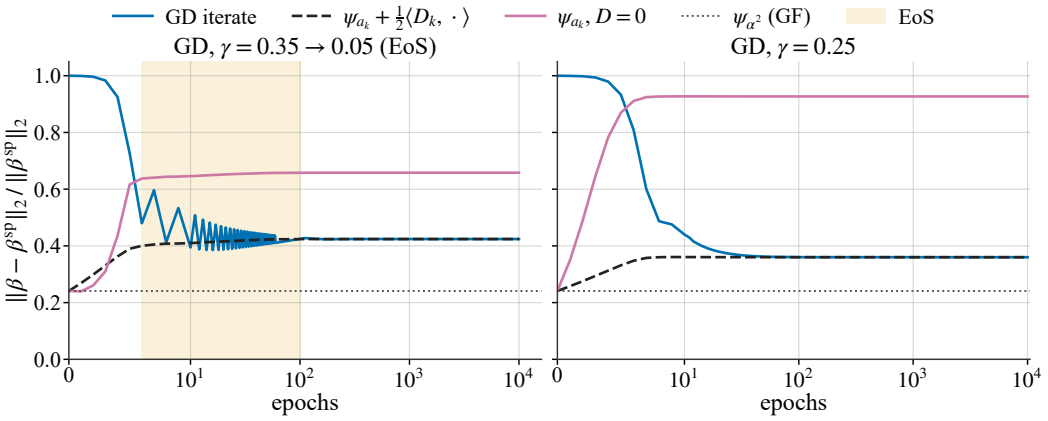


Figure 8: The Drift is necessary for sparse recovery. At every checkpoint k we minimise the implicit-bias potential over the interpolation manifold $\mathcal{M}$, using the Gain Q_k and Drift D_k accumulated so far, and plot the distance of that minimizer to the sparse teacher. With the Drift, $\psi_{a_k} + \frac{1}{2}\langle D_k, \cdot \rangle$ (dashed) converges exactly to the GD iterate (blue), as predicted by Prop. 2.3. Keeping the same Gain but setting $D = 0$ (purple) selects a much denser interpolator. The gradient-flow potential ψ_{α^2} (dotted, $Q = D = 0$) is shown for reference. GF recovers the sparsest solution. The shaded region marks the oscillatory phase ($\gamma_k S_k \geq 2$)

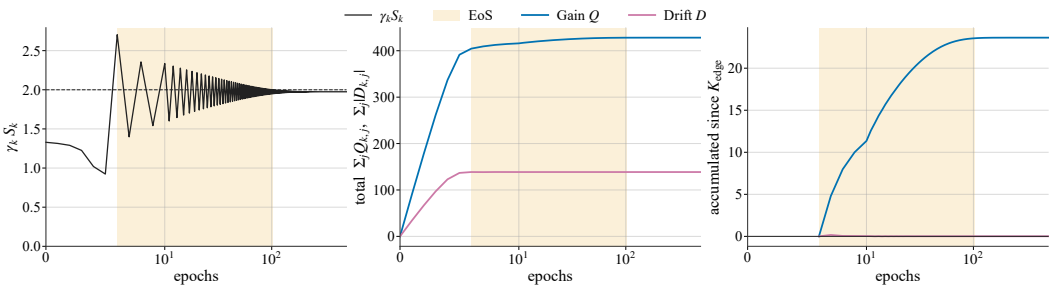


Figure 9: Following the onset of EoS oscillations, Gain continues to accumulate while the net change in Drift remains small. Measuring both quantities relative to K_{edge} makes this separation explicit, supporting the observed saturation of Drift during the oscillatory phase.

B Extended Related Works

Implicit regularization. Optimization algorithms can favor particular interpolators even without explicit regularization. In linear networks, this phenomenon has been studied through matrix factorization, sparse recovery, and the transition between kernel and feature-learning regimes (Gunasekar et al., 2017; Arora et al., 2019; Vaškevičius et al., 2019; Woodworth et al., 2020). Reparameterization, normalization, and optimization geometry provide further ways to influence solution selection (Chou et al., 2021; 2023; 2024; Gunasekar et al., 2018; Jacobs & Burkholz, 2025; Jacobs et al., 2025; Jacobs & Montúfar, 2026; Jacobs et al., 2026; Liang & Montúfar, 2025; Liang et al., 2026; Kolb et al., 2025; Ziyin & Wang, 2022). Our work studies how discrete optimization modifies the implicit bias of diagonal linear networks, producing competing tendencies toward sparse and dense interpolators.

Stochasticity and discretization. Stochasticity and large learning rates can substantially alter implicit regularization and convergence (Pesme et al., 2021; Andriushchenko et al., 2022; Liang et al., 2026; Galli et al., 2026; Shalova et al., 2024), while small-initialization analyses reveal saddle-to-saddle feature learning (Pesme & Flammarion, 2023). Most closely related, Even et al. (2023) characterize the implicit bias of discrete GD and SGD in diagonal networks through an effective initialization and a shifted Bregman reference point. We build on this trajectory-dependent characterization by analyzing the competition between Gain and Drift across progressive sharpening and oscillatory training. Backward error analysis also connects discretization to implicit regularization (Barrett & Dherin, 2021), while studies of sharpness-aware minimization illustrate how modified updates affect solution selection (Long & Bartlett, 2024). Our Gain Control algorithm uses the identified mechanism to actively steer the recovered interpolator.

Dynamics at the Edge of Stability. Cohen et al. (2021) document progressive sharpening followed by persistent oscillations near the stability threshold. Subsequent analyses explain curvature-dependent dynamics and self-stabilization (Arora et al., 2022; Damian et al., 2022), with extensions to adaptive methods and non-Euclidean geometries (Cohen et al., 2022; Islamov et al., 2026). Central flow and Rod Flow approximate the evolution of oscillatory GD trajectories (Cohen et al., 2024; Regis & Chewi, 2026). Related work connects large learning rates to balancing and catapult dynamics (Wang et al., 2025), and examines chaos, stochastic loss spikes, and nonlinear stability (Liang & Montúfar, 2025; Gess & Heydecker, 2026; Mulayoff & Stich, 2026). Our analysis connects these training phases to the accumulated quantities determining sparsity: Gain grows monotonically and favors density in our setting, whereas Drift can counteract this tendency before the oscillatory phase begins.

Sharpness and competing biases. Kreisler et al. (2023) show that GD monotonically reduces the sharpness of the corresponding gradient-flow solution in scalar networks, providing a connection between discrete updates and the sharpness of the eventual solution. Complementarily, Fojtík et al. (2025) demonstrate that norm and sharpness biases can compete in diagonal linear networks, with neither alone explaining generalization. We identify how Gain and Drift govern this competition in our setting, using a minimal network with explicit sparse and dense interpolators to determine where each effect dominates. We further show that controlling Drift can recover sharper sparse solutions. Related work studies generalization through long-term EoS dynamics (Tuci et al., 2026) and below EoS (Liang et al., 2025); a broader overview of convergence, implicit bias, and EoS appears in Min et al. (2026).

C Omitted Proofs

C.1 Proofs of Section 2

Proof of Proposition 2.3.

Proof. Since $q_\pm = q \pm d$, the effective branch scales satisfy $a_{\pm,\infty} = a_\infty \odot \exp(\mp D_\infty)$. Consequently, the perturbed reference point is

$$\widetilde{\beta}_0 = \frac{1}{2}(a_{+,\infty} - a_{-,\infty}) = -a_\infty \odot \sinh(D_\infty).$$

Since $\nabla\psi_a(\beta) = \frac{1}{2}\operatorname{arsinh}(\beta/a)$, we get $\nabla\psi_{a_\infty}(\widetilde{\beta}_0) = -\frac{1}{2}D_\infty$. Expanding the Bregman divergence therefore gives

$$\mathcal{D}_{\psi_{a_\infty}}(\beta, \widetilde{\beta}_0) = \psi_{a_\infty}(\beta) + \frac{1}{2}\langle D_\infty, \beta\rangle + C,$$

where C is independent of β. Dropping this constant proves the claim. □

Proof of Lemma 2.5.

Proof. Let the accumulated drift in the progressive sharpening phase be $D_{K_{\mathrm{ps}}}$. By equation 9, every coordinate of every z_k lies in $[-\rho, \rho] \subset (-1, 1)$, so $D_{K_{\mathrm{ps}}}$ is well defined.

The function d is smooth and odd on $(-1, 1)$, with $d(0) = 0$ and

$$d'(z) = \frac{1}{1+z} + \frac{1}{1-z} - 2 = \frac{2z^2}{1-z^2}, \qquad |z| < 1.$$

Hence, by the fundamental theorem of calculus, for every $s \in [0, \rho]$,

$$d(s) = 2\int_0^s \frac{t^2}{1-t^2}\, dt.$$

For $t \in [0, s] \subseteq [0, \rho]$ we have $1 \le (1-t^2)^{-1} \le (1-\rho^2)^{-1}$. Integrating these bounds against $2t^2$ over $[0, s]$ yields

$$\frac{2}{3}\, s^3 \;\le\; d(s) \;\le\; \frac{2}{3(1-\rho^2)}\, s^3, \qquad s \in [0, \rho]. \tag{13}$$

Fix $j \in \{1, \dots, d\}$ and $0 \le k < K_{\mathrm{edge}}$. Since $\sigma_j \in \{-1, 1\}$ and $\sigma_j z_{k,j} \ge 0$, we have $\sigma_j z_{k,j} = |z_{k,j}|$. Oddness of d gives $d(\sigma_j z) = \sigma_j\, d(z)$, and therefore

$$\sigma_j\, d(z_{k,j}) = d(\sigma_j z_{k,j}) = d\big(|z_{k,j}|\big) \ge 0,$$

where nonnegativity follows from equation 13. Summing over k,

$$\sigma_j D_{K_{\mathrm{edge}},j} = \sum_{k=0}^{K_{\mathrm{edge}}-1} d\big(|z_{k,j}|\big) \;\ge\; 0,$$

and since $|\sigma_j| = 1$, it follows that $|D_{K_{\mathrm{edge}},j}| = |\sigma_j D_{K_{\mathrm{edge}},j}| = \sigma_j D_{K_{\mathrm{edge}},j}$.

Applying equation 13 to each summand with $s = |z_{k,j}| \in [0, \rho]$, and using $|z_{k,j}| = \big|\gamma_k\, \partial_j L(\beta_k)\big|$, gives equation 10. □

Proof of Lemma 2.6.

Proof. Fix ℓ and set $k = k_0 + 2\ell$. Since d is odd,

$$D_{k+2} - D_k = d(z_{k+1}) + d(z_k) = d(z_{k+1}) - d(-z_k),$$

which vanishes when $z_{k+1} = -z_k$. In general, $z_{k+1,j}$ and $-z_{k,j}$ both lie in $[-\rho, \rho]$, on which

$$|d'(z)| = \frac{2z^2}{1-z^2} \le \frac{2\rho^2}{1-\rho^2}.$$

The mean value theorem, applied coordinatewise, therefore gives

$$\|D_{k+2} - D_k\|_\infty \le \frac{2\rho^2}{1-\rho^2}\, \|z_{k+1} + z_k\|_\infty = \frac{2\rho^2}{1-\rho^2}\, \|e_\ell\|_\infty. \tag{14}$$

Telescoping over the m pairs and applying the triangle inequality,

$$\begin{aligned}
\|D_{k_0+2m} - D_{k_0}\|_\infty &= \Big\| \sum_{\ell=0}^{m-1} \big(D_{k_0+2\ell+2} - D_{k_0+2\ell}\big) \Big\|_\infty \\
&\le \sum_{\ell=0}^{m-1} \big\| D_{k_0+2\ell+2} - D_{k_0+2\ell} \big\|_\infty \\
&\le \frac{2\rho^2}{1-\rho^2} \sum_{\ell=0}^{m-1} \|e_\ell\|_\infty,
\end{aligned}$$

where the last inequality applies equation 14 to each pair. This is equation 11. □

Proof of Theorem 2.4.

Proof. Let $h = \hat{\beta} - \beta^*$ and $t = \|h\|_{S^C}$. Let $\mathcal{M}$ be the solution manifold i.e. $\mathcal{M} = \{\beta \in \mathbb{R}^d : X\beta = y\}$. Since both β^* and $\hat{\beta}$ satisfy regression $X\beta = y$, this implies $X(\beta^* - \hat{\beta}) = Xh = 0$ i.e. $h \in \ker X$. From the NSP assumption Eq. (NSP), it follows that

$$\|h\|_1 = \|h_S\|_1 + \|h\|_{S^C} \leq (1 + c)t$$

For $t = 0$, the assumption holds $h = 0$ i.e. $\beta^* = \hat{\beta}$ and the theorem holds trivially. We assume $t > 0$.

Since h is a direction along which the interpolator can move without leaving the solution manifold $\mathcal{M}$, by first order optimality

$$\langle \nabla\psi + \frac{1}{2}D, h\rangle = 0 \implies \langle \operatorname{arcsinh}\left(\frac{\hat{\beta}}{a}\right) + D, h\rangle = 0 \tag{15}$$

For every $j \in S$, by monotonicity of $\operatorname{arcsinh}$,

$$m_j := \left[\operatorname{arcsinh}\left(\frac{\hat{\beta}_j}{a_j}\right) - \operatorname{arcsinh}\left(\frac{\beta_j^\star}{a_j}\right)\right] h_j \geq 0, \tag{16}$$

since $\operatorname{sign} h_j = \operatorname{sign}(\hat{\beta}_j - \beta_j^\star) = \operatorname{sign}\left[\operatorname{arcsinh}\left(\frac{\hat{\beta}_j}{a_j}\right) - \operatorname{arcsinh}\left(\frac{\beta_j^\star}{a_j}\right)\right]$. For $j \in S^C$, $\beta_j^* = 0$ hence $h_j = \hat{\beta}_j$. Using this, from equation 15, we get

$$\operatorname{arcsinh}\left(\frac{\hat{\beta}_j}{a_j}\right) h_j = \operatorname{arcsinh}\left(\frac{h_j}{a_j}\right) h_j = |h_j| \operatorname{arcsinh}\left(\frac{|h_j|}{a_j}\right). \tag{17}$$

Splitting the inner product in equation 15 over S and S^c, adding and subtracting $\operatorname{arcsinh}(\beta_j^\star/a_j)\, h_j$ for each $j \in S$, and using equation 16 and equation 17, we obtain

$$\begin{aligned}
0 &= \sum_{j\in S} \operatorname{arcsinh}\left(\frac{\hat{\beta}_j}{a_j}\right) h_j + \langle D_S, h_S\rangle + \sum_{j\in S^c} \operatorname{arcsinh}\left(\frac{\hat{\beta}_j}{a_j}\right) h_j + \langle D_{S^c}, h_{S^c}\rangle \\
&= \sum_{j\in S} m_j + \left\langle \operatorname{arcsinh}\left(\frac{\beta_S^\star}{a_S}\right) + D_S,\, h_S \right\rangle \\
&\quad + \sum_{j\in S^c} |h_j| \operatorname{arcsinh}\left(\frac{|h_j|}{a_j}\right) + \langle D_{S^c}, h_{S^c}\rangle.
\end{aligned} \tag{18}$$

Solving equation 18 for the S^c sum and dropping $-\sum_{j\in S} m_j \leq 0$ gives

$$\begin{aligned}
\sum_{j\in S^c} |h_j| \operatorname{arcsinh}\left(\frac{|h_j|}{a_j}\right) &= -\sum_{j\in S} m_j - \left\langle \operatorname{arcsinh}\left(\frac{\beta_S^\star}{a_S}\right) + D_S,\, h_S \right\rangle - \langle D_{S^c},\, h_{S^c}\rangle \\
&\leq -\left\langle \operatorname{arcsinh}\left(\frac{\beta_S^\star}{a_S}\right) + D_S,\, h_S \right\rangle - \langle D_{S^c},\, h_{S^c}\rangle. && (19) \\
&\overset{\text{Hölder}}{\leq} R_S\|h_S\|_1 + \|D_{S^c}\|_\infty \|h_{S^c}\|_1 && (20) \\
&\leq (cR_S + \|D_{S^C}\|_\infty)\, t && (21)
\end{aligned}$$

Here, Hölder's inequality is applied separately on S and S^C.

$$|\langle u, v\rangle| \leq \|u\|_\infty \|v\|_1.$$

We now wish to find a lower bound on $\sum_{j\in S^c} |h_j| \operatorname{arcsinh}\left(\frac{|h_j|}{a_j}\right)$ in terms of t. Consider

$$\varphi(r) = r \operatorname{arcsinh}(r), \qquad r \geq 0.$$

This function is convex: since $\varphi'(r) = \operatorname{arcsinh}(r) + r/\sqrt{1+r^2}$,

$$\varphi''(r) = \frac{1}{\sqrt{1+r^2}} + \frac{1}{(1+r^2)^{3/2}} = \frac{r^2+2}{(1+r^2)^{3/2}} > 0.$$

Since $S \subsetneq \{1, \dots, d\}$, the set S^c is nonempty, and each $a_j > 0$, so $A_{S^c} > 0$. Hence the weights $w_j := a_j / A_{S^c}$, $j \in S^c$, are positive and sum to one, and the weighted mean of the points $r_j := |h_j|/a_j$ is

$$\sum_{j \in S^c} w_j r_j = \sum_{j \in S^c} \frac{a_j}{A_{S^c}} \cdot \frac{|h_j|}{a_j} = \frac{1}{A_{S^c}} \sum_{j \in S^c} |h_j| = \frac{t}{A_{S^c}}.$$

Jensen's inequality with these weights therefore gives

$$\begin{aligned} \sum_{j \in S^c} |h_j| \operatorname{arcsinh}\left(\frac{|h_j|}{a_j}\right) &= \sum_{j \in S^c} a_j \, \varphi\left(\frac{|h_j|}{a_j}\right) = A_{S^c} \sum_{j \in S^c} \frac{a_j}{A_{S^c}} \, \varphi\left(\frac{|h_j|}{a_j}\right) \\ &\geq A_{S^c} \, \varphi\left(\frac{t}{A_{S^c}}\right) \\ &= A_{S^c} \cdot \frac{t}{A_{S^c}} \operatorname{arcsinh}\left(\frac{t}{A_{S^c}}\right) = t \operatorname{arcsinh}\left(\frac{t}{A_{S^c}}\right). \end{aligned} \tag{22}$$

From equation 22, equation 19 and equation 21, we get

$$t \operatorname{arcsinh}\left(\frac{t}{A_{S^c}}\right) \leq \sum_{j \in S^c} |h_j| \operatorname{arcsinh}\left(\frac{|h_j|}{a_j}\right) \leq \left(c\, R_S + \|D_{S^c}\|_\infty\right) t.$$

Dividing by $t > 0$, we obtain

$$\operatorname{arcsinh}\left(\frac{t}{A_{S^c}}\right) \leq c\, R_S + \|D_{S^c}\|_\infty. \tag{23}$$

Since $\sinh$ is strictly increasing on $\mathbb{R}$ ($\sinh' = \cosh > 0$) and $\sinh(\operatorname{arcsinh}(x)) = x$ for every $x \in \mathbb{R}$, applying $\sinh$ to both sides of equation 23 preserves the inequality:

$$\frac{t}{A_{S^c}} = \sinh\left(\operatorname{arcsinh}\left(\frac{t}{A_{S^c}}\right)\right) \leq \sinh\left(c\, R_S + \|D_{S^c}\|_\infty\right).$$

Multiplying both sides by $A_{S^c} > 0$ gives

$$t \leq A_{S^c} \, \sinh\left(c\, R_S + \|D_{S^c}\|_\infty\right). \tag{24}$$

Since $\beta^\star_{S^c} = 0$, we have $h_{S^c} = \hat{\beta}_{S^c}$, so by the definition of t and equation 24,

$$\|\hat{\beta}_{S^c}\|_1 = \|h_{S^c}\|_1 = t \leq A_{S^c} \, \sinh\left(c\, R_S + \|D_{S^c}\|_\infty\right),$$

Finally, since $h = \hat{\beta} - \beta^\star$, combining $\|h\|_1 \leq (1+c)\, t$ with equation 24 gives

$$\|\hat{\beta} - \beta^\star\|_1 = \|h\|_1 \leq (1+c)\, t \leq (1+c)\, A_{S^c} \, \sinh\left(c\, R_S + \|D_{S^c}\|_\infty\right),$$

which is the bound of Theorem 2.4. □

Extended Discussion We summarize the factors that influence the sparsity of the solution

1. We define the set of feasible directions as $\mathcal{S}(\beta^*) = \{h \in \ker X : \operatorname{sign}(\beta^*_j) h_j \leq 0 \, \forall j \in \operatorname{supp}(\beta^*)\}$, such that moving along any $h \in \mathcal{S}$ reduces coefficient in the support coordinates and increases it in the offsupport coordinate. The alignment of the Drift with these set of directions determines its behavior.
2. For small initialization α, the entropy function resembles the following

$$\widetilde{\psi}(\beta, a, D) := \psi_a(\beta) + \frac{1}{2}\langle D, \beta \rangle \tag{25}$$

$$= \log\left(\frac{1}{\alpha}\right) \|\beta\|_1 + \frac{1}{2} \sum_{i=1}^{d} \operatorname{Gain}_\gamma(i) |\beta_i| + \frac{1}{2} \sum_{i=1}^{d} \operatorname{Drift}_\gamma(i) \beta_i + O_\beta(1). \tag{26}$$

Therefore, the magnitude of the well-aligned Drift in the support coordinates is also important so as to be able to appropriately offset the effects of the Gain.

C.2 Example case for Drift and Gain

In this appendix, we provide detailed calculations for the example case in §2.2. We restate the regression problem here for completeness

$$X = \begin{pmatrix} \sqrt{2} & 0 & 1 \\ 0 & \sqrt{2} & 1 \end{pmatrix} \in \mathbb{R}^{2\times 3}, \qquad y = B\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad B \in \mathbb{R}_+ \tag{27}$$

where the two solutions of interest are the sparse $\beta_s^* = B\,(0,0,1)^\top$ and the dense solution $\beta_d^* = B\,(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0)$. We consider the space of interpolators that lie on the line joining these two solutions as $\mathcal{S} = \{\beta(t) : t \in [0,1]\}$, where

$$\beta(t) = (1-t)\beta_s^* + t\beta_d^* = B\left(\frac{t}{\sqrt{2}}, \frac{t}{\sqrt{2}}, 1-t\right)^\top \qquad h = \left(\tfrac{1}{\sqrt{2}}, \tfrac{1}{\sqrt{2}}, -1\right). \tag{28}$$

Equivalently, $\beta(t) = \beta_s^* + Bth$. The vector $h \in \ker X$ since

$$Xh = \begin{pmatrix} \sqrt{2} & 0 & 1 \\ 0 & \sqrt{2} & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Norm The norm of the interpolator along the line is

$$\|\beta(t)\|_1 = B\big(\frac{t}{\sqrt{2}} + \frac{t}{\sqrt{2}} + 1 - t\big) = B\big(1 + (\sqrt{2}-1)t\big)$$

Therefore, along the line the norm increases, with the sparser $\|\beta_s^*\| = B$ and the denser $\|\beta_d^*\| = \sqrt{2}B$.

Sharpness Under the squared loss function $\mathcal{L} = \frac{1}{4}\|X\beta - y\|^2$, the Hessian is as follows

$$H = X\,\mathrm{diag}(\beta(t))X^\top = B\begin{pmatrix} 1-t+\sqrt{2}t & 1-t \\ 1-t & 1-t+\sqrt{2}t \end{pmatrix}. \tag{29}$$

The eigenvalues of the Hessian are

$$\lambda_1(t) = B\big(2(1-t) + \sqrt{2}t\big), \qquad \lambda_2(t) = \sqrt{2}B$$

Since sharpness $\bar{S} = \lambda_{max}(H)$, and $\lambda_1(t) \geq \lambda_2(t)$ on $t \in [0,1]$, the minimum sharpness is

$$\bar{S}_\infty(t) = B\big(2(1-t) + \sqrt{2}t\big) = 2B(1-t) + \sqrt{2}Bt$$

Since $\bar{S}'(t) = B(\sqrt{2}-2) < 0$, sharpness decreases linearly from the sparse solution ($\bar{S}(0) = 2B$ to the dense solution $\bar{S}(1) = \sqrt{2}B$.

In the following sections, we consider the standard initialization of $\beta = u \odot v$, with $u_0 = \sqrt{2}\alpha$ and $v = \mathbf{0}$. Hence the residual $r_0 = \sqrt{2}\beta_{0,1} + \beta_{0,3} - B = -B < 0$. Let $z_{k,j} = \gamma\partial_{\beta_j}\mathcal{L}(\beta_k)$. It follows that for the off-support coordinates $z_{k,o} = \frac{1}{\sqrt{2}}\gamma r_k$, while for the support coordinates $z_{k,s} = \gamma r_k$.

Alignment induced by Gain We measure the alignment of the Gain by considering the direction towards which the change in the potential induced by the gain points at. Taking the derivative of the potential function $\psi_a(\beta)$

$$\partial_{\beta_j}\psi_a(\beta) = \frac{1}{2}\mathrm{arcsinh}\big(\frac{\beta_j}{a_j}\big)$$

From Eq. 8, an additional Gain ΔQ diverges the branch initialization as follows

$$a_j' = a_j e^{-\Delta Q_j}$$

The change in the potential gradient is

$$C_j := \partial_{\beta_j}(\psi_{a'} - \psi_a) = \frac{1}{2}\big[\mathrm{arcsinh}\big(\frac{\beta_j e^{\Delta Q_j}}{a_j}\big) - \mathrm{arcsinh}\big(\frac{\beta_j}{a_j}\big)\big]$$

. The inductive bias points towards decreasing potential. Consider moving along $+h$ by a factor of $\varepsilon > 0$. The induced change in the potential is

$$\Delta\psi_Q(\beta + \varepsilon h) - \Delta\psi_Q(\beta) = \varepsilon\langle h, C\rangle + O(\varepsilon^2) = -\varepsilon A_Q + O(\varepsilon^2)$$

Hence the alignment A_Q is

$$A_Q = -\langle h, C\,\rangle = C_s - \sqrt{2}C_o$$

Thus $A_Q > 0$ means the bias induced by Q points towards the dense solution, while $A_Q < 0$ means the bias is towards the sparse solution.

Alignment induced by the Drift The alignment of the Drift is simpler compared to the Gain, since it linearly affects the inductive bias formulation (Prop 2.3). Hence the alignment of the bias induced by the Drift is

$$A_D = -\frac{1}{2}\langle h, \Delta D\rangle = \frac{1}{2}(\Delta D_s - \sqrt{2}\Delta D_o)$$

Since the drift increment Eq. (6) is odd and the residual $r < 0$ in the progressive sharpening phase,

$$A_d(r) = -\frac{1}{2}(\sqrt{2}d(\frac{\gamma r}{\sqrt{2}}) - d(\gamma r))$$

Similar to the Gain alignment, $A_D > 0$ indicates that the Drift biases towards the dense solution and vice versa. In Fig. 2, we plot $-A_Q$ and $-A_D$ across the different residual regimes.

	Gain Bias		**Drift Bias**	
Regime	Direction	Solution	Direction	Solution
Above both poles	$-h$	β_s^*	$+h$	β_d^*
Between the poles	$-h \to +h$	-	$+h \to -h$	-
Below both poles	$+h$	β_d^*	$-h$	β_s^*

Table 1: Directional biases of Gain and Drift toward the sparse solution β_s^* and dense solution β_d^*. Between the poles, arrows indicate alignment reversals as the residual magnitude decreases.

C.2.1 Below both poles i.e. $\gamma|r_k| < 1$

Proof of Gain biasing towards the dense solution Let

$$\psi_0^* = \operatorname*{argmin}_{t\in[0,1]} \psi_\alpha(\beta(t)), \qquad \psi_Q^* = \operatorname*{argmin}_{t\in[0,1]} \psi_a(\beta(t))$$

where $a = \alpha^2 e^{-Q}$ from Eq. 8. Then $\psi_Q^* \geq \psi_0^*$, with strict inequality whenever $\gamma_k r_k \neq 0$ for at least one $k < K$. Thus, the accumulated gain shifts the minimizer towards the dense interpolator, without requiring the residual to retain its sign.

Proof. Under the averaged loss, the support and off-support gradients are r_k and $r_k/\sqrt{2}$, respectively. Hence,

$$Q_s - 2Q_o = \sum_{k=0}^{K-1} \log\left(\frac{(1-\gamma_k^2 r_k^2/2)^2}{1-\gamma_k^2 r_k^2}\right) \geq 0. \tag{30}$$

If $\gamma_k r_k = 0$ for every $k < K$, both gains vanish and the claim is immediate. Otherwise, $Q_s \geq 2Q_o > 0$.

Both objectives are strictly convex along the segment, with negative derivative at $t = 0$ and positive derivative at $t = 1$. Their minimizers are therefore unique and interior. Using $\partial_{\beta_j}\psi_\alpha(\beta) = \frac{1}{2}\operatorname{arsinh}(\beta_j/\alpha)$, stationarity of the initial potential gives

$$\operatorname{arcsinh}\left(\frac{\beta_s(\psi_0^*)}{\alpha^2}\right) = \sqrt{2}\operatorname{arcsinh}\left(\frac{\beta_o(\psi_0^*)}{\alpha^2}\right). \tag{31}$$

In particular, $\beta_s(\psi_0^*) > \beta_o(\psi_0^*) > 0$.

For $b, q \geq 0$, define the gain-induced change in the potential gradient

$$F(b,q) := \frac{1}{2}\left[\operatorname{arcsinh}\left(\frac{be^q}{\alpha^2}\right) - \operatorname{arcsinh}\left(\frac{b}{\alpha^2}\right)\right]. \tag{32}$$

Direct differentiation shows that F is nondecreasing in both arguments and convex in q. Since $F(b,0) = 0$, we obtain

$$F(\beta_s(\psi_0^*), Q_s) \geq F(\beta_o(\psi_0^*), 2Q_o) \tag{33}$$

$$\geq 2F(\beta_o(\psi_0^*), Q_o). \tag{34}$$

Using stationarity of the initial potential, the derivative of the final potential at its initial minimizer satisfies

$$\left.\frac{d}{dt}\psi_a(\beta(t))\right|_{t=\psi_0^*} = B\left[\sqrt{2}F(\beta_o(\psi_0^*), Q_o) - F(\beta_s(\psi_0^*), Q_s)\right] \tag{35}$$

$$\leq B(\sqrt{2}-2)F(\beta_o(\psi_0^*), Q_o) < 0. \tag{36}$$

Strict convexity now implies $\psi_Q^* > \psi_0^*$. □

Proof of Drift biasing towards the sparse solution Let $D = (D_o, D_o, D_s)^\top$ and define

$$\psi_0^* := \operatorname*{argmin}_{t\in[0,1]} \psi_\alpha(\beta(t)), \qquad \psi_D^* := \operatorname*{argmin}_{t\in[0,1]} \left\{\psi_\alpha(\beta(t)) + \frac{1}{2}\langle D, \beta(t)\rangle\right\}. \tag{37}$$

Then $\psi_D^* \leq \psi_0^*$, with strict inequality whenever $\gamma_k r_k < 0$ for at least one $k < K$. Thus, the accumulated drift shifts the minimizer towards the sparse interpolator.

Proof. For $|z| < 1$, the drift admits the expansion

$$d(z) = 2\sum_{m=1}^{\infty} \frac{z^{2m+1}}{2m+1}. \tag{38}$$

Consequently, for $-1 < z \leq 0$,

$$\sqrt{2}\, d(z/\sqrt{2}) - d(z) = 2\sum_{m=1}^{\infty} \frac{(2^{-m}-1)z^{2m+1}}{2m+1} \geq 0, \tag{39}$$

with strict inequality for $z < 0$. Summing over the phase gives

$$\sqrt{2}D_o - D_s \geq 0, \tag{40}$$

strictly whenever at least one update has $\gamma_k r_k < 0$.

Along the interpolation segment, the drift contributes the affine term

$$\langle D, \beta(t)\rangle = B\left[D_s + t(\sqrt{2}D_o - D_s)\right]. \tag{41}$$

If all updates have zero residual, then $D = 0$ and the minimizers coincide. Otherwise, this term is strictly increasing in t.

The initial potential is strictly convex along the segment and has an interior minimizer ψ_0^*. By its stationarity,

$$\left.\frac{d}{dt}\left[\psi_\alpha(\beta(t)) + \frac{1}{2}\langle D, \beta(t)\rangle\right]\right|_{t=\psi_0^*} = B(\sqrt{2}D_o - D_s) > 0. \tag{42}$$

Adding an affine term preserves strict convexity, so the unique minimizer satisfies $\psi_D^* < \psi_0^*$, possibly attaining the boundary $t = 0$. □

The same argument applies with ψ_α replaced by ψ_a for any fixed positive effective scale a. Thus, the drift also shifts the minimizer sparseward when the accumulated gain is held fixed. For the above-pole and between-pole regimes, use the real-valued extensions

$$q(z) := -\log|1-z^2|, \qquad d(z) := \log\left|\frac{1+z}{1-z}\right| - 2z, \qquad z \neq \pm 1. \tag{43}$$

We retain the definitions of ψ_0^*, ψ_Q^* and ψ_D^* above, with $a = \alpha e^{-Q}$.

C.2.2 ABOVE BOTH POLES I.E. $\gamma|r_k| > \sqrt{2}$

Gain biases the interpolator towards the sparse solution above the poles. Suppose that $\sqrt{2} < |\gamma_k r_k| \le 2$ for every $k < K$, where $K \ge 1$. Then $\psi_Q^* < \psi_0^*$.

Proof. Throughout this interval,

$$q(\gamma_k r_k) < 0, \qquad q(\gamma_k r_k/\sqrt{2}) \ge 0. \tag{44}$$

Consequently, $Q_s < 0$ and $Q_o \ge 0$, so the effective scales satisfy

$$a_s = \alpha^2 e^{-Q_s} > \alpha^2, \qquad a_o = \alpha^2 e^{-Q_o} \le \alpha^2. \tag{45}$$

Using positivity of the coordinates at ψ_0^* and stationarity of the initial potential, we obtain

$$\frac{d}{dt}\psi_a(\beta(t))\Big|_{t=\psi_0^*} = \frac{B}{2}\left[\sqrt{2}\,\operatorname{arcsinh}\left(\frac{\beta_o(\psi_0^*)}{a_o}\right) - \operatorname{arcsinh}\left(\frac{\beta_s(\psi_0^*)}{a_s}\right)\right] \tag{46}$$

$$> \frac{B}{2}\left[\sqrt{2}\,\operatorname{arcsinh}\left(\frac{\beta_o(\psi_0^*)}{\alpha}\right) - \operatorname{arcsinh}\left(\frac{\beta_s(\psi_0^*)}{\alpha}\right)\right] = 0. \tag{47}$$

Strict convexity therefore implies $\psi_Q^* < \psi_0^*$. □

Drift biases the interpolator towards the dense solution above the poles. Suppose that $\gamma_k r_k < -\sqrt{2}$ for every $k < K$, where $K \ge 1$. Then $\psi_D^* > \psi_0^*$.

Proof. Write $x_k = -\gamma_k r_k > \sqrt{2}$. By oddness of d and cancellation of its linear terms,

$$\sqrt{2}D_o - D_s = \sum_{k=0}^{K-1}\Big[d(x_k) - \sqrt{2}\,d(x_k/\sqrt{2})\Big] \tag{48}$$

$$= \sum_{k=0}^{K-1} 2\Big[\operatorname{arctanh}(1/x_k) - \sqrt{2}\,\operatorname{arctanh}(\sqrt{2}/x_k)\Big] < 0. \tag{49}$$

The last inequality follows because $\operatorname{arctanh}$ is positive and strictly increasing on $(0,1)$. Stationarity of the initial potential therefore gives

$$\frac{d}{dt}\left[\psi_\alpha(\beta(t)) + \frac{1}{2}\langle D, \beta(t)\rangle\right]\Big|_{t=\psi_0^*} = B(\sqrt{2}D_o - D_s) < 0. \tag{50}$$

Strict convexity implies $\psi_D^* > \psi_0^*$, possibly attaining the boundary $t = 1$. □

C.2.3 IN BETWEEN THE POLES

Gain reverses its bias between the poles. To characterize the transition, consider the gain contributed by one update with $x := |\gamma_k r_k| \in (1, \sqrt{2})$:

$$Q_o(x) = -\log(1 - x^2/2), \qquad Q_s(x) = -\log(x^2 - 1). \tag{51}$$

Let $\psi_Q^*(x)$ denote the corresponding minimizer. There exists a unique $x_Q \in (1, \sqrt{2})$ such that

$$\begin{cases} \psi_Q^*(x) > \psi_0^*, & 1 < x < x_Q, \\ \psi_Q^*(x) = \psi_0^*, & x = x_Q, \\ \psi_Q^*(x) < \psi_0^*, & x_Q < x < \sqrt{2}. \end{cases} \tag{52}$$

Thus, as the residual magnitude decreases from the off-support pole towards the support pole, the gain changes from sparseward to denseward.

Proof. The effective scales are

$$a_o(x) = \alpha^2(1 - x^2/2), \qquad a_s(x) = \alpha^2(x^2 - 1). \tag{53}$$

Hence,

$$\frac{d}{dt}\psi_{a(x)}(\beta(t))\bigg|_{t=\psi_0^*} = \frac{B}{2}\left[\sqrt{2}\,\operatorname{arcsinh}\left(\frac{\beta_o(\psi_0^*)}{\alpha^2(1-x^2/2)}\right) - \operatorname{arcsinh}\left(\frac{\beta_s(\psi_0^*)}{\alpha^2(x^2-1)}\right)\right]. \tag{54}$$

This derivative is continuous and strictly increasing in x: the off-support scale decreases, while the support scale increases. Moreover,

$$\lim_{x\downarrow 1}\frac{d}{dt}\psi_{a(x)}(\beta(t))\bigg|_{t=\psi_0^*} = -\infty, \tag{55}$$

$$\lim_{x\uparrow\sqrt{2}}\frac{d}{dt}\psi_{a(x)}(\beta(t))\bigg|_{t=\psi_0^*} = +\infty. \tag{56}$$

At the support pole, the support contribution diverges and produces a denseward bias. At the off-support pole, the off-support contribution diverges and produces a sparseward bias. The derivative therefore vanishes at exactly one x_Q. Strict convexity yields the stated ordering of the minimizers. □

Drift reverses its bias between the poles. Consider one update with negative residual and $x := -\gamma_k r_k \in (1, \sqrt{2})$, and define

$$D_o(x) = d(-x/\sqrt{2}), \qquad D_s(x) = d(-x). \tag{57}$$

Let $\psi_D^*(x)$ denote the corresponding minimizer. There exists a unique $x_D \in (1, \sqrt{2})$ such that

$$\begin{cases} \psi_D^*(x) < \psi_0^*, & 1 < x < x_D, \\ \psi_D^*(x) = \psi_0^*, & x = x_D, \\ \psi_D^*(x) > \psi_0^*, & x_D < x < \sqrt{2}. \end{cases} \tag{58}$$

Thus, during the same downward passage, the drift changes from denseward to sparseward.

Proof. By oddness of d, the slope of the drift contribution is

$$\frac{d}{dt}\langle D(x), \beta(t)\rangle = B\left[d(x) - \sqrt{2}\,d(x/\sqrt{2})\right]. \tag{59}$$

Since $d'(z) = z^2/(1-z^2)$ away from the poles,

$$\frac{d}{dx}\left[d(x) - \sqrt{2}\,d(x/\sqrt{2})\right] = \frac{2x^2}{(1-x^2)(2-x^2)} < 0. \tag{60}$$

Furthermore,

$$\lim_{x\downarrow 1}\left[d(x) - \sqrt{2}\,d(x/\sqrt{2})\right] = +\infty, \tag{61}$$

$$\lim_{x\uparrow\sqrt{2}}\left[d(x) - \sqrt{2}\,d(x/\sqrt{2})\right] = -\infty. \tag{62}$$

The support pole therefore produces a sparseward drift bias, whereas the off-support pole produces a denseward drift bias. Continuity and strict monotonicity give a unique zero x_D. By stationarity of the initial potential and strict convexity of the tilted potential, the sign of this slope gives the stated ordering of the minimizers. □

The crossover values x_Q and x_D need not coincide. For an interval containing several updates between the poles, the net bias is determined by the accumulated Q and D; the single-update reversals alone do not fix its direction.

C.3 Proofs of Gain Control Scheme

C.3.1 Convergence Proof

We prove convergence through two ingredients: an exact representation of the update in mirror coordinates, and a descent estimate for the associated Bregman divergence. The theorem then combines this estimate with coercivity to establish boundedness, loss summability, and convergence of the full sequence.

Let $X \in \mathbb{R}^{n\times d}$ satisfy $X \neq 0$, and let $y \in \mathrm{range}(X)$. Define

$$\mathcal{L}(\beta) = \frac{1}{2n}\|X\beta - y\|_2^2, \qquad \kappa = \max_j \frac{\|X_{:j}\|_2^2}{n} > 0.$$

For $a \in (0,\infty)^d$, write

$$\psi_a(x) = \frac{1}{2}\sum_j \left[x_j \operatorname{arcsinh}(x_j/a_j) - \sqrt{x_j^2 + a_j^2} + a_j\right],$$

$$\mathcal{D}_a(c,x) = \psi_a(c) - \psi_a(x) - \langle \nabla\psi_a(x), c - x\rangle.$$

Direct differentiation gives

$$\nabla_j \psi_a(x) = \frac{1}{2}\operatorname{arcsinh}(x_j/a_j), \qquad \nabla^2\psi_a(x) = \operatorname{diag}\left(\frac{1}{2\sqrt{x_j^2 + a_j^2}}\right). \tag{63}$$

In particular, ψ_a is strictly convex and nonnegative, with its minimum at the origin.

Given $\beta_k = u_k \odot v_k$, set

$$g_k = \nabla\mathcal{L}(\beta_k), \qquad z_k = \gamma g_k, \qquad p(t) = 1 + t^2 + t^4, \qquad M_k = \max_j p(z_{k,j}).$$

The normalized quadratic update is

$$u_{k+1,j} = \frac{(1+z_{k,j}^2)u_{k,j} - z_{k,j}v_{k,j}}{\sqrt{M_k}}, \qquad v_{k+1,j} = \frac{(1+z_{k,j}^2)v_{k,j} - z_{k,j}u_{k,j}}{\sqrt{M_k}}. \tag{64}$$

Lemma C.1 (Exact mirror and scale recursions)**.** *Suppose $u_{0,j} > |v_{0,j}|$ for every j, and let $\gamma > 0$. Then the iterates in equation 64 are well defined and satisfy $u_{k,j} > |v_{k,j}|$. With*

$$a_{k,j} = \frac{u_{k,j}^2 - v_{k,j}^2}{2}, \qquad \phi(t) = \frac{1}{2}\log\frac{1+t+t^2}{1-t+t^2},$$

we have

$$a_{k+1,j} = a_{k,j}\frac{p(z_{k,j})}{M_k}, \qquad 0 < a_{k+1,j} \le a_{k,j}, \tag{65}$$

$$\nabla\psi_{a_{k+1}}(\beta_{k+1}) = \nabla\psi_{a_k}(\beta_k) - \phi(z_k). \tag{66}$$

Moreover, for every $t \in \mathbb{R}$,

$$|\phi(t)| \le |t|, \qquad |\phi(t) - t| \le |t|^3. \tag{67}$$

Proof. Introduce the branches $w_{\pm,k} = (u_k \pm v_k)/\sqrt{2}$. Their updates are

$$w_{\pm,k+1,j} = \frac{1 \mp z_{k,j} + z_{k,j}^2}{\sqrt{M_k}} w_{\pm,k,j}.$$

Since $1 \pm t + t^2 = (t \pm 1/2)^2 + 3/4 > 0$ and $M_k \ge 1$, both branches remain positive. Multiplying their recursions gives the effective initialization Eq. (4),

$$a_{k+1,j} = w_{+,k+1,j}w_{-,k+1,j} = \frac{(1+z_{k,j}+z_{k,j}^2)(1-z_{k,j}+z_{k,j}^2)}{M_k}a_{k,j} = a_{k,j}\frac{p(z_{k,j})}{M_k}$$

because $a_{k,j} = w_{+,k,j}w_{-,k,j}$ and $(1-t+t^2)(1+t+t^2) = p(t)$.

Taking their ratio gives at each j, equation 66, using the relationship between $\operatorname{arcsinh}(x) := \log(x + \sqrt{x^2+1})$:

$$\nabla_j \psi_{a_k}(\beta_k) = \frac{1}{2}\operatorname{arcsinh}(\beta_{k,j}/a_{k,j}) = \frac{1}{2}\operatorname{arcsinh}(\frac{2u_{k,j}v_{k,j}}{u_{k,j}^2 - v_{k,j}^2}) = \frac{1}{2}\log\frac{w_{+,k,j}}{w_{-,k,j}}.$$

Finally, for $t \geq 0$, $\phi(t) \geq 0$ and

$$0 \leq t - \phi(t) = \int_0^t \frac{2r^2 + r^4}{1 + r^2 + r^4}\, dr \leq \int_0^t 3r^2\, dr = t^3.$$

This also implies $\phi(t) \leq t$. Oddness of ϕ proves equation 67 for all t. □

The mirror recursion suggests measuring progress by $\mathcal{D}_{a_k}(c, \beta_k)$, where c is an interpolator. The next lemma shows that the progress from the gradient term dominates the costs of a finite step and a changing potential.

Lemma C.2 (One-step Bregman descent)**.** *Consider one update equation 64, and suppress the iteration index. Suppose $u_j > |v_j|$, let $\beta = u \odot v$, and set $a_j = (u_j^2 - v_j^2)/2$. Let c satisfy $Xc = y$. If, for some $R > 0$,*

$$\|\beta\|_1 \leq R, \qquad \|c\|_1 \leq R, \qquad \sum_j a_j \leq R, \qquad 0 < \gamma \leq \frac{1}{32\kappa R},$$

then

$$\mathcal{D}_{a^+}(c, \beta^+) \leq \mathcal{D}_a(c, \beta) - \gamma\mathcal{L}(\beta). \tag{68}$$

Proof. Let $g = \nabla\mathcal{L}(\beta)$, $z = \gamma g$, and $s = \|z\|_\infty$. Since $Xc = y$, the least-squares identities and Hölders inequality give

$$\langle g, \beta - c\rangle = 2\mathcal{L}(\beta), \qquad \|g\|_\infty \leq \kappa\|\beta - c\|_1, \qquad \|g\|_\infty^2 \leq 2\kappa\mathcal{L}(\beta). \tag{69}$$

Indeed, $g := H(\beta - c)$ with $H := X^\top X/n$ and

$$|H_{ij}| = \frac{|x_i^\top x_j|}{n} \leq \frac{\|x_i\|_2\|x_j\|_2}{n} = \sqrt{H_{ii}H_{jj}} \leq \kappa.$$

For the last inequality, write $r = X\beta - y$. By Cauchy–Schwarz,

$$|g_i|^2 = \left|\frac{x_i^\top r}{n}\right|^2 \leq \frac{\|x_i\|_2^2\|r\|_2^2}{n^2} = 2H_{ii}\mathcal{L}(\beta) \leq 2\kappa\mathcal{L}(\beta).$$

Taking the maximum over i yields $\|g\|_\infty^2 \leq 2\kappa\mathcal{L}(\beta)$. Consequently,

$$s := \gamma\|g\|_\infty \leq 2\gamma\kappa R \leq \frac{1}{16}, \qquad s^2 \leq 2\kappa\gamma^2\mathcal{L}(\beta). \tag{70}$$

having used the pointwise inequality in the first part $|\beta_i - c_i| \leq |\beta_i| + |c_i|$.

We first control the next predictor using the current radius alone. Writing $m = \nabla\psi_a(\beta)$, Lemma C.1 gives $\beta_j^+ = a_j^+ \sinh(2m_j - 2\phi(z_j))$. Since $a_j^+ \leq a_j$ and $a_j\cosh(2m_j) = \sqrt{\beta_j^2 + a_j^2}$,

$$\|\beta^+\|_1 \leq e^{2s}\|\beta\|_1 + \sinh(2s)\sum_j a_j \leq 3R. \tag{71}$$

Indeed, using $a_j^+ \leq a_j$, $|\phi(z_j)| \leq s$, and the hyperbolic addition formula,

$$\begin{aligned}
|\beta_j^+| &= a_j^+\big|\sinh(2m_j - 2\phi(z_j))\big| \\
&\leq a_j|\sinh(2m_j)|\cosh(2s) + a_j\cosh(2m_j)\sinh(2s) \\
&= |\beta_j|\cosh(2s) + \sqrt{\beta_j^2 + a_j^2}\sinh(2s) \\
&\leq |\beta_j|\big(\cosh(2s) + \sinh(2s)\big) + a_j\sinh(2s) \\
&= e^{2s}|\beta_j| + a_j\sinh(2s),
\end{aligned}$$

where we used $\sqrt{\beta_j^2 + a_j^2} \le |\beta_j| + a_j$. Summing over j and using $s \le 1/16$ gives

$$\|\beta^+\|_1 \le e^{2s}\|\beta\|_1 + \sinh(2s)\sum_j a_j \le \big(e^{1/8} + \sinh(1/8)\big)R < 3R.$$

Next let $\ell_j = \log(a_j/a_j^+)$ and $\delta\psi = \psi_{a^+} - \psi_a$. As $M = p(s)$,

$$0 \le \ell_j = \log\frac{M}{p(z_j)} \le \log p(s) \le 2\log(1+s^2) \le 2s^2.$$

For the scalar summand ψ_a of ψ_a, direct differentiation yields

$$\frac{d}{dt}\psi_{ae^{-t}}(x) = \frac{1}{2}\left(\sqrt{x^2 + a^2e^{-2t}} - ae^{-t}\right), \qquad \frac{d}{dt}\psi'_{ae^{-t}}(x) = \frac{x}{2\sqrt{x^2 + a^2e^{-2t}}}.$$

The first derivative lies between 0 and $|x|/2$, and the second has absolute value at most $1/2$. Integrating over the logarithmic scale change therefore gives

$$0 \le \delta\psi(x) \le s^2\|x\|_1, \qquad \|\nabla\psi_{a^+}(x) - \nabla\psi_a(x)\|_\infty \le s^2. \tag{72}$$

Expanding the divergences and using equation 66 gives the exact identity

$$\begin{aligned}\mathcal{D}_{a^+}(c,\beta^+) - \mathcal{D}_a(c,\beta) = &-\langle\phi(z), \beta - c\rangle + \mathcal{D}_{a^+}(\beta,\beta^+)\\ &+ \delta\psi(c) - \delta\psi(\beta).\end{aligned} \tag{73}$$

The first and last terms satisfy

$$\begin{aligned}-\langle\phi(z), \beta - c\rangle &\le -2\gamma\mathcal{L}(\beta) + 2Rs^3,\\ \delta\psi(c) - \delta\psi(\beta) &\le Rs^2,\end{aligned} \tag{74}$$

by equation 67, equation 69, and equation 72.

To bound the remaining divergence, measure both predictors in the updated dual coordinates. Their difference satisfies

$$\delta m := \nabla\psi_{a^+}(\beta^+) - \nabla\psi_{a^+}(\beta) = -\phi(z) - \nabla\delta\psi(\beta), \qquad \|\delta m\|_\infty \le s + s^2.$$

For any fixed scale $\widetilde{a} > 0$, the inverse scalar mirror map is $q_j(m) = \widetilde{a}_j \sinh(2m)$, with $q_j'(m) = 2\sqrt{q_j(m)^2 + \widetilde{a}_j^2}$. If $m = \nabla\psi_{\widetilde{a}}(x)$ and $m' = \nabla\psi_{\widetilde{a}}(x')$, a change of variables in the scalar Bregman integral gives

$$\mathcal{D}_{\widetilde{a}}(x,x') = \sum_j (m_j - m_j')^2 \int_0^1 t\, q_j'\big(m_j' + t(m_j - m_j')\big)\, dt \le \sum_j (|x_j| + |x_j'| + \widetilde{a}_j)|m_j - m_j'|^2.$$

Here q_j stays between x_j and x_j' on the integration interval. Applying this bound with scale a^+ and using equation 71,

$$\mathcal{D}_{a^+}(\beta,\beta^+) \le \left(\|\beta\|_1 + \|\beta^+\|_1 + \sum_j a_j^+\right)(s+s^2)^2 \le 5R(s+s^2)^2 \le 8Rs^2.$$

Substituting into equation 73 and using equation 70 yields

$$\begin{aligned}\mathcal{D}_{a^+}(c,\beta^+) - \mathcal{D}_a(c,\beta) &\le -2\gamma\mathcal{L}(\beta) + (2s+9)Rs^2\\ &\le -2\gamma\mathcal{L}(\beta) + 10Rs^2\\ &\le -\big(2 - 20\kappa R\gamma\big)\gamma\mathcal{L}(\beta)\\ &\le -\gamma\mathcal{L}(\beta).\end{aligned}$$

□

The descent estimate requires a bound on the current predictor. Coercivity of the divergence supplies the complementary implication: a bound on the energy bounds the predictor. The following theorem closes this argument by induction and then establishes convergence to a single interpolator.

Theorem C.3 (Global convergence)**.** *Suppose $u_{0,j} > |v_{0,j}|$ for every j. Fix b with $Xb = y$, and define*

$$a_{0,j} = \frac{u_{0,j}^2 - v_{0,j}^2}{2}, \qquad A = \sum_j a_{0,j}, \qquad a_{\min} = \min_j a_{0,j},$$
$$S = \|b\|_1, \qquad V_0 = \mathcal{D}_{a_0}(b, \beta_0).$$

Set

$$B = 4V_0 + 2A + 2S\left[1 + \log\left(1 + \frac{4S}{a_{\min}}\right)\right]. \tag{75}$$

If $0 < \gamma \le \frac{1}{(32\kappa B)}$, then the normalized quadratic iterates are well defined and satisfy

$$\|\beta_k\|_1 \le B, \qquad \|u_k\|_2^2 + \|v_k\|_2^2 \le 4B, \qquad \gamma \sum_{k=0}^{\infty} \mathcal{L}(\beta_k) \le V_0.$$

Moreover, there exist $u_\infty, v_\infty, \beta_\infty$ such that

$$u_k \to u_\infty, \qquad v_k \to v_\infty, \qquad \beta_k \to \beta_\infty = u_\infty \odot v_\infty, \qquad X\beta_\infty = y.$$

Proof. We first establish the coercivity bound needed to apply Lemma C.2 at every iteration. By equation 63, shrinking the scale increases the Hessian and hence the Bregman divergence. Thus $\mathcal{D}_a(b, \beta) \ge \mathcal{D}_{a_0}(b, \beta)$ whenever $0 < a \le a_0$. Let $r = \|\beta\|_1$ and $V = \mathcal{D}_{a_0}(b, \beta)$. Expanding the divergence, using $\psi_{a_0}(b) \ge 0$, and bounding $\operatorname{arcsinh}(t) \le \log(1 + 2t)$ for $t \ge 0$, we obtain

$$\begin{aligned} 2V &= 2\psi_{a_0}(b) + \sum_j \sqrt{\beta_j^2 + a_{0,j}^2} - A - \sum_j b_j \operatorname{arcsinh}(\beta_j / a_{0,j}) \\ &\ge r - A - S \log\left(1 + \frac{2r}{a_{\min}}\right). \end{aligned}$$

For $S > 0$, the tangent bound for the logarithm at $r = 2S$ gives

$$S \log\left(1 + \frac{2r}{a_{\min}}\right) \le S \log\left(1 + \frac{4S}{a_{\min}}\right) + \frac{r}{2}.$$

For $S = 0$, the logarithmic term is absent. In both cases, this proves

$$\|\beta\|_1 \le 4\mathcal{D}_a(b, \beta) + 2A + 2S \log\left(1 + \frac{4S}{a_{\min}}\right), \qquad 0 < a \le a_0. \tag{76}$$

Write $V_k = \mathcal{D}_{a_k}(b, \beta_k)$. Since $A, S \le B$, equation 76 gives $\|\beta_0\|_1 \le B$. Suppose $\|\beta_k\|_1 \le B$ and $V_k \le V_0$. Lemma C.1 gives $0 < a_k \le a_0$, so $\sum_j a_{k,j} \le A \le B$. Applying Lemma C.2 with $c = b$ and $R = B$, followed by equation 76, yields

$$V_{k+1} \le V_k - \gamma \mathcal{L}(\beta_k) \le V_0, \qquad \|\beta_{k+1}\|_1 \le B. \tag{77}$$

Induction proves these bounds for every k. Telescoping equation 77 and using $V_k \ge 0$ gives

$$\gamma \sum_{k=0}^{\infty} \mathcal{L}(\beta_k) \le V_0, \qquad \mathcal{L}(\beta_k) \to 0. \tag{78}$$

We next show that the scales have positive limits. They are positive and nonincreasing by equation 65. Set $s_k = \|\gamma g_k\|_\infty$. The estimates used in the descent proof imply

$$0 \le \log \frac{a_{k,j}}{a_{k+1,j}} \le 2s_k^2, \qquad \sum_{k=0}^{\infty} s_k^2 \le 2\kappa\gamma^2 \sum_{k=0}^{\infty} \mathcal{L}(\beta_k) \le 2\kappa\gamma V_0.$$

Consequently,

$$a_{k,j} \longrightarrow a_{\infty,j} \ge a_{0,j} \exp(-4\kappa\gamma V_0) > 0. \tag{79}$$

By boundedness, a subsequence satisfies $\beta_{k_\ell} \to \bar{\beta}$. Equation equation 78 implies $X\bar{\beta} = y$, and $\|\bar{\beta}\|_1 \le B$. We can therefore apply Lemma C.2 with $c = \bar{\beta}$ and $R = B$. The sequence

$$W_k = \mathcal{D}_{a_k}(\bar{\beta}, \beta_k)$$

is nonnegative and nonincreasing. By equation 79 and continuity, $W_{k_\ell} \to \mathcal{D}_{a_\infty}(\bar{\beta}, \bar{\beta}) = 0$; therefore $W_k \to 0$ along the full sequence. On the segment joining β_k and $\bar{\beta}$, we have $|x_j| \le B$ and $a_{k,j} \le A \le B$. Eq. 63 gives $\nabla^2 \psi_{a_k}(x) \succeq (4B)^{-1} I$, so Taylor's integral formula yields

$$W_k \ge \frac{1}{8B} \|\beta_k - \bar{\beta}\|_2^2.$$

Hence $\beta_k \to \beta_\infty := \bar{\beta}$.

Finally, positivity of the branches uniquely determines the factors from (β_k, a_k):

$$u_{k,j} = \sqrt{\sqrt{\beta_{k,j}^2 + a_{k,j}^2} + a_{k,j}}, \qquad v_{k,j} = \frac{\beta_{k,j}}{u_{k,j}}.$$

These expressions are continuous at (β_∞, a_∞) because $a_{\infty,j} > 0$, proving factor convergence and $u_\infty \odot v_\infty = \beta_\infty$. Moreover,

$$\|u_k\|_2^2 + \|v_k\|_2^2 = 2 \sum_j \sqrt{\beta_{k,j}^2 + a_{k,j}^2} \le 2 \left(\|\beta_k\|_1 + \sum_j a_{k,j} \right) \le 2(B + A) \le 4B.$$

□

Lemma C.4. *Gain and Drift for the Gain Control Mechanism (Restated) Let $(u_k, v_k)_{k \ge 0}$ be the iterates generated by the update in Definition 3.1 For every $K \ge 1$ and $i \in \{1, \dots, d\}$, and $z = \gamma \nabla \mathcal{L}(\beta)$, the accumulated Gain and Drift are as follows*

$$Q_{K,i}^{\mathrm{GC}} = \sum_{k=0}^{K-1} \log \left(\frac{M_k}{1 + z_{k,i}^2 + z_{k,i}^4} \right). \tag{80}$$

$$D_{K,i}^{\mathrm{GC}} = \sum_{k=0}^{K-1} \left[\log \left(\frac{1 + z_{k,i} + z_{k,i}^2}{1 - z_{k,i} + z_{k,i}^2} \right) - 2 z_{k,i} \right]. \tag{81}$$

Proof. We consider the standard initialization of the layer weights $u_0 = \sqrt{2}\alpha \mathbf{1}, v_0 = \mathbf{0}$. Similar to C.1, we reparameterize the layer weights

$$w_{\pm,k} = (u_k \pm v_k)/\sqrt{2}$$

From the update step in equation 3.1

$$w_{+,k+1,i} = \frac{1 - z_{k,i} + z_{k,i}^2}{\sqrt{M_k}} \, w_{+,k,i} \tag{82}$$

$$w_{-,k+1,i} = \frac{1 + z_{k,i} + z_{k,i}^2}{\sqrt{M_k}} \, w_{-,k,i} \tag{83}$$

where $z = \gamma \nabla \mathcal{L}$.

Since

$$1 \pm z + z^2 = \left(z \pm \tfrac{1}{2}\right)^2 + \tfrac{3}{4} > 0,$$

and $w_{+,0,i} = w_{-,0,i} = \alpha_i > 0$, both branches remain positive. Their logarithms are therefore well defined for all finite $z_{k,i}$.

We further introduce the quantities

$$a_{k,j} = w_{+,k,j} \, w_{-k,j} \,, \qquad \theta_{k,j} = \ln \frac{w_{+,k,j}}{w_{-,k,j}} \tag{84}$$

Rewriting β in terms of these quantities

$$\beta_{k,j} = a_{k,j} \sinh(\theta_{k,j}) \tag{85}$$

with $a_{0,i} = \alpha_i^2$ and $\theta_0, j = 0$.

Multiplying the two recursive factors in equation 82, we get

$$a_{k+1,i} = \frac{(1 - z_{k,i} + z_{k,i}^2)(1 + z_{k,i} + z_{k,i}^2)}{M_k} a_{k,i} \tag{86}$$

$$= \frac{1 + z_{k,i}^2 + z_{k,i}^4}{M_k} a_{k,i}. \tag{87}$$

Iterating this identity yields

$$a_{K,i} = \alpha_i^2 \prod_{k=0}^{K-1} \frac{1 + z_{k,i}^2 + z_{k,i}^4}{M_k}.$$

The accumulated gain is the logarithmic contraction of this scale, namely $a_{K,i} = \alpha_i^2 e^{-Q_{K,i}^{\text{GC}}}$. Consequently,

$$Q_{K,i}^{\text{GC}} = \sum_{k=0}^{K-1} \log\left(\frac{M_k}{1 + z_{k,i}^2 + z_{k,i}^4}\right).$$

Similarly, dividing the recurrences in equation 82, we get

$$\theta_{k+1,i} - \theta_{k,i} = -\log\left(\frac{1 + z_{k,i} + z_{k,i}^2}{1 - z_{k,i} + z_{k,i}^2}\right). \tag{88}$$

Expanding the logarithmic increment around $z_{k,i} = 0$ gives $-2z_{k,i} + O(z_{k,i}^3)$. The linear term lies in the range of X and does not create any bias towards any interpolator i.e. $\langle \sum_k z_k, \beta \rangle$ is constant over all β satisfying $X\beta = y$. We therefore separate the leading gradient contribution $-2z_{k,i}$ from the remaining correction, retaining the latter exactly.

$$\theta_{k+1,i} - \theta_{k,i} = -2z_{k,i} - \left[\log\left(\frac{1 + z_{k,i} + z_{k,i}^2}{1 - z_{k,i} + z_{k,i}^2}\right) - 2z_{k,i}\right].$$

Summing over k, and using $\theta_{0,i} = 0$, we obtain

$$\theta_{K,i} = -2 \sum_{k=0}^{K-1} z_{k,i} - D_{K,i}^{\text{GC}}, \tag{89}$$

where

$$D_{K,i}^{\text{GC}} = \sum_{k=0}^{K-1} \left[\log\left(\frac{1 + z_{k,i} + z_{k,i}^2}{1 - z_{k,i} + z_{k,i}^2}\right) - 2z_{k,i}\right].$$

□

C.3.2 Proof of Proposition 3.3

Proof. We aim to show that β_∞ is the unique solution of

$$\min_{\beta \in \mathbb{R}^d} \left\{ \psi_{a_\infty}(\beta) + \frac{1}{2}\langle D_\infty, \beta \rangle \right\} \qquad \text{subject to} \qquad X\beta = y. \tag{90}$$

The associated Lagrangian is

$$\mathcal{L}(\beta, \lambda) = \psi_{a_\infty}(\beta) + \frac{1}{2}\langle D_\infty, \beta \rangle + \langle \lambda, X\beta - y \rangle, \qquad \lambda \in \mathbb{R}^n. \tag{91}$$

The KKT conditions at β_∞ are therefore

$$X\beta_\infty = y, \tag{92}$$

$$\nabla\psi_{a_\infty}(\beta_\infty) + \frac{1}{2}D_\infty + X^\top\lambda_\infty = 0 \tag{93}$$

for some $\lambda_\infty \in \mathbb{R}^n$.

Since $a_{\infty,i} > 0$ for every i, the objective is strictly convex, with Hessian

$$\nabla^2\psi_{a_\infty}(\beta) = \operatorname{diag}\left(\frac{1}{2\sqrt{\beta_i^2 + a_{\infty,i}^2}}\right)_{i=1}^{d} \succ 0. \tag{94}$$

Consequently, these conditions are sufficient to establish that β_∞ is the unique minimizer. Feasibility holds by the assumed convergence to an interpolator. It therefore suffices to prove that

$$\nabla\psi_{a_\infty}(\beta_\infty) + \frac{1}{2}D_\infty \in \operatorname{range}(X^\top). \tag{95}$$

From Lemma 3.2, the Gain and Drift for the gain control scheme are as follows

$$Q_{K,i}^{\mathrm{GC}} = \sum_{k=0}^{K-1} \log\left(\frac{M_k}{1 + z_{k,i}^2 + z_{k,i}^4}\right). \tag{96}$$

$$D_{K,i}^{\mathrm{GC}} = \sum_{k=0}^{K-1}\left[\log\left(\frac{1 + z_{k,i} + z_{k,i}^2}{1 - z_{k,i} + z_{k,i}^2}\right) - 2z_{k,i}\right]. \tag{97}$$

For $K \geq 1$, from equation 84 and equation 85, the predictor $\beta_{K,i}$ is as follows

$$\beta_{K,i} = a_{K,i}\sinh\left(\ln\frac{w_{+,K,i}}{w_{-,K,i}}\right) \tag{98}$$

$$= a_{K,i}\sinh\left(\theta_{K,i}\right) \tag{99}$$

Simplifying, we get

$$\operatorname{arcsinh}\left(\frac{\beta_{K,i}}{a_{K,i}}\right) = \theta_{K,i} \tag{100}$$

Substituting equation 89, we get

$$\operatorname{arcsinh}\left(\frac{\beta_{K,i}}{a_{K,i}}\right) = -2\sum_{k=0}^{K-1} z_{k,i} - D_{K,i}^{\mathrm{GC}} \tag{101}$$

Additionally, we observe from equation 2 that

$$[\nabla\psi_a(\beta)]_i = \frac{1}{2}\operatorname{arcsinh}\left(\frac{\beta_i}{a_i}\right) \tag{102}$$

Therefore,

$$\nabla\psi_{a_K}(\beta_K) + \frac{1}{2}D_K^{\mathrm{GC}} = -\sum_{k=0}^{K-1} z_k. \tag{103}$$

Since $z_k \in \operatorname{range}(X^\top)$, hence

$$\nabla\psi_{a_K}(\beta_K) + \frac{1}{2}D_K^{\mathrm{GC}} \in \operatorname{range}(X^\top). \tag{104}$$

The assumed convergence implies $a_K \to a_\infty > 0$. Taking limits and using that $\operatorname{range}(X^\top)$ is closed proves equation 95

$$\nabla\psi_{a_\infty}(\beta_\infty) + \frac{1}{2}D_\infty \in \operatorname{range}(X^\top). \tag{105}$$

□

Batch size	Learning rate	QK effective rank	Val. accuracy
128	0.02	1959.05 ± 38.49	79.68 ± 0.096
128	0.04	1815.71 ± 19.38	79.24 ± 0.161
256	0.04	1980.98 ± 31.96	79.56 ± 0.064

Table 2: Effective rank and validation accuracy on ImageNet. Error bars denote 95% confidence intervals over three seeds. The maximum effective rank possible is 4608.

D Vision Experiment Details

We train a DeiT-Small (Touvron et al., 2020) on ImageNet (Deng et al., 2009) for 300 epochs using Muon (Jordan et al., 2024). We use the experimental setup from (Nelaturu et al.). Training is distributed across four GPUs and uses label smoothing, weight decay 0.01, CutMix, and MixUp. We use a triangular learning-rate schedule with 10 warmup epochs. The batch size and learning rate are chosen from

$$B \in \{128, 256\}, \qquad \eta \in \{0.02, 0.04\}.$$

For the Drift–Gain analysis in the main text, model checkpoints are stored every 10 epochs. We express the Query–Key dynamics in a rotational frame fixed at initialization, so that changes in the tracked quantities reflect the evolution relative to a common reference frame rather than rotations of the singular vectors.

Effective rank. To quantify spectral sparsity, we additionally report the effective rank of the learned linear map product. Given singular values $\sigma_1, \ldots, \sigma_r$, we define

$$p_i = \frac{\sigma_i}{\sum_j \sigma_j}, \qquad \mathrm{EffRank}(W) = \exp\left(- \sum_i p_i \log(p_i + \varepsilon) \right),$$

with $\varepsilon = 10^{-12}$ for numerical stability. A smaller effective rank corresponds to a stronger concentration of spectral mass on a few singular directions and therefore serves as a smooth proxy for spectral sparsity. We report the sum of the effective ranks across the analyzed linear maps. For Query–Key matrices, the effective rank is computed at the head level before aggregation, matching the spectral decomposition used in the Drift–Gain analysis. We report this in Table 2 together with the resulting validation accuracy.