# IMPLICIT BIAS OF FACTORWISE SPECTRAL OPTIMIZATION IN SELF-ATTENTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

In self-attention, logits depend on the product $W = KQ^\top$, yet per-head Muon, used in recent LLMs, updates $K$ and $Q$ separately. The training dynamics therefore depend on the factorization, not only on the product and its gradient. We ask which normalized product these updates select, and study this question through the momentum-free factorwise polar flow, a continuous-time idealization of per-head Muon. For a fixed gradient, sufficiently aligned factors keep full-rank gradients and, above the largest omitted singular value, converge to the leading singular directions. When factor alignment accumulates without bound, the minimum margin of the normalized product converges to the optimal width-$h$ margin characterized by the Ky Fan norm. For exact softmax attention, when each query has a prescribed target key, we prove local convergence of the product near the target, while the factors approach an orbit of common rotations. If the gradient weights on these target–competitor comparisons are also locally identifiable, they converge as well, at a polynomial rate in optimizer time. Numerical experiments illustrate that the product can converge even when its limit does not determine the gradient weights.

## 1 INTRODUCTION

Muon (Jordan et al., 2024) has recently been used to pretrain large language models such as Moonlight (Liu et al., 2025) and, as MuonClip, Kimi K2 (Kimi Team, 2025). It accelerates small-scale training (Jordan et al., 2024) and achieves roughly twice the compute efficiency of AdamW in scaling-law experiments (Liu et al., 2025). Since Muon orthogonalizes the momentum of each weight matrix, the partition of parameters into matrices affects the optimization dynamics. This is especially relevant in the bilinear attention model studied here, where a head's logits depend on the key $\widetilde{K}$ and query $\widetilde{Q}$ projections only through $\widetilde{W} = \widetilde{K}\widetilde{Q}^\top$ (without rotary position encodings, Vaswani et al. 2017). Recent LLMs treat these factors differently: Kimi K2 applies per-head QK-Clip after the Muon update (Kimi Team, 2025), whereas GLM-5 and Kimi K3 orthogonalize attention projections head by head (GLM-5 Team, 2026; Kimi Team, 2026). These choices motivate us to study the dynamics of $\widetilde{K}$ and $\widetilde{Q}$ rather than only their product. Following Kimi Team (2026), we call separate spectral updates of each head's key and query factors *per-head Muon*.

The scale of $\widetilde{W}$ does not determine which keys a head prefers: once a target outranks its competitors, positive rescaling preserves this ordering and only sharpens the softmax. We study the limiting direction of $\widetilde{W}$, when it exists, which is called the optimizer's *implicit bias*. On separable problems, such limiting directions are often characterized by max-margin problems, with the relevant norm determined by the optimization geometry (Soudry et al., 2018; Gunasekar et al., 2018; Sun et al., 2023). For attention, along regularization paths, direct parameterization of $\widetilde{W}$ leads to a Frobenius-norm margin and the factorization $\widetilde{W} = \widetilde{K}\widetilde{Q}^\top$ to a nuclear-norm margin (Ataee Tarzanagh et al., 2023a); the nuclear-norm result also holds for Euclidean gradient flow under additional assumptions (Sheen et al., 2024). For directly parameterized multiclass linear classifiers, spectral updates instead select the operator-norm max-margin solution (Fan et al., 2025). These results leave the factorized spectral case open, where $\widetilde{K}$ and $\widetilde{Q}$ follow separate spectral updates; we study it as an idealization of per-head Muon.

In this setting, the relevant margin problem (3.4) is built on the normalized products of a width-$h$ head, $\{KQ^\top : K, Q \in \mathbb{R}^{d\times h},\ \|K\|_{\mathrm{op}}, \|Q\|_{\mathrm{op}} \le 1\}$, or the set of matrices with rank at most $h$ and



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.

operator norm at most one. This set is nonconvex, but its convex hull is the unit ball of the dual Ky Fan norm $\|W\|_{(h)}^* = \max\{\|W\|_{\mathrm{op}}, \|W\|_*/h\}$ (Bhatia, 1997; Grussler & Giselsson, 2018), which reduces to the nuclear norm at $h = 1$ and the operator norm at $h = d$. Solving this problem, however, does not tell us which product the factorwise dynamics select. The factorized problem has balanced rank-one stationary points, even when $h > 1$, whose margins can be arbitrarily worse than optimal (Appendix F.1), so we have to follow the trajectory.

To isolate these dynamics, we drop Muon's momentum and Newton–Schulz approximation, which are studied in Kim & Oh (2026); Choudhury et al. (2026); Li et al. (2026b), and keep only the exact orthogonalization step, which is steepest descent in the operator norm (Bernstein & Newhouse, 2024; Jordan et al., 2024). We call its continuous-time limit the *factorwise polar flow* (more details on the background can be found in Appendix A.1). Two features distinguish it from dynamics defined directly on $\widetilde{W}$. First, the product does not evolve on its own: with $\widetilde{W} = I_2$ and $-\nabla L(\widetilde{W}) = I_2$, the factor pairs $(I_2, I_2)$ and $(\mathrm{diag}(2,1), \mathrm{diag}(1/2,1))$ give different product velocities, $\dot{\widetilde{W}} = 2I_2$ and $\mathrm{diag}(5/2, 2)$. Euclidean gradient flow conserves the balance $\widetilde{K}^\top \widetilde{K} - \widetilde{Q}^\top \widetilde{Q}$ (Du et al., 2018; Arora et al., 2018), so balanced factors stay balanced, and their product follows its own flow. The polar flow does not conserve this quantity, so we track both factors. Second, the gradient itself moves during training: near favored targets it is a positive combination of target–competitor comparisons (3.3), whose normalized coefficients, the gradient weights, change with the product.

**Contributions.** Our analysis focuses on a property of the factors that the product alone does not reveal. Given normalized factors $K, Q$ and the normalized negative gradient $Z$, the symmetric cross-Gram matrix $G = \mathrm{sym}(K^\top Z Q)$ measures how strongly the latent directions of $K$ and $Q$ align through the current gradient, so two pairs $(K, Q)$ with the same product can evolve differently under factorwise spectral updates.

- **Fixed gradient direction.** We first hold $Z$ fixed. If the initial alignment exceeds the largest omitted singular value, $\lambda_{\min}(G(t_0)) > \sigma_{h+1}(Z)$, then $ZQ$ and $Z^\top K$ keep full column rank, and the factors converge to a leading singular-vector pair up to a common rotation (Theorem 4.1).
- **Moving gradient direction.** We then let the gradient direction evolve with the product. For separable comparisons with positive, regular base weights, the gap to the optimal margin decays exponentially in the integrated coercivity. If this quantity diverges, the minimum margin converges to the optimal width-$h$ value and the convex relaxation is tight, even without convergence of the gradient direction or the gradient weights (Theorem 5.2).
- **Exact attention.** For exact softmax attention, we replace this trajectory-wide condition with local ones: a dual certificate for the optimal product, and prescribed targets favored by the loss and readouts. At a large enough logit scale, trajectories that start near this product converge to it, and the factors approach its orbit of common rotations (Theorem 6.2). Under strict complementarity, dual nondegeneracy, and suitable local initialization, the gradient weights also converge, at a rate polynomial in optimizer time (Theorem 6.3).

A commuting example shows that product convergence does not require divergent integrated coercivity, and that the product limit does not fix the gradient weights. We illustrate these effects on controlled attention tasks and on a survey of 1,536 initializations (Section 7). Our analysis covers a single head with prescribed targets and fixed representations, values, and readouts, without momentum, clipping, weight decay, or interacting heads.

## 2 Related work

**From stationarity to optimality.** Tsilivis et al. (2026, Theorem 4) show that normalized accumulation points of steepest descent on homogeneous networks are KKT directions, and Gronich & Vardi (2026, Theorem 3.2 and Corollary 3.3) show that, for momentum methods including Muon, any limiting direction is a KKT direction. For factorized products, KKT directions can be arbitrarily suboptimal (Appendix F.1). Fan et al. (2025) prove an optimal spectral-norm margin for directly parameterized multiclass classifiers, where no factorization arises. We give conditions under which a width-constrained factorized product attains the optimal margin and converges; Table 10 in Appendix H compares the hypotheses.

**Implicit bias of attention.** Ataee Tarzanagh et al. (2023b) relate token selection in attention to support vector machines. For a directly parameterized product, gradient descent converges quickly (Vasudeva et al., 2025) and mirror descent selects entrywise $\ell_p$ margins (Julistiono et al., 2026). For a factorized product, a nuclear-norm bias holds along regularization paths (Ataee Tarzanagh et al., 2023a) and under Euclidean gradient flow with structural alignment and an assumed product limit (Sheen et al., 2024, Theorem 5.1), and factorization rescales the effective learning rates of the query–key and output–value circuits (Vashisht & Ramaswamy, 2026). These analyses use Euclidean or entrywise geometries; we study spectral updates of the factors and prove the product limit instead of assuming it.

**Spectral optimizers and factorized parameters.** In attention, Muon can orthogonalize a full projection, which couples the heads through one polar factor, or each head's block separately. Zhang et al. (2026) compare the two choices through one-step descent bounds, whereas we ask what a single head selects in the long run. Muon has also been analyzed for squared-loss matrix factorization, where the data fix the target product (Kang et al., 2026; Rhee et al., 2026). Other work builds matrix spectral updates from duals of Ky Fan norms (Kravatskiy et al., 2025), and low-rank or Riemannian variants from gauge or manifold geometry (Cesista et al., 2026; Li et al., 2026c). Our rectangular-factor identities also carry over algebraically to LoRA (Hu et al., 2021) (Appendix H).

## 3 Problem Setting

We study a single attention head trained on fixed token representations, following attention-margin analyses (Ataee Tarzanagh et al., 2023b;a).

**Notation.** Write $[n] = \{1, \dots, n\}$ and $\langle A, B\rangle = \operatorname{tr}(A^\top B)$. We use $\|\cdot\|_{\mathrm{op}}$, $\|\cdot\|_*$, and $\|\cdot\|_F$ for the operator, nuclear, and Frobenius norms. For $Z \in \mathbb{R}^{m\times n}$, order the singular values as $\sigma_1(Z) \ge \sigma_2(Z) \ge \cdots \ge \sigma_{\min\{m,n\}}(Z) \ge 0$, with $\sigma_{\min\{m,n\}+1}(Z) = 0$. The Ky Fan $h$-norm is $\|Z\|_{(h)} = \sum_{q=1}^{h} \sigma_q(Z)$. For a vector $a$ and an index set $\mathcal{M}$, $\operatorname{softmax}_{\mathcal{M}}(a)$ denotes the softmax of $a$ restricted to $\mathcal{M}$, with zeros outside $\mathcal{M}$.

**Model.** We consider a single-layer, single-head attention model trained on $N$ sequences. Sequence $i$ has fixed token embeddings $X_i = [x_{i1}, \dots, x_{iT_i}]^\top \in \mathbb{R}^{T_i\times d}$ and fixed values $V_i = [v_{i1}, \dots, v_{iT_i}]^\top \in \mathbb{R}^{T_i\times d_v}$. The positions in $\mathcal{R}_i \subseteq [T_i]$ are supervised queries, and query $r \in \mathcal{R}_i$ attends to the keys in a mask $\mathcal{M}_{ir} \subseteq [T_i]$, such as a causal mask. The raw key and query matrices are $\widetilde{K}, \widetilde{Q} \in \mathbb{R}^{d\times h}$ with $1 \le h \le d$; the logits depend on them only through $\widetilde{W} = \widetilde{K}\widetilde{Q}^\top$. The output at query $r$ is

$$u_{ir}(\widetilde{W}) = c_{ir}^\top V_i^\top \operatorname{softmax}_{\mathcal{M}_{ir}}\big(\beta X_i \widetilde{W} x_{ir}\big), \tag{3.1}$$

where $c_{ir} \in \mathbb{R}^{d_v}$ is a fixed readout and $\beta > 0$ is an inverse temperature. We write

$$a_{irk}(\widetilde{W}) = x_{ik}^\top \widetilde{W} x_{ir}, \qquad \alpha_{irk}(\widetilde{W}) = \big[\operatorname{softmax}_{\mathcal{M}_{ir}}(\beta X_i \widetilde{W} x_{ir})\big]_k, \qquad g_{irk} = c_{ir}^\top v_{ik}$$

for the attention logits, the attention probabilities, and the token utilities, so that $u_{ir}(\widetilde{W}) = \sum_{k\in\mathcal{M}_{ir}} \alpha_{irk}(\widetilde{W})\, g_{irk}$. The utility $g_{irk}$ is the output that query $r$ would produce if it attended only to key $k$. With a single query per sequence, $\beta = 1$, $V_i = X_i$, and $c_{ir} = Y_i v$ for a label $Y_i$ and a linear head $v$, Equation (3.1) becomes $Y_i f(X_i)$ with $f(X) = v^\top X^\top \operatorname{softmax}(X\widetilde{W}z)$, the model studied by Ataee Tarzanagh et al. (2023a). Only $\widetilde{K}$ and $\widetilde{Q}$ are trained.

**Training objective.** Training minimizes the empirical loss over the factors,

$$\mathcal{L}(\widetilde{K}, \widetilde{Q}) = L(\widetilde{K}\widetilde{Q}^\top), \qquad L(\widetilde{W}) = \sum_{i=1}^{N} \ell_i\big(\mathbf{u}_i(\widetilde{W})\big), \qquad \mathbf{u}_i = (u_{ir})_{r\in\mathcal{R}_i}, \tag{3.2}$$

where $\ell_i : \mathbb{R}^{|\mathcal{R}_i|} \to \mathbb{R}$ is the loss of sequence $i$ and may couple the outputs of all its queries. Standard examples are the per-query logistic loss $\ell_i(\mathbf{u}) = \sum_r \log(1 + e^{-u_r})$ and the soft-minimum loss $\ell_i(\mathbf{u}) = \log \sum_r e^{-u_r}$, which is dominated by the worst query in the sequence.

**Assumption 3.1** (Loss)**.** Each $\ell_i$ is $C^3$ and locally decreasing in every coordinate.

**Targets and margins.** Each query has a prescribed target key $k_{ir}^\star \in \mathcal{M}_{ir}$, the key it should attend to. For every visible competitor $k \neq k_{ir}^\star$, we define the target–competitor matrix and its linear margin

$$\Delta_{irk} = (x_{i,k_{ir}^\star} - x_{ik})x_{ir}^\top, \qquad m_{irk}(W) = \langle W, \Delta_{irk}\rangle.$$

The raw margin $m_{irk}(\widetilde{W})$ is the target's logit advantage. Since $\alpha_{ir,k_{ir}^\star}(\widetilde{W})/\alpha_{irk}(\widetilde{W}) = e^{\beta m_{irk}(\widetilde{W})}$, attention concentrates on the target as all these margins grow. We enumerate the comparisons as $\Delta_1, \dots, \Delta_M$, with $M \geq 1$, and write $m_j(W) = \langle W, \Delta_j\rangle$. The constraints $m_j \geq 1$ are the token-separation constraints of Att-SVM (Ataee Tarzanagh et al., 2023a) for these targets; Section 6 adds conditions that tie the targets to the readouts and the loss.

**Gradient and normalized gradient weights.** Where the gradient prefactors $\rho_j$ are positive, the negative loss gradient has the representation

$$-\nabla L(\widetilde{W}) = \sum_{j=1}^{M} \omega_j(\widetilde{W})\Delta_j, \qquad \omega_j(\widetilde{W}) = \rho_j(\widetilde{W})e^{-\beta m_j(\widetilde{W})}, \qquad \rho_j(\widetilde{W}) > 0. \tag{3.3}$$

Let $\mathcal{P}_M = \{p \in \mathbb{R}_+^M : \mathbf{1}^\top p = 1\}$, $p_j = \omega_j / \sum_\ell \omega_\ell$, and $\Delta(p) = \sum_j p_j \Delta_j$, so that $-\nabla L = (\sum_j \omega_j)\Delta(p)$. The normalized gradient weights $p$ sum to one over all comparisons; the attention probabilities $\alpha$ sum to one within each query. The raw product and loss determine $p$ (Appendix D.1).

**The width-$h$ max-margin problem.** For factorwise spectral updates, the natural norm is the dual Ky Fan norm $\|W\|_{(h)}^* = \max\{\|W\|_{\mathrm{op}}, \|W\|_*/h\}$, whose unit ball is the convex hull of the width-$h$ products with operator norm at most one (Appendix A.1). This leads to the max-margin problem

$$\text{(Att-SVM}_h\text{)} \qquad \min_W \|W\|_{(h)}^* \quad \text{subject to} \quad \langle W, \Delta_j\rangle \geq 1 \quad (j \in [M]). \tag{3.4}$$

Its objective is the nuclear norm at $h = 1$, as in the factorized Att-SVM of Ataee Tarzanagh et al. (2023a), and the operator norm at $h = d$. In normalized form, the rank-constrained problem and its convex relaxation read

$$\begin{aligned} \gamma_h^{\mathrm{rk}} &= \max_{\substack{\mathrm{rank}(W)\leq h \\ \|W\|_{\mathrm{op}}\leq 1}} \min_j \langle W, \Delta_j\rangle, \\ \gamma_h &= \max_{\|W\|_{(h)}^*\leq 1} \min_j \langle W, \Delta_j\rangle = \min_{p\in\mathcal{P}_M} \|\Delta(p)\|_{(h)}, \qquad \gamma_h^{\mathrm{rk}} \leq \gamma_h. \end{aligned} \tag{3.5}$$

We call the comparison family *separable* when $\gamma_h > 0$. In that case, (Att-SVM$_h$) has value $1/\gamma_h$, and its solutions are the maximizers of $\gamma_h$ divided by $\gamma_h$. We call the relaxation *tight* when some product of rank at most $h$ attains $\gamma_h$; separability alone does not guarantee this. Every $p \in \mathcal{P}_M$ gives the dual bound $\gamma_h \leq \|\Delta(p)\|_{(h)}$, in particular along any trajectory.

**Update rule.** In optimizer time $s$, each raw factor moves along a maximizer of its inner product with the negative gradient over the operator-norm unit ball:

$$\dot{\widetilde{K}} \in \partial\|\cdot\|_*\big((-\nabla L(\widetilde{W}))\widetilde{Q}\big), \qquad \dot{\widetilde{Q}} \in \partial\|\cdot\|_*\big((-\nabla L(\widetilde{W}))^\top \widetilde{K}\big). \tag{3.6}$$

Here $\widetilde{W} = \widetilde{K}\widetilde{Q}^\top$, dots denote $d/ds$, and $\partial\|\cdot\|_*(X) = \arg\max_{\|S\|_{\mathrm{op}}\leq 1}\langle S, X\rangle$. If $X$ has full column rank, the maximizer is $\mathrm{polar}(X) = X(X^\top X)^{-1/2} = UV^\top$ for a thin SVD $X = U\Sigma V^\top$; at rank deficiency, any maximizer is allowed.

**Normalized dynamics.** Since both velocities have operator norm at most one, $\tau(s) = s + \tau_0$ bounds the raw factor norms whenever $\tau_0 > 0$ and $\tau_0 \geq \max\{\|\widetilde{K}(0)\|_{\mathrm{op}}, \|\widetilde{Q}(0)\|_{\mathrm{op}}\}$. Set $t = \log(\tau(s)/\tau_0)$, with inverse $s(t) = \tau_0(e^t - 1)$, and define

$$K(t) = \frac{\widetilde{K}(s(t))}{\tau(s(t))}, \qquad Q(t) = \frac{\widetilde{Q}(s(t))}{\tau(s(t))}, \qquad W(t) = K(t)Q(t)^\top = \frac{\widetilde{W}(s(t))}{\tau(s(t))^2}.$$

Thus $\|K(t)\|_{\mathrm{op}}, \|Q(t)\|_{\mathrm{op}} \leq 1$, but $W(t)$ need not equal $\widetilde{W}/\|\widetilde{W}\|_{\mathrm{op}}$. Positive rescaling leaves maximizing directions unchanged, so, with primes denoting $d/dt$, the normalized flow is

$$Z(t) = \frac{-\nabla L(\widetilde{W}(s(t)))}{\|-\nabla L(\widetilde{W}(s(t)))\|_{(h)}}, \qquad \begin{aligned} K' &\in \partial\|\cdot\|_*(Z(t)Q) - K, \\ Q' &\in \partial\|\cdot\|_*(Z(t)^\top K) - Q. \end{aligned} \tag{3.7}$$

Thus $\beta m_j(\widetilde{W}(s(t))) = \Lambda(t) m_j(W(t))$, where $\Lambda(t) = \beta\tau(s(t))^2 = \Lambda_0 e^{2(t-t_0)}$ and $\Lambda_0 = \Lambda(t_0) > 0$. We ask when the minimum margin of $W(t)$ tends to $\gamma_h$ and its limit, divided by $\gamma_h$, solves (Att-SVM$_h$).

All results below concern solutions in the following sense.

**Assumption 3.2** (Trajectory)**.** On its interval of existence starting at $t_0$, $(K, Q)$ is an absolutely continuous solution of Equation (3.7), with $Z(t)$ specified in each result and $\|K(t_0)\|_{\mathrm{op}}, \|Q(t_0)\|_{\mathrm{op}} \leq 1$. When a polar argument is rank deficient, any maximizing direction is allowed.

The factor balls are forward invariant, and every solution extends to all $t \geq t_0$ as long as the normalization defining $Z(t)$ stays well defined along the trajectory, that is, the gradient does not vanish (Appendix A.1).

## 4 Fixed-matrix dynamics

Fix $Z$ with $\|Z\|_{(h)} = 1$ and define the cross-Gram matrix $\Gamma = \Gamma_Z(K, Q) = K^\top Z Q$ and its symmetric part $G = G_Z(K, Q) = \mathrm{sym}\,\Gamma$, where $\mathrm{sym}(A) = (A + A^\top)/2$. For a unit vector $v \in \mathbb{R}^h$, $v^\top G v = (Kv)^\top Z (Qv)$ is the score of the latent direction $v$ through $Z$. Thus $\lambda_{\min}(G)$ is the minimum score over unit latent directions, whereas $\mathrm{tr}(G) = \langle W, Z\rangle$ records their total score.

The cross-Gram matrix obeys an exact evolution identity. Writing $|X| = (X^\top X)^{1/2}$, every maximizing direction $S \in \partial\|\cdot\|_*(X)$ satisfies $S^\top X = |X|$, even at rank deficiency, so, since $\|K\|_{\mathrm{op}}, \|Q\|_{\mathrm{op}} \leq 1$, almost everywhere,

$$\Gamma' = |ZQ| + |Z^\top K| - 2\Gamma, \qquad G' \succeq 0.$$

In particular, $G$ is nondecreasing in the Loewner order, so $G(t_0) \succeq \mu I_h$ implies $G(t) \succeq \mu I_h$ for all $t \geq t_0$. We write $G = G_+ - G_-$ for the Jordan decomposition of $G$, where $G_\pm = \frac{1}{2}(|G| \pm G) \succeq 0$; the negative part $G_-$ satisfies $\|G_-\|_{\mathrm{op}} = \max\{-\lambda_{\min}(G), 0\}$. When $\sigma_h(Z) > \sigma_{h+1}(Z)$, the selected product is the rank-$h$ polar factor $\mathrm{polar}_h(Z) = U_h V_h^\top$ of an SVD $Z = U\Sigma V^\top$, together with its *orthogonal orbit* $\mathcal{O}_h(Z) = \{(U_h O, V_h O) : O \in O(h)\}$ of factor pairs that produce it through a common rotation. Distances use the Frobenius norm, with $\|(K, Q)\|^2 = \|K\|_F^2 + \|Q\|_F^2$.

**Theorem 4.1** (Fixed-matrix dynamics)**.** *Suppose Assumption 3.2 holds and fix $Z = U\Sigma V^\top$ with $\|Z\|_{(h)} = 1$.*

(i) *The negative part of the symmetric cross-Gram matrix decays exponentially: for all $t \geq t_0$,*

$$\|G_-(t)\|_{\mathrm{op}} \leq e^{-4(t-t_0)} \|G_-(t_0)\|_{\mathrm{op}}. \tag{4.1}$$

(ii) *If $G(t_0) \succeq \mu I_h$ for some $\mu > 0$, then $\sigma_h(ZQ(t))$ and $\sigma_h(Z^\top K(t))$ remain bounded below by $\mu$ for all $t \geq t_0$. Both matrices therefore remain full column rank, and the trajectory is smooth and unique.*

(iii) *If $\sigma_h(Z) > \sigma_{h+1}(Z)$ and $\lambda_{\min}(G(t_0)) > \sigma_{h+1}(Z)$, then $(K(t), Q(t)) \to (U_h O_\infty, V_h O_\infty)$ for some $O_\infty \in O(h)$, and $W(t) \to U_h V_h^\top$. Moreover, for any rate $\nu \in (0, 1 - \sigma_{h+1}(Z)/\sigma_h(Z))$, there exists a constant $C_\nu > 0$ such that, for all $t \geq t_0$,*

$$\mathrm{dist}((K(t), Q(t)), \mathcal{O}_h(Z)) + \|W(t) - U_h V_h^\top\|_F \leq C_\nu e^{-\nu(t-t_0)}. \tag{4.2}$$

Decay of $G_-$ alone does not guarantee selection, but the threshold in (iii) is sharp among equilibria with full-column-rank $ZQ$ and $Z^\top K$ (Appendix B).

**Proof sketch.** The cross-Gram identity gives $\lambda_{\min}(G)' \geq 4(-\lambda_{\min}(G))_+$ and preserves positive lower bounds. Under the spectral threshold, the matrices $ZQ$ and $Z^\top K$ retain singular values above $\sigma_{h+1}(Z)$. Lemma 5.1 then makes the support-function residual decay exponentially and gives convergence to the orthogonal orbit. Smoothness near that orbit yields finite tail length and a single factor limit. The transverse linearization gives every rate below $1 - \sigma_{h+1}(Z)/\sigma_h(Z)$ (Appendix B).

## 5 MAX-MARGIN CONVERGENCE WITH A TIME-DEPENDENT GRADIENT

When the gradient direction $Z(t)$ varies, the cross-Gram lower bound may no longer hold. We instead compare the product score with the support function of the feasible products. For a fixed $Z$, the time derivative of this score satisfies, almost everywhere,

$$\frac{d}{dt}\langle W, Z\rangle = \mathcal{D}_Z(K,Q), \qquad \mathcal{D}_Z(K,Q) = \|ZQ\|_* + \|Z^\top K\|_* - 2\langle KQ^\top, Z\rangle.$$

The dissipation $\mathcal{D}_Z(K,Q)$ is the sum of the two factorwise operator–nuclear duality gaps. To quantify convergence, we introduce the support-function residual $r_Z(K,Q) = \|Z\|_{(h)} - \langle KQ^\top, Z\rangle \geq 0$ alongside the spectral cutoff gap $\mathrm{gap}_h(Z) = \sigma_h(Z) - \sigma_{h+1}(Z)$.

**Lemma 5.1** (Dissipation and product error)**.** *Let* $\|K\|_{\mathrm{op}}, \|Q\|_{\mathrm{op}} \leq 1$ *and* $W = KQ^\top$.

(i) Duality bounds. *For every matrix* $Z$,

$$\langle Z, W\rangle \leq \tfrac{1}{2}\big(\|ZQ\|_* + \|Z^\top K\|_*\big) \leq \|Z\|_{(h)}, \qquad \langle Z, W\rangle \leq \|Z\|_{(h)}\|W\|_{\mathrm{op}}. \tag{5.1}$$

(ii) Coercivity. *Suppose* $\|Z\|_{(h)} = 1$*, and let* $a = \min\{\sigma_h(ZQ), \sigma_h(Z^\top K)\}$*. If the separation* $\eta = a - \sigma_{h+1}(Z)$ *is strictly positive, then*

$$\mathcal{D}_Z(K,Q) \geq \frac{2\eta^2}{\eta^2 + 2\sigma_1(Z)a}\, r_Z(K,Q) \geq \frac{2}{3}\eta^2 r_Z(K,Q). \tag{5.2}$$

(iii) Geometric error. *If* $\mathrm{gap}_h(Z) > 0$*, the support residual bounds the orbit and product distances:*

$$\mathrm{dist}((K,Q), \mathcal{O}_h(Z))^2 \leq \frac{4r_Z(K,Q)}{\mathrm{gap}_h(Z)}, \qquad \|KQ^\top - \mathrm{polar}_h(Z)\|_F^2 \leq \frac{8r_Z(K,Q)}{\mathrm{gap}_h(Z)}. \tag{5.3}$$

Part (ii) bounds the dissipation below by the residual under spectral separation, while part (iii) converts this residual into product and orbit errors. Part (iii) aligns the factors by a single orthogonal transformation and does not invert the singular values arising in the polar decomposition. Hence the estimate also covers rank-deficient factors and polar inputs (Appendix A.2).

**Time-dependent Gibbs weights.** In softmax attention, both the margins and the base weights $\pi_j = \rho_j / \sum_k \rho_k$ change with the product. We therefore let the weight trajectory $\pi(t) \in \mathcal{P}_M$ be locally absolutely continuous, with

$$\pi_j(t) \geq \pi_- > 0, \qquad |(\log \pi_j(t))'| \leq B_\pi \quad \text{a.e.} \tag{5.4}$$

With the exponential schedule $\Lambda(t) = \Lambda_0 e^{2(t-t_0)}$ from Section 3, $\Lambda_0 > 0$, we study the normalized flow Equation (3.7) driven by

$$p_j(t) = \frac{\pi_j(t) e^{-\Lambda(t) m_j(W(t))}}{\sum_k \pi_k(t) e^{-\Lambda(t) m_k(W(t))}}, \qquad Z(t) = \frac{\Delta(p(t))}{\|\Delta(p(t))\|_{(h)}}. \tag{5.5}$$

Since $\Delta(p)$ never vanishes for separable comparisons, this flow is globally defined and stays in the factor balls (Appendix A.1). We measure the separation required in Lemma 5.1(ii) by the polar-input separation

$$\eta_Z(K,Q) = \min\{\sigma_h(ZQ), \sigma_h(Z^\top K)\} - \sigma_{h+1}(Z), \qquad \eta(t) = \eta_{Z(t)}(K(t), Q(t)).$$

A cross-Gram lower bound $G \succeq \mu I_h$ with $\mu > \sigma_{h+1}(Z)$ implies $\eta_Z \geq \mu - \sigma_{h+1}(Z) > 0$, but positive separation alone does not determine the relative orientation of the two factors. Since $\eta(t)$ may be nonpositive for a while, we integrate the dissipation it guarantees. For $\rho > 0$, let

$$\kappa(\rho) = \min\left\{1,\ \inf \frac{\mathcal{D}_Z(K,Q)}{r_Z(K,Q)}\right\}, \tag{5.6}$$

where the infimum is over matrices in the stated dimensions with $\|Z\|_{(h)} = 1$, $\|K\|_{\mathrm{op}}, \|Q\|_{\mathrm{op}} \leq 1$, $r_Z(K,Q) > 0$, and $\eta_Z(K,Q) \geq \rho$. An empty infimum is $+\infty$, and $\kappa(\rho) = 0$ for $\rho \leq 0$. By Lemma 5.1(ii), $\kappa(\rho) \geq \frac{2}{3}\rho^2$ for $0 < \rho \leq 1$, so the integrated coercivity satisfies

$$\Xi(t) = \int_{t_0}^{t} \kappa(\eta(u))\, du \geq \frac{2}{3}\int_{t_0}^{t} \eta_+(u)^2\, du, \qquad \eta_+ = \max\{\eta, 0\}. \tag{5.7}$$

**Theorem 5.2** (Max-margin convergence)**.** *Suppose the comparison family $\{\Delta_j\}_{j=1}^M$ is separable, and let Assumption 3.2 hold with the Gibbs weights in Equation (5.5), whose base weights $\pi$ satisfy Equation (5.4). Define the margin shortfall $e_m(t) = \gamma_h - \min_j m_j(W(t))$ and set $C_\pi = \log(1/\pi_-)$. The following hold:*

(a) *For $C_0 = e_m(t_0) + (3C_\pi + B_\pi)/\Lambda_0$ and all $t \geq t_0$, the shortfall decays exponentially in the integrated coercivity:*

$$0 \leq e_m(t) \leq C_0 e^{-\Xi(t)} \leq C_0 \exp\left[-\frac{2}{3}\int_{t_0}^{t} \eta_+(u)^2\, du\right]. \tag{5.8}$$

(b) *If the integrated coercivity diverges ($\Xi(t) \to \infty$), the convex relaxation is tight ($\gamma_h^{\mathrm{rk}} = \gamma_h$), and every accumulation point of $W(t)$ solves both margin problems in Equation (3.5).*

(c) *If $\Xi(t) \to \infty$ and the convex optimizer $W^\star$ is unique, both the product $W(t)$ and its operator-norm normalization $\widetilde{W}(s(t))/\|\widetilde{W}(s(t))\|_{\mathrm{op}}$ converge to $W^\star$.*

Neither the gradient direction nor the Gibbs weights are assumed to converge. A dual minimizer $p^\star$ with $\mathrm{gap}_h(\Delta(p^\star)) > 0$ gives the unique product $W^\star = \mathrm{polar}_h(\Delta(p^\star))$ and a quantitative error bound (Appendix C). Divergence of $\Xi(t)$ is sufficient but not necessary for product convergence under uniqueness (Appendix F.2).

**Proof sketch.** Differentiating the soft minimum $m_\Lambda = -\Lambda^{-1}\log\sum_j \pi_j e^{-\Lambda m_j(W)}$ gives $\mathcal{D}_{\Delta(p)} + O(\Lambda^{-1})$, without differentiating $p$ or $Z$. The corrected gap $Y = \gamma_h - m_\Lambda + (2C_\pi + B_\pi)/\Lambda$ satisfies $0 \leq e_m \leq Y$, $Y(t_0) \leq C_0$, and $Y' \leq -\kappa(\eta)Y$. Integration gives the rate. If $\Xi \to \infty$, compactness and closedness of the rank constraint give tightness (Appendix C).

## 6 Softmax attention: product and weight convergence

For exact softmax attention, local target and initialization conditions replace the requirement $\Xi(t) \to \infty$. Throughout this section, trajectories satisfy Assumption 3.2 for the loss in Equation (3.2).

**Assumption 6.1** (Target conditions)**.** Fix target keys and an active support set $\mathcal{S} \subseteq [M]$. Write $\mathcal{P}_\mathcal{S} = \{p \in \mathcal{P}_M : \mathrm{supp}(p) \subseteq \mathcal{S}\}$ for the active simplex and $T_\mathcal{S} = \{v \in \mathbb{R}^M : \mathrm{supp}(v) \subseteq \mathcal{S},\ \mathbf{1}^\top v = 0\}$ for its tangent space. Choose weights $p^\star \in \mathrm{relint}\,\mathcal{P}_\mathcal{S}$ and set $\Delta^\star = \sum_{j\in\mathcal{S}} p_j^\star \Delta_j$.

(MM) *Max-margin certificate.* The target matrix has a strict cutoff gap $\sigma_h(\Delta^\star) > \sigma_{h+1}(\Delta^\star)$ and $\|\Delta^\star\|_{(h)} > 0$. For $W^\star = \mathrm{polar}_h(\Delta^\star)$, the active comparisons attain the margin and the inactive ones satisfy it: $\langle W^\star, \Delta_j\rangle = \|\Delta^\star\|_{(h)} \quad (j \in \mathcal{S}), \langle W^\star, \Delta_k\rangle \geq \|\Delta^\star\|_{(h)} \quad (k \notin \mathcal{S})$.

(TC) *Target compatibility.* The readouts and the loss favor the prescribed targets: for every supervised query $(i, r)$ and visible competitor $k \neq k_{ir}^\star$, we have $g_{ir,k_{ir}^\star} > g_{irk}$, and the loss still rewards larger outputs at the target state, $-\partial_r \ell_i(\mathbf{u}_i^\star) > 0$ for $\mathbf{u}_i^\star = (g_{ir,k_{ir}^\star})_{r\in\mathcal{R}_i}$.

(SC) *Strict complementarity.* The margin inequalities in (MM) are strict for every inactive comparison $k \notin \mathcal{S}$.

(ND) *Dual nondegeneracy.* The Hessian of the dual objective $p \mapsto \|\Delta(p)\|_{(h)}$ at $p^\star$ is positive definite on $T_\mathcal{S}$.

Condition (MM) makes the rank-constrained and convex margins coincide, $\gamma_h^{\mathrm{rk}} = \gamma_h = \|\Delta^\star\|_{(h)}$, and identifies $W^\star$ as the unique optimal product. Together with (SC) and (ND), it also identifies $p^\star$ as the unique dual minimizer (Appendices A.2 and E.1).

To track convergence, we use the inverse logit scale $\varepsilon(t) = \Lambda(t)^{-1}$, which satisfies $\varepsilon' = -2\varepsilon$. The prefactors $\rho_j(t)$ and softmax probabilities $\alpha_{irk}(t)$ are evaluated at the unnormalized product $\widetilde{W}(s(t))$, and we extend $p^\star$ by zero outside $\mathcal{S}$. Wherever all $\rho_j(t) > 0$, we define the local energy

$$\mathcal{E}(t) = \gamma_h - \langle W(t), \Delta^\star\rangle + \varepsilon(t) D_{\mathrm{KL}}(p^\star \| p(t)). \tag{6.1}$$

The first term controls the squared distance to the orthogonal orbit $\mathcal{O}_h(\Delta^\star)$. Because the Kullback–Leibler term is scaled by $\varepsilon(t)$, the energy can be small even when the weight error is large. All constants below depend only on the problem data and the target conditions.

**Theorem 6.2** (Local product convergence). *Suppose conditions* (MM) *and* (TC) *hold. There exist constants $r_0, \varepsilon_*, c_0 > 0$ such that every trajectory with $0 < \varepsilon(t_0) \le \varepsilon_*$ and $1 - \langle W(t_0), \Delta^\star \rangle / \gamma_h \le r_0$ satisfies the following:*

(a) *For all $t \ge t_0$, the prefactors stay positive, $\rho_j(t) > 0$ for every $j$, and $\min_j m_j(W(t)) \ge c_0$.*

(b) *As $t \to \infty$,*

$$W(t) \to W^\star, \qquad \operatorname{dist}((K(t), Q(t)), \mathcal{O}_h(\Delta^\star)) \to 0, \tag{6.2}$$

*and the operator-norm normalization $\widetilde{W}(s(t))/\|\widetilde{W}(s(t))\|_{\mathrm{op}}$ also converges to $W^\star$.*

(c) *For every supervised query $(i, r)$ and all $t \ge t_0$, $1 - \alpha_{ir,k^\star_{ir}}(t) \le (|\mathcal{M}_{ir}| - 1) e^{-c_0/\varepsilon(t)}$.*

The conclusion identifies the orbit, not a fixed factor limit. Neither weight convergence nor full column rank of $ZQ$ and $Z^\top K$ is assumed.

**Proof sketch.** The corrected energy $\mathcal{V} = \mathcal{E} + C\varepsilon$ satisfies $e_m \le \mathcal{V} = e_m + O(\varepsilon)$ and $\mathcal{V}' \le -\mathcal{D}_{\Delta(p)}$, even at rank deficiency. Hence the dissipation is integrable and the minimum margin converges. Along a subsequence where the dissipation vanishes, the Gibbs ratios put the limiting weights on the minimum-margin comparisons, which gives a point of the stationary set. This set is compact and semialgebraic, and the margin is constant on each of its connected components, so there are finitely many stationary margin values. A small enough initial error rules out the suboptimal ones, and Lemma 5.1 then gives product and orbit convergence (Appendix D).

Under the initialization conditions below, adding (SC) and (ND) gives joint convergence.

**Theorem 6.3** (Local joint convergence). *Suppose* (MM), (TC), (SC), *and* (ND) *hold, with $|\mathcal{S}| \ge 2$. There exist constants $\varepsilon_0, \mathcal{E}_0, c_0, c, C > 0$ such that every trajectory with $\rho_j(t_0) > 0$ for all $j$, $0 < \varepsilon(t_0) \le \varepsilon_0$, and $\mathcal{E}(t_0) \le \mathcal{E}_0$ satisfies the following:*

(a) *For all $t \ge t_0$, $\rho_j(t) > 0$ for every $j$ and $\min_j m_j(W(t)) \ge c_0$.*

(b) *There is a time $t_1 \ge t_0$ with $t_1 - t_0 \le C(\mathcal{E}(t_0) + \varepsilon(t_0))$ such that, for all $t \ge t_1$,*

$$\operatorname{dist}((K(t), Q(t)), \mathcal{O}_h(\Delta^\star)) + \|W(t) - W^\star\|_F + \|p(t) - p^\star\|_1 \le C e^{-c(t - t_1)}. \tag{6.3}$$

(c) *$\widetilde{W}(s(t))/\|\widetilde{W}(s(t))\|_{\mathrm{op}} \to W^\star$, and for every supervised query $(i, r)$ and all $t \ge t_0$, $1 - \alpha_{ir,k^\star_{ir}}(t) \le C e^{-c_0/\varepsilon(t)}$.*

Since $t = \log(\tau(s)/\tau_0)$, the exponential rate in $t$ is polynomial in optimizer time: $e^{-c(t-t_1)} = ((s_1 + \tau_0)/(s + \tau_0))^c$, where $s_1$ corresponds to $t_1$. The condition $|\mathcal{S}| \ge 2$ excludes the singleton case, where $T_{\mathcal{S}} = \{0\}$ and there are no active weights to identify.

**Proof sketch.** At the target factors, (ND) makes the dissipation grow quadratically in the active-weight error on the active simplex. A small initial energy keeps the factors near the target orbit, and (SC) suppresses the inactive weights. A persistent active-weight error would therefore make the nonnegative energy decrease at a fixed positive rate, so the trajectory enters a small neighborhood of the target in finite time. There, a Procrustes slice separates the factor coordinates, which contract modulo common rotations, from the log-weight errors, which relax on the faster $O(\varepsilon)$ scale. A coupled Lyapunov estimate accounting for the drift of the quasi-steady state gives the joint exponential rate (Appendix E).

## 7 Numerical illustrations and scope

**Two-mode rank-one example.** We illustrate Theorem 6.3 when a nonzero singular direction lies beyond the selected width. Take $d = 2, h = 1$, $v_\pm = (1, \pm a)^\top$, and $\Delta_\pm = v_\pm v_\pm^\top$, realized by two masked queries. At $p^\star = (1/2, 1/2)$, the certificate has $\Delta^\star = \operatorname{diag}(1, a^2)$ and $W^\star = e_1 e_1^\top$, so $\operatorname{rank}(\Delta^\star) = 2 > h$. We use eight paired, perturbed initializations at each $a \in \{0.5, 0.8, 0.95\}$ and compute the weights from the current softmax loss. All 24 trajectories reach $t = 8$, with maximum product error $2.07 \times 10^{-4}$ and weight error $2.47 \times 10^{-10}$ (Figure 1).

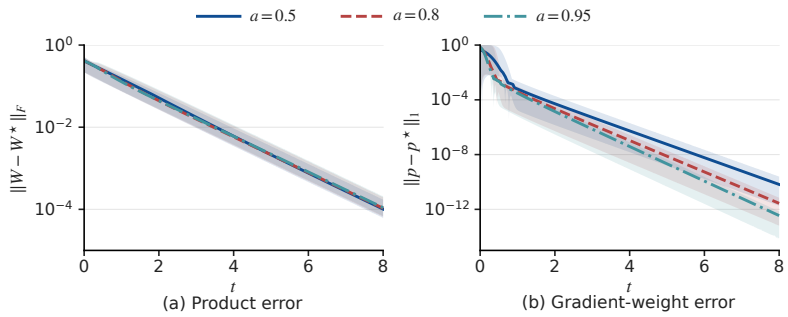
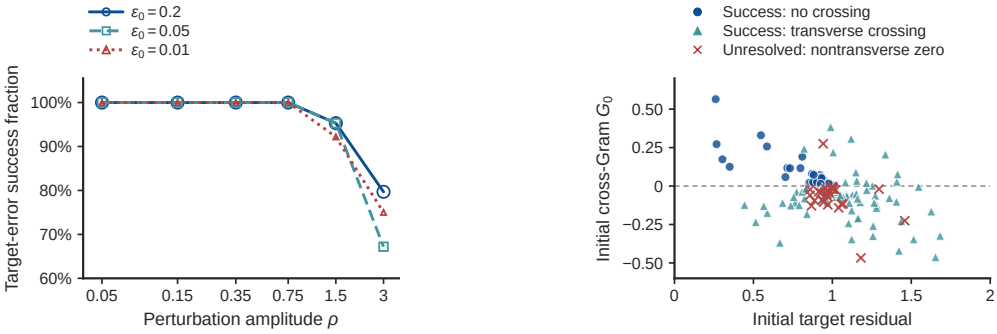



**Figure 1:** Product and weight errors with a nonzero omitted mode, $\mathrm{rank}(\Delta^\star) = 2 > h = 1$. Lines show medians; bands show full ranges of eight paired initializations per $a$, not confidence intervals. Task and checks: Appendix G.2.

**Product convergence and weight limits.** We illustrate the distinction between product and weight convergence in Theorem 6.2 with three initializations of a commuting softmax-attention example satisfying (MM) and (TC) but not (ND). A second system has externally oscillating base weights (Appendices F.2 and G.3). The softmax-attention products converge to $I_2$, while $p_1$ converges to $1/2$, 1, or 0 depending on initialization, and the two imbalanced paths have $\Xi(\infty) < \infty$. The externally varying system converges in product while its weights keep oscillating.

**Initialization and rank boundaries.** We vary 512 factor pairs in the coupled softmax-attention example over three initial inverse logit scales, giving 1,536 initial states. At $T = 10$, 1,392 runs meet the stated product, gradient-weight, and positive-margin error criteria; the remaining 144 stop at nontransverse zeros of an input to the polar map and are left unresolved. All 768 perturbations with $r_{\mathrm{init}} \leq 0.75$ meet the target-error criterion, and 492 successful trajectories cross a transverse rank boundary (Figure 2). These tests probe behavior away from the target, including transverse rank crossings. At nontransverse rank loss we stop rather than impose a continuation; this does not establish nonconvergence.




**(a)** Perturbed initializations ($\rho = r_{\mathrm{init}}$). **(b)** Disk initializations at $\varepsilon_0 = 0.05$.

**Figure 2:** Finite-horizon initialization survey for the coupled softmax-attention example. Success means meeting the numerical target-error criterion defined in Appendix G.4. In (b), triangles denote successful paths that cross a transverse rank boundary; crosses are unresolved nontransverse zeros, not observed convergence failures.

Appendix G also reports fixed-matrix, Newton–Schulz, commuting-weight, coupled-attention, and frozen-feature certificate diagnostics.

**Scope and limitations.** The softmax-attention guarantees are local to prescribed targets with fixed representations, values, masks, and readouts. They exclude discrete or approximate updates, momentum, clipping, weight decay, and interacting heads, and give neither a computable basin nor a rate under the assumptions of Theorem 6.2 alone. Finite-horizon diagnostics do not certify integrated coercivity or rigorous error bounds, explain QK-Clip, or establish faster convergence or end-to-end performance.

## AI USE STATEMENT

Generative AI tools assisted with literature retrieval, manuscript organization and editing, proof checking, experimental design, and implementation and checking of numerical diagnostics. The authors checked cited references against their sources and manually reviewed every AI-assisted claim, derivation, proof, and numerical conclusion against the stated assumptions and definitions. Numerical diagnostics supplemented rather than replaced the mathematical arguments. The authors take full responsibility for all text, claims, proofs, experiments, and reported results.

## REPRODUCIBILITY STATEMENT

Appendices A–F contain the proofs and analytical constructions. Appendix G specifies the synthetic models, initializations, solver tolerances, event handling, error metrics, and numerical consistency checks. It also reports numerical certificate checks on frozen features; the selected sentences, feature tensors, and readouts needed to reconstruct those examples are not included in the source archive. Each main-text figure points to its experimental description.

## REFERENCES

Nicola Aladrah, Emanuele Ballarin, Matteo Biagetti, Alessio Ansuini, Alberto d'Onofrio, and Fabio Anselmi. Understanding and inverse design of implicit bias in stochastic learning: A geometric perspective. arXiv preprint arXiv:2601.06597, 2026. URL https://arxiv.org/abs/2601.06597.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 244–253, 2018.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, volume 32, pp. 7413–7424. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/c0c783b5fc0d7d808f1d14a6e9c8280d-Abstract.html.

Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023a. URL https://openreview.net/forum?id=gLwzzmh79K.

Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism. In *Advances in Neural Information Processing Systems*, volume 36, pp. 48314–48362. Curran Associates, Inc., 2023b. doi: 10.52202/075280-2096. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/970f59b22f4c72aec75174aae63c7459-Abstract-Conference.html.

Jean-Pierre Aubin and Arrigo Cellina. *Differential Inclusions: Set-Valued Maps and Viability Theory*, volume 264 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin, Heidelberg, 1984. doi: 10.1007/978-3-642-69512-4.

Pierfrancesco Beneventano, Mahmoud Abdelmoneum, and Tomaso Poggio. The spectral dynamics and noise geometry of Muon. arXiv preprint arXiv:2606.08388, 2026. URL https://arxiv.org/abs/2606.08388.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. In *OPT 2024: 16th Annual Workshop on Optimization for Machine Learning*, 2024. URL https://openreview.net/forum?id=ux18f5nOpD.

Rajendra Bhatia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer New York, 1997. doi: 10.1007/978-1-4612-0653-8.

Stephen Boyd. Linear quadratic Lyapunov theory. EE363 course notes, Stanford University, 2009. URL https://ee363.stanford.edu/archive/notes/lq-lyap-notes.pdf.

Franz Louis Cesista, Katherine Crowson, Cédric Simal, and Stella Biderman. LoRA-Muon: Spectral steepest descent on the low-rank manifold. arXiv preprint arXiv:2606.12921, 2026. URL https://arxiv.org/abs/2606.12921.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. arXiv preprint arXiv:2506.15054, 2025. URL https://arxiv.org/abs/2506.15054.

Sayantan Choudhury, Xiaoran Cheng, Martin Takáč, Sen Na, and Mladen Kolar. Muon with Nesterov momentum: Heavy-tailed noise and (randomized) inexact polar decomposition. arXiv preprint arXiv:2605.06884, 2026. URL https://arxiv.org/abs/2605.06884.

Jorge Cortés. Discontinuous dynamical systems: A tutorial on solutions, nonsmooth analysis, and stability. *IEEE Control Systems Magazine*, 28(3):36–73, 2008. doi: 10.1109/MCS.2008.919306. URL https://arxiv.org/abs/0901.3583. Extended author version: arXiv:0901.3583.

Michel Coste. An introduction to semialgebraic geometry. Lecture notes, Institut de Recherche Mathématique de Rennes, 2002. URL https://perso.univ-rennes1.fr/michel.coste/polyens/SAG.pdf.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and Muon on multiclass separable data. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-1322. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/386432c7534eec9a1cd7cbeea90d7e9f-Abstract-Conference.html.

Evan S. Gawlik and Melvin Leok. Iterative computation of the Fréchet derivative of the polar decomposition. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1354–1379, 2017. doi: 10.1137/16M108971X.

GLM-5 Team. GLM-5: From vibe coding to agentic engineering. arXiv preprint arXiv:2602.15763, 2026. URL https://arxiv.org/abs/2602.15763.

Eitan Gronich and Gal Vardi. The implicit bias of Adam and Muon on smooth homogeneous neural networks. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=DpIc1cpNKG.

Christian Grussler and Pontus Giselsson. Low-rank inducing norms with optimality interpretations. *SIAM Journal on Optimization*, 28(4):3057–3078, 2018. doi: 10.1137/17M1115770.

Suriya Gunasekar, Blake E. Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6151–6159. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/58191d2a914c6dae66371c9dcdc91b41-Abstract.html.

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, 2018. URL https://proceedings.mlr.press/v80/gunasekar18a.html.

Nicholas J. Higham. Computing the polar decomposition—with applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1160–1174, 1986. doi: 10.1137/0907079.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021. URL https://arxiv.org/abs/2106.09685.

Ziwei Ji and Matus J. Telgarsky. Directional convergence and alignment in deep learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17176–17186. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c76e4b2fa54f8506719a5c0dc14c2eb9-Abstract.html.

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/.

Addison Kristanto Julistiono, Davoud Ataee Tarzanagh, and Navid Azizan. Optimizing attention with mirror descent: Generalized max-margin token selection. *Journal of Machine Learning Research*, 27(36):1–61, 2026. URL https://www.jmlr.org/papers/v27/25-0549.html.

Changmin Kang, Jihun Yun, Baekrok Shin, Yeseul Cho, and Chulhee Yun. Uniform spectral growth and convergence of Muon in LoRA-style matrix factorization. arXiv preprint arXiv:2602.06385, 2026. URL https://arxiv.org/abs/2602.06385.

Gyu Yeol Kim and Min-hwan Oh. Convergence of Muon with Newton–Schulz. arXiv preprint arXiv:2601.19156, 2026. URL https://arxiv.org/abs/2601.19156.

Kimi Team. Kimi K2: Open agentic intelligence. arXiv preprint arXiv:2507.20534, 2025. URL https://arxiv.org/abs/2507.20534.

Kimi Team. Kimi K3: Open frontier intelligence. arXiv preprint arXiv:2607.24653, 2026. URL https://arxiv.org/abs/2607.24653.

Steven G. Krantz and Harold R. Parks. *The Implicit Function Theorem: History, Theory, and Applications*. Modern Birkhäuser Classics. Birkhäuser, New York, 2013. doi: 10.1007/978-1-4614-5981-1.

Alexey Kravatskiy, Ivan Kozyrev, Nikolai Kozlov, Alexander Vinogradov, Daniil Merkulov, and Ivan Oseledets. Ky Fan norms and beyond: Dual norms and combinations for matrix optimization. arXiv preprint arXiv:2512.09678, 2025. URL https://arxiv.org/abs/2512.09678.

Adrian S. Lewis. Derivatives of spectral functions. *Mathematics of Operations Research*, 21(3):576–588, 1996. doi: 10.1287/moor.21.3.576. URL https://people.orie.cornell.edu/aslewis/publications/96-derivatives.pdf.

Jichu Li, Xuan Tang, and Difan Zou. The implicit bias of steepest descent with mini-batch stochastic gradient. arXiv preprint arXiv:2602.11557, 2026a. URL https://arxiv.org/abs/2602.11557.

Xianliang Li, Zihan Zhang, Weiyang Liu, and Han Bao. Denoise first, orthogonalize later: Understanding momentum in Muon via spectral filtering. arXiv preprint arXiv:2606.03899, 2026b. URL https://arxiv.org/abs/2606.03899.

Yibang Li, Bihari Lal Pandey, Ravi Sah, Andi Han, Cyrus Mostajeran, Pratik Jawanpuria, and Bamdev Mishra. Intrinsic Muon: Spectral optimization on Riemannian matrix manifolds. arXiv preprint arXiv:2605.09238, 2026c. URL https://arxiv.org/abs/2605.09238.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=AHOs7Sm5H7R.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training. arXiv preprint arXiv:2502.16982, 2025. URL https://arxiv.org/abs/2502.16982.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJeLIgBKPS.

Jianhao Ma, Yu Huang, Yuejie Chi, and Yuxin Chen. Preconditioning benefits of spectral orthogonalization in Muon. arXiv preprint arXiv:2601.13474, 2026. URL https://arxiv.org/abs/2601.13474.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *International Conference on Learning Representations Workshop Track*, 2015. URL https://arxiv.org/abs/1412.6614.

Mark Rhee, Jamie Simon, and Dhruva Karkada. Muon learns balanced solutions in matrix factorization without slow saddle-to-saddle dynamics. arXiv preprint arXiv:2606.30509, 2026. URL https://arxiv.org/abs/2606.30509.

Heejune Sheen, Siyu Chen, Tianhao Wang, and Harrison H. Zhou. Implicit regularization of gradient flow on one-layer softmax attention. arXiv preprint arXiv:2403.08699, 2024. URL https://arxiv.org/abs/2403.08699.

Devender Singh. The loss does not see the basis, but Adam does. arXiv preprint arXiv:2608.05136, 2026. URL https://arxiv.org/abs/2608.05136.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958. doi: 10.2140/pjm.1958.8.171.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642. Association for Computational Linguistics, 2013. URL https://aclanthology.org/D13-1170/.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018. URL https://jmlr.org/papers/v19/18-188.html.

Haoyuan Sun, Khashayar Gatmiry, Kwangjun Ahn, and Navid Azizan. A unified approach to controlling implicit regularization via mirror descent. *Journal of Machine Learning Research*, 24(393):1–58, 2023. URL https://www.jmlr.org/papers/v24/23-0836.html.

Nikolaos Tsilivis, Eitan Gronich, Julia Kempe, and Gal Vardi. Flavors of margin: Implicit bias of steepest descent in homogeneous neural networks. *Journal of Machine Learning Research*, 27(104):1–37, 2026. URL https://www.jmlr.org/papers/v27/25-0634.html.

Rahul Vashisht and Harish G. Ramaswamy. Faster query-key learning sharpens attention in self-attention models. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2608.06776.

Bhavya Vasudeva, Puneesh Deora, and Christos Thrampoulidis. Implicit bias and fast convergence rates for self-attention. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=pKilnjQsb0.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Hongtao Zhang, Wenjie Zhou, Wei Chen, and Xueqi Cheng. When and why grouping attention heads accelerates Muon optimization. arXiv preprint arXiv:2605.08933, 2026. URL https://arxiv.org/abs/2605.08933.

# Supplementary Material

Implicit Bias of Factorwise Spectral Optimization in Self-Attention


**A Shared geometry and proof of Lemma 5.1** . . . 15
A.1 Flow conventions and margin duality . . . 15
A.2 Proof of Lemma 5.1 and the spectral certificate . . . 16
**B Proof of Theorem 4.1: fixed-matrix selection** . . . 18
**C Proof of Theorem 5.2: time-dependent Gibbs weights** . . . 21
**D Proof of Theorem 6.2: local product convergence** . . . 22
D.1 Gradient weights and a local Lyapunov function . . . 23
D.2 The stationary-margin gap and product convergence . . . 24
**E Proof of Theorem 6.3: local joint convergence** . . . 26
E.1 Dual nondegeneracy and finite-time entry . . . 26
E.2 Local factor and log-weight coordinates . . . 28
E.3 Coupled Lyapunov estimate . . . 31
**F Examples and limits of the conclusions** . . . 33
F.1 Stationarity, orientation, and rank loss . . . 33
F.2 Product convergence does not identify normalized gradient weights . . . 34
F.3 Attention comparisons with a nondegenerate dual minimum . . . 35
**G Experimental details** . . . 36
G.1 Common numerical protocol and continuous-flow metrics . . . 36
G.2 Two-mode rank-one example: Figure 1 . . . 37
G.3 Additional controlled and frozen-token diagnostics . . . 38
G.4 Initialization and rank-boundary survey . . . 42
**H Comparison with prior implicit-bias results** . . . 43
**I Auxiliary lemmas** . . . 44


**Conventions.** Solutions are absolutely continuous and satisfy Equation (3.7) almost everywhere. At rank loss, the results apply to every absolutely continuous solution using allowed maximizing directions. Primes denote logarithmic time $t = \log((s+\tau_0)/\tau_0)$; in softmax attention, $\varepsilon = \Lambda^{-1}$ and $\varepsilon' = -2\varepsilon$. Throughout, $K, Q, W$ are normalized; $\widetilde{K}, \widetilde{Q}, \widetilde{W}$ are raw continuous-time variables. We use $O(h) = \{O \in \mathbb{R}^{h\times h} : O^\top O = I_h\}$; a common rotation may include a reflection. Pair norms are product Frobenius norms. Constants in local estimates depend on the fixed problem and target conditions, but not on the particular sufficiently small initial inverse scale.

Appendix A proves the geometric estimates used throughout. The fixed- and moving-gradient proofs follow in Appendices B and C. The attention proofs in Appendices D and E share the prefactor and energy estimates, but use different arguments to identify the limit: stationary margin values for the product, and dual curvature for the weights. Neither theorem is needed to prove the other. Appendices F–H contain the examples, experimental details, and comparison with prior work. Appendix I collects the matrix and analytic facts used in these proofs.

# A SHARED GEOMETRY AND PROOF OF LEMMA 5.1

For $K \in \mathbb{R}^{m\times h}$, $Q \in \mathbb{R}^{n\times h}$, and $1 \le h \le \min\{m,n\}$, we compute the support function of the feasible products. Under a strict cutoff gap, the support-function residual bounds the factor and product errors.

## A.1 FLOW CONVENTIONS AND MARGIN DUALITY

**Relation to Muon.** For discrete factors $K_n, Q_n$ with $W_n = K_n Q_n^\top$, the negative factor gradients are

$$g_{K,n} = (-\nabla L(W_n))Q_n, \qquad g_{Q,n} = (-\nabla L(W_n))^\top K_n.$$

Muon forms a momentum direction and approximately orthogonalizes it before updating each matrix block (Jordan et al., 2024). For ordinary momentum, a schematic factorwise update is

$$\begin{aligned} M_{X,n+1} &= \beta_{\mathrm{m}} M_{X,n} + (1-\beta_{\mathrm{m}}) g_{X,n}, \\ X_{n+1} &= X_n + \alpha_n \,\mathrm{NS}_q(M_{X,n+1}), \qquad X \in \{K,Q\}, \end{aligned}$$

where $M_{X,0} = 0$, $0 \le \beta_{\mathrm{m}} < 1$, $\alpha_n > 0$, and $\mathrm{NS}_q$ denotes $q$ Newton–Schulz steps after Frobenius normalization, with $\mathrm{NS}_q(0) = 0$. Both gradients use the same pre-update factors. The Nesterov variant changes the momentum direction supplied to this iteration; matrix-dependent update scaling, weight decay, and clipping are omitted here.

We instead remove momentum, replace approximate orthogonalization by an exact maximizer over the operator-norm unit ball, and pass to raw continuous-time variables $\widetilde{K}(s), \widetilde{Q}(s)$, giving Equation (3.6). For a full-column-rank input the maximizer is its polar factor; at rank deficiency, the inclusion admits every maximizer. This flow isolates the factorwise spectral geometry rather than the full practical algorithm. It does not assume that the tuned Newton–Schulz iteration converges to the polar factor as $q \to \infty$.

**Normalized flow.** The change of time satisfies $ds/dt = \tau$. Thus, with dots denoting raw optimizer-time derivatives,

$$\frac{d}{dt}\frac{\widetilde{K}(s(t))}{\tau(s(t))} = \dot{\widetilde{K}}(s(t)) - K(t).$$

The negative factor gradient is a positive scalar multiple of $ZQ$ whenever $Z$ is defined. Positive rescaling leaves the set of support maximizers unchanged, giving Equation (3.7); the calculation for $Q$ is identical. Write $\mathcal{B} = \{(K,Q) : \|K\|_{\mathrm{op}}, \|Q\|_{\mathrm{op}} \le 1\}$. A measurable choice of maximizing directions in Equation (3.7) gives $K' = S_K - K$ and $Q' = S_Q - Q$, with $\|S_K\|_{\mathrm{op}}, \|S_Q\|_{\mathrm{op}} \le 1$. Thus

$$K(t) = e^{-(t-t_0)}K(t_0) + \int_{t_0}^{t} e^{-(t-u)} S_K(u)\, du, \tag{A.1}$$

and likewise for $Q$. Taking operator norms gives

$$\|K(t)\|_{\mathrm{op}} \le e^{-(t-t_0)} + \int_{t_0}^{t} e^{-(t-u)}\, du = 1.$$

The factor balls are therefore invariant for every maximizing selection. Since $\|K'\|_{\mathrm{op}}, \|Q'\|_{\mathrm{op}} \le 2$, each factor also has a limit at any finite endpoint of its existence interval. For fixed $Z$, the right-hand side of the inclusion has nonempty compact convex values and a closed graph, so local solutions exist (Aubin & Cellina, 1984; Cortés, 2008). Restarting at a finite endpoint extends a solution globally. The invariance calculation also applies to a time-dependent matrix $Z(t)$. For the Gibbs-weighted and softmax-attention flows, the same continuation argument applies wherever the continuous state-dependent normalization is defined.

The support-maximizer identity in Lemma I.1 gives

$$S^\top X = |X| := (X^\top X)^{1/2}, \qquad S \in \partial\|\cdot\|_*(X). \tag{A.2}$$

To prove this at a possibly singular $X$, take a compact SVD $X = U_r \Sigma_r V_r^\top$. Each $u_i^\top S v_i \le 1$, and equality in $\langle S, X\rangle = \|X\|_*$ forces every deficit $1 - u_i^\top S v_i$ with $\sigma_i > 0$ to vanish. Since $S$ is a contraction, equality in Cauchy–Schwarz gives $Sv_i = u_i$ and $S^\top u_i = v_i$. Hence

$$S^\top X = \sum_{i=1}^{r} \sigma_i (S^\top u_i) v_i^\top = \sum_{i=1}^{r} \sigma_i v_i v_i^\top = |X|.$$

The calculation uses only the positive-singular-value subspaces and therefore holds for every support maximizer.

**Product geometry.** The feasible products are exactly

$$\mathcal{B}_h = \{KQ^\top : (K, Q) \in \mathcal{B}\} = \{W : \operatorname{rank} W \le h,\ \|W\|_{\mathrm{op}} \le 1\}.$$

The forward inclusion uses $\operatorname{rank}(KQ^\top) \le h$ and submultiplicativity. Conversely, write a rank-$r$ product as $W = U_r \Sigma_r V_r^\top$, where $r \le h$ and $0 \preceq \Sigma_r \preceq I_r$. The factors $U_r \Sigma_r^{1/2}$ and $V_r \Sigma_r^{1/2}$, padded with zero columns, are contractions and multiply to $W$. For an arbitrary matrix $\widetilde{W}$ of rank at most $h$, the same factorization gives

$$\inf_{\widetilde{K}\widetilde{Q}^\top = \widetilde{W}} \max\{\|\widetilde{K}\|_{\mathrm{op}}^2, \|\widetilde{Q}\|_{\mathrm{op}}^2\} = \|\widetilde{W}\|_{\mathrm{op}}.$$

To compute the support function, let $s_i = \sigma_i(W)$. For $W \in \mathcal{B}_h$, only the first $h$ of these can be nonzero and each is at most one. Von Neumann's trace inequality (Lemma I.1) therefore bounds $\langle W, Z\rangle$ by $\sum_{i=1}^{h} \sigma_i(Z)$. The product $U_h V_h^\top$ attains this bound. The same bound holds under $0 \le s_i \le 1$ and $\sum_i s_i \le h$: allocating the singular-value budget to the largest coefficients maximizes $\sum_i s_i \sigma_i(Z)$. This computes the support function of both $\mathcal{B}_h$ and $\mathcal{C}_h = \{W : \|W\|_{\mathrm{op}} \le 1,\ \|W\|_* \le h\}$:

$$\sup_{W \in \mathcal{B}_h} \langle W, Z\rangle = \sup_{W \in \mathcal{C}_h} \langle W, Z\rangle = \|Z\|_{(h)}, \qquad \operatorname{conv} \mathcal{B}_h = \mathcal{C}_h. \tag{A.3}$$

Here $\operatorname{conv} \mathcal{B}_h$ is compact and has the same support function as $\mathcal{C}_h$; equality of compact convex sets follows. Hence $\mathcal{C}_h$ is the unit ball of $\|W\|_{(h)}^* = \max\{\|W\|_{\mathrm{op}}, \|W\|_*/h\}$ (Bhatia, 1997; Grussler & Giselsson, 2018).

The pairing $\langle W, \Delta(p)\rangle$ is bilinear on the compact convex sets $\mathcal{C}_h$ and $\mathcal{P}_M$. Since $\min_j \langle W, \Delta_j\rangle = \min_{p \in \mathcal{P}_M} \langle W, \Delta(p)\rangle$, Sion's minimax theorem (Lemma I.8; Sion, 1958) gives

$$\gamma_h = \max_{W \in \mathcal{C}_h} \min_{p \in \mathcal{P}_M} \langle W, \Delta(p)\rangle = \min_{p \in \mathcal{P}_M} \|\Delta(p)\|_{(h)}.$$

In particular, separability gives $\|\Delta(p)\|_{(h)} \ge \gamma_h > 0$ for every $p$. This prevents cancellation in the normalization of $\Delta(p)$ and, together with the preceding continuation argument, gives global solutions with time-dependent Gibbs weights. For the unit-margin formulation, a normalized optimizer divided by $\gamma_h$ is feasible with norm at most $1/\gamma_h$. Conversely, any unit-margin feasible $W$ gives the normalized feasible point $W/\|W\|_{(h)}^*$ with minimum margin at least $1/\|W\|_{(h)}^*$. Hence $\|W\|_{(h)}^* \ge 1/\gamma_h$, proving Equation (3.4) and its value. The convex-hull identity does not by itself imply $\gamma_h^{\mathrm{rk}} = \gamma_h$: the minimum-margin objective is concave, not linear, in $W$.

### A.2 Proof of Lemma 5.1 and the spectral certificate

*Proof of Lemma 5.1.* **Part (i).** Operator–nuclear duality gives $\langle Z, KQ^\top\rangle \le \|ZQ\|_*$ and $\langle Z, KQ^\top\rangle \le \|Z^\top K\|_*$. The singular-value bounds in Lemma I.2 give

$$\|ZQ\|_* \le \|Q\|_{\mathrm{op}} \|Z\|_{(h)}, \qquad \|Z^\top K\|_* \le \|K\|_{\mathrm{op}} \|Z\|_{(h)}.$$

Averaging proves the first chain in Equation (5.1). The last inequality follows from Equation (A.3) applied to $W/\|W\|_{\mathrm{op}}$ when $W \neq 0$; both sides are zero when $W = 0$.

**Part (ii).** We bound the support-function residual using the distances to the two polar updates, the factor norm deficits, and the components outside the leading subspaces. Set $b = \sigma_{h+1}(Z)$, so $a > b \ge 0$. Both $ZQ$ and $Z^\top K$ have full column rank. Let $S_K = \operatorname{polar}(ZQ)$, $S_Q = \operatorname{polar}(Z^\top K)$, and

$$d^2 = \|S_K - K\|_F^2 + \|S_Q - Q\|_F^2, \qquad d_0 = 2h - \|K\|_F^2 - \|Q\|_F^2 \ge 0.$$

Writing $ZQ = S_K R_Q$ with $R_Q \succeq aI_h$, we have $I_h - \mathrm{sym}(S_K^\top K) \succeq 0$ and

$$\|ZQ\|_* - \langle K, ZQ\rangle = \mathrm{tr}\big(R_Q[I_h - \mathrm{sym}(S_K^\top K)]\big) \geq \frac{a}{2}\big(\|S_K - K\|_F^2 + h - \|K\|_F^2\big).$$

The matrix inequality follows from $v^\top \mathrm{sym}(S_K^\top K)v = \langle S_K v, Kv\rangle \leq \|v\|^2$, and Lemma I.3 gives $\mathrm{tr}(R_Q P) \geq a\,\mathrm{tr}\,P$ for $P \succeq 0$. Since $S_K^\top S_K = I_h$,

$$2\,\mathrm{tr}[I_h - \mathrm{sym}(S_K^\top K)] = \|S_K - K\|_F^2 + h - \|K\|_F^2.$$

Adding the corresponding bound for $Q$ gives

$$\mathcal{D}_Z(K, Q) \geq \frac{a}{2}(d^2 + d_0). \tag{A.4}$$

Because $a \leq \sigma_h(Z)$, the inequality $a > b$ also gives a strict cutoff gap. Let $U, V$ contain the leading $h$ singular vectors, and put $K_\perp = (I - UU^\top)K$, $Q_\perp = (I - VV^\top)Q$, and $T = \|K_\perp\|_F^2 + \|Q_\perp\|_F^2$. The omitted block of $Z$ has operator norm $b$, and

$$(I - UU^\top)S_K = (I - UU^\top)ZQ_\perp R_Q^{-1}.$$

Thus the polar formulas and $\|R_Q^{-1}\|_{\mathrm{op}} \leq 1/a$ imply

$$\|(I - UU^\top)S_K\|_F \leq (b/a)\|Q_\perp\|_F, \qquad \|(I - VV^\top)S_Q\|_F \leq (b/a)\|K_\perp\|_F.$$

Projecting $(K, Q) = (S_K, S_Q) + (K - S_K, Q - S_Q)$ onto the two omitted spaces, then taking the norm of the pair, yields

$$\sqrt{T} \leq (b/a)\sqrt{T} + d, \qquad T \leq \frac{d^2}{(1 - b/a)^2}. \tag{A.5}$$

To control the remaining nuclear-norm deficits, write $K_h = U^\top K$ and $Q_h = V^\top Q$. Since $\|UQ_h\|_{\mathrm{op}} \leq 1$,

$$\|ZQ\|_* \geq \langle UQ_h, ZQ\rangle = \mathrm{tr}(Q_h^\top \Sigma_h Q_h), \qquad 1 - \|ZQ\|_* \leq \sigma_1(Z)(h - \|Q_h\|_F^2).$$

The last inequality uses $Q_h Q_h^\top \preceq I_h$ and $\mathrm{tr}\,\Sigma_h = 1$. The two norm deficits satisfy

$$(h - \|Q_h\|_F^2) + (h - \|K_h\|_F^2) = d_0 + T.$$

Since $0 \leq b/a < 1$, Equation (A.5) also gives

$$d_0 + T \leq \frac{d_0 + d^2}{(1 - b/a)^2} \leq \frac{2a}{(a - b)^2}\mathcal{D}_Z.$$

Adding the corresponding nuclear-norm estimate for $K$ therefore gives

$$[1 - \|ZQ\|_*] + [1 - \|Z^\top K\|_*] \leq \sigma_1(Z)(d_0 + T) \leq \frac{2\sigma_1(Z)a}{(a - b)^2}\mathcal{D}_Z.$$

Now $2r_Z = \mathcal{D}_Z + [1 - \|ZQ\|_*] + [1 - \|Z^\top K\|_*]$, proving the first bound in Equation (5.2). Since $a \leq \sigma_1(Z) \leq 1$ and $0 < \eta = a - b \leq 1$, its coefficient is at least $2\eta^2/(\eta^2 + 2) \geq (2/3)\eta^2$.

**Part (iii).** Let $Z$ have a strict cutoff gap and let $K, Q$ be contractions. Decompose the score in the leading singular-vector bases $U, V$, write $K_h = U^\top K$, $Q_h = V^\top Q$, and retain the definitions of $T, d_0$. Set $S_0 = \|K_h - Q_h\|_F^2$ and $\sigma = \sigma_h(Z)$. The leading block satisfies

$$\begin{aligned}\|Z\|_{(h)} - \mathrm{tr}(K_h^\top \Sigma_h Q_h) &= \tfrac{1}{2}\,\mathrm{tr}\big(\Sigma_h[I_h - K_h K_h^\top]\big) + \tfrac{1}{2}\,\mathrm{tr}\big(\Sigma_h[I_h - Q_h Q_h^\top]\big) \\ &\quad + \tfrac{1}{2}\,\mathrm{tr}\big((K_h - Q_h)^\top \Sigma_h (K_h - Q_h)\big) \geq \tfrac{\sigma}{2}(d_0 + T + S_0).\end{aligned}$$

For the omitted block, Cauchy–Schwarz and $2ab \leq a^2 + b^2$ give

$$|\langle K_\perp, ZQ_\perp\rangle| \leq \sigma_{h+1}(Z)\|K_\perp\|_F\|Q_\perp\|_F \leq \frac{\sigma_{h+1}(Z)}{2}T.$$

Combining the two blocks gives

$$r_Z(K,Q) \geq \frac{\mathrm{gap}_h(Z)}{2} T + \frac{\sigma}{2}(d_0 + S_0). \tag{A.6}$$

For $M = (K_h + Q_h)/2$, choose a nearest orthogonal matrix $O$ by a full SVD, as in Lemma I.5. The contraction bound $\|M\|_{\mathrm{op}} \leq 1$ gives

$$\|M - O\|_F^2 = \sum_i (1 - \sigma_i(M))^2 \leq \sum_i (1 - \sigma_i(M)^2) = h - \|M\|_F^2.$$

For any common $O$, the two leading-space errors satisfy

$$\|K_h - O\|_F^2 + \|Q_h - O\|_F^2 = \tfrac{1}{2}\|K_h - Q_h\|_F^2 + 2\|M - O\|_F^2.$$

Using $2\|M\|_F^2 + S_0/2 = \|K_h\|_F^2 + \|Q_h\|_F^2$ and minimizing over the same $O$ for both factors yields

$$\mathrm{dist}((K,Q), \mathcal{O}_h(Z))^2 = T + \tfrac{1}{2}S_0 + 2\,\mathrm{dist}(M, O(h))^2 \leq d_0 + 2T + S_0 \leq \frac{4 r_Z(K,Q)}{\mathrm{gap}_h(Z)}.$$

To pass to the product, use $KQ^\top - UV^\top = (K - UO)Q^\top + UO(Q - VO)^\top$. Since $Q$ and $UO$ have operator norm at most one, for the same $O$,

$$\|KQ^\top - UV^\top\|_F \leq \|K - UO\|_F + \|Q - VO\|_F \leq \sqrt{2}\,\mathrm{dist}((K,Q), \mathcal{O}_h(Z)),$$

which proves the product bound.

Under a strict cutoff, this distance estimate also identifies the equality cases in part (i). If $\|ZQ\|_* = \|Z\|_{(h)}$, choose a nuclear-norm maximizer $S$ for $ZQ$. Then $r_Z(S,Q) = 0$, so $Q = V_h O_Q$. Similarly, $\|Z^\top K\|_* = \|Z\|_{(h)}$ forces $K = U_h O_K$. Both nuclear-norm bounds are saturated by these factors, while equality with the product score requires

$$\mathrm{tr}(\Sigma_h O_Q O_K^\top) = \mathrm{tr}\,\Sigma_h.$$

Since $\Sigma_h \succ 0$, this holds exactly when $O_Q = O_K$. More generally, equality in $\|ZQ\|_* \leq r_Q \|Z\|_{(h)}$, where $r_Q = \|Q\|_{\mathrm{op}}$, is equivalent to $QQ^\top = r_Q^2 V_h V_h^\top$, including $r_Q = 0$; the transposed characterization holds for $K$. Finally, for $W \neq 0$, equality in the last bound of Equation (5.1) is equivalent to $W/\|W\|_{\mathrm{op}} = U_h V_h^\top$. □

**The spectral certificate.** Under (MM), $W^\star = \mathrm{polar}_h(\Delta^\star)$ is feasible for the rank-constrained problem with minimum margin $\|\Delta^\star\|_{(h)}$, and $p^\star$ has the same dual value. Thus $\gamma_h^{\mathrm{rk}} = \gamma_h = \|\Delta^\star\|_{(h)}$. Every convex optimizer $W$ satisfies

$$\gamma_h \leq \langle W, \Delta^\star \rangle \leq \|\Delta^\star\|_{(h)} = \gamma_h.$$

Part (iii) makes $W^\star$ the unique support maximizer on $\mathcal{B}_h$. To extend uniqueness to its convex hull, write $W = \sum_\ell \theta_\ell W_\ell$ with $W_\ell \in \mathcal{B}_h$, $\theta_\ell \geq 0$, and $\sum_\ell \theta_\ell = 1$. The nonnegative deficits $\gamma_h - \langle W_\ell, \Delta^\star \rangle$ have weighted sum zero, so every $W_\ell$ with $\theta_\ell > 0$ equals $W^\star$. Thus $W^\star$ is the unique convex optimizer, with rank $h$ and operator norm one. Conversely, if $p^\star$ is a dual minimizer with a strict cutoff gap, the strict-gap derivative in Lemma I.7 gives gradient $W^\star$. The feasible directions $e_j - p^\star$ imply $\langle W^\star, \Delta_j \rangle \geq \gamma_h$, with equality on the support after averaging with $p^\star$. This identifies the product from (MM); identification of the weights is treated in Appendix E.1.

## B Proof of Theorem 4.1: fixed-matrix selection

We use cross-Gram monotonicity to preserve rank, then contract the support-function residual by Lemma 5.1. An integrable velocity gives a factor limit, and linearization at that limit gives the spectral rate in part (iii).

**The symmetric part is nondecreasing.** Write $f = (K,Q)$, $\Gamma = K^\top ZQ$, and $G = \mathrm{sym}\,\Gamma$. The support identity Equation (A.2) gives, almost everywhere,

$$\Gamma' = |ZQ| + |Z^\top K| - 2\Gamma. \tag{B.1}$$

Since $\|K\|_{\mathrm{op}}, \|Q\|_{\mathrm{op}} \le 1$, $\Gamma^\top \Gamma \preceq |ZQ|^2$ and $\Gamma\Gamma^\top \preceq |Z^\top K|^2$. The positive square root preserves these inequalities by Lemma I.3. For any square matrix $B = U\Sigma V^\top$, the block matrix has the factorization

$$\begin{bmatrix} |B^\top| & B \\ B^\top & |B| \end{bmatrix} = \begin{bmatrix} U \\ V \end{bmatrix} \Sigma \begin{bmatrix} U^\top & V^\top \end{bmatrix} \succeq 0.$$

Testing on $(v, -v)$ and $(v, v)$ yields

$$B + B^\top \preceq |B^\top| + |B|, \qquad v^\top(|B^\top| + |B|)v \ge 2|v^\top \operatorname{sym}(B) v|.$$

Consequently,

$$G' \succeq 0, \qquad \operatorname{skew}\Gamma(t) = e^{-2(t-t_0)} \operatorname{skew}\Gamma(t_0), \qquad v^\top(|ZQ| + |Z^\top K|)v \ge 2|v^\top G v|. \tag{B.2}$$

Let $\lambda = \lambda_{\min}(G)$. By Lemma I.4, $\lambda$ is locally absolutely continuous whenever $G$ is. At a differentiability time, its derivative is determined by the quadratic forms of $G'$ on the minimum eigenspace. Every unit vector $v$ in that eigenspace satisfies

$$v^\top G' v \ge 2|v^\top G v| - 2v^\top G v = 2|\lambda| - 2\lambda = 4(-\lambda)_+.$$

Because the bound holds for every vector in the minimum eigenspace, it also applies at repeated eigenvalues. Hence $\lambda' \ge 4(-\lambda)_+$ almost everywhere. The negative part $y = (-\lambda)_+ = \|G_-\|_{\mathrm{op}}$ satisfies $y' \le -4y$ almost everywhere (and $y' = 0$ almost everywhere on $\{y = 0\}$). Multiplying by $e^{4(t-t_0)}$ and integrating proves part (i).

If $G(t_0) \succeq \mu I_h$ with $\mu > 0$, monotonicity gives the same lower bound at every later time. For any $v$,

$$\mu\|v\|^2 \le v^\top \Gamma v \le \|\Gamma v\|\,\|v\|.$$

For unit $v$, this gives $\|\Gamma v\| \ge \mu$, and the same bound holds for $\Gamma^\top$ because it has the same symmetric part. Since both factors are contractions,

$$\|ZQv\| \ge \|K^\top ZQv\| = \|\Gamma v\| \ge \mu, \qquad \|Z^\top K v\| \ge \|Q^\top Z^\top K v\| \ge \mu.$$

Taking the minimum over unit vectors gives

$$\sigma_h(ZQ(t)),\ \sigma_h(Z^\top K(t)) \ge \mu \qquad (t \ge t_0). \tag{B.3}$$

Both polar maps are smooth on this uniformly full-rank region, proving part (ii), including uniqueness.

**Residual decay gives a factor limit.** For part (iii), set $\mu = \lambda_{\min}(G(t_0))$ and $\eta_0 = \mu - \sigma_{h+1}(Z) > 0$. The polar-input separation is at least $\eta_0$ by Equation (B.3). The support-function residual is $r_Z(f(t)) = 1 - \langle W(t), Z\rangle$. Since $Z$ is fixed, Lemma 5.1(ii) gives

$$\frac{d}{dt} r_Z(f(t)) = -\mathcal{D}_Z(f(t)) \le -\tfrac{2}{3}\eta_0^2 r_Z(f(t)), \qquad r_Z(f(t)) \le r_Z(f(t_0)) e^{-(2/3)\eta_0^2(t-t_0)}. \tag{B.4}$$

Substituting this estimate into Lemma 5.1(iii) gives

$$\operatorname{dist}(f(t), \mathcal{O}_h(Z))^2 \le \frac{4 r_Z(f(t_0))}{\operatorname{gap}_h(Z)} e^{-(2/3)\eta_0^2(t-t_0)},$$
$$\|W(t) - \operatorname{polar}_h(Z)\|_F^2 \le \frac{8 r_Z(f(t_0))}{\operatorname{gap}_h(Z)} e^{-(2/3)\eta_0^2(t-t_0)}.$$

The smooth vector field $b_Z$ vanishes on the compact orbit $\mathcal{O}_h(Z)$. In a sufficiently small tube it is Lipschitz, so

$$\|f'(t)\| = \|b_Z(f(t))\| \le C \operatorname{dist}(f(t), \mathcal{O}_h(Z)).$$

The factor distance decays at least as $Ce^{-\eta_0^2(t-t_0)/3}$ by Equation (B.4). Thus for sufficiently large $T$ and $t_2 > t_1 \ge T$,

$$\|f(t_2) - f(t_1)\| \le \int_{t_1}^{t_2} \|f'(u)\|\,du \le C \int_{t_1}^{\infty} e^{-\eta_0^2(u-t_0)/3}\,du \longrightarrow 0.$$

The factors are therefore Cauchy and converge to one pair $(U_h O_\infty, V_h O_\infty)$ on the closed orbit. Integrability of the velocity is needed here: decay of the distance to the orbit alone would not rule out motion along it.

**The normal spectrum gives the rate.** The exponent above depends on the initial eigenvalue bound. For the rate stated in the theorem, linearize at $(U_h, V_h)$; common right rotations give the other representatives. At a full-column-rank polar decomposition $X = U_X R_X$, where $R_X = (X^\top X)^{1/2} \succ 0$, the polar derivative (Lemma I.6) is

$$D\,\mathrm{polar}(X)[E] = U_X\Omega + (I - U_X U_X^\top)ER_X^{-1}, \qquad \Omega R_X + R_X\Omega = U_X^\top E - E^\top U_X, \quad \Omega^\top = -\Omega. \tag{B.5}$$

Differentiating $X = U_X R_X$ and $U_X^\top U_X = I_h$ gives $E = \dot{U}_X R_X + U_X \dot{R}_X$ with $\Omega = U_X^\top \dot{U}_X$ skew-symmetric. Subtracting the transpose of $U_X^\top E$ gives the Sylvester equation. The orthogonal component is $(I - U_X U_X^\top)\dot{U}_X = (I - U_X U_X^\top)ER_X^{-1}$. In a basis diagonalizing $R_X$, the skew entries satisfy

$$\Omega_{ab} = \frac{(U_X^\top E)_{ab} - (U_X^\top E)_{ba}}{s_a + s_b}.$$

The denominators are positive, so the solution is unique (Higham, 1986; Gawlik & Leok, 2017). Write the selected-block perturbations as $E_\pm = (E_K \pm E_Q)/2 = S_\pm + \Omega_\pm$, with symmetric $S_\pm$ and skew-symmetric $\Omega_\pm$. For $(T_\Sigma S)_{ab} = (\sigma_a - \sigma_b)S_{ab}/(\sigma_a + \sigma_b)$, the linearized equations are

$$S_+' = -S_+, \quad \Omega_+' = T_\Sigma S_+, \qquad S_-' = -S_-, \quad \Omega_-' = -2\Omega_- - T_\Sigma S_-.$$

The symmetric modes have eigenvalue $-1$ and the relative-skew modes have eigenvalue $-2$. Symmetric forcing can give a relative-skew perturbation a rate-one component. The common-skew modes are tangent to the orbit and have eigenvalue zero. The coupling from $S_+$ to $\Omega_+$ connects distinct eigenvalues, so the zero modes are semisimple. An omitted paired direction $i$ coupled to a selected direction $j$ gives

$$\begin{pmatrix} b_K \\ b_Q \end{pmatrix}' = \begin{pmatrix} -1 & \sigma_i(Z)/\sigma_j(Z) \\ \sigma_i(Z)/\sigma_j(Z) & -1 \end{pmatrix} \begin{pmatrix} b_K \\ b_Q \end{pmatrix}. \tag{B.6}$$

Its eigenvalues are $-1 \pm \sigma_i(Z)/\sigma_j(Z)$. Unpaired rectangular null directions have eigenvalue $-1$. Hence all normal eigenvalues have real part at most $-\lambda_*$, where $\lambda_* = 1 - \sigma_{h+1}(Z)/\sigma_h(Z) > 0$, and the kernel consists exactly of orbit tangents.

To remove the neutral directions explicitly, put $M = U_h^\top K + V_h^\top Q$ and $O = \mathrm{polar}(M)$. Near the orbit $M$ is invertible. The aligned pair $(KO^\top, QO^\top)$ minimizes the distance to $(U_h, V_h)$, since maximizing $\langle O, M\rangle$ is precisely the common Procrustes problem. Its displacement $x$ satisfies

$$\mathrm{skew}(U_h^\top x_K + V_h^\top x_Q) = 0, \qquad \|x\| = \mathrm{dist}((K, Q), \mathcal{O}_h(Z)).$$

The aligned field is smooth and vanishes at $x = 0$, which represents the equilibrium $(U_h, V_h)$. Its Taylor expansion is therefore $x' = A_\perp x + O(\|x\|^2)$. The slice tangent space is the orthogonal complement of the orbit tangents. The ambient linearization kills those tangents, so the induced operator on this slice has exactly the nonzero spectrum computed above. For any $0 < \nu < \lambda_*$, Lemma I.11 applied to $A_\perp + \nu I$ shows that the matrix

$$X_\nu = \int_0^\infty e^{(A_\perp + \nu I)^\top u} e^{(A_\perp + \nu I)u}\, du$$

is positive definite and satisfies $A_\perp^\top X_\nu + X_\nu A_\perp = -I - 2\nu X_\nu$. Thus, once $x$ is small,

$$\frac{d}{dt}(x^\top X_\nu x) = -\|x\|^2 - 2\nu x^\top X_\nu x + O(\|x\|^3) \le -2\nu x^\top X_\nu x.$$

Shrinking the neighborhood absorbs the cubic remainder into $-\|x\|^2$. The factor convergence already proved ensures eventual entry; $\nu < \lambda_*$ ensures that the integral defining $X_\nu$ converges. Norm equivalence and the product estimate in Lemma 5.1 give Equation (4.2); enlarging $C_\nu$ covers the finite initial interval. □

A one-sided polar initialization makes the threshold directly checkable: if $\|Q(t_0)\|_{\mathrm{op}} \le 1$ and $ZQ(t_0)$ has full column rank, then $K(t_0) = \mathrm{polar}(ZQ(t_0))$ gives $G(t_0) = |ZQ(t_0)|$. In particular, $\sigma_h(ZQ(t_0)) > \sigma_{h+1}(Z)$ implies selection. The threshold is sharp among equilibria with full-column-rank $ZQ$ and $Z^\top K$: if $\sigma_{h+1}(Z) > 0$, selecting directions $\{1, \dots, h-1, h+1\}$ gives equality and a nonleading product. When $\sigma_{h+1}(Z) = 0$, equality is on the rank-deficient boundary.

## C Proof of Theorem 5.2: time-dependent Gibbs weights

We differentiate a soft minimum through $W$, $\Lambda$, and $\pi$ to separate factor dissipation from the scale and base-weight terms, without derivatives of $p$ or $Z$. Correcting the resulting error gives a scalar contraction. Write $m_{\min}(W) = \min_j m_j(W)$ and define

$$m_\Lambda = -\Lambda^{-1} \log \sum_j \pi_j e^{-\Lambda m_j(W)}. \tag{C.1}$$

The Gibbs representation and $\pi_j \geq \pi_-$ imply

$$\begin{gathered} m_{\min}(W) \leq m_\Lambda \leq m_{\min}(W) + C_\pi/\Lambda, \qquad m_\Lambda - \langle W, \Delta(p)\rangle = \Lambda^{-1} D_{\mathrm{KL}}(p\|\pi), \\ 0 \leq D_{\mathrm{KL}}(q\|\pi) \leq C_\pi \quad (q \in \mathcal{P}_M), \qquad C_\pi = \log(1/\pi_-). \end{gathered} \tag{C.2}$$

Factoring $e^{-\Lambda m_{\min}}$ out of the sum leaves a quantity between $\pi_-$ and one, which proves the first two bounds. Taking logarithms in the Gibbs formula gives

$$\log(p_j/\pi_j) = \Lambda(m_\Lambda - m_j(W)).$$

Averaging with $p$ proves the entropy identity, while $D_{\mathrm{KL}}(q\|\pi) = \sum_j q_j \log q_j - \sum_j q_j \log \pi_j \leq \log(1/\pi_-)$ follows because the first sum is nonpositive. At almost every time, the maximizing directions of the two factors satisfy

$$\langle W', \Delta(p)\rangle = \|\Delta(p)Q\|_* + \|\Delta(p)^\top K\|_* - 2\langle W, \Delta(p)\rangle = \mathcal{D}_{\Delta(p)}(K, Q). \tag{C.3}$$

Let $F = \sum_j \pi_j e^{-\Lambda m_j(W)}$ and $\mu_j = m_j(W)$. On each finite interval all these functions are absolutely continuous and $F > 0$. At almost every time,

$$\frac{F'}{F} = \sum_j p_j \big((\log \pi_j)' - 2\Lambda\mu_j - \Lambda\mu_j'\big),$$

where we used $\Lambda' = 2\Lambda$. Therefore

$$\begin{aligned} m_\Lambda' &= \frac{2}{\Lambda} \log F - \frac{1}{\Lambda}\frac{F'}{F} \\ &= \sum_j p_j \mu_j' - 2\big(m_\Lambda - \sum_j p_j \mu_j\big) - \frac{1}{\Lambda}\sum_j p_j (\log \pi_j)'. \end{aligned}$$

Substituting Equation (C.3) and the entropy identity in Equation (C.2) gives

$$m_\Lambda' = \mathcal{D}_{\Delta(p)} - \Lambda^{-1}\left[2D_{\mathrm{KL}}(p\|\pi) + \sum_j p_j (\log \pi_j)'\right]. \tag{C.4}$$

The stated bounds on $\pi$ give an absolute bound of $2C_\pi + B_\pi$ on the bracket.

**The corrected gap contracts.** Put $A_\pi = 2C_\pi + B_\pi$ and define

$$Y = \gamma_h - m_\Lambda + \frac{A_\pi}{\Lambda}, \qquad 0 \leq e_m \leq Y. \tag{C.5}$$

By Equation (C.2),

$$Y - e_m = m_{\min} - m_\Lambda + A_\pi/\Lambda \geq (A_\pi - C_\pi)/\Lambda \geq 0,$$

and $e_m \geq 0$ because $W \in \mathcal{B}_h \subseteq \mathcal{C}_h$. Write $a = \|\Delta(p)\|_{(h)}$, so $\Delta(p) = aZ$ and $a \geq \gamma_h > 0$. Both quantities are homogeneous in the matrix argument:

$$\mathcal{D}_{aZ}(K, Q) = a\mathcal{D}_Z(K, Q), \qquad r_{aZ}(K, Q) = a r_Z(K, Q).$$

For normalized $Z$, the definition of $\kappa$ gives $\mathcal{D}_Z \geq \kappa(\eta) r_Z$ when $\eta > 0$ and $r_Z > 0$. Multiplying by $a$ yields

$$\mathcal{D}_{\Delta(p)} \geq \kappa(\eta)\big(\|\Delta(p)\|_{(h)} - \langle W, \Delta(p)\rangle\big).$$

For $\eta \le 0$ or zero support-function residual, the right side is zero and the bound follows from nonnegative dissipation. Dual feasibility and Equation (C.2) also give

$$\|\Delta(p)\|_{(h)} - \langle W, \Delta(p)\rangle \ge \gamma_h - m_\Lambda = Y - \frac{A_\pi}{\Lambda}.$$

Since $(A_\pi/\Lambda)' = -2A_\pi/\Lambda$, the bound on the error in $-m'_\Lambda$ gives

$$Y' \le -\mathcal{D}_{\Delta(p)} - A_\pi/\Lambda.$$

Combining this with the preceding lower bound on dissipation yields

$$Y' \le -\mathcal{D}_{\Delta(p)} - \frac{A_\pi}{\Lambda} \le -\kappa(\eta)Y - (1-\kappa(\eta))\frac{A_\pi}{\Lambda} \le -\kappa(\eta)Y, \tag{C.6}$$

where the last inequality uses the cap $\kappa \le 1$. Increasing the argument of $\kappa$ shrinks its admissible set, so $\kappa$ is nondecreasing and Borel measurable. Since $\eta(t)$ is continuous, the coefficient $\kappa(\eta(t))$ is measurable and bounded. The integrating-factor estimate in Lemma I.12 gives

$$e_m(t) \le Y(t_0)e^{-\Xi(t)}, \qquad Y(t_0) \le e_m(t_0) + \frac{A_\pi}{\Lambda_0} = C_0 - \frac{C_\pi}{\Lambda_0} \le C_0.$$

This uses the theorem's stated constant $C_0 = e_m(t_0) + (3C_\pi + B_\pi)/\Lambda_0$. The lower bound on $\Xi$ in Equation (5.7) gives the explicit integral bound in part (a).

If $\Xi(t) \to \infty$, the minimum margin tends to $\gamma_h$. Compactness of $\mathcal{B}_h$ and closedness of the rank constraint imply that every accumulation point of $W$ has rank at most $h$ and attains this value. Hence $\gamma_h^{\mathrm{rk}} = \gamma_h$, and every such point solves both margin problems. If the convex optimizer $W^\star$ is unique, the entire product trajectory converges to it: otherwise a sequence staying a fixed positive distance from $W^\star$ would have a different accumulation point by compactness. The limit has operator norm one: otherwise positive rescaling would give a rank-feasible product with margin larger than $\gamma_h > 0$. Thus

$$\frac{\widetilde{W}(s(t))}{\|\widetilde{W}(s(t))\|_{\mathrm{op}}} = \frac{W(t)}{\|W(t)\|_{\mathrm{op}}} \longrightarrow W^\star,$$

which proves parts (b) and (c). □

**Product error and sufficient conditions.** If a dual minimizer $p^\star$ has $\mathrm{gap}_h(\Delta^\star) > 0$, the certificate argument in Appendix A.2 identifies $W^\star = \mathrm{polar}_h(\Delta^\star)$. Since $0 \le \gamma_h - \langle W, \Delta^\star\rangle \le e_m$, Lemma 5.1(iii) gives

$$\mathrm{dist}((K,Q), \mathcal{O}_h(\Delta^\star))^2 \le \frac{4e_m}{\mathrm{gap}_h(\Delta^\star)}, \qquad \|W - W^\star\|_F^2 \le \frac{8e_m}{\mathrm{gap}_h(\Delta^\star)} \le \frac{8C_0e^{-\Xi(t)}}{\mathrm{gap}_h(\Delta^\star)}. \tag{C.7}$$

The quadratic-integral bound follows from Equation (5.7); the product estimate uses the strict cutoff of the chosen dual minimizer.

Because $\kappa$ is nondecreasing and positive at positive arguments, spending infinite total time above a fixed positive separation suffices for $\Xi \to \infty$; a uniform bound $\eta \ge \eta_0 > 0$ gives $\Xi(t) \ge \kappa(\eta_0)(t - t_0)$. A sufficiently small total variation of $Z(t)$ relative to the initial separation gives another sufficient condition. If $Z$ is locally absolutely continuous, differentiating $\Gamma = K^\top ZQ$ adds one term to the fixed-matrix identity:

$$\Gamma' = |ZQ| + |Z^\top K| - 2\Gamma + K^\top Z'Q. \tag{C.8}$$

The fixed-matrix contribution to $G'$ is positive semidefinite, whereas $\mathrm{sym}(K^\top Z'Q) \succeq -\|Z'\|_{\mathrm{op}} I_h$. Integration and Weyl's bound from Lemma I.2 give

$$\lambda_{\min}(G(t)) - \sigma_{h+1}(Z(t)) \ge \lambda_{\min}(G(t_0)) - \sigma_{h+1}(Z(t_0)) - 2\int_{t_0}^{t} \|Z'(u)\|_{\mathrm{op}}\, du. \tag{C.9}$$

If the right side remains above $s_0 > 0$, the pointwise rank bound in Appendix B gives $\eta \ge s_0$ and hence $\Xi(t) \to \infty$. Appendix F.2 gives product convergence with finite $\Xi(\infty)$.

## D Proof of Theorem 6.2: Local product convergence

Under (MM) and (TC), we use the local energy to preserve positive margins and make the dissipation integrable, then identify the limit from the finite set of stationary margin values. Appendix E uses the same prefactor and energy estimates.

### D.1 Gradient weights and a local Lyapunov function

All gradients here are with respect to $\widetilde{W}$, and key sums are restricted to $\mathcal{M}_{ir}$. Differentiating the output of one supervised query gives

$$\nabla u_{ir} = \beta \sum_{k \in \mathcal{M}_{ir}} \alpha_{irk}(g_{irk} - u_{ir}) x_{ik} x_{ir}^\top.$$

Since $\sum_k \alpha_{irk}(g_{irk} - u_{ir}) = 0$, subtracting $x_{i,k_{ir}^\star} x_{ir}^\top$ inside the sum removes the target term. Changing both signs in the remaining terms gives

$$\nabla u_{ir} = \beta \sum_{k \neq k_{ir}^\star} \alpha_{irk}(u_{ir} - g_{irk}) \Delta_{irk}.$$

Substitute $\alpha_{irk} = \alpha_{ir,k_{ir}^\star} e^{-\beta m_{j(i,r,k)}(\widetilde{W})}$ and apply the chain rule for the loss. This gives

$$\begin{aligned} -\nabla L(\widetilde{W}) &= \sum_j \rho_j(\widetilde{W}) e^{-\beta m_j(\widetilde{W})} \Delta_j, \\ \rho_{j(i,r,k)}(\widetilde{W}) &= \beta[-\partial_r \ell_i(\mathbf{u}_i(\widetilde{W}))] \alpha_{ir,k_{ir}^\star}(\widetilde{W})[u_{ir}(\widetilde{W}) - g_{irk}]. \end{aligned} \tag{D.1}$$

The sum includes all supervised queries, including when the loss couples their outputs; both factors use this same gradient.

**The prefactors are positive and regular near the target.** Regard $\rho_j$ as a function of $f = (K, Q)$ and $\varepsilon$ through $\widetilde{W}(s(t)) = (\beta\varepsilon)^{-1} K Q^\top$. We work on a compact factor neighborhood where every normalized target margin is at least $\delta > 0$, so that (TC) will give $u_{ir} - g_{irk} > 0$ at sufficiently small inverse scale. Dividing by the target exponential gives

$$\alpha_{ir,k_{ir}^\star} = \left(1 + \sum_{k \neq k_{ir}^\star} e^{-m_{irk}(KQ^\top)/\varepsilon}\right)^{-1}.$$

Since every exponent is at most $-\delta/\varepsilon$, the softmax ratios give

$$1 - \alpha_{ir,k_{ir}^\star} = O(e^{-\delta/\varepsilon}), \qquad \mathbf{u}_i - \mathbf{u}_i^\star = O(e^{-\delta/\varepsilon}),$$

with positive limiting prefactors by (TC):

$$\rho_{j(i,r,k)}^\star = \beta[-\partial_r \ell_i(\mathbf{u}_i^\star)](g_{ir,k_{ir}^\star} - g_{irk}) > 0.$$

After decreasing the upper bound on $\varepsilon$, uniformly on this neighborhood,

$$0 < \rho_{\min} \le \rho_j(f, \varepsilon) \le \rho_{\max}, \tag{D.2}$$

$$\left\| D_f^{k_1} \partial_\varepsilon^{k_2} (\rho_j(f, \varepsilon) - \rho_j^\star) \right\| \le C_{k_1,k_2} e^{-\delta/(2\varepsilon)}, \qquad k_1 + k_2 \le 2. \tag{D.3}$$

For the derivative bound, put $m = m_j(KQ^\top)$. We have

$$D_f e^{-m/\varepsilon} = -\varepsilon^{-1} e^{-m/\varepsilon} D_f m, \qquad \partial_\varepsilon e^{-m/\varepsilon} = m\varepsilon^{-2} e^{-m/\varepsilon}.$$

Second derivatives introduce further powers of $\varepsilon^{-1}$ and bounded derivatives of the bilinear margin. For each fixed integer $q \ge 0$,

$$\varepsilon^{-q} e^{-m/\varepsilon} \le \left(\sup_{0 < z \le \bar{\varepsilon}} z^{-q} \exp[-\delta/(2z)]\right) e^{-\delta/(2\varepsilon)}.$$

The supremum is finite, and the rescaled softmax denominator is at least one. Applying the quotient and chain rules in Equation (D.1) therefore proves Equation (D.3). The loss assumption $\ell_i \in C^3$ supplies the required bounded derivatives of $\partial_r \ell_i$ through order two. Together with (TC), the estimates give positivity and a $C^2$ extension $\rho_j(f, 0) = \rho_j^\star$, with all derivatives of positive order at most two tending to zero. Along a local trajectory,

$$\frac{d}{dt} \log \rho_j = \frac{D_f \rho_j[f'] - 2\varepsilon \partial_\varepsilon \rho_j}{\rho_j}.$$

The derivative estimates, bounded factor velocities, and lower bound $\rho_j \geq \rho_{\min}$ make this logarithmic derivative uniformly bounded.

**The corrected energy controls the margin error.** Normalize these prefactors by $\pi_j = \rho_j / \sum_k \rho_k$. On the same neighborhood there are constants $\pi_- > 0$ and $B_\pi < \infty$ with $\pi_j \geq \pi_-$ and $|(\log \pi_j)'| \leq B_\pi$ almost everywhere. Let $C_\pi = \log(1/\pi_-)$ and use the soft minimum from Equation (C.1). The certificate argument in Appendix A.2 gives $\|\Delta^\star\|_{(h)} = \gamma_h$ and $\|W^\star\|_{\mathrm{op}} = 1$. For $j \in \operatorname{supp} p^\star$, the Gibbs formula gives

$$\log \frac{p_j^\star}{p_j} = \log \frac{p_j^\star}{\pi_j} + \frac{m_j(W) - m_\Lambda}{\varepsilon}.$$

Averaging with $p^\star$ yields

$$\varepsilon D_{\mathrm{KL}}(p^\star \| p) = \varepsilon D_{\mathrm{KL}}(p^\star \| \pi) + \langle W, \Delta^\star \rangle - m_\Lambda.$$

The product score cancels when this is substituted into Equation (6.1), giving

$$\mathcal{E} = \gamma_h - m_\Lambda + \varepsilon D_{\mathrm{KL}}(p^\star \| \pi), \qquad |\mathcal{E} - e_m| \leq C_\pi \varepsilon. \tag{D.4}$$

Coordinates outside $\operatorname{supp} p^\star$ contribute zero to each relative entropy. Since $p^\star$ is fixed and $\varepsilon' = -2\varepsilon$,

$$\frac{d}{dt}\left[\varepsilon D_{\mathrm{KL}}(p^\star \| \pi)\right] = -2\varepsilon D_{\mathrm{KL}}(p^\star \| \pi) - \varepsilon \sum_j p_j^\star (\log \pi_j)'.$$

Combining this derivative with Equation (C.4) gives

$$\mathcal{E}' = -\mathcal{D}_{\Delta(p)}(f) + \varepsilon \left[ \sum_j (p_j - p_j^\star)(\log \pi_j)' + 2D_{\mathrm{KL}}(p \| \pi) - 2D_{\mathrm{KL}}(p^\star \| \pi) \right]. \tag{D.5}$$

The bracket is bounded in absolute value by $2B_\pi + 2C_\pi$: use $\|p - p^\star\|_1 \leq 2$ for the first term and the bounds $0 \leq D_{\mathrm{KL}}(\cdot \| \pi) \leq C_\pi$ for the other two. Since $\varepsilon' = -2\varepsilon$, adding $(C_\pi + B_\pi)\varepsilon$ to $\mathcal{E}$ absorbs this error. The resulting function $\mathcal{V} = \mathcal{E} + (C_\pi + B_\pi)\varepsilon$ satisfies

$$e_m \leq \mathcal{V} \leq e_m + (2C_\pi + B_\pi)\varepsilon, \qquad \mathcal{V}' \leq -\mathcal{D}_{\Delta(p)}(f) \leq 0. \tag{D.6}$$

These estimates use Equation (C.3) and hold throughout the region with the stated local prefactor bounds.

Write $r_\star(f) = \gamma_h - \langle KQ^\top, \Delta^\star \rangle$ and $\mathcal{O}^\star = \mathcal{O}_h(\Delta^\star)$. Nonnegativity of relative entropy and Lemma 5.1(iii) give

$$0 \leq r_\star(f) \leq \mathcal{E} \leq \mathcal{V}, \qquad \operatorname{dist}(f, \mathcal{O}^\star)^2 \leq \frac{4 r_\star(f)}{\operatorname{gap}_h(\Delta^\star)}. \tag{D.7}$$

Small $\mathcal{V}$ therefore keeps the factors near $\mathcal{O}^\star$, while the weight error enters $\mathcal{E}$ through $\varepsilon D_{\mathrm{KL}}(p^\star \| p)$.

### D.2 The stationary-margin gap and product convergence

Recall $m_{\min}(W) = \min_j m_j(W)$. A stationary configuration is a point $(K, Q, p) \in \mathcal{B} \times \mathcal{P}_M$ with

$$\operatorname{supp} p \subseteq \arg\min_j \langle KQ^\top, \Delta_j \rangle, \qquad \mathcal{D}_{\Delta(p)}(K, Q) = 0. \tag{D.8}$$

At such a configuration, each factor maximizes its linear objective with the other factor and $p$ held fixed. At rank loss another admissible maximizing selection may still give a nonzero velocity. We apply semialgebraic geometry to this stationary set.

**Lemma D.1** (Finitely many stationary margin values)**.** *For any finite family of real comparison matrices, the set $\mathcal{M}_{\mathrm{stat}}$ of values of $m_{\min}(KQ^\top)$ at the configurations in Equation (D.8) is finite. Neither the factors nor the polar inputs need have full column rank.*

*Proof.* We show that the stationary set has finitely many connected components and that the margin is constant on each. The factor constraints $I_h - K^\top K \succeq 0$ and $I_h - Q^\top Q \succeq 0$ are semialgebraic because they are equivalent to polynomial inequalities on principal minors. We use Lemma I.9 for projections and paths. The graph of the nuclear norm can be described by introducing $H \succeq 0$ with $H^2 = X^\top X$

and taking $\operatorname{tr} H$; projection preserves semialgebraicity. The simplex, bilinear margins, and their finite minimum are semialgebraic as well. The support condition is equivalent to $p_j(\langle KQ^\top, \Delta_j\rangle - m_{\min}(KQ^\top)) = 0$ for every $j$. Hence the stationary set is compact and semialgebraic, with finitely many connected components. Each component is connected by semialgebraic paths. Partition such a path into finitely many intervals on which $K, Q, p$, and $\varphi(u) = m_{\min}(K(u)Q(u)^\top)$ are $C^1$ and each $p_j$ is either positive or identically zero (Coste, 2002).

Fix one interval. Here primes denote derivatives in the path parameter $u$, not training time. Each positive-weight margin equals $\varphi$; differentiating these equalities and then averaging with $p_j(u)$ gives

$$\varphi' = \langle K', \Delta(p)Q\rangle + \langle Q', \Delta(p)^\top K\rangle.$$

Here the weights enter only after differentiation. Zero dissipation makes both duality gaps zero, so $K(u)$ maximizes $X \mapsto \langle X, \Delta(p(u))Q(u)\rangle$ over the factor ball. With this linear objective frozen at $u$, feasibility gives, for both signs of sufficiently small $v$,

$$\langle K(u+v) - K(u), \Delta(p(u))Q(u)\rangle \le 0.$$

The scalar function has a differentiable local maximum at $v = 0$, so $\langle K'(u), \Delta(p(u))Q(u)\rangle = 0$. Applying the same argument to $Q$ gives $\langle Q'(u), \Delta(p(u))^\top K(u)\rangle = 0$, hence $\varphi' = 0$ on each interval. Continuity across the finitely many endpoints makes the margin constant along the path. Each of the finitely many connected components therefore contributes one margin value. □

Under (MM), the target factors and $p^\star$ give stationary value $\gamma_h > 0$, while the zero factors give value zero. All stationary values are at most $\gamma_h$. Therefore

$$\operatorname{gap}_{\text{stat}} = \gamma_h - \max\{v \in \mathcal{M}_{\text{stat}} : v < \gamma_h\} > 0. \tag{D.9}$$

The initialization below ensures $\mathcal{V}(t_0) < \operatorname{gap}_{\text{stat}}$.

**Proof of Theorem 6.2.** Choose a neighborhood with positive margins before fixing the initialization thresholds. Let $L_\Delta = \max_j \|\Delta_j\|_F$. Lemma 5.1 gives

$$m_j(KQ^\top) \ge \gamma_h - L_\Delta\sqrt{8r_\star(f)/\operatorname{gap}_h(\Delta^\star)}.$$

Choose $\bar\delta > 0$ with $L_\Delta\sqrt{8\bar\delta/\operatorname{gap}_h(\Delta^\star)} < \gamma_h/2$. On the compact sublevel $r_\star(f) \le \bar\delta$, every margin then exceeds $c_0 = \gamma_h/2$, and (TC) and Appendix D.1 give positive prefactors and fixed constants $C_\pi, B_\pi$ for $0 < \varepsilon \le \bar\varepsilon$. The theorem's initialization implies

$$r_\star(f(t_0)) \le \gamma_h r_0, \qquad e_m(t_0) \le L_\Delta\sqrt{8\gamma_h r_0/\operatorname{gap}_h(\Delta^\star)}.$$

With this neighborhood and its constants fixed, choose $r_0$ and then $\varepsilon_* \le \bar\varepsilon$ so that $\gamma_h r_0 < \bar\delta$ and

$$L_\Delta\sqrt{8\gamma_h r_0/\operatorname{gap}_h(\Delta^\star)} + (2C_\pi + B_\pi)\varepsilon_* < \min\{\bar\delta, \operatorname{gap}_{\text{stat}}\}.$$

Before a first exit, Equations (D.6) and (D.7) give $r_\star(f(t)) \le \mathcal{V}(t) \le \mathcal{V}(t_0) < \bar\delta$. Exit through the target-residual boundary is impossible, and $\varepsilon' = -2\varepsilon$ prevents exit through the scale boundary. The factors remain in their operator-norm unit balls. The continuation argument of Appendix A.1 applies on every finite time interval, where $\varepsilon > 0$, the prefactors are positive, and $\|\Delta(p)\|_{(h)} \ge \gamma_h$. Thus every admissible continuation exists globally and retains the same positive margin bound.

Since $\mathcal{V} \ge 0$ and $\mathcal{V}' \le -\mathcal{D}_{\Delta(p)}$, integration gives

$$\int_{t_0}^{\infty} \mathcal{D}_{\Delta(p(t))}(f(t))\,dt \le \mathcal{V}(t_0) < \infty,$$

and the nonincreasing function $\mathcal{V}$ has a limit. The sandwich in Equation (D.6), together with $\varepsilon \to 0$, shows that $e_m$ and the minimum margin also have limits. Integrability gives times $t_k \to \infty$ with $\mathcal{D}_{\Delta(p(t_k))}(f(t_k)) \to 0$. Compactness then gives a subsequence such that

$$(K(t_k), Q(t_k), p(t_k)) \longrightarrow (K_\infty, Q_\infty, p_\infty).$$

If comparison $j$ is not minimizing at $W_\infty = K_\infty Q_\infty^\top$, choose a minimizing comparison $i$. Their positive limiting margin difference gives $m_j(W(t_k)) - m_i(W(t_k)) \ge c > 0$ for all large $k$. The Gibbs formula gives $p_j/p_i = (\pi_j/\pi_i)\exp[-(m_j - m_i)/\varepsilon]$. Using $p_i \le 1$, we obtain

$$p_j(t_k) \le \frac{\pi_j(t_k)}{\pi_i(t_k)} e^{-c/\varepsilon(t_k)} \le \pi_-^{-1} e^{-c/\varepsilon(t_k)} \longrightarrow 0.$$

Thus $p_\infty$ is supported on the minimum-margin comparisons. Continuity of the nuclear-norm expression for dissipation makes the limit a stationary configuration. Its margin $v_\infty$ satisfies

$$0 \le \gamma_h - v_\infty = \lim_{t\to\infty} e_m(t) \le \mathcal{V}(t_0) < \mathrm{gap}_{\mathrm{stat}},$$

so Equation (D.9) forces $v_\infty = \gamma_h$. Since $e_m(t)$ has a limit and tends to zero along this subsequence, $e_m(t) \to 0$ on the whole trajectory.

Since $\langle W, \Delta^\star\rangle$ is a $p^\star$-weighted average of the margins,

$$0 \le r_\star(f(t)) \le e_m(t) \longrightarrow 0.$$

Lemma 5.1(iii) gives the product and orbit limits in part (b). Since $\|W^\star\|_{\mathrm{op}} = 1$, also $\widetilde{W}(s(t))/\|\widetilde{W}(s(t))\|_{\mathrm{op}} = W/\|W\|_{\mathrm{op}} \to W^\star$. For every query, the attention ratios and the positive margin floor give

$$1 - \alpha_{ir,k^\star_{ir}} = \frac{\sum_{k\neq k^\star_{ir}} e^{-m_{irk}(W)/\varepsilon}}{1 + \sum_{k\neq k^\star_{ir}} e^{-m_{irk}(W)/\varepsilon}} \le (|\mathcal{M}_{ir}| - 1)e^{-c_0/\varepsilon}.$$

This proves part (c); the positive margin bound in part (a) was established by the trapping argument.□

# E Proof of Theorem 6.3: Local joint convergence

Assume (MM), (TC), (SC), and (ND), with $|\mathcal{S}| \ge 2$. Small energy controls the factors but only the scaled weight error $\varepsilon D_{\mathrm{KL}}(p^\star\|p)$, so we first use dissipation to bring the weights near $p^\star$. After removing the neutral common rotations, a coupled Lyapunov estimate gives the rate. The entry argument uses the prefactor and energy estimates of Appendix D.1.

The geometric calculations use the rectangular notation from Appendix A, with $m = n = d$ for attention. Active-weight logarithms and ratios are taken on $\mathcal{S}$. All inverses of active Hessians act on $T_{\mathcal{S}}$.

## E.1 Dual nondegeneracy and finite-time entry

Fix $f^\star = (K^\star, Q^\star) = (U_h, V_h) \in \mathcal{O}^\star$, and for $v \in T_{\mathcal{S}}$ write $\Delta(v) = \sum_{j\in\mathcal{S}} v_j\Delta_j$. Let $\Psi(p) = \|\Delta(p)\|_{(h)}$ and define

$$H_\Psi = D^2\Psi(p^\star)|_{T_{\mathcal{S}}}, \qquad H_{\mathrm{diss}} = D_p^2\mathcal{D}_{\Delta(p)}(f^\star)|_{p=p^\star,T_{\mathcal{S}}}.$$

The cutoff gap in (MM) gives local smoothness of both objectives by Lemmas I.6 and I.7. In singular-vector bases of $\Delta^\star = U\Sigma V^\top$, define

$$\mathcal{K}_{\mathrm{pol}} = \left\{ E \in \mathbb{R}^{m\times n} : U^\top EV = \begin{pmatrix} S & 0 \\ 0 & E_\perp \end{pmatrix},\ S = S^\top \right\}. \tag{E.1}$$

**Lemma E.1** (Dual nondegeneracy and dissipation)**.** *Under* (MM)*,*

$$0 \preceq H_{\mathrm{diss}} \preceq 2H_\Psi, \qquad \ker H_{\mathrm{diss}} = \ker H_\Psi = \{v \in T_{\mathcal{S}} : \Delta(v) \in \mathcal{K}_{\mathrm{pol}}\}. \tag{E.2}$$

*If* (ND) *also holds, then* $H_{\mathrm{diss}} \succ 0$ *on* $T_{\mathcal{S}}$ *and, for some* $c > 0$*,*

$$\mathcal{D}_{\Delta(q)}(f^\star) \ge c\|q - p^\star\|_2^2 \qquad (q \in \mathcal{P}_{\mathcal{S}}). \tag{E.3}$$

*This bound includes the boundary of the active simplex.*

*Proof.* For $X = U_X R_X$ with $R_X \succ 0$, diagonalize $R_X$ with eigenvalues $s_a$ and put $E_\| = U_X^\top E$, $E_\perp = (I - U_X U_X^\top)E$. By Lemma I.6, the differential in Equation (B.5) is self-adjoint and positive semidefinite, with

$$\langle E, D\,\mathrm{polar}(X)[E]\rangle = \sum_{a<b} \frac{((E_\|)_{ab} - (E_\|)_{ba})^2}{s_a + s_b} + \mathrm{tr}(E_\perp^\top E_\perp R_X^{-1}). \tag{E.4}$$

It vanishes exactly when $E = U_X S$ for symmetric $S$, equivalently when the polar differential vanishes. Since the nuclear-norm gradient at a full-column-rank matrix is its polar factor,

$$\begin{aligned}\langle v, H_{\mathrm{diss}} v\rangle = \langle \Delta(v)Q^\star, D\,\mathrm{polar}(\Delta^\star Q^\star)[\Delta(v)Q^\star]\rangle \\ + \langle \Delta(v)^\top K^\star, D\,\mathrm{polar}((\Delta^\star)^\top K^\star)[\Delta(v)^\top K^\star]\rangle.\end{aligned}$$

Write $E = \Delta(v)$ in these bases as

$$U^\top E V = \begin{pmatrix} E_{11} & E_{12} \\ E_{21} & E_{22}\end{pmatrix}.$$

The first term vanishes exactly when $E_{21} = 0$ and $E_{11}$ is symmetric; the second requires $E_{12} = 0$ and symmetric $E_{11}^\top$. Since both terms are nonnegative, their sum vanishes exactly when $E \in \mathcal{K}_{\mathrm{pol}}$. The block $E_{22}$ is unrestricted.

To compare with $H_\Psi$, apply the Ky Fan bound to $2\|Z\|_{(h)} - \|ZQ^\star\|_* - \|Z^\top K^\star\|_*$. This expression is nonnegative and zero at $Z = \Delta^\star$. Its second derivative along $Z = \Delta^\star + a\Delta(v)$ is therefore nonnegative at $a = 0$, giving $0 \le 2v^\top H_\Psi v - v^\top H_{\mathrm{diss}} v$ and $\ker H_\Psi \subseteq \ker H_{\mathrm{diss}}$. For the reverse inclusion, suppose $H_{\mathrm{diss}} v = 0$. Then the block form in Equation (E.1) keeps $\Delta^\star + a\Delta(v)$ block diagonal in these singular-vector bases. For sufficiently small $|a|$, $\Sigma_h + aS$ remains symmetric positive definite, and the strict cutoff gap ensures that its $h$ eigenvalues are still the $h$ leading singular values of the full matrix. Therefore

$$\|\Delta^\star + a\Delta(v)\|_{(h)} = \mathrm{tr}(\Sigma_h + aS).$$

The right side is affine, so $v^\top H_\Psi v = 0$ and hence $H_\Psi v = 0$. This proves equality of the kernels and gives $H_{\mathrm{diss}} \succ 0$ under (ND). The required injectivity is that of the two polar differentials together; either component may have a kernel.

To extend the bound to the closed simplex, use the convex nonnegative function $d(q) = \mathcal{D}_{\Delta(q)}(f^\star)$, a sum of norms of linear maps minus a linear functional. Its tangential gradient vanishes at the relative-interior minimizer $p^\star$, so the positive-definite Hessian gives $d(q) \ge c_1\|q - p^\star\|_2^2$ locally by Taylor's formula.

A second zero would, by convexity, give a segment of zeros through $p^\star$, contradicting this local bound. On the compact complement of a smaller relative neighborhood, let $d_{\min} > 0$ be the minimum of $d$, and put $D = \max_{q\in\mathcal{P}_\mathcal{S}} \|q - p^\star\|_2$. Taking $c = \min\{c_1, d_{\min}/D^2\}$ proves the bound on the whole simplex; if the complement is empty, take $c = c_1$. Continuity on the closed simplex justifies the minimum $d_{\min}$. □

We use strict complementarity to control inactive weights. If $\mathcal{S} \neq [M]$, set $\mathrm{gap}_{\mathrm{in}} = \min_{k\notin\mathcal{S}}(m_k(W^\star) - \gamma_h) > 0$. When $\mathcal{S} = [M]$, choose any positive value and omit inactive terms. Then

$$\|\Delta(p)\|_{(h)} \ge \langle W^\star, \Delta(p)\rangle \ge \gamma_h + \mathrm{gap}_{\mathrm{in}} \sum_{k\notin\mathcal{S}} p_k.$$

Every dual minimizer is supported on $\mathcal{S}$. Convexity and $H_\Psi \succ 0$ rule out a second minimizer, since the objective would be constant on the segment to $p^\star$. Together with Appendix A.2, this identifies both the product and the weights.

**Finite-time entry.** The conditional active distribution is

$$p_j^\mathcal{S} = \frac{p_j}{\sum_{k\in\mathcal{S}} p_k} \quad (j \in \mathcal{S}),$$

extended by zero elsewhere. For a fixed sufficiently small $r_1 > 0$, let

$$\mathcal{K}_{r_1} = \{q \in \mathrm{relint}\,\mathcal{P}_\mathcal{S} : \|\Pi_{T_\mathcal{S}} \log(q/p^\star)\| \le r_1\}.$$

For $z \in T_\mathcal{S}$, the inverse centered log-ratio map is

$$q_j(z) = \frac{p_j^\star e^{z_j}}{\sum_{k\in\mathcal{S}} p_k^\star e^{z_k}}.$$

The normalizing term disappears under centering, so $\Pi_{T_{\mathcal{S}}} \log(q(z)/p^\star) = z$. Thus $\mathcal{K}_{r_1}$ is a compact neighborhood of $p^\star$ contained in the relative interior. By Equation (E.3),

$$\kappa_0 = \inf_{q \in \mathcal{P}_{\mathcal{S}} \setminus \operatorname{int}_{\mathcal{P}_{\mathcal{S}}} \mathcal{K}_{r_1}} \mathcal{D}_{\Delta(q)}(f^\star) > 0.$$

The entry estimate holds for each fixed $r_1$; Appendix E.3 chooses this radius before the factor tube and initialization thresholds.

Choose $d_{\mathcal{N}} > 0$ small enough that the compact tube $\mathcal{N} = \{f \in \mathcal{B} : \operatorname{dist}(f, \mathcal{O}^\star) \le d_{\mathcal{N}}\}$ has positive target margins, with every inactive margin exceeding every active one by at least $3\,\mathrm{gap}_{\mathrm{in}}/4$. Appendix D.1 gives uniform prefactor bounds on $\mathcal{N}$ for $0 < \varepsilon \le \bar{\varepsilon}$. Its relative boundary in $\mathcal{B}$ is separated from the orbit. By Lemma 5.1(iii),

$$r_\star(f) \ge \tfrac{1}{4}\,\mathrm{gap}_h(\Delta^\star)\operatorname{dist}(f, \mathcal{O}^\star)^2,$$

so the boundary has a positive residual floor $r_\partial$. Lemma 5.1(iii) places $f(t_0)$ inside the tube when $\mathcal{E}(t_0)$ is sufficiently small. Choose $\mathcal{E}_0, \varepsilon_0$ such that

$$\mathcal{V}(t_0) = \mathcal{E}(t_0) + (C_\pi + B_\pi)\varepsilon(t_0) < r_\partial, \qquad \varepsilon_0 \le \bar{\varepsilon}.$$

The inequality $r_\star(f(t)) \le \mathcal{V}(t) \le \mathcal{V}(t_0)$ excludes a first factor exit, while $\varepsilon' = -2\varepsilon$ keeps the inverse scale below $\bar{\varepsilon}$.

For inactive $k$ and active $j$, the Gibbs ratio gives $p_k/p_j \le C\exp[-3\,\mathrm{gap}_{\mathrm{in}}/(4\varepsilon)]$. Summing yields

$$\sum_{k \notin \mathcal{S}} p_k(t) \le C e^{-\mathrm{gap}_{\mathrm{in}}/(2\varepsilon(t))}, \qquad \|p(t) - p^{\mathcal{S}}(t)\|_1 = 2\sum_{k \notin \mathcal{S}} p_k(t). \tag{E.5}$$

For $q \in \mathcal{P}_{\mathcal{S}}$, the dissipation satisfies

$$\big|\mathcal{D}_{\Delta(p)}(f) - \mathcal{D}_{\Delta(q)}(f^\star)\big| \le C\big(\operatorname{dist}(f, \mathcal{O}^\star) + \|p - q\|_1\big).$$

A common rotation sends a nearest orbit point to $f^\star$ without changing the dissipation. The nuclear-norm inequality $|\|X\|_* - \|Y\|_*| \le \|X - Y\|_*$, bounded factors, and the bilinear score then give the bound. With $q = p^{\mathcal{S}}$, shrink the tube and scale threshold until this difference is less than $\kappa_0/2$. Then $\mathcal{D}_{\Delta(p)}(f) \ge \kappa_0/2$ before the first entry time $t_1$ into $\mathcal{K}_{r_1}$. Integrating up to any time before entry gives

$$0 \le \mathcal{V}(t) \le \mathcal{V}(t_0) - \tfrac{1}{2}\kappa_0(t - t_0).$$

Nonnegativity of $\mathcal{V}$ therefore forces entry by

$$t_1 - t_0 \le \frac{2\mathcal{V}(t_0)}{\kappa_0} \le C\big(\mathcal{E}(t_0) + \varepsilon(t_0)\big). \tag{E.6}$$

Set $t_1 = t_0$ if $p^{\mathcal{S}}(t_0) \in \mathcal{K}_{r_1}$. Bounded velocities and the compact-valued inclusion give continuation through every finite time before entry by Appendix A.1.

### E.2 Local factor and log-weight coordinates

At $(\mathcal{O}^\star, p^\star)$, the matrices $\Delta(p)Q$ and $\Delta(p)^\top K$ have smallest singular value $\sigma_h(\Delta^\star) > 0$, so they remain uniformly full rank in a small neighborhood. We work there until the first-exit argument in Appendix E.3. Positive scalar normalization does not change polar factors, so the factor field is

$$b(f, p) = \big(\operatorname{polar}(\Delta(p)Q) - K,\ \operatorname{polar}(\Delta(p)^\top K) - Q\big).$$

For $p = p^\star$ it vanishes on $\mathcal{O}^\star$. At finite inverse scale, the weights at the target factors need not equal $p^\star$, so those factors need not be equilibria of the full system.

**Common-rotation slice.** Let $\mathcal{T}_{\mathrm{orb}} = \{(K^\star\Omega, Q^\star\Omega) : \Omega^\top = -\Omega\}$ and $\mathcal{X}_{\mathrm{sl}} = \mathcal{T}_{\mathrm{orb}}^{\perp}$. For a nearby pair, $M = (K^\star)^\top K + (Q^\star)^\top Q$ is invertible. Set $O = \operatorname{polar}(M)$ and $f_{\mathrm{sl}} = (KO^\top, QO^\top)$. Then $M_{\mathrm{sl}} = (K^\star)^\top K_{\mathrm{sl}} + (Q^\star)^\top Q_{\mathrm{sl}}$ is symmetric positive definite and $f_{\mathrm{sl}} - f^\star \in \mathcal{X}_{\mathrm{sl}}$. The common Procrustes identity in Lemma I.5 gives $\|f_{\mathrm{sl}} - f^\star\| = \operatorname{dist}(f, \mathcal{O}^\star)$, and using the same rotation for both factors preserves $KQ^\top$. Its induced field on the slice is

$$b_{\mathrm{sl}}(f_{\mathrm{sl}}, p) = b(f_{\mathrm{sl}}, p) - (K_{\mathrm{sl}}\Omega, Q_{\mathrm{sl}}\Omega) \in \mathcal{X}_{\mathrm{sl}},$$

where the unique skew-symmetric $\Omega$ solves

$$M_{\mathrm{sl}}\Omega + \Omega M_{\mathrm{sl}} = 2\,\mathrm{skew}((K^\star)^\top b_K + (Q^\star)^\top b_Q).$$

Differentiating $\mathrm{skew}\, M_{\mathrm{sl}} = 0$ gives this Sylvester equation; positive definiteness of $M_{\mathrm{sl}}$ makes $\Omega$ unique and smooth by Lemma I.6. The removed term contributes zero product velocity:

$$(K_{\mathrm{sl}}\Omega)Q_{\mathrm{sl}}^\top + K_{\mathrm{sl}}(Q_{\mathrm{sl}}\Omega)^\top = 0.$$

It therefore leaves margins and gradient weights unchanged. The target is on the boundary of the factor balls, but the full-rank field and coordinate map extend to an open ambient neighborhood. We apply the inverse-function theorem there (Lemma I.10); the trajectory itself remains in the factor balls.

Define the centered margin map $m_{\mathcal{S}}(f) = \Pi_{T_{\mathcal{S}}}(\langle KQ^\top, \Delta_j\rangle)_{j\in\mathcal{S}}$. At $(f^\star, p^\star)$, let $A = D_f b_{\mathrm{sl}}$, $B = D_p b_{\mathrm{sl}}|_{T_{\mathcal{S}}}$, and $D_m = D_f m_{\mathcal{S}}|_{\mathcal{X}_{\mathrm{sl}}}$. To compute these derivatives and check stability, use temporarily the ambient maps

$$\mathcal{U}v = (\Delta(v)Q^\star, \Delta(v)^\top K^\star), \qquad \mathcal{T}\zeta = (\Delta^\star\zeta_Q, (\Delta^\star)^\top\zeta_K),$$

and the block polar differential $J_{\mathrm{pol}}$ at $\Delta^\star Q^\star$ and $(\Delta^\star)^\top K^\star$. Both $J_{\mathrm{pol}}$ and $\mathcal{T}$ are self-adjoint, and $J_{\mathrm{pol}} \succeq 0$. With $\Pi_{\mathrm{sl}}$ the orthogonal projection onto $\mathcal{X}_{\mathrm{sl}}$,

$$A = \Pi_{\mathrm{sl}}(J_{\mathrm{pol}}\mathcal{T} - I)|_{\mathcal{X}_{\mathrm{sl}}}, \qquad B = \Pi_{\mathrm{sl}}J_{\mathrm{pol}}\mathcal{U}, \qquad D_m = \mathcal{U}^*|_{\mathcal{X}_{\mathrm{sl}}}. \tag{E.7}$$

Differentiating $b(f^\star O, p^\star) = 0$ in an orbit-tangent direction gives $(J_{\mathrm{pol}}\mathcal{T} - I)\mathcal{T}_{\mathrm{orb}} = 0$. Margin invariance gives $\mathcal{U}^*\mathcal{T}_{\mathrm{orb}} = 0$ and $\mathcal{U}^*\Pi_{\mathrm{sl}} = \mathcal{U}^*$. Therefore

$$D_m B = \mathcal{U}^* J_{\mathrm{pol}}\mathcal{U} = H_{\mathrm{diss}}, \qquad \Pi_0 = I - BH_{\mathrm{diss}}^{-1}D_m, \qquad A_0 = \Pi_0 A|_{\ker D_m}. \tag{E.8}$$

Since $H_{\mathrm{diss}} \succ 0$, the identity $D_m B = H_{\mathrm{diss}}$ implies that $D_m$ is onto and $BH_{\mathrm{diss}}^{-1}$ is a right inverse. Direct multiplication gives

$$D_m\Pi_0 = 0, \qquad \Pi_0 B = 0, \qquad \Pi_0\zeta = \zeta \quad (\zeta \in \ker D_m).$$

Thus every slice perturbation splits uniquely as

$$\zeta = \underbrace{\Pi_0\zeta}_{\text{zero linearized active-margin difference}} + \underbrace{BH_{\mathrm{diss}}^{-1}D_m\zeta}_{\text{active-margin change}}.$$

The operator $A_0 = \Pi_0 A|_{\ker D_m}$ governs the remaining factor motion. Since $\Pi_0$ is oblique, stability of $A$ alone does not establish stability of $A_0$.

**The constrained operator is Hurwitz.** Restrict to $\mathcal{H}_{\mathrm{pol}} = (\ker J_{\mathrm{pol}})^\perp = \mathrm{range}\, J_{\mathrm{pol}}$. On this space, $J_{\mathrm{pol}}^{1/2}$ is invertible and

$$\mathcal{A} = J_{\mathrm{pol}}^{1/2}\mathcal{T}J_{\mathrm{pol}}^{1/2} - I, \qquad (J_{\mathrm{pol}}\mathcal{T} - I)J_{\mathrm{pol}}^{1/2} = J_{\mathrm{pol}}^{1/2}\mathcal{A}.$$

Because $J_{\mathrm{pol}}\mathcal{T}$ takes values in $\mathcal{H}_{\mathrm{pol}}$, this space is invariant under $J_{\mathrm{pol}}\mathcal{T} - I$, and the identity gives a similarity on it. Every orbit tangent lies in this space: $(J_{\mathrm{pol}}\mathcal{T} - I)g = 0$ implies $g = J_{\mathrm{pol}}\mathcal{T}g$. The spectrum computed in Appendix B applies here, since positive rescaling of $\Delta^\star$ leaves the factor field unchanged. Thus the self-adjoint $\mathcal{A}$ is negative semidefinite, with kernel $\mathcal{T}_{\mathrm{pol}} = \{u \in \mathcal{H}_{\mathrm{pol}} : J_{\mathrm{pol}}^{1/2}u \in \mathcal{T}_{\mathrm{orb}}\}$, and is negative definite on its orthogonal complement in $\mathcal{H}_{\mathrm{pol}}$.

The margin constraint is compatible with the quotient by orbit tangents because $\mathcal{U}^*\mathcal{T}_{\mathrm{orb}} = 0$. On $\ker\mathcal{U}^*/\mathcal{T}_{\mathrm{orb}}$, $A_0$ represents

$$[\zeta] \longmapsto [(I - J_{\mathrm{pol}}\mathcal{U}H_{\mathrm{diss}}^{-1}\mathcal{U}^*)(J_{\mathrm{pol}}\mathcal{T} - I)\zeta].$$

For a nonzero eigenclass with eigenvalue $\lambda$, set $v = H_{\mathrm{diss}}^{-1}\mathcal{U}^*(J_{\mathrm{pol}}\mathcal{T} - I)\zeta$. The eigenvalue equation reads

$$J_{\mathrm{pol}}(\mathcal{T}\zeta - \mathcal{U}v) = (\lambda + 1)\zeta + g, \qquad g \in \mathcal{T}_{\mathrm{orb}}, \qquad \mathcal{U}^*\zeta = 0.$$

For $\lambda = -1$ the eigenclass is already stable. Otherwise, the left side and $g$ lie in $\mathcal{H}_{\mathrm{pol}}$, so $\zeta \in \mathcal{H}_{\mathrm{pol}}$. Write $\zeta = J_{\mathrm{pol}}^{1/2}u$ and remove the component of $u$ in $\mathcal{T}_{\mathrm{pol}}$. This changes $\zeta$ only by an orbit tangent

and preserves both its quotient class and the constraint. Applying the inverse square root on $\mathcal{H}_{\mathrm{pol}}$ gives

$$\mathcal{A}u - J_{\mathrm{pol}}^{1/2}\mathcal{U}v = \lambda u + n, \qquad n \in \mathcal{T}_{\mathrm{pol}}, \qquad \mathcal{U}^* J_{\mathrm{pol}}^{1/2} u = 0.$$

Now $u \neq 0$ and $u \perp \mathcal{T}_{\mathrm{pol}}$. The constraint gives $\langle u, J_{\mathrm{pol}}^{1/2}\mathcal{U}v\rangle = \langle \mathcal{U}^* J_{\mathrm{pol}}^{1/2} u, v\rangle = 0$, while $\langle u, n\rangle = 0$ by orthogonality. The Hermitian inner product with $u$ therefore gives, also for a complex eigenvector,

$$\lambda \|u\|^2 = \langle u, \mathcal{A}u\rangle < 0. \tag{E.9}$$

Every eigenvalue of $A_0$ is real and negative, so $A_0$ is Hurwitz.

**Log-weight coordinates.** The Gibbs formula relates weight ratios to margin differences divided by $\varepsilon$. Let $\ell_\rho(f,\varepsilon) = \Pi_{T_\mathcal{S}}(\log \rho_\mathcal{S}(f,\varepsilon) - \log p^\star)$ and set

$$\theta = \varepsilon^{-1} m_\mathcal{S}(f_{\mathrm{sl}}) - \ell_\rho(f_{\mathrm{sl}},\varepsilon), \qquad p_j^\mathcal{S} = \frac{p_j^\star e^{-\theta_j}}{\sum_{k\in\mathcal{S}} p_k^\star e^{-\theta_k}} \quad (j \in \mathcal{S}). \tag{E.10}$$

The second identity follows by taking centered logarithms: $\Pi_{T_\mathcal{S}} \log(p^\mathcal{S}/p^\star) = -\theta$. Write $p^\mathcal{S}(\theta)$ for this distribution. Its derivative at zero is $-J_{\mathrm{sm}}^\star$, where $J_{\mathrm{sm}}^\star = \mathrm{diag}(p^\star) - p^\star (p^\star)^\top$. For $v \in T_\mathcal{S}$,

$$v^\top J_{\mathrm{sm}}^\star v = \sum_{j\in\mathcal{S}} p_j^\star v_j^2 - \left(\sum_{j\in\mathcal{S}} p_j^\star v_j\right)^2.$$

Since $p_j^\star > 0$, this variance vanishes only on constant vectors, whose intersection with $T_\mathcal{S}$ is zero. Thus $J_{\mathrm{sm}}^\star \succ 0$ on $T_\mathcal{S}$ and

$$(J_{\mathrm{sm}}^\star)^{1/2}(H_{\mathrm{diss}} J_{\mathrm{sm}}^\star)(J_{\mathrm{sm}}^\star)^{-1/2} = (J_{\mathrm{sm}}^\star)^{1/2} H_{\mathrm{diss}} (J_{\mathrm{sm}}^\star)^{1/2} \succ 0.$$

This similarity gives positive eigenvalues for $H_{\mathrm{diss}} J_{\mathrm{sm}}^\star$.

The slice map $f_{\mathrm{sl}} \mapsto (\xi, \delta m) = (\Pi_0(f_{\mathrm{sl}} - f^\star), m_\mathcal{S}(f_{\mathrm{sl}}))$ has derivative $\zeta \mapsto (\Pi_0\zeta, D_m\zeta)$, with inverse $(\xi,\delta m) \mapsto \xi + B H_{\mathrm{diss}}^{-1}\delta m$. The inverse follows from $D_m B H_{\mathrm{diss}}^{-1} = I$ and $\Pi_0 B = 0$. Lemma I.10 gives a $C^2$ chart $f_{\mathrm{sl}} = \chi(\xi,\delta m)$ with $\xi \in \ker D_m \subseteq \mathcal{X}_{\mathrm{sl}}$. In this chart the log-weight identity reads

$$\delta m = \varepsilon[\theta + \ell_\rho(\chi(\xi,\delta m),\varepsilon)].$$

For fixed $(\xi,\theta,\varepsilon)$, the derivative of this equation with respect to $\delta m$ is

$$I - \varepsilon D_f \ell_\rho \, D_{\delta m}\chi = I + O(\varepsilon).$$

All factors after $\varepsilon$ are uniformly bounded on a fixed neighborhood, so this operator has a uniformly bounded inverse once $\varepsilon$ is small. The $C^2$ prefactor extension in Appendix D.1 and the implicit-function theorem give a solution $\delta m = \delta m(\xi,\theta,\varepsilon)$ on a fixed neighborhood, with

$$\|\delta m\| \le C\varepsilon(1 + \|\theta\|), \qquad \|D_\xi \delta m\| + \|D_\theta \delta m\| \le C\varepsilon. \tag{E.11}$$

For example, with $A_m = I - \varepsilon D_f \ell_\rho D_{\delta m}\chi$, differentiation gives

$$D_\theta \delta m = \varepsilon A_m^{-1}, \qquad D_\xi \delta m = \varepsilon A_m^{-1} D_f \ell_\rho D_\xi \chi.$$

These formulas give the derivative bounds in Equation (E.11); derivatives through order two remain uniformly bounded. At $\varepsilon = 0$, apply the implicit-function theorem after extending the flat prefactors by their constant limits to $\varepsilon < 0$. Only the coordinate construction uses this extension.

The chart gives $f_{\mathrm{sl}} - f^\star = \xi + B H_{\mathrm{diss}}^{-1}\delta m + O((\|\xi\| + \|\delta m\|)^2)$, hence $\|f_{\mathrm{sl}} - f^\star\| \le C(\|\xi\| + \varepsilon(1 + \|\theta\|))$. Extend $p^\mathcal{S}(\theta)$ by zero outside $\mathcal{S}$. The margin separation in Equation (E.5) gives $p - p^\mathcal{S}(\theta) = O_{C^2}(e^{-\mathrm{gap}_{\mathrm{in}}/(2\varepsilon)})$: derivatives introduce only bounded chart coefficients and powers of $\varepsilon^{-1}$, absorbed by the slack between $3\,\mathrm{gap}_{\mathrm{in}}/4$ and $\mathrm{gap}_{\mathrm{in}}/2$. This difference extends by zero to $\varepsilon \le 0$. Expanding the field,

$$b_{\mathrm{sl}}(f_{\mathrm{sl}}, p) = A(f_{\mathrm{sl}} - f^\star) + B(p^\mathcal{S}(\theta) - p^\star) + O\big(\|f_{\mathrm{sl}} - f^\star\|^2 + \|\theta\|^2 + e^{-\mathrm{gap}_{\mathrm{in}}/(2\varepsilon)}\big).$$

Since $\xi = \Pi_0(f_{\mathrm{sl}} - f^\star)$, the identity $\Pi_0 B = 0$ removes the linear weight term from $\xi'$. To differentiate $\theta = \varepsilon^{-1} m_\mathcal{S} - \ell_\rho$, use $(\varepsilon^{-1})' = 2\varepsilon^{-1}$ and $\varepsilon' = -2\varepsilon$. Multiplication by $\varepsilon$ gives

$$\varepsilon\theta' = D m_\mathcal{S}(f_{\mathrm{sl}}) b_{\mathrm{sl}} + 2 m_\mathcal{S}(f_{\mathrm{sl}}) - \varepsilon D_f \ell_\rho \, b_{\mathrm{sl}} + 2\varepsilon^2 \partial_\varepsilon \ell_\rho.$$

To identify the linear terms, write $\zeta = f_{\mathrm{sl}} - f^\star$. The chart and active softmax give

$$\zeta = \xi + O\big(\|\xi\|^2 + \varepsilon(1+\|\theta\|)\big), \qquad p^{\mathcal{S}}(\theta) - p^\star = -J^\star_{\mathrm{sm}}\theta + O(\|\theta\|^2).$$

Using $Dm_{\mathcal{S}}(f_{\mathrm{sl}}) = D_m + O(\|\zeta\|)$ and $D_m B = H_{\mathrm{diss}}$, the first term in the fast equation becomes

$$Dm_{\mathcal{S}}(f_{\mathrm{sl}})b_{\mathrm{sl}} = D_m A\xi - H_{\mathrm{diss}}J^\star_{\mathrm{sm}}\theta + O\big((\|\xi\| + \|\theta\|)^2 + \varepsilon\big).$$

The remaining terms carry explicit powers of the inverse scale:

$$\begin{aligned} 2m_{\mathcal{S}}(f_{\mathrm{sl}}) &= 2\varepsilon[\theta + \ell_\rho(f_{\mathrm{sl}}, \varepsilon)] = O\big(\varepsilon(1+\|\theta\|)\big), \\ -\varepsilon D_f \ell_\rho\, b_{\mathrm{sl}} &= O\big(\varepsilon(\|\xi\| + \|\theta\| + \varepsilon)\big), \\ 2\varepsilon^2 \partial_\varepsilon \ell_\rho &= O(\varepsilon^2). \end{aligned}$$

Here bounded derivatives of $\ell_\rho$ give the last two estimates, and the flat inactive mass is absorbed into the $\varepsilon$ terms. Applying $\Pi_0$ to the factor equation uses $\Pi_0 B = 0$ and $\Pi_0 A\xi = A_0\xi$. Together these estimates give

$$\begin{aligned} \xi' &= A_0\xi + \mathcal{R}_\xi(\xi,\theta,\varepsilon), \\ \varepsilon\theta' &= -H_{\mathrm{diss}}J^\star_{\mathrm{sm}}\theta + D_m A\xi + \mathcal{R}_\theta(\xi,\theta,\varepsilon), \\ \|\mathcal{R}_\xi\| &\le C\big((\|\xi\| + \|\theta\|)^2 + \varepsilon\big), \\ \|\mathcal{R}_\theta\| &\le C\big((\|\xi\| + \|\theta\|)^2 + \varepsilon(1 + \|\xi\| + \|\theta\|)\big). \end{aligned} \tag{E.12}$$

The smooth factor field and $C^2$ chart give quadratic Taylor remainders with first derivatives bounded by the distance to the origin. The terms containing derivatives of $\ell_\rho$ are $C^1$ and carry factors $\varepsilon$ or $\varepsilon^2$. The $C^3$ loss assumption supplies the $C^2$ bound on $\ell_\rho$, which controls the first derivatives of these terms. Differentiating the remainders with respect to $\theta$ therefore gives

$$\|D_\theta \mathcal{R}_\theta\| \le C(\|\xi\| + \|\theta\| + \varepsilon).$$

The singular scale remains in the coefficient of $\theta'$.

At $(\xi,\theta,\varepsilon) = (0,0,0)$, the derivative of the fast right-hand side with respect to $\theta$ is $-H_{\mathrm{diss}}J^\star_{\mathrm{sm}}$. Its invertibility and the $C^1$ implicit-function theorem give a unique local zero for $(\xi,\varepsilon)$ near zero,

$$\theta_{\mathrm{qs}}(\xi,\varepsilon) = (H_{\mathrm{diss}}J^\star_{\mathrm{sm}})^{-1} D_m A\xi + O(\|\xi\|^2 + \varepsilon), \tag{E.13}$$

with bounded first derivatives. The expansion follows from Equation (E.12) by absorbing the small multiples of $\|\theta_{\mathrm{qs}}\|$. Centering at this zero removes the forcing $D_m A\xi/\varepsilon$, which need not vanish. Since $(\xi,\varepsilon)$ evolve, the graph of $\theta_{\mathrm{qs}}$ is not invariant; its drift must still be included.

### E.3 Coupled Lyapunov estimate

Set $e = \theta - \theta_{\mathrm{qs}}(\xi,\varepsilon)$. Substitution into the slow equation gives

$$\xi' = A_0\xi + \mathcal{R}_\xi(\xi, \theta_{\mathrm{qs}} + e, \varepsilon), \qquad \|\mathcal{R}_\xi(\xi,\theta_{\mathrm{qs}}+e,\varepsilon)\| \le C(\|\xi\|^2 + \|e\|^2 + \varepsilon). \tag{E.14}$$

Subtracting the derivative of the quasi-steady state from the fast equation yields

$$\begin{aligned} \varepsilon e' = &- H_{\mathrm{diss}}J^\star_{\mathrm{sm}} e + \mathcal{R}_\theta(\xi, \theta_{\mathrm{qs}} + e, \varepsilon) - \mathcal{R}_\theta(\xi, \theta_{\mathrm{qs}}, \varepsilon) \\ &- \varepsilon D_\xi \theta_{\mathrm{qs}}\, \xi' - \varepsilon \partial_\varepsilon \theta_{\mathrm{qs}}\, \varepsilon'. \end{aligned} \tag{E.15}$$

Integrating $D_\theta \mathcal{R}_\theta$ along the segment from $\theta_{\mathrm{qs}}$ to $\theta_{\mathrm{qs}} + e$ bounds the remainder difference by $C(\|\xi\| + \|e\| + \varepsilon)\|e\|$. The last two terms in Equation (E.15) come from differentiating $\theta_{\mathrm{qs}}(\xi,\varepsilon)$ along the trajectory.

Since $A_0$ is Hurwitz and $H_{\mathrm{diss}}J^\star_{\mathrm{sm}}$ has positive spectrum, Lemma I.11 gives positive-definite matrices $X_0, Y_{\mathrm{f}}$ satisfying

$$A_0^\top X_0 + X_0 A_0 = -I, \qquad (H_{\mathrm{diss}}J^\star_{\mathrm{sm}})^\top Y_{\mathrm{f}} + Y_{\mathrm{f}} H_{\mathrm{diss}}J^\star_{\mathrm{sm}} = I.$$

For example,

$$X_0 = \int_0^\infty e^{A_0^\top u} e^{A_0 u}\, du, \qquad Y_{\mathrm{f}} = \int_0^\infty e^{-(H_{\mathrm{diss}}J^\star_{\mathrm{sm}})^\top u} e^{-H_{\mathrm{diss}}J^\star_{\mathrm{sm}} u}\, du.$$

These convergent integrals give positive-definite matrices depending only on the fixed problem. Define

$$\mathscr{L} = \xi^\top X_0 \xi + e^\top Y_{\mathrm{f}} e. \tag{E.16}$$

On a small coordinate ball $\|\xi\| + \|e\| \le r$, differentiate the slow quadratic form:

$$\frac{d}{dt}(\xi^\top X_0 \xi) = -\|\xi\|^2 + 2\langle X_0\xi, \mathcal{R}_\xi\rangle \le -\|\xi\|^2 + C\|\xi\|(\|\xi\|^2 + \|e\|^2 + \varepsilon).$$

Shrinking $r$ absorbs the cubic term. The mixed quadratic term is at most $Cr\|e\|^2$, and Young's inequality bounds $C\varepsilon\|\xi\|$ by $\|\xi\|^2/4 + C\varepsilon^2$. Thus

$$\frac{d}{dt}(\xi^\top X_0 \xi) \le -\tfrac{1}{2}\|\xi\|^2 + Cr\|e\|^2 + C\varepsilon^2.$$

Dividing Equation (E.15) by $\varepsilon$, the equation for $Y_{\mathrm{f}}$ gives the linear contribution $-\|e\|^2/\varepsilon$. Choose $r$ and then an upper bound on $\varepsilon$ to absorb the nonlinear term $C(r+\varepsilon)\|e\|^2/\varepsilon$. The drift terms have no inverse power of $\varepsilon$ left. Bounded derivatives of $\theta_{\mathrm{qs}}$ and the estimate $\|\xi'\| \le C(\|\xi\| + \|e\| + \varepsilon)$ bound their contribution by $C\|e\|(\|\xi\| + \|e\| + \varepsilon)$. Thus

$$\frac{d}{dt}(e^\top Y_{\mathrm{f}} e) \le -\frac{c_1}{\varepsilon}\|e\|^2 + C\|e\|\|\xi\| + C\|e\|^2 + C\varepsilon\|e\|.$$

Bound the mixed term by $\|\xi\|^2/4 + C\|e\|^2$ and the last term by $c_1\|e\|^2/(4\varepsilon) + C\varepsilon^3$. For sufficiently small $\varepsilon \le \bar{\varepsilon}$, the fast dissipation absorbs all remaining multiples of $\|e\|^2$, including $Cr\|e\|^2$. The resulting estimate has fixed positive coefficients $a_0, b_0$. Choose

$$\lambda_0 = \min\left\{\frac{a_0}{\lambda_{\max}(X_0)}, \frac{b_0}{\bar{\varepsilon}\lambda_{\max}(Y_{\mathrm{f}})}\right\} > 0.$$

For $0 < \varepsilon(t) \le \bar{\varepsilon}$, this choice gives

$$\lambda_0\mathscr{L} \le a_0\|\xi\|^2 + \frac{b_0}{\bar{\varepsilon}}\|e\|^2 \le a_0\|\xi\|^2 + \frac{b_0}{\varepsilon(t)}\|e\|^2.$$

Thus

$$\mathscr{L}' \le -a_0\|\xi\|^2 - \frac{b_0}{\varepsilon}\|e\|^2 + C\varepsilon^2 \le -\lambda_0\mathscr{L} + C\varepsilon^2. \tag{E.17}$$

All constants are uniform for $0 < \varepsilon(t_1) \le \bar{\varepsilon}$.

Because $\varepsilon(t)^2 = \varepsilon(t_1)^2 e^{-4(t-t_1)}$, integration gives, with $T = t - t_1$,

$$\mathscr{L}(t) \le e^{-\lambda_0 T}\mathscr{L}(t_1) + C\varepsilon(t_1)^2\int_0^T e^{-\lambda_0(T-u)}e^{-4u}\,du.$$

For $0 < \lambda < \min\{\lambda_0, 4\}$, Lemma I.12 bounds the integral by $e^{-\lambda T}/(4-\lambda)$, also when $\lambda_0 = 4$. Hence

$$\mathscr{L}(t) \le C\big(\mathscr{L}(t_1) + \varepsilon(t_1)^2\big)e^{-\lambda(t-t_1)}.$$

The bound on $\theta_{\mathrm{qs}}$ and norm equivalence imply, for some $c > 0$,

$$\|\xi(t)\| + \|\theta(t)\| + \varepsilon(t) \le C\big(\|\xi(t_1)\| + \|\theta(t_1)\| + \varepsilon(t_1)\big)e^{-c(t-t_1)}. \tag{E.18}$$

This estimate also bounds $\|\xi\| + \|e\|$, since $e = \theta - \theta_{\mathrm{qs}}$ and $\|\theta_{\mathrm{qs}}\| \le C(\|\xi\| + \varepsilon)$. Choose the initial coordinate size at $t_1$ so that the bound is less than half the coordinate-ball radius. At a first exit, the same estimate would contradict the boundary value. It therefore holds for all $t \ge t_1$. Continuation is possible at every finite time because $\varepsilon > 0$ and both inputs to the polar maps remain full rank.

The constants are chosen in this order: fix the coordinate ball and admissible initial size, choose $r_1$ so that $p^{\mathcal{S}}(t_1) \in \mathcal{K}_{r_1}$ gives small $\|\theta(t_1)\|$, then shrink the factor tube and choose $\mathcal{E}_0, \varepsilon_0$ to bound $\xi(t_1)$ and $\varepsilon(t_1)$. The earlier Lyapunov estimate keeps the factors in that tube before entry, and all choices depend only on the problem and target conditions.

Returning to the original errors,

$$\begin{aligned}
\operatorname{dist}(f, \mathcal{O}^\star) &= \|f_{\mathrm{sl}} - f^\star\| \le C(\|\xi\| + \varepsilon(1 + \|\theta\|)),\\
\|W - W^\star\|_F &\le \sqrt{2}\,\operatorname{dist}(f, \mathcal{O}^\star),\\
\|p - p^\star\|_1 &\le \|p - p^{\mathcal{S}}\|_1 + \|p^{\mathcal{S}}(\theta) - p^\star\|_1 \le Ce^{-\operatorname{gap}_{\mathrm{in}}/(2\varepsilon)} + C\|\theta\|.
\end{aligned}$$

The last bound uses the smooth active softmax and Equation (E.5). On a sufficiently small fixed scale interval, the exponentially small inactive term is bounded by $C\varepsilon$, so Equation (E.18) controls all three errors. Absorbing the bounded initial coordinate size into $C$ gives Equation (6.3), while Equation (E.6) gives its entry-time bound. The same positive margin floor $c_0$ holds before and after entry, so the attention ratios imply $1 - \alpha_{ir,k_{ir}^\star}(t) \le Ce^{-c_0/\varepsilon(t)}$ for all $t \ge t_0$. Since $W(t) \to W^\star$ and $\|W^\star\|_{\mathrm{op}} = 1$, also $\widetilde{W}(s(t))/\|\widetilde{W}(s(t))\|_{\mathrm{op}} \to W^\star$. This proves Theorem 6.3. □

The initialization condition follows in particular from a small target support-function residual. If $\delta_0 = 1 - \langle W(t_0), \Delta^\star\rangle/\gamma_h$, Lemma 5.1 bounds the orbit and product errors by $C\sqrt{\delta_0}$. For small $\delta_0, \varepsilon(t_0)$, the prefactors are positive and the Gibbs identity gives $\mathcal{E}(t_0) \le e_m(t_0) + C_\pi\varepsilon(t_0) \le C(\sqrt{\delta_0} + \varepsilon(t_0))$. Alternatively, a global softmax-attention solution with positive gradient prefactors, uniformly regular base weights, and $\Xi(t) \to \infty$ has $e_m(t) \to 0$ by Theorem 5.2; under (MM), it eventually meets the same local thresholds.

# F Examples and limits of the conclusions

The first examples separate factor stationarity from margin optimality and product convergence from identification of the gradient weights. The last construction gives attention comparisons with dual nondegeneracy.

## F.1 Stationarity, orientation, and rank loss

Start with the abstract comparison $\Delta = \mathrm{diag}(a_1, \dots, a_d)$, where $a_i > 0$, and impose $\langle \widetilde{K}\widetilde{Q}^\top, \Delta\rangle \ge 1$. An attention-realizable rank-one comparison family follows below. If $a_{(1)} \ge \cdots \ge a_{(d)}$, the optimal factor objective $\max\{\|\widetilde{K}\|_{\mathrm{op}}^2, \|\widetilde{Q}\|_{\mathrm{op}}^2\}$ has value $(\sum_{i=1}^h a_{(i)})^{-1}$ by the geometry in Appendix A. For a unit latent vector $u$, put $\widetilde{K}_j = \widetilde{Q}_j = a_j^{-1/2} e_j u^\top$. The constraint is tight, and with $\lambda = 1/a_j$,

$$(\lambda\Delta\widetilde{Q}_j, \lambda\Delta^\top\widetilde{K}_j) = (\widetilde{K}_j, \widetilde{Q}_j) \in \partial\max\{\|\widetilde{K}_j\|_{\mathrm{op}}^2, \|\widetilde{Q}_j\|_{\mathrm{op}}^2\}.$$

Each factor has a simple largest singular value. The two active block gradients of the maximum are therefore $(2\widetilde{K}_j, 0)$ and $(0, 2\widetilde{Q}_j)$. Their average $(\widetilde{K}_j, \widetilde{Q}_j)$ is a subgradient and equals the multiplier times the margin gradient. Thus the pair satisfies the factor KKT conditions and is balanced. Its objective can nevertheless be far from optimal:

$$\gamma_h\|\widetilde{K}_j\widetilde{Q}_j^\top\|_{\mathrm{op}} = \frac{\sum_{i=1}^h a_{(i)}}{a_j}, \tag{F.1}$$

which can be arbitrarily large.

These points are not local minima when $h \ge 2$. Choose $i \ne j$ and a unit $g \perp u$, and let $c_\vartheta = (1 + a_i\vartheta^2)^{-1/2}$. The factors

$$\widetilde{K}(\vartheta) = c_\vartheta(\widetilde{K}_j + \vartheta e_i g^\top), \qquad \widetilde{Q}(\vartheta) = c_\vartheta(\widetilde{Q}_j + \vartheta e_i g^\top)$$

remain feasible with equality. The orthogonality $u^\top g = 0$ removes the cross terms in the product, so its margin is $c_\vartheta^2(1 + a_i\vartheta^2) = 1$. The squared singular values of either factor are $c_\vartheta^2/a_j$ and $c_\vartheta^2\vartheta^2$. For $0 < \vartheta^2 < a_j^{-1}$, their maximum is $c_\vartheta^2/a_j < 1/a_j$. When $h = 1$ and $a_i > a_j$, use $\widetilde{K}(\vartheta) = \widetilde{Q}(\vartheta) = c_\vartheta(a_j^{-1/2}e_j + \vartheta e_i)$ instead. Its objective is $(a_j^{-1} + \vartheta^2)/(1 + a_i\vartheta^2) < a_j^{-1}$. At width one, only a leading coordinate among these stationary points is optimal.

For an attention-realizable example, take the two masked-query comparisons $\Delta_\pm = v_\pm v_\pm^\top$, $v_\pm = (1, \pm a)^\top$, $0 < a < 1$, from Appendix G.2, now with width $h = 2$. The uniform dual distribution gives $\Delta^\star = \mathrm{diag}(1, a^2)$, and $W = I_2$ attains both margins $1 + a^2$, so $\gamma_2 = \gamma_2^{\mathrm{rk}} = 1 + a^2$. For a unit $u \in \mathbb{R}^2$, set $\widetilde{K} = \widetilde{Q} = a^{-1}e_2u^\top$. Both unit-margin constraints are tight. With $\lambda_+ = \lambda_- = 1/(2a^2)$,

$$\lambda_+\Delta_+ + \lambda_-\Delta_- = \mathrm{diag}(a^{-2}, 1), \qquad \big((\lambda_+\Delta_+ + \lambda_-\Delta_-)\widetilde{Q}, (\lambda_+\Delta_+ + \lambda_-\Delta_-)^\top\widetilde{K}\big) = (\widetilde{K}, \widetilde{Q}).$$

The subgradient calculation above gives the KKT conditions, with objective ratio $(1 + a^2)/a^2$. In the symmetric softmax-attention example, the normalized pair $K = Q = e_2u^\top$ also admits a stationary continuation. Its two margins are equal, so the weights are equal, and each factor maximizes the linear objective of its rank-one input to the polar map. The stationary margin is $a^2$, arbitrarily small relative to $1 + a^2$.

**Polar flow need not preserve Gram balance.** For a prescribed gradient, consider the following full-rank state in raw optimizer time:

$$-\nabla L = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad \widetilde{K} = \widetilde{Q} = \operatorname{diag}(2, 1).$$

The negative factor gradients are $\left(\begin{smallmatrix} 2 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 2 & 0 \\ 2 & 1 \end{smallmatrix}\right)$. Their polar factors give

$$\dot{\widetilde{K}} = \frac{1}{\sqrt{10}} \begin{pmatrix} 3 & 1 \\ -1 & 3 \end{pmatrix}, \qquad \dot{\widetilde{Q}} = \frac{1}{\sqrt{13}} \begin{pmatrix} 3 & -2 \\ 2 & 3 \end{pmatrix}.$$

Although the two Gram matrices initially agree,

$$\left[\frac{d}{ds}(\widetilde{K}^\top \widetilde{K} - \widetilde{Q}^\top \widetilde{Q})\right]_{11} = 12\left(\frac{1}{\sqrt{10}} - \frac{1}{\sqrt{13}}\right) > 0.$$

The introduction's two factorizations of $\widetilde{W} = I_2$ also have the same raw cross-Gram matrix $\widetilde{K}^\top(-\nabla L)\widetilde{Q} = I_2$ but different product velocities, so that matrix alone does not determine the dynamics.

**Rank loss.** For a fixed invertible $Z$ and full-width square factors, suppose $\det(K(t_0)^\top Z Q(t_0)) < 0$. If both $ZQ$ and $Z^\top K$ stayed uniformly full rank, then $K, Q$ would remain invertible and this determinant could not change sign. The score for fixed $Z$ is nondecreasing and bounded, so its dissipation is integrable. Choose a vanishing-dissipation subsequence and a convergent subsequence of the factors. Uniform full rank makes the two limiting polar maximizers unique; zero duality gaps then force the limiting factors themselves to be those maximizers. They are square orthogonal matrices, and $K^\top Z Q = |ZQ| \succ 0$ there, contradicting the preserved negative determinant. Thus the trajectory encounters or approaches rank deficiency. In the scalar example below it reaches the boundary in finite time. For $Z = 1$ and $K(t_0) = a$, $Q(t_0) = -a$, with $0 < a \le 1$, put

$$c(t) = \max\{(1 + a)e^{-(t-t_0)} - 1, 0\}, \qquad (K(t), Q(t)) = (c(t), -c(t)). \tag{F.2}$$

Before $t_0 + \log(1 + a)$ this solves the unique full-rank equations; afterward it is the admissible zero continuation. Thus max-margin convergence cannot hold for every initialization and every allowed continuation.

**Initialization criteria.** The cross-Gram condition follows from the scalar sufficient condition $\langle KQ^\top, Z\rangle > 1 - \operatorname{gap}_h(Z)$, but the converse need not hold. Indeed, for $\Gamma = K^\top Z Q$ and $G = \operatorname{sym}\Gamma$, let $E_k$ span the right singular vectors of $\Gamma$ indexed by $k, \dots, h$. For a unit $v \in E_k$, $v^\top G v \le \|\Gamma v\| \le \sigma_k(\Gamma)$. Courant–Fischer and the factor-norm bounds give $\lambda_k(G) \le \sigma_k(\Gamma) \le \sigma_k(Z)$. Thus, when $\|Z\|_{(h)} = 1$,

$$\langle KQ^\top, Z\rangle > 1 - \operatorname{gap}_h(Z) \quad \Longrightarrow \quad \lambda_h(G) \ge \operatorname{tr} G - \sum_{i<h} \sigma_i(Z) > \sigma_{h+1}(Z).$$

The fixed-matrix diagnostic in Appendix G.3 gives an initialization satisfying the cross-Gram criterion but not the scalar one.

### F.2 Product convergence does not identify normalized gradient weights

Let $\Delta_i = \zeta u_i v_i^\top$ for $i \in [h]$, where $\zeta > 0$ is a common scale and $(u_i), (v_i)$ are orthonormal families. In an attention model in a common ambient space, query $v_i$, target-key representation $\zeta u_i$, and a zero distractor realize this comparison by a two-key mask. Consider positive diagonal factor coordinates

$$K(t) = U \operatorname{diag}(a_i(t)), \qquad Q(t) = V \operatorname{diag}(c_i(t)), \qquad 0 < a_i(t_0), c_i(t_0) \le 1.$$

Every positive comparison distribution gives the normalized matrix $Z = U \operatorname{diag}(p) V^\top$. Indeed, $ZQ = U \operatorname{diag}(p_i c_i)$ and $Z^\top K = V \operatorname{diag}(p_i a_i)$ have positive diagonal coefficients, so their polar updates are $U, V$, independent of the positive weights. The factor equations reduce to $a_i' = 1 - a_i$ and $c_i' = 1 - c_i$. With $r = t - t_0$, $d_i^K = 1 - a_i(t_0)$, $d_i^Q = 1 - c_i(t_0)$, $A_i = d_i^K + d_i^Q$, and $B_i = d_i^K d_i^Q$, the exact solution is

$$a_i(t) = 1 - d_i^K e^{-r}, \qquad c_i(t) = 1 - d_i^Q e^{-r}, \qquad m_i(W(t)) = \zeta(1 - A_i e^{-r} + B_i e^{-2r}). \tag{F.3}$$

Hence $W \to UV^\top$ regardless of the weight evolution. The uniform certificate has $\Delta^\star = (\zeta/h)UV^\top$, $W^\star = UV^\top$, and $\gamma_h = \zeta$. Reweighting changes only positive singular values, so $H_{\mathrm{diss}} = 0$ and $\Psi(p) \equiv \zeta$ on the simplex; the dual minimizer is not unique when $h \geq 2$.

For $\Lambda(t) = \Lambda_0 e^{2r}$, the Gibbs odds satisfy the exact identity

$$\log \frac{p_i(t)}{p_j(t)} = \log \frac{\rho_i(t)}{\rho_j(t)} + \zeta\Lambda_0\big[(A_i - A_j)e^r - (B_i - B_j)\big]. \tag{F.4}$$

When the initial deficits $A_i$ differ, the normalized margin differences decay as $e^{-r}$ but are multiplied by a logit scale of order $e^{2r}$. These differences can therefore determine the weights even though all normalized margins have the same limit. In exact two-key attention satisfying (TC), the positive normalized margins imply $\rho_i(t) \to \rho_i^\star > 0$ by Appendix D.1. Such trajectories are realized, for example, by target/distractor utilities $1, 0$ and a loss decreasing in every query output; for a two-key row the gradient prefactors are then positive. With $I_{\max} = \arg\max_i A_i$,

$$p_i(\infty) = \begin{cases} \dfrac{\rho_i^\star e^{-\zeta\Lambda_0 B_i}}{\sum_{j\in I_{\max}} \rho_j^\star e^{-\zeta\Lambda_0 B_j}}, & i \in I_{\max}, \\ 0, & i \notin I_{\max}. \end{cases}$$

Each trajectory has convergent weights, but their limit varies with the initial deficits while the product limit remains $UV^\top$. In the symmetric two-row task of Appendix G.3, $p_1$ tends to $1/2$, $1$, or $0$.

These trajectories can also have $\Xi(\infty) < \infty$. To prove finiteness, we need an upper bound on $\kappa$. For $0 < \rho < 1/h$, take $Z_0 = UV^\top/h$ and $K = cU, Q = cV$ with $c = h\rho < 1$. The polar-input separation is $\rho$, the support-function residual is $1-c^2$, and the dissipation is $2c(1-c)$. This admissible point in the defining infimum gives

$$\kappa(\rho) \leq \frac{2h\rho}{1+h\rho} \leq 2h\rho, \qquad 0 < \rho < 1/h. \tag{F.5}$$

If the $A_i$ are not all equal, Equation (F.4) gives $\min_i p_i(t) \leq C\exp(-ce^r)$. Here $0 < \eta(t) \leq \min_i p_i(t)$, since the selected matrix has singular values $p_i$ and both factor diagonals are at most one. The polar-input separation is integrable and eventually less than $1/h$. Equation F.5, together with $\kappa \leq 1$ on the finite initial interval, yields $\Xi(\infty) < \infty$ although the product converges.

For externally prescribed base weights, the broader Gibbs model also admits nonconvergent weights. With $d = 3$, $h = 2$, take

$$\Delta_1 = \mathrm{diag}(2, 0.5, 0.1), \qquad \Delta_2 = \mathrm{diag}(0.5, 2, 0.1), \qquad \pi_1(t) = \tfrac{1}{2} + \tfrac{1}{5}\sin t, \quad \pi_2 = 1 - \pi_1.$$

For $t_0 = 0$ and equal factors initially $c_0[e_1, e_2]$, $0 < c_0 \leq 1$, the two leading polar directions remain $[e_1, e_2]$. Therefore

$$K(t) = Q(t) = \big(1 - (1-c_0)e^{-t}\big)[e_1, e_2], \qquad p(t) = \pi(t). \tag{F.6}$$

The last identity holds because the two margins are equal along the path. Thus the product converges to $\mathrm{diag}(1, 1, 0)$ while the weights oscillate. Appendix G.3 reports the corresponding prescribed-base-weight diagnostic.

### F.3 Attention comparisons with a nondegenerate dual minimum

The following construction reduces dual nondegeneracy to connectivity of a Gram graph. Let $M_0 = [a_1, \ldots, a_h] \in \mathbb{R}^{h\times h}$ be invertible and $z_i = M_0^{-\top} e_i$. In one sequence, create $3h$ token positions: queries $z_i$, target keys $ha_i$, and zero distractors. Mask query $i$ to its target and distractor. The resulting comparisons and certificate are

$$\Delta_i = ha_i z_i^\top, \qquad p_i^\star = 1/h, \qquad \Delta^\star = W^\star = I_h. \tag{F.7}$$

All margins at $W^\star$ equal $h$, so (MM) holds and (SC) is vacuous. Target/distractor utilities $1, 0$ and a smooth loss with strictly negative query derivatives give (TC). It remains to check (ND).

Let $R_0 = M_0^\top M_0$, connect $i, j$ when $(R_0)_{ij} \neq 0$, and define $L_G$ by $v^\top L_G v = \sum_{i<j}(R_0)_{ij}^2(v_i - v_j)^2$. For $v \in \mathbf{1}^\perp$, put $D_v = \mathrm{diag}(v)$ and $\Delta(v) = hM_0 D_v M_0^{-1}$. At the identity, $D\,\mathrm{polar}(I)[E] = \mathrm{skew}(E)$, so

$$\begin{aligned} \langle v, H_{\mathrm{diss}} v\rangle &= 2\|\,\mathrm{skew}\,\Delta(v)\|_F^2 = \tfrac{1}{2}\|\Delta(v) - \Delta(v)^\top\|_F^2, \\ \Delta(v) - \Delta(v)^\top &= hM_0^{-\top}[R_0, D_v]M_0^{-1}. \end{aligned} \tag{F.8}$$

Entrywise, $[R_0, D_v]_{ij} = (R_0)_{ij}(v_j - v_i)$. Since $R_0$ is symmetric, summing over both orders of each pair gives

$$\|[R_0, D_v]\|_F^2 = 2\sum_{i<j}(R_0)_{ij}^2(v_i - v_j)^2 = 2v^\top L_{\mathrm{G}} v.$$

Since $M_0$ is invertible, the Hessian vanishes exactly when the commutator does: $v_i = v_j$ on every edge. Its kernel therefore consists of vectors constant on each connected component. The multiplication bounds in Lemma I.2, applied to $M_0^{-\top}$ and $M_0^{-1}$, give

$$\frac{h^2}{\|M_0\|_{\mathrm{op}}^4} v^\top L_{\mathrm{G}} v \le \langle v, H_{\mathrm{diss}} v\rangle \le \frac{h^2}{\sigma_{\min}(M_0)^4} v^\top L_{\mathrm{G}} v. \tag{F.9}$$

A connected graph leaves only constant kernel vectors, whose intersection with $\mathbf{1}^\perp$ is zero. For a disconnected graph, different constants on its components can have zero total sum. Thus $H_{\mathrm{diss}} \succ 0$ on $\mathbf{1}^\perp$ exactly when this graph is connected, and Lemma E.1 then gives (ND). For $h \ge 2$, the smallest eigenvalue is at least $h^2\lambda_2(L_{\mathrm{G}})/\|M_0\|_{\mathrm{op}}^4$. Within this construction, an invertible $M_0$ drawn from a distribution with a density has a complete Gram graph almost surely: each off-diagonal Gram entry is a nonzero polynomial in its entries, whose zero set has Lebesgue measure zero.

For example, $M_0 = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ gives the connected width-two family

$$\Delta_1 = \begin{pmatrix} 2 & -2 \\ 0 & 0 \end{pmatrix}, \qquad \Delta_2 = \begin{pmatrix} 0 & 2 \\ 0 & 2 \end{pmatrix}, \qquad (\Delta_1 + \Delta_2)/2 = I_2.$$

# G Experimental details

The experiments examine spectral selection, weight limits, and sensitivity to initialization. Each diagnostic specifies its continuous-flow model; solver comparisons measure numerical consistency, not certified error.

## G.1 Common numerical protocol and continuous-flow metrics

**Solvers and event handling.** We integrate the continuous-time tasks in double precision (FP64), using the settings in Table 1. Fixed-matrix refinements tighten both tolerances by a factor of ten and halve the maximum step. Solutions of the normalized softmax-attention flow use the coordinate systems given below. The frozen-token study switches from Radau to BDF at $t = 4$ and reports endpoints at $t = 6$.

| Task / integration | Solver | Relative / absolute tolerance | Maximum step |
|---|---|---|---|
| Fixed matrix, primary | DOP853 | $2 \times 10^{-11}$ / $2 \times 10^{-13}$ | 0.1 |
| Two-mode, primary $[0, 8]$ | Radau | $10^{-9}$ / $10^{-12}$ | 0.05 |
| Two-mode, refinement $[0, 8]$ | Radau | $10^{-10}$ / $10^{-13}$ | 0.025 |
| Two-mode, independent $[0, 8]$ | BDF | $10^{-10}$ / $10^{-13}$ | 0.025 |
| Two-mode, original factors $[0, 3]$ | Radau | $10^{-10}$ / $10^{-13}$ | 0.025 |
| Coupled task, primary | Radau | $10^{-9}$ / $10^{-12}$ | 0.03 |
| Coupled task, refinement | Radau | $10^{-10}$ / $10^{-13}$ | 0.015 |
| Frozen tokens, primary $[0, 4]$ | Radau | $2 \times 10^{-10}$ / $2 \times 10^{-12}$ | $10^{-2}$ |
| Frozen tokens, refinement $[0, 4]$ | Radau | $2 \times 10^{-11}$ / $2 \times 10^{-13}$ | $5 \times 10^{-3}$ |
| Initialization survey, primary | Radau | $2 \times 10^{-9}$ / $2 \times 10^{-12}$ | 0.05 |
| Initialization survey, refinement | Radau | $2 \times 10^{-10}$ / $2 \times 10^{-13}$ | 0.025 |
| Initialization survey, 24 checks | BDF | $2 \times 10^{-10}$ / $2 \times 10^{-13}$ | 0.025 |

**Table 1:** Numerical integration settings. Every two-mode rank-one problem has all four integrations. The survey's BDF checks use seed 0 in each perturbation/scale cell and disk seeds 0 and 64 at each scale.

We include every attempted initialization in the reported counts and distinguish failed integrations from completed trajectories. For a nonzero input to the polar map we use a maximizing SVD direction. At zero, the scalar diagnostic takes the admissible zero direction. The two-mode rank-one integrations stop when either polar input is zero or the coordinate chart becomes invalid. The survey continues a

transverse zero only when both one-sided normal derivatives point in the crossing direction, the norm of the other input exceeds $10^{-8}$, and the relative transversality test exceeds $10^{-9}$. Nontransverse events are boundary-unresolved.

**Continuous-flow metrics.** We measure normalized product error by $\|W - W^\star\|_F$ and normalized-gradient-weight error by $\|p - p^\star\|_1$. For normalized factors and target singular-vector bases $U_h, V_h$, the distance to the orthogonal orbit is

$$d_{\mathrm{orb}}(K, Q) = \min_{O \in O(h)} \left(\|K - U_h O\|_F^2 + \|Q - V_h O\|_F^2\right)^{1/2},$$

where a common orthogonal rotation fits both factors. With $M = U_h^\top K + V_h^\top Q$, we obtain

$$d_{\mathrm{orb}}(K, Q)^2 = \|K\|_F^2 + \|Q\|_F^2 + 2h - 2\|M\|_*.$$

**Reporting.** We report synthetic trajectories without smoothing or probability flooring. The two-mode curves show medians and full min–max ranges across eight paired initializations; rate fits concern individual trajectories. Solver comparisons use the same saved times. Their discrepancies are reported separately from errors relative to the target.

### G.2 Two-mode rank-one example: Figure 1

**Task and certificate.** Here the certificate has a nonzero singular mode beyond the head width. For $0 < a < 1$, let $v_+ = (1, a)^\top$, $v_- = (1, -a)^\top$, and

$$X = (v_+, 0, v_+, v_-, 0, v_-)^\top.$$

Queries 3 and 6 see keys $\{1, 2\}$ and $\{4, 5\}$, with the first key as target. Values are $(1, 0, 0, 1, 0, 0)^\top$, both readouts are one, $\beta = 1$, and $K, Q \in \mathbb{R}^{2\times 1}$ are shared. The coupled loss is

$$\ell(u_+, u_-) = \exp[-(u_+ + u_-)/2].$$

The target utility is one and competitor utility zero. Thus (TC) holds and both gradient coefficients are positive at every finite scale. With $p = p_+$, the comparisons $\Delta_\pm = v_\pm v_\pm^\top$ give

$$\Delta(p) = \begin{pmatrix} 1 & a(2p-1) \\ a(2p-1) & a^2 \end{pmatrix}, \qquad p^\star = (1/2, 1/2), \quad \Delta^\star = \mathrm{diag}(1, a^2), \quad W^\star = e_1 e_1^\top. \tag{G.1}$$

At $W^\star$, both margins equal one. Thus $\gamma_h = 1$, the cutoff gap is $1 - a^2 > 0$, and there are no inactive comparisons to check for (SC). The dual objective is

$$\|\Delta(p)\|_{(1)} = \frac{1 + a^2 + \sqrt{(1-a^2)^2 + 4a^2(2p-1)^2}}{2}.$$

Its Hessian on the unit simplex tangent $(1, -1)/\sqrt{2}$, together with the dissipation Hessian, is

$$H_\Psi = \frac{4a^2}{1 - a^2} > 0, \qquad H_{\mathrm{diss}} = 4a^2 > 0. \tag{G.2}$$

The perturbation of each polar input is $\sqrt{2}ae_2$, contributing $2a^2$ to $H_{\mathrm{diss}}$. Thus all certificate conditions hold analytically.

**Scale-resolved coordinates.** For $K = (x, y)^\top$, $Q = (u, v)^\top$, use the equivalent coordinates

$$x_c = (x + u)/2, \quad d_x = (x - u)/2, \quad d_y = (y - v)/2, \quad z = a(xv + yu)/\varepsilon.$$

For $x_c > 0$,

$$y_c = (y + v)/2 = \frac{d_x d_y + \varepsilon z/(2a)}{x_c}.$$

The logits and weight odds are

$$\begin{aligned} l_\pm &= \frac{x_c^2 - d_x^2 + a^2(y_c^2 - d_y^2)}{\varepsilon} \pm z, \\ \alpha_\pm &= \mathrm{sigmoid}(l_\pm), \qquad \log(p_+/p_-) = 2(\log \alpha_+ - \log \alpha_-) - 2z. \end{aligned} \tag{G.3}$$

The gradient coefficients are $\omega_\pm = \frac{1}{2}\ell\alpha_\pm(1-\alpha_\pm)$, with $2p_+ - 1 = \tanh(\log\alpha_+ - \log\alpha_- - z)$. Put $b = a(2p_+ - 1)$, $c_+ = \Delta(p)(x_c, y_c)^\top$, $c_- = \Delta(p)(d_x, d_y)^\top$, $n_K = \|c_+ - c_-\|_2$, and $n_Q = \|c_+ + c_-\|_2$. Then

$$S = n_K^{-1} + n_Q^{-1}, \qquad R = -\frac{4\langle c_+, c_-\rangle}{(n_K + n_Q)n_K n_Q} = n_Q^{-1} - n_K^{-1}.$$

The normalized flow is

$$\begin{aligned} x_c' &= (c_+S + c_-R)_1/2 - x_c, \\ d_x' &= (-c_+R - c_-S)_1/2 - d_x, \qquad d_y' = (-c_+R - c_-S)_2/2 - d_y, \\ z' &= \frac{a}{\varepsilon}\Big[(1+a^2)\Big\{\Big(2d_xd_y + \frac{\varepsilon z}{2a}\Big)S + (x_cd_y + d_xy_c)R\Big\} \\ &\qquad + b\big\{(x_c^2 + y_c^2 + d_x^2 + d_y^2)S + 2(x_cd_x + y_cd_y)R\big\}\Big]. \end{aligned} \tag{G.4}$$

All four factor coordinates are retained. In the equation for $z'$, the linear normalization terms cancel against $\varepsilon' = -2\varepsilon$.

**Perturbed initialization.** Seeds $j \in \{0, \dots, 7\}$ select independent $r_K, r_Q \sim \mathrm{Unif}[0.8, 0.95]$ and $\theta_K, \theta_Q \sim \mathrm{Unif}[-0.4, 0.4]$ radians. Set

$$K(0) = r_K(\cos\theta_K, \sin\theta_K)^\top, \qquad Q(0) = r_Q(\cos\theta_Q, \sin\theta_Q)^\top.$$

We reuse these draws at $a = 0.5, 0.8, 0.95$. The 24 initial product errors range from $0.219$ to $0.516$, with $\varepsilon_0 = 0.1$, $\varepsilon(t) = \varepsilon_0 e^{-2t}$, and $T = 8$. Evaluation times combine 801 equally spaced points on $[0, 8]$, 65 geometrically spaced points on $[10^{-6}, 0.1]$, and zero, after removing duplicates.

**Results and numerical checks.** All 96 integrations complete, comprising four integrations for each of the 24 initializations. The smallest sampled target margin is $0.253$. Endpoint errors appear in Table 2; Table 3 reports discrepancies between integrations. The plotted bands show full ranges across the eight paired initializations per $a$, not confidence intervals. On 192 diagnostic states, the loss-gradient relative error is below $5.30 \times 10^{-16}$, the coordinate residual below $4.65 \times 10^{-16}$, and the Jacobian relative error below $1.07 \times 10^{-10}$. We also check the certificate Hessian and the family in Equation (F.5).

| $a$ | $1 - a^2$ | $H_\Psi$ | $H_{\mathrm{diss}}$ | Maximum product error | Maximum weight error |
|---|---|---|---|---|---|
| 0.50 | 0.7500 | 1.3333 | 1.00 | $1.72 \times 10^{-4}$ | $2.47 \times 10^{-10}$ |
| 0.80 | 0.3600 | 7.1111 | 2.56 | $1.97 \times 10^{-4}$ | $7.06 \times 10^{-12}$ |
| 0.95 | 0.0975 | 37.0256 | 3.61 | $2.07 \times 10^{-4}$ | $7.10 \times 10^{-13}$ |

**Table 2:** Largest endpoint errors at $T = 8$ over eight initializations for each $a$, rounded upward. The omitted singular value is $a^2$. Changing $a$ also changes the dual curvature $H_\Psi$.

| Comparison | Interval | Quantity | Maximum discrepancy |
|---|---|---|---|
| Primary–refinement | $[0, 8]$ | Product / factors / weights | $(2.39, 2.07, 2.00) \times 10^{-10}$ |
| Primary–BDF | $[0, 8]$ | Product / weights | $(1.24, 5.96) \times 10^{-9}$ |
| Resolved–original factors | $[0, 3]$ | Product / weights | $(2.58, 1.98) \times 10^{-10}$ |

**Table 3:** Reported two-mode rank-one integration discrepancies at common saved times. Parenthesized values follow the quantity order in the preceding column; the power of ten applies to each value. These are not errors relative to $W^\star$ or $p^\star$.

### G.3 Additional controlled and frozen-token diagnostics

The fixed-matrix and scalar tasks illustrate the distinctions in Appendix F.1. The commuting and prescribed-base-weight examples use the explicit solutions proved once in Appendix F.2. The coupled task has an analytic certificate, whereas the frozen-SST-2 tasks report numerical certificate checks.

**Fixed-matrix selection.** This example tests the cross-Gram condition when the stronger scalar criterion fails. Take $d = 6, h = 2$ and

$$Z = U\,\mathrm{diag}(1, 0.8, 0.56, 0.35, 0.18, 0.08)V^\top/1.8,$$

where independent Gaussian QR factorizations with seed 20260907 give $U, V$. Define

$$A_1 = \cos(0.14)e_1 + \sin(0.14)e_4, \quad A_2 = \cos(0.18)e_2 + \sin(0.18)e_3,$$
$$B_1 = \cos(0.11)e_1 - \sin(0.11)e_4, \quad B_2 = \cos(0.16)e_2 + \sin(0.16)e_3,$$

and set $K(0) = 0.92U[A_1, A_2]$, $Q(0) = 0.90V[B_1, B_2]R(0.08)$, with planar rotation $R$. For a separate rate test use

$$Z_\nu = \mathrm{diag}(1.3, 1, 1-\nu, 0.03)/2.3, \quad K(0) = Q(0) = 0.995[e_1, \cos\theta\, e_2 + \sin\theta\, e_3],$$

with $\nu \in \{0.12, 0.25, 0.5, 0.75, 0.9\}$. Seeds 11 and 29 choose $\theta \in [0.12, 0.2]$, each run ends at $16/\nu$, and least squares uses product errors in $[10^{-5}, 10^{-2}]$.

Table 4 compares the initial criteria and the target errors. For the first fixed-matrix task, the fitted rate is 0.3000, matching $1 - 0.56/0.8 = 0.3$. Dissipation is checked against Equation (5.2) through $t = 18$, before subtraction reaches roundoff. The inequality tests comprise 20,000 random cases, including 10,092 with positive separation, and 216 deterministic cases, including 81 rank-deficient and 105 with positive separation. No applicable inequality is violated beyond $10^{-11}$.

| Quantity | Reported value |
|---|---|
| $\lambda_{\min}(G(0))$ / $\sigma_3(Z)$ | 0.363474 / 0.311111 |
| Initial score / scalar threshold | 0.812451 / 0.866667 |
| Product error at $t = 45$ | $3.22 \times 10^{-7}$ |
| Maximum primary–reference factor difference | below $4.0 \times 10^{-13}$ |
| Maximum skew-identity residual | below $1.8 \times 10^{-13}$ |
| Maximum fitted-rate error, ten runs | $5.95 \times 10^{-4}$ |
| Minimum fit $R^2$ | 0.9999999895 |

**Table 4:** Fixed-matrix diagnostics. The cross-Gram criterion holds although the stronger scalar criterion fails.

The scalar diagnostic uses $a = 0.6$ and $t_0 = 0$ in Equation (F.2). The boundary time is $\log 1.6$, followed by the zero continuation; its squared scalar factor satisfies $c(t)^2 \le 0.36e^{-4t}$.

**Newton–Schulz approximation.** In the fixed-matrix task we replace the polar direction by the classical cubic iteration

$$X_0 = M/\|M\|_F, \qquad X_{j+1} = \tfrac{3}{2}X_j - \tfrac{1}{2}X_jX_j^\top X_j,$$

and test depths $1, 2, 3, 5, 8$ and the exact polar factor through $t = 20$. For directions $S_K, S_Q$, the duality defect is

$$\big(\|ZQ\|_* - \langle S_K, ZQ\rangle\big) + \big(\|Z^\top K\|_* - \langle S_Q, Z^\top K\rangle\big).$$

Table 5 shows a negative separation at depth one and close agreement with the exact polar factor at depths five and eight.

| Polar / depth | Product error at $t = 20$ | Duality defect | Separation parameter |
|---|---|---|---|
| Exact polar factor | $5.83 \times 10^{-4}$ | $4.44 \times 10^{-16}$ | 0.13333 |
| 1 | $5.72 \times 10^{-1}$ | $2.23 \times 10^{-1}$ | -0.01980 |
| 2 | $1.14 \times 10^{-1}$ | $5.44 \times 10^{-2}$ | 0.10731 |
| 3 | $7.01 \times 10^{-3}$ | $3.17 \times 10^{-3}$ | 0.13178 |
| 5 | $5.83 \times 10^{-4}$ | $4.07 \times 10^{-10}$ | 0.13333 |
| 8 | $5.83 \times 10^{-4}$ | 0 | 0.13333 |

**Table 5:** Approximate-polar endpoints on each method's own continuous-time trajectory at $t = 20$. This diagnostic uses cubic directions, not practical Muon's quintic approximation; a displayed zero is at the reported numerical precision.

At depths five and eight, the product errors agree with the result using the exact polar factor to the precision reported in the table.

**Commuting softmax attention.** This task checks the initialization-dependent weight limits derived in Appendix F.2. Use $X = (2e_1, 0, e_1, 2e_2, 0, e_2)^\top$, queries 3 and 6, masks $\{1, 2\}$ and $\{4, 5\}$, values $(1, 0, 0, 1, 0, 0)^\top$, and unit readouts. The loss is $\ell(u_1, u_2) = \exp[-(u_1 + u_2)/2]$. The $2 \times 2$ factors realize $\Delta_1 = 2e_1e_1^\top$, $\Delta_2 = 2e_2e_2^\top$, with $p^\star = (1/2, 1/2)$, $\Delta^\star = W^\star = I_2$, and $\gamma_h = 2$. The dual Hessian is zero. The factors are given by Equation (F.3) with $\zeta = 2$, $t_0 = 0$, $d_i^K = a_i$, and $d_i^Q = b_i$. We use $\Lambda(t) = 20e^{2t}$ and $\log \alpha_i = -\log(1 + e^{-2\Lambda W_{ii}})$. The gradient coefficients have a common loss multiplier and $\rho_i \propto \alpha_i^2$, so the odds are evaluated using Equation (F.4) with $\log(\rho_1/\rho_2) = 2\log(\alpha_1/\alpha_2)$.

We evaluate 801 points on $[0, 8]$. The symmetric initialization is $a = b = (0.003, 0.003)$, the imbalanced one has $a = (0.012, 0.004)$, $b = (0.010, 0.006)$, and the third exchanges coordinates.

Table 6 records the limits derived in Appendix F.2. Both imbalanced paths have finite $\Xi(\infty)$ and the same limiting product $I_2$; the exchanged initializations also have identical product-error curves.

| System / initialization | Product limit | Normalized gradient weights |
|---|---|---|
| Commuting / equal deficits | $I_2$ | $p_1 \to 1/2$ |
| Commuting / row 1 slower | $I_2$ | $p_1 \to 1$ |
| Commuting / row 2 slower | $I_2$ | $p_1 \to 0$ |
| Time-dependent base weights | $\mathrm{diag}(1, 1, 0)$ | $p(t) = \pi(t)$ oscillates |

**Table 6:** Product convergence with different weight behavior. The exchanged commuting initializations have identical product-error curves. Only the last row uses externally prescribed base weights.

**Time-dependent base weights.** For the externally varying system in Appendix F.2, use $c_0 = 0.7$ in Equation (F.6) and $\Lambda(t) = 20e^{2t}$. The comparison matrices and externally prescribed oscillating base weights are those specified there.

Here $\eta(t) = c(t) \min\{Z_{11}(t), Z_{22}(t)\} - 0.04 \geq 0.226$. With $C_\pi = \log(1/0.3)$ and $B_\pi = 2/3$, the theorem gives

$$e_m(t) = 2.5(1 - c(t)^2) \leq C_0 \exp\left[-\tfrac{2}{3}\int_0^t \eta(u)^2\, du\right],$$

$$C_0 = 2.5(1 - 0.7^2) + \frac{3\log(1/0.3) + 2/3}{20} \approx 1.4889293.$$

Composite trapezoidal quadrature evaluates the integral, and full factor integration differs from the closed-form path by less than $5 \times 10^{-16}$. Table 6 records the limiting behavior.

**Coupled softmax attention.** For a loss with unequal query multipliers, take $d = 2, h = 1$, $X = ((1, 1), 0, e_1, (1, -2), 0, e_1)^\top$, with queries 3 and 6 and masks $\{1, 2\}$ and $\{4, 5\}$. Target utilities are $g_1 = 1, g_2 = 2$, competitor utilities are zero, and $\beta = 1$, $\kappa_\ell = 3$. Use

$$\ell(\mathbf{u}) = \kappa_\ell^{-1} \log \sum_{r=1}^{2} e^{-\kappa_\ell u_r}, \qquad -\partial_r \ell = \frac{e^{-\kappa_\ell u_r}}{\sum_j e^{-\kappa_\ell u_j}}.$$

The comparisons and certificate are

$$\Delta_1 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \qquad \Delta_2 = \begin{pmatrix} 1 & 0 \\ -2 & 0 \end{pmatrix}, \qquad p^\star = (2/3, 1/3), \quad W^\star = \Delta^\star = e_1e_1^\top.$$

The optimal value and gap are one, and $H_\Psi = H_{\mathrm{diss}} = 9/2$. Both factors start at $e_1$. For $K = (x, y)^\top$, $Q = e_1$, $z = 3y/\varepsilon$, and $\varepsilon = \varepsilon_0 e^{-2t}$, the logits are $l_1 = x/\varepsilon + z/3$ and $l_2 = x/\varepsilon - 2z/3$. Set $\alpha_r = \mathrm{sigmoid}(l_r)$ and

$$\widetilde{\rho}_r = g_r \alpha_r^2 e^{-\kappa_\ell g_r \alpha_r}, \qquad \rho_r = \frac{\widetilde{\rho}_r}{\sum_{j=1}^2 e^{-\kappa_\ell g_j \alpha_j}}.$$

Then

$$\log \widetilde{\rho}_r = -\kappa_\ell g_r \alpha_r + \log g_r + 2\log \alpha_r, \quad \log \frac{p_1}{p_2} = \log \widetilde{\rho}_1 - \log \widetilde{\rho}_2 - z, \quad b = 3p_1 - 2,$$

and

$$x' = \frac{1}{\sqrt{1+b^2}} - x, \qquad z' = \frac{3b}{\varepsilon\sqrt{1+b^2}} + z. \tag{G.5}$$

The condition $x + by > 0$ holds throughout, preserving $Q = e_1$. We test $\varepsilon_0 \in \{0.2, 0.05, 0.01\}$ through $t = 6$ on 721 evaluation times, including a logarithmic grid near zero.

Table 7 gives final target errors and the Lyapunov-identity residual. Directional Jacobian checks give relative errors below $2.7 \times 10^{-9}$, original-coordinate checks give discrepancies below $4.8 \times 10^{-15}$ on nonsaturated states, and the loss-gradient relative error is $9.73 \times 10^{-11}$. The energy comparison differentiates $\mathcal{E} = 1 - x + \varepsilon D_{\mathrm{KL}}(p^\star \| p)$ and the prefactors independently.

| $\varepsilon(0)$ | $\|p(6) - p^\star\|_1$ | $\|W(6) - W^\star\|_F$ | Maximum $p$ discrepancy | Maximum identity residual |
|---|---|---|---|---|
| 0.2 | $4.41 \times 10^{-7}$ | $4.68 \times 10^{-5}$ | $1.40 \times 10^{-10}$ | $6.22 \times 10^{-16}$ |
| 0.05 | $1.10 \times 10^{-7}$ | $1.28 \times 10^{-5}$ | $1.28 \times 10^{-10}$ | $6.80 \times 10^{-16}$ |
| 0.01 | $2.20 \times 10^{-8}$ | $2.62 \times 10^{-6}$ | $1.09 \times 10^{-10}$ | $5.92 \times 10^{-16}$ |

**Table 7:** Coupled softmax attention at $t = 6$. Distance to the orthogonal orbit equals product error. The last columns report the largest primary–reference weight difference and pointwise Lyapunov-identity residual.

**Frozen-feature certificate examples.** We check the target conditions numerically on three comparison problems built from two SST-2 sentences (Socher et al., 2013), with $d = 64$, $h = 4$, four queries, and causal-prefix masks. Fixed 768-dimensional BERT states (Devlin et al., 2019) are projected by a seeded Gaussian $768 \times 64$ reduced-QR factor with fixed signs and divided by $\sqrt{64}$. Cases A/B use target utility one and competitor utility zero, with projection seeds 0/2. Case C uses a frozen task readout and seed 0. Cases B/C share a sentence but differ in both projection and utilities. The loss is

$$\ell(\mathbf{u}) = \frac{1}{10} \log \sum_{r=1}^{4} e^{-10u_r}, \qquad u_r = \sum_{k \in \mathcal{M}_r} \alpha_{rk} g_{rk}.$$

One factor pair is shared across the four queries, with $\widetilde{W}(s(t)) = \tau(s(t))^2 K Q^\top$, $\beta = 1$, and $\varepsilon(t) = 10^{-2} e^{-2t}$. Factors start at the numerically certified target. The comparison combinations have rank at most four, equal to their width.

Table 8 reports endpoint and certificate checks. The product errors peak at approximately 0.125, 0.121, and 0.398, then decrease. Weight errors at $t = 6$ range from $8.77 \times 10^{-6}$ to $1.36 \times 10^{-4}$. Refinement differences are below $6 \times 10^{-10}$ for weights and $7.2 \times 10^{-11}$ for both product and same-gauge factors. Condition (ND) passes each numerical check. The full-matrix check gives $\lambda_{\min}(2H_\Psi - H_{\mathrm{diss}}) \geq -1.55 \times 10^{-8}$ at tolerance $2 \times 10^{-6}$.

| Case | Utility / seed | $\lvert\mathcal{S}\rvert$ | $\|p - p^\star\|_1$ | $\|W - W^\star\|_F$ | Factor distance |
|---|---|---|---|---|---|
| A | Controlled / 0 | 20 | $1.53 \times 10^{-5}$ | $2.83 \times 10^{-4}$ | $2.80 \times 10^{-4}$ |
| B | Controlled / 2 | 10 | $8.77 \times 10^{-6}$ | $1.40 \times 10^{-4}$ | $1.35 \times 10^{-4}$ |
| C | Readout / 0 | 8 | $1.36 \times 10^{-4}$ | $5.04 \times 10^{-4}$ | $3.65 \times 10^{-4}$ |

| Case | $\mathrm{gap}_{\mathrm{in}}$ | $\lambda_{\min}(H_\Psi)$ | $\lambda_{\min}(H_{\mathrm{diss}})$ | Active residual | $\|p_{\mathrm{P}} - p_{\mathrm{R}}\|_1$ |
|---|---|---|---|---|---|
| A | $4.39 \times 10^{-4}$ | 0.06537 | 0.09206 | $9.23 \times 10^{-9}$ | $5.98 \times 10^{-10}$ |
| B | $7.11 \times 10^{-3}$ | 0.08331 | 0.11162 | $7.93 \times 10^{-9}$ | $4.78 \times 10^{-10}$ |
| C | $6.32 \times 10^{-3}$ | 0.09992 | 0.19985 | $5.06 \times 10^{-8}$ | $2.77 \times 10^{-10}$ |

**Table 8:** Frozen-token endpoints and certificate checks at $t = 6$. Factor distance is to the orthogonal orbit. Here $\mathrm{gap}_{\mathrm{in}} = \min_{j \notin \mathcal{S}}(m_j(W^\star) - \gamma_h)$, the active residual is $\max_{j \in \mathcal{S}} |m_j(W^\star) - \gamma_h|$, and P/R denote primary/refinement integrations.

The initial increase in product error is consistent with the finite-scale behavior in Appendix E.2: the target factors need not be equilibria while the weights differ from $p^\star$. The certificate checks here are numerical. Reconstructing these examples requires the selected sentences, feature tensors, and fixed readouts, which are not included in the accompanying source.

### G.4 Initialization and rank-boundary survey

To test sensitivity to initialization, we vary both factors in the coupled task and record target errors and rank-boundary events. We retain $d = 2, h = 1, \beta = 1, \kappa_\ell = 3$ and

$$W^\star = \Delta^\star = e_1 e_1^\top, \qquad p^\star = (2/3, 1/3), \qquad \gamma_h = 1.$$

Let $\Pi_B(v) = v/\max\{1, \|v\|_2\}$. Normalizing 64 Gaussian vectors in $\mathbb{R}^4$ gives $(u_K, u_Q) \in S^3 = \{u \in \mathbb{R}^4 : \|u\|_2 = 1\}$, and

$$\begin{gathered} K(0) = \Pi_B(e_1 + r_{\text{init}} u_K), \qquad Q(0) = \Pi_B(e_1 + r_{\text{init}} u_Q), \\ r_{\text{init}} \in \{0.05, 0.15, 0.35, 0.75, 1.5, 3\}. \end{gathered} \tag{G.6}$$

We reuse directions across amplitudes and add 128 independent pairs uniform in the unit disks. Each of $64 \times 6 + 128 = 512$ pairs is tested at $\varepsilon_0 \in \{0.2, 0.05, 0.01\}$, giving 1,536 problems with $\varepsilon = \varepsilon_0 e^{-2t}$. A run meets the target-error criterion at $T = 10$ when

$$\|W(T) - W^\star\|_F \le 10^{-3}, \qquad \|p(T) - p^\star\|_1 \le 10^{-4}, \qquad \min_j m_j(W(T)) > 0.$$

We record completed runs outside these thresholds, solver failures, and boundary-unresolved runs separately. At an unresolved boundary the procedure stops without exploring all admissible continuations, so the event does not establish nonconvergence.

For $K = (x, y)^\top$, $Q = (q, v)^\top$, the logits are $l_1 = q(x + y)/\varepsilon$ and $l_2 = q(x - 2y)/\varepsilon$. With $\alpha_r = \text{sigmoid}(l_r)$,

$$\log \omega_r = -\kappa_\ell g_r \alpha_r + \log g_r + \log \alpha_r + \log(1 - \alpha_r) + \text{a common term}.$$

For $b = 3p_1 - 2$, $n = \sqrt{1 + b^2}$, and $Z = (1, b)^\top e_1^\top / n$,

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \text{sign}(q) \frac{(1, b)^\top}{n} - \begin{pmatrix} x \\ y \end{pmatrix}, \qquad q' = \text{sign}(x + by) - q, \qquad v' = -v. \tag{G.7}$$

The two polar-input norms are $|q|$ and $|x + by|/n$, and $G = q(x + by)/n$. At $a = x + by = 0$, the one-sided derivatives are $a' = x' + by' + yb'$ with $b' = 3p_1 p_2 \frac{d}{dt} \log(\omega_1/\omega_2)$ and $q' = \pm 1 - q$. After $t \ge 3$, once both margins exceed $0.05$ and the update directions have the same sign, a common sign change makes $q > 0$. With $z = 3qy/\varepsilon$ and transition time $t_s$,

$$\begin{gathered} q(t) = 1 + (q(t_s) - 1)e^{-(t - t_s)}, \qquad v(t) = v(t_s) e^{-(t - t_s)}, \\ x' = \frac{1}{n} - x, \qquad z' = \frac{3qb}{\varepsilon n} + \frac{z}{q}, \qquad l_1 = \frac{qx}{\varepsilon} + \frac{z}{3}, \quad l_2 = \frac{qx}{\varepsilon} - \frac{2z}{3}. \end{gathered}$$

Table 9 retains all 1,536 attempts: 1,392 meet the target-error criterion and 144 stop at nontransverse zeros. All 768 cases with $r_{\text{init}} \le 0.75$ meet the criterion. Among successful runs, 501 start with a negative margin, 411 have $G(0) < 0$, and 492 cross a transverse rank boundary. The smallest transverse derivative ratio is $3.6 \times 10^{-3}$, while all 144 unresolved events have opposite one-sided signs. The other-argument norm always exceeds $2.3 \times 10^{-6}$. Among successful runs, product errors have maximum $4.18 \times 10^{-4}$ and median $3.04 \times 10^{-5}$, gradient-weight errors have maximum $1.49 \times 10^{-10}$, and minimum margins exceed $0.99958$. Figure 2(a) summarizes the perturbed families, while panel (b) plots the disk initializations at $\varepsilon_0 = 0.05$ against $1 - \langle K(0)Q(0)^\top, \Delta^\star \rangle / \gamma_h$ and the scalar cross-Gram value $G(0)$.

All primary/refinement pairs agree on outcomes and crossing sequences. Maximum differences are $8.21 \times 10^{-9}$ for factors, $8.47 \times 10^{-9}$ for products, $8.66 \times 10^{-7}$ for weights, and $3.49 \times 10^{-9}$ for event times. The 24 BDF checks agree on categories and crossings, with product/weight differences $4.07 \times 10^{-8}$/$2.39 \times 10^{-6}$. On 192 states per diagnostic, the maximum loss-gradient relative error is $1.25 \times 10^{-10}$; checks also cover both coordinate Jacobians and event derivatives.

The completed runs meet the target-error criterion even away from the target initialization.

| Initialization | Pairs per scale | $\varepsilon_0 = 0.2$ | $\varepsilon_0 = 0.05$ | $\varepsilon_0 = 0.01$ |
|---|---|---|---|---|
| $r_{\text{init}} = 0.05$ | 64 | 64 / 0 | 64 / 0 | 64 / 0 |
| $r_{\text{init}} = 0.15$ | 64 | 64 / 0 | 64 / 0 | 64 / 0 |
| $r_{\text{init}} = 0.35$ | 64 | 64 / 0 | 64 / 0 | 64 / 0 |
| $r_{\text{init}} = 0.75$ | 64 | 64 / 0 | 64 / 0 | 64 / 0 |
| $r_{\text{init}} = 1.5$ | 64 | 61 / 3 | 61 / 3 | 59 / 5 |
| $r_{\text{init}} = 3$ | 64 | 51 / 13 | 43 / 21 | 48 / 16 |
| Uniform disks | 128 | 103 / 25 | 99 / 29 | 99 / 29 |

**Table 9:** Target-error success / boundary-unresolved counts at $T = 10$. Every initialization remains in the denominator. There are no solver failures or completed runs outside the stated target-error thresholds.

## H Comparison with prior implicit-bias results

Table 10 records whether each result assumes directional convergence and whether it establishes KKT stationarity or an optimal margin.

| **Result** | **Model; update** | **Limit assumed** | **Conclusion** |
|---|---|---|---|
| *Prior work* | | | |
| Fan et al. (2025) | Linear multiclass; spectral descent, Muon | No | Optimal spectral-norm margin |
| Tsilivis et al. (2026) | Homogeneous; steepest flow | No | Accumulation points are KKT |
| Gronich & Vardi (2026) | Homogeneous; momentum, incl. Muon | Yes | Limit is KKT |
| Sheen et al. (2024) | Factorized attention; gradient flow | Yes[†] | Nuclear-norm optimum |
| *This paper* | | | |
| Theorem 5.2 | Factors $(K, Q)$; factorwise polar flow | No | If $\Xi \to \infty$: optimal width-$h$ margin and a tight relaxation; product limit if the optimizer is unique |
| Theorems 6.2 and 6.3 | Attention $(K, Q)$; factorwise polar flow | No | Local product and orthogonal-orbit convergence under (MM) and (TC); joint convergence with (SC), (ND) and the stated energy/scale initialization |

**Table 10:** Implicit-bias results for steepest-descent-type methods. *Limit assumed*: the result assumes that the normalized parameters or product converge. KKT points are stationary for the margin problem but need not be optimal. Each conclusion holds under that result's own hypotheses. [†]Also assumes structural alignment of the data and initialization.

**Implicit bias and factorization.** General implicit-bias results connect directional limits to margin stationarity (Neyshabur et al., 2015; Gunasekar et al., 2018; Sun et al., 2023; Lyu & Li, 2020; Ji & Telgarsky, 2020; Tsilivis et al., 2026; Gronich & Vardi, 2026). For matrix factorization, initialization affects the spectral dynamics (Gunasekar et al., 2017; Arora et al., 2019; Li et al., 2021); for spectral updates, squared-loss analyses study singular-value growth and relative factor orientation (Kang et al., 2026; Rhee et al., 2026). In our attention model, the gradient matrix itself changes through the normalized gradient weights $p_j$.

Dual-norm and manifold formulations give other matrix update geometries (Kravatskiy et al., 2025; Cesista et al., 2026; Li et al., 2026c), with further analyses of spectral constraints, preconditioning, and spectral bias (Chen et al., 2025; Ma et al., 2026; Beneventano et al., 2026). Stochasticity, momentum, and inexact orthogonalization introduce effects absent from our deterministic polar flow (Aladrah et al., 2026; Li et al., 2026a;b; Kim & Oh, 2026; Choudhury et al., 2026). Basis dependence for Adam is studied separately by Singh (2026). The headwise updates in GLM-5 and Kimi K3 motivate

our factorwise geometry, but their full architectures and training rules are not covered (GLM-5 Team, 2026; Kimi Team, 2026).

**Algebraic transfer to additive low-rank parameters.** For $\widetilde{W} = W_0 + BA$ with fixed $W_0$ (Hu et al., 2021), identify the raw factors as $\widetilde{K} = B$ and $\widetilde{Q} = A^\top$. The negative factor gradients are $(-\nabla L(\widetilde{W}))A^\top$ and $B^\top(-\nabla L(\widetilde{W}))$. Transposition commutes with polar maximizing directions, so the same two-factor identities and Lemma 5.1 apply to $BA$ under the stated factor-norm bounds. The fixed-matrix theorem applies when the normalized gradient direction $Z$ is fixed; for time-dependent $Z(t)$, the term in Equation (C.8) remains. A LoRA implicit-bias theorem would additionally require control of $Z(t)$, the loss, and initialization, and any discrete, momentum, or manifold-specific update would require its own analysis.

# I AUXILIARY LEMMAS

We collect the matrix inequalities and analytic facts used in Appendices A–F. The statements use real matrices and the trace inner product. All dimensions are finite; singular values and eigenvalues are ordered decreasingly.

**Lemma I.1** (von Neumann's trace inequality and norm duality; Bhatia, 1997)**.** *For $A, B \in \mathbb{R}^{m\times n}$ and $r = \min\{m, n\}$,*

$$|\langle A, B\rangle| \le \sum_{j=1}^{r} \sigma_j(A)\sigma_j(B).$$

*Consequently, for every $X \in \mathbb{R}^{m\times n}$,*

$$\max_{\|S\|_{\mathrm{op}}\le 1} \langle S, X\rangle = \|X\|_*, \qquad |\langle S, X\rangle| \le \|S\|_{\mathrm{op}}\|X\|_*.$$

*If $X = U_q\Sigma_q V_q^\top$ is a compact SVD with $q = \operatorname{rank} X$, every maximizing $S$ satisfies*

$$Sv_j = u_j, \quad S^\top u_j = v_j \quad (j \le q), \qquad S^\top X = (X^\top X)^{1/2}.$$

*When $m \ge n$ and $X$ has full column rank, the maximizer is unique and equals $\operatorname{polar}(X)$. At rank deficiency it need not be unique.*

The duality bound follows by bounding each singular value of $S$ by $\|S\|_{\mathrm{op}}$; $U_qV_q^\top$ attains the maximum. Equality forces $u_j^\top Sv_j = 1$ for each positive singular value. Since $S$ is a contraction, this implies the two vector identities. This is the support identity used in Appendix A.1, including at rank loss.

**Lemma I.2** (Courant–Fischer, Weyl, and multiplication bounds; Bhatia, 1997)**.** *For a symmetric $H \in \mathbb{R}^{d\times d}$,*

$$\lambda_j(H) = \max_{\dim V = j} \min_{\substack{v\in V \\ \|v\|_2=1}} v^\top Hv = \min_{\dim V = d-j+1} \max_{\substack{v\in V \\ \|v\|_2=1}} v^\top Hv.$$

*For symmetric $H, \widehat{H}$ of the same size and arbitrary matrices $X, \widehat{X}$ of the same size,*

$$|\lambda_j(H) - \lambda_j(\widehat{H})| \le \|H - \widehat{H}\|_{\mathrm{op}}, \qquad |\sigma_j(X) - \sigma_j(\widehat{X})| \le \|X - \widehat{X}\|_{\mathrm{op}}.$$

*For compatible products, with singular-value lists extended by zeros,*

$$\sigma_j(AXB) \le \|A\|_{\mathrm{op}}\,\sigma_j(X)\,\|B\|_{\mathrm{op}}, \qquad \|AXB\|_F \le \|A\|_{\mathrm{op}}\|X\|_F\|B\|_{\mathrm{op}}.$$

*If $A$ and $B$ are square and invertible, then also*

$$\|AXB\|_F \ge \sigma_{\min}(A)\sigma_{\min}(B)\|X\|_F.$$

*In particular, if $1 \le h \le \min\{m, n\}$, $Z \in \mathbb{R}^{m\times n}$, and $Q \in \mathbb{R}^{n\times h}$, then $\|ZQ\|_* \le \|Q\|_{\mathrm{op}}\|Z\|_{(h)}$.*

The eigenvalue perturbation bound follows from the variational formulas. Applying them to $XX^\top$ and using $XBB^\top X^\top \preceq \|B\|_{\mathrm{op}}^2 XX^\top$ gives the singular-value multiplication bound. The lower Frobenius bound follows by applying the upper one to $X = A^{-1}(AXB)B^{-1}$. These bounds are used in Appendices A.2, C, and F.3.

**Lemma I.3** (Löwner–Heinz square-root monotonicity and positive traces; Bhatia, 1997)**.** *For symmetric matrices of the same size, $0 \preceq A \preceq B$ implies $A^{1/2} \preceq B^{1/2}$. For positive-definite matrices, $0 \prec A \preceq B$ also implies $B^{-1} \preceq A^{-1}$. If $P \succeq 0$ and $R \succeq aI$, then $\operatorname{tr}(RP) \geq a \operatorname{tr} P$. None of these statements requires the matrices to commute.*

For inverse order, conjugate by $B^{-1/2}$, invert the resulting positive-definite contraction, and conjugate back. Square-root monotonicity follows from the spectral-calculus identity

$$A^{1/2} = \frac{1}{\pi} \int_0^\infty A(A + tI)^{-1} t^{-1/2} \, dt.$$

For $t > 0$, inverse order gives $(A+tI)^{-1} \succeq (B+tI)^{-1}$, hence $A(A+tI)^{-1} = I - t(A+tI)^{-1} \preceq B(B+tI)^{-1}$. Integration proves the assertion, also for singular matrices. The trace inequality follows from $\operatorname{tr}((R - aI)P) = \operatorname{tr}(P^{1/2}(R - aI)P^{1/2}) \geq 0$. Appendix B uses the square-root inequality; Appendix A.2 uses the trace inequality.

**Lemma I.4** (Minimum eigenvalue along an absolutely continuous path)**.** *Let $H(t)$ be a locally absolutely continuous symmetric matrix path, and let $\lambda(t) = \lambda_{\min}(H(t))$. Then $\lambda$ is locally absolutely continuous. At any time where $H$ is differentiable, its right derivative satisfies*

$$\lim_{s \downarrow 0} \frac{\lambda(t+s) - \lambda(t)}{s} = \min_{\substack{v \in \ker(H(t) - \lambda(t) I) \\ \|v\|_2 = 1}} v^\top H'(t) v.$$

*Thus, if $v^\top H'(t) v \geq b(t)$ for every unit vector in the minimum eigenspace at almost every $t$, then $\lambda'(t) \geq b(t)$ almost everywhere. The eigenvalue need not be simple.*

*Proof.* The Lipschitz eigenvalue bound in Lemma I.2 gives absolute continuity. At a differentiability time, write $H(t+s) = H(t) + sH'(t) + o(s)$. Testing the Rayleigh quotient on a unit minimum eigenvector of $H(t)$ bounds the upper right derivative by the displayed minimum. Conversely, let $v_s$ minimize the perturbed Rayleigh quotient. Every accumulation point of $v_s$ lies in the minimum eigenspace of $H(t)$, and

$$\lambda(t+s) - \lambda(t) \geq s \, v_s^\top H'(t) v_s + o(s).$$

Compactness of the unit sphere gives the reverse bound. At almost every time, both $H$ and $\lambda$ are differentiable, so the claimed differential inequality follows. □

This is the eigenvalue step in the cross-Gram estimate of Appendix B.

**Lemma I.5** (Orthogonal Procrustes; Higham, 1986)**.** *For $M \in \mathbb{R}^{h \times h}$ with a full SVD $M = L\Sigma R^\top$, the orthogonal matrix $O = LR^\top$ maximizes $\langle O, M \rangle$ over $O(h)$, and*

$$\min_{O \in O(h)} \|M - O\|_F^2 = \sum_{j=1}^{h} (\sigma_j(M) - 1)^2.$$

*The optimizer is unique when $M$ is invertible, and then equals $\operatorname{polar}(M)$. If $\|M\|_{\mathrm{op}} \leq 1$, this minimum is at most $h - \|M\|_F^2$.*

*For $K, U \in \mathbb{R}^{m \times h}$ and $Q, V \in \mathbb{R}^{n \times h}$ with $U^\top U = V^\top V = I_h$, set $M = U^\top K + V^\top Q$. Then*

$$\min_{O \in O(h)} \left( \|K - UO\|_F^2 + \|Q - VO\|_F^2 \right) = \|K\|_F^2 + \|Q\|_F^2 + 2h - 2\|M\|_*.$$

*The same maximizing $O$ gives the common alignment of both factors.*

These are the real-matrix forms of the polar best-approximation property in Higham (1986, Theorem 2.2 and Corollary 2.3). Expanding the squared distances reduces each problem to maximizing $\langle O, M \rangle$; Lemma I.1 gives the value $\|M\|_*$. The contraction bound uses $(1-s)^2 \leq 1 - s^2$ for $0 \leq s \leq 1$. Appendices A.2, B, and E.2 use a common orthogonal matrix, not independent alignments of the factors.

**Lemma I.6** (Sylvester equation and polar derivative; Gawlik & Leok, 2017)**.** *If $H \succ 0$ is symmetric, then $H\Omega + \Omega H = F$ has a unique solution for every $F$. If $F$ is skew-symmetric, so is $\Omega$, and*

$$\|\Omega\|_F \le \frac{\|F\|_F}{2\lambda_{\min}(H)}.$$

*Let $X \in \mathbb{R}^{m\times h}$ have full column rank, with $m \ge h$, and write $X = UH$, where $U^\top U = I_h$ and $H = (X^\top X)^{1/2} \succ 0$. The polar map is smooth near $X$, and*

$$D\,\mathrm{polar}(X)[E] = U\Omega + (I - UU^\top)EH^{-1}, \qquad H\Omega + \Omega H = U^\top E - E^\top U.$$

*As a linear operator on matrices with the Frobenius inner product, $D\,\mathrm{polar}(X)$ is self-adjoint and positive semidefinite. Its kernel is*

$$\ker D\,\mathrm{polar}(X) = \{US : S = S^\top\}.$$

In an eigenbasis of $H$, the Sylvester equation reads $\Omega_{ij} = F_{ij}/(s_i + s_j)$, proving uniqueness and the norm bound. Differentiating $X = UH$ and $U^\top U = I_h$ gives the polar formula. Self-adjointness follows because the polar map is the gradient of the nuclear norm on the full-column-rank set. With $E_\perp = (I - UU^\top)E$, its quadratic form is

$$\langle E, D\,\mathrm{polar}(X)[E]\rangle = \tfrac{1}{2}\langle U^\top E - E^\top U, \Omega\rangle + \mathrm{tr}(E_\perp^\top E_\perp H^{-1}) \ge 0.$$

Both terms vanish exactly when $E_\perp = 0$ and $U^\top E$ is symmetric. This derivative is used in Appendices B and E; unlike the support identity in Lemma I.1, it requires full column rank.

**Lemma I.7** (Ky Fan smoothness at a strict cutoff; Bhatia, 1997; Lewis, 1996)**.** *Let $Z \in \mathbb{R}^{m\times n}$ and $1 \le h \le \min\{m,n\}$ satisfy $\sigma_h(Z) > \sigma_{h+1}(Z)$, with $\sigma_{\min\{m,n\}+1}(Z) = 0$. Then $X \mapsto \|X\|_{(h)}$ is smooth in a neighborhood of $Z$, and*

$$D\|\cdot\|_{(h)}(Z)[E] = \langle U_h V_h^\top, E\rangle, \qquad \nabla\|\cdot\|_{(h)}(Z) = \mathrm{polar}_h(Z).$$

*Repeated singular values within the selected block or within the omitted block are allowed.*

To see the role of the cutoff, form the symmetric dilation

$$\mathcal{H}(Z) = \begin{pmatrix} 0 & Z \\ Z^\top & 0 \end{pmatrix}.$$

Its top $h$ eigenvalues are $\sigma_1(Z), \dots, \sigma_h(Z)$ and form an isolated cluster. The spectral projection onto that cluster is smooth locally (Bhatia, 1997, Chapter VII). The sum of the cluster eigenvalues is therefore smooth. Its gradient with respect to $\mathcal{H}(Z)$ is that projection (Lewis, 1996, Theorem 1.1), whose off-diagonal block is $U_h V_h^\top/2$. Pairing with $\mathcal{H}(E)$ gives the formula above. Appendix A.2 uses the first derivative, and Appendix E.1 uses local second derivatives.

**Lemma I.8** (Sion's minimax theorem, compact form; Sion, 1958)**.** *Let $C$ and $P$ be nonempty compact convex subsets of finite-dimensional real vector spaces. Suppose $f : C \times P \to \mathbb{R}$ is continuous, concave in its first argument, and convex in its second. Then*

$$\max_{x\in C}\min_{p\in P} f(x,p) = \min_{p\in P}\max_{x\in C} f(x,p).$$

*Both outer extrema are attained.*

This is the continuous concave–convex case of Sion (1958, Theorem 3.4). In Appendix A.1, $C = \mathcal{C}_h$, $P = \mathcal{P}_M$, and $f(W,p) = \langle W, \Delta(p)\rangle$ is bilinear. The compact convex domain is $\mathcal{C}_h$, not the nonconvex rank-constrained set $\mathcal{B}_h$.

**Lemma I.9** (Semialgebraic projection, components, and paths; Coste, 2002)**.** *A subset of $\mathbb{R}^d$ is* semialgebraic *if it is a finite Boolean combination of polynomial equalities and inequalities. A function is semialgebraic when its graph is semialgebraic. The following facts hold:*

(i) *Coordinate projections of semialgebraic sets are semialgebraic. Finite unions, intersections, complements, and closures of semialgebraic sets are semialgebraic. Compositions of semialgebraic maps are semialgebraic.*

(ii) *Every semialgebraic set has finitely many connected components, each semialgebraic. Two points in one component can be joined by a continuous semialgebraic path.*

(iii) *Such a path admits a finite subdivision on whose open intervals it is $C^1$. For finitely many continuous semialgebraic functions on the set, the subdivision can be refined so that their compositions with the path are $C^1$ and each has constant sign or is identically zero on each open interval.*

Part (i) is the Tarski–Seidenberg property (Coste, 2002, Theorem 2.3). For components and paths, see Coste (2002, Theorem 2.23 and Exercise 3.14); finite smooth subdivisions follow from semialgebraic cell decomposition and stratification (Coste, 2002, Corollary 3.8). Sign changes can be isolated by a further finite subdivision because a semialgebraic subset of the line is a finite union of intervals and points. Appendix D.2 applies these facts to the stationary configurations. It does not assume that the softmax trajectory is semialgebraic.

**Lemma I.10** (Inverse and implicit function theorems; Krantz & Parks, 2013)**.** *Let $F : \mathcal{U} \to \mathbb{R}^b$ be $C^k$, $k \geq 1$, on an open set $\mathcal{U} \subset \mathbb{R}^a \times \mathbb{R}^b$. If $F(x_0, y_0) = 0$ and $D_y F(x_0, y_0)$ is invertible, then, near $(x_0, y_0)$, its zero set is the graph of a unique $C^k$ function $y = g(x)$, with*

$$Dg(x) = -[D_y F(x, g(x))]^{-1} D_x F(x, g(x)).$$

*In particular, a $C^k$ map between open subsets of equal-dimensional Euclidean spaces with invertible derivative at a point has a local $C^k$ inverse. Parameters may be included among the $x$ coordinates.*

Appendix E.2 applies this result in an open ambient neighborhood of the factor slice. It uses $k = 2$ for the margin chart and $k = 1$ for the fast quasi-steady state. Uniform inverse bounds there follow from

$$\|I - A\|_{\mathrm{op}} \leq q < 1 \quad \Longrightarrow \quad A^{-1} = \sum_{j=0}^{\infty} (I - A)^j, \qquad \|A^{-1}\|_{\mathrm{op}} \leq (1 - q)^{-1}.$$

This explains why an operator of the form $I + O(\varepsilon)$ stays uniformly invertible for sufficiently small $\varepsilon$.

**Lemma I.11** (Positive products and Lyapunov equations; Bhatia, 1997; Boyd, 2009)**.** *If $H, J$ are symmetric positive definite, the possibly nonsymmetric product $HJ$ has positive real eigenvalues, since*

$$J^{1/2}(HJ)J^{-1/2} = J^{1/2} H J^{1/2} \succ 0.$$

*A real matrix $A$ is* Hurwitz *if all its eigenvalues have strictly negative real parts. For Hurwitz $A$ and symmetric $Q \succ 0$, the equation*

$$A^\top X + XA = -Q$$

*has the unique solution*

$$X = \int_0^\infty e^{A^\top u} Q e^{Au}\, du \succ 0.$$

*Conversely, the existence of symmetric $X, Q \succ 0$ satisfying this equation implies that $A$ is Hurwitz.*

The similarity gives the first assertion. For Hurwitz $A$, the matrix exponential decays exponentially, so the integral converges. Differentiating its integrand gives the Lyapunov equation. Its quadratic form is positive for each nonzero vector. If $A^\top Y + YA = 0$, then $e^{A^\top u} Y e^{Au}$ is constant and tends to zero, giving uniqueness. For the converse, a complex eigenvector $Av = \lambda v$ satisfies $2\,\mathrm{Re}(\lambda) v^* X v = -v^* Q v < 0$. Appendices B and E.3 apply the result to their stable linear operators. In Appendix E.2, the positive-definite operators are restricted to $T_{\mathcal{S}}$; they need not be invertible on the full weight space.

**Lemma I.12** (Scalar comparison by an integrating factor)**.** *Let $y$ be locally absolutely continuous on $[t_0, \infty)$, and let $a, b$ be locally integrable. If $y' \leq -ay + b$ almost everywhere, then*

$$y(t) \leq y(t_0) e^{-\int_{t_0}^t a(v)\, dv} + \int_{t_0}^t e^{-\int_u^t a(v)\, dv} b(u)\, du.$$

*In particular, $b = 0$ gives $y(t) \leq y(t_0) e^{-\int_{t_0}^t a}$ without requiring $a$ to be continuous. If $a(t) = \lambda_0 > 0$ and $b(t) \leq B e^{-\mu(t - t_0)}$ with $B \geq 0$ and $\mu > 0$, then, for every $0 < \lambda < \min\{\lambda_0, \mu\}$,*

$$y(t) \leq \left( |y(t_0)| + \frac{B}{\mu - \lambda} \right) e^{-\lambda(t - t_0)}.$$

*Proof.* Multiply the differential inequality by $\exp(\int_{t_0}^{t} a)$ and integrate. For the last claim, put $T = t - t_0$ and use

$$\int_0^T e^{-\lambda_0(T-u)} e^{-\mu u}\, du \le e^{-\lambda T} \int_0^T e^{-(\mu-\lambda)u}\, du \le \frac{e^{-\lambda T}}{\mu - \lambda}.$$

This also covers $\lambda_0 = \mu$. □

Appendix C takes $a = \kappa(\eta)$ and $b = 0$. Appendix E.3 uses the forced estimate with $\mu = 4$.