OVERFITTING OF SPECTRAL GRADIENT DESCENT: HOW MATRIX GEOMETRY SHAPES GENERALIZATION AND IMPLICIT BIAS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the generalization of spectral gradient descent (SpecGD) in overparameterized matrix classification with corrupted labels. Each input combines a shared low-rank signal with a rank-one sample-specific perturbation, referred to as a shortcut, that enables memorization but does not generalize. We contrast collapsed shortcuts, which share a singular direction, with dispersed shortcuts, which occupy distinct singular directions. Changing only this geometry can reverse the relative generalization of GD and SpecGD: collapsed shortcuts can favor SpecGD, while dispersed shortcuts can favor GD. In the dispersed regime, exact shortcut orthogonality eliminates the signal from the late-stage SpecGD direction, while vanishing random correlations collectively generate a small but generalization-relevant signal through a second-order effect. To identify the direction selected by SpecGD, which the spectral max-margin problem alone does not determine, we combine a refined analysis of its dual with the exponentiated-gradient dynamics of normalized loss weights. Finally, we show that a single SpecGD step can already interpolate and generalize well, while continued training converges to a direction with substantially worse generalization.

1 INTRODUCTION

The remarkable empirical success of Muon (Jordan et al., 2024; Liu et al., 2025; AI et al., 2025) has made it one of the most prominent alternatives to standard first-order optimizers for training large neural networks. Unlike stochastic gradient descent or Adam (Kingma & Ba, 2015), which treat the parameters of a neural network as vectors, Muon explicitly exploits their matrix structure. At an idealized level, it replaces each gradient matrix by its polar factor, thereby equalizing its nonzero singular values rather than preserving their relative magnitudes.

By changing the geometry of the update, spectral gradient methods can substantially reshape training dynamics. In several tractable models, they mitigate the effects of anisotropy (Braun et al., 2026) or ill-conditioning (Ma et al., 2026), and promote more uniform learning across spectral components (Kang et al., 2026; Vasudeva et al., 2026), which may in turn favor more robust and transferable feature learning (Ruan et al., 2026). In a two-layer linear regression model, Dragutinović et al. (2026) identify a closely related phenomenon: spectral normalization suppresses the sequential simplicity bias of gradient descent by learning singular modes more uniformly. Across these works, the theoretical benefits of spectral normalization are primarily attributed to faster or more balanced learning of different components, rather than to the selection of a different predictor.

These results leave unresolved whether spectral normalization can affect generalization through the late-stage solution itself, rather than only through differences in the training trajectory. The closest work for this unresolved point is Fan et al. (2025) for separable classification, which show that SpecGD asymptotically maximizes the margin under the spectral norm, revealing an implicit bias that can differ from that of GD. However, their characterization does not in general identify the particular direction selected by the dynamics, since the spectral max-margin problem need not have a unique solution, and therefore does not determine the generalization properties of the resulting classifier. This motivates our central question: when does SpecGD select a classifier with different

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.

generalization behavior from GD, and which properties of the matrix-valued data determine which optimizer generalizes better?

To address this question, we study GD and SpecGD in an overparameterized matrix classification model with corrupted labels. Each observation contains a common class signal and a sample-specific rank-one noise, which we refer to as a *shortcut*. We vary how these shortcuts are organized, especially, mainly consider two setups: (i) shortcuts collapse along a common singular direction (CSM), and (ii) shortcuts disperse across many directions (DSM). This allows us to isolate the effect of matrix structure on generalization and to characterize regimes in which spectral normalization improves or degrades the population risk relative to GD.

1.1 Contributions & Technical novelty

Our main contributions are as follows:

- **Shortcut geometry can reverse the late-stage generalization ordering of GD and SpecGD.** We show that the organization of the shortcuts across singular directions can reverse the relative generalization performance of GD and SpecGD, measured by the population 0–1 risk in the late-stage training. In particular, while the risk of SpecGD $R^\star_{\mathrm{SpecGD}}$ is generally smaller than $R^\star_{\mathrm{GD}}$ in the collapsed shortcut model (CSM), the performance is reversed in a specific regime of shortcut strength with the dispersed shortcut model (DSM). Figure 1 shows the summary. This result implies that the spectral normalization can sometimes have a negative effect to its generalization.

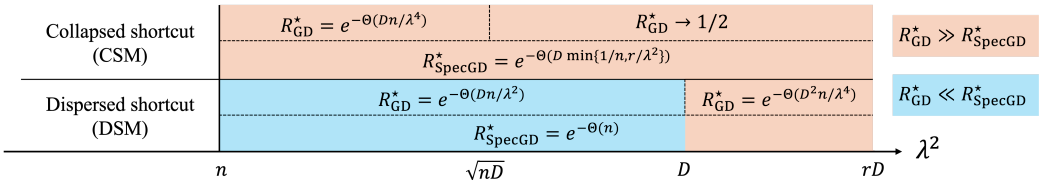

Figure 1: Late-stage test risk across shortcut geometries (CSM/DSM). Orange indicates $R^\star_{\mathrm{SpecGD}} \ll R^\star_{\mathrm{GD}}$, while blue indicates $R^\star_{\mathrm{GD}} \ll R^\star_{\mathrm{SpecGD}}$. Here n, D, r, λ denote sample size, shortcut dimension, signal rank, and shortcut strength, respectively.

- **Early-stopping avoids the deterioration of SpecGD in the late-stage.** In both shortcut geometries, we show that a single SpecGD step can already interpolate the corrupted training labels while generalizing well. The risk of SpecGD with early stopping can be smaller than $R^\star_{\mathrm{SpecGD}}$, despite already interpolating the corrupted training data. Thus, early stopping can avoid the deterioration induced by the late-stage implicit bias.
- **Weak non-orthogonality of dispersed shortcuts improves SpecGD.** In dispersed geometry, we prove that tiny correlations among dispersed shortcuts collectively generate a signal component to help generalization of SpecGD, whereas exact shortcut orthogonality eliminates the signal from the late-stage SpecGD direction. We identify that this difference is a genuinely second-order effect: a second-order expansion of the nuclear norm of the shortcut aggregate reveals curvature that is absent under exact orthogonality.

The technical novelty of our work is two-fold:

- **A dual-dynamical framework gives directional selection beyond static max-margin theory.** Recent work characterizes late-stage implicit of SpecGD in multiclass linear classification by spectral margin maximization. However, a static spectral max-margin characterization is not sufficient to identify the selected solution since the primal max-margin problem is non-unique. Consequently, it does not determine the signal coefficient that governs generalization. We develop a dual-dynamical approach that establishes directional convergence of SpecGD without requiring uniqueness of the primal problem. The key observation is that the normalized exponential-loss weights follow exponentiated-gradient dynamics on the max-margin dual, allowing convergence of the dual iterates to determine the direction selected by SpecGD.
- **Second-order geometry distinguishes exact from near orthogonality.** Under exact shortcut orthogonality, the dual objective is degenerate and fails to identify a unique dual optimizer. We show that the first nontrivial effect of random near-orthogonality appears at second order. A second-order expansion of the nuclear norm reveals a deterministic curvature term of order D^{-1} together

with random fluctuations and higher-order remainders. Sharp uniform concentration bounds show that this curvature dominates, leading to uniqueness and interiority of the dual minimizer, which are the key ingredients to establish directional convergence using the dual-dynamical framework.

1.2 Related Work

Implicit bias. On separable data, gradient descent converges in direction toward the Euclidean max-margin classifier (Soudry et al., 2018); more generally, the selected margin geometry can depend on the optimization norm (Gunasekar et al., 2018). Most closely related to our work, Fan et al. (2025) show that SpecGD exhibits spectral-norm margin maximization in multiclass linear classification. However, the associated max-margin problem need not have a unique optimizer, so this characterization does not in general identify the direction selected by SpecGD or its generalization. Using a dual-dynamical perspective related to the primal-dual analysis of GD (Ji & Telgarsky, 2021), we characterize the selected direction and its generalization despite non-uniqueness of the spectral max-margin solution.

Benign overfitting in classification. A broad literature studies when classifiers that interpolate corrupted labels nevertheless achieve small population error, particularly for high-dimensional linear max-margin classifiers (Chatterji & Long, 2021; Cao et al., 2021; Wang & Thrampoulidis, 2022; Hashimoto et al., 2026), with extensions to neural networks trained by gradient descent (Frei et al., 2023; Hashimoto, 2026). Related phenomena have also been studied in attention models and Transformers (Magen et al., 2025). We instead study matrix-linear predictors trained by SpecGD with intrinsically matrix-valued sample-specific features, showing that their organization across singular directions can determine whether interpolation is benign or catastrophic and can reverse the generalization ordering of GD and SpecGD.

Memorization features. Overparameterized neural networks can interpolate noisy or even randomly labeled training data (Zhang et al., 2017; Arpit et al., 2017), motivating theoretical work on how statistically uninformative features can be used for memorization (Bombari & Mondelli, 2024). Our model separates the shared predictive signal from sample-specific memorization features, allowing us to study how optimization exploits the latter to fit the training labels.

2 Statistical Framework

Data model. We consider binary classification with matrix-valued covariates. For each sample $i = 1, \dots, n$, the ground truth label $y_i \in \{+1, -1\}$ is drawn independently and uniformly at random. The observed label is $\tilde{y}_i = \xi_i y_i$ where ξ_i are i.i.d. random variables satisfying $\mathbb{P}(\xi_i = -1) = \rho = 1 - \mathbb{P}(\xi_i = 1)$ with $\rho \in (0, 1/2)$. The matrix-valued covariates are generated as

$$X_i = y_i S + Z_i \in \mathbb{R}^{d \times d}, \tag{2.1}$$

where $S \in \mathbb{R}^{d \times d}$ is a deterministic rank-r signal matrix, and $Z_i \in \mathbb{R}^{d \times d}$ is a random rank-one sample-specific feature, which we refer to as a *shortcut*. The component $y_i S$ carries the class information shared across samples, whereas Z_i has no predictive value for unseen data but can nevertheless be exploited by an overparameterized classifier to memorize corrupted labels. We denote the observed training sample by $\mathcal{D}_n := \{(X_i, \tilde{y}_i)\}_{i=1}^n$ and the uncorrupted and corrupted sample index sets by $I_+ = \{i \mid \xi_i = 1, i = 1, \dots, n\}$ and $I_- = \{i \mid \xi_i = -1, i = 1, \dots, n\}$.

Our model can be viewed as a matrix-valued counterpart of a signal-plus-sample-specific-noise model $\boldsymbol{x}_i = y_i \boldsymbol{\mu} + \boldsymbol{z}_i$ for vectors $\boldsymbol{x}_i, \boldsymbol{\mu}, \boldsymbol{z}_i \in \mathbb{R}^d$. This model for vectors has been widely studied in the benign-overfitting literature for binary classification (Chatterji & Long, 2021; Cao et al., 2021; Wang & Thrampoulidis, 2022; Shamir, 2022; Frei et al., 2023; Xu & Gu, 2023; Hashimoto, 2026).

Shortcut geometry. As a simple motivation, consider images with one spurious modified pixel per sample. If its location varies across the image, the shortcuts are dispersed; if it is restricted to a fixed column, they share a common right direction and collapse to a rank-one structure.

Let $S = \sum_{l \le r} \sigma_l \boldsymbol{p}_\ell \boldsymbol{q}_\ell^\top$ be the SVD of S. The polar factor of S will be denoted by $P_S := \mathrm{Polar}(S) = \sum_{\ell=1}^r \boldsymbol{p}_\ell \boldsymbol{q}_\ell^\top$. We write $s := \|S\|_1$ for the nuclear norm of S. Set $\mathcal{U} =$

$\mathrm{span}\{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_r\}^\perp$, $\mathcal{V} = \mathrm{span}\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_r\}^\perp$. Both subspaces have dimension $D = d - r$. For any subspace $\mathcal{A}$, let $\mathbb{S}(\mathcal{A})$ denote its unit sphere. Let $K \in \{1, \ldots, D\}$ and consider a subspace $\mathcal{V}_K \subset \mathcal{V}$ of dimension K. The K-dimensional shortcut model is

$$Z_i = \lambda \boldsymbol{u}_i \boldsymbol{v}_i^\top, \qquad \boldsymbol{u}_i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\big(\mathbb{S}(\mathcal{U})\big), \qquad \boldsymbol{v}_i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\big(\mathbb{S}(\mathcal{V}_K)\big), \tag{2.2}$$

where the two families are mutually independent and $\lambda > 0$ controls the shortcut strength.

Our theoretical analysis focuses on the two extreme cases $K = 1$ and $K = D$, while simulations complement the theory by illustrating the behavior in intermediate regimes $1 < K < D$. We consider the following setups:

- **Collapsed shortcut model** (CSM). We set $K = 1$: all shortcuts share a common right singular direction, so any weighted sum of the shortcut matrices remains rank one.
- **Dispersed shortcut model** (DSM). We set $K = D$: the right shortcut directions are sampled independently from the entire subspace $\mathcal{V}$, so it typically has rank n.
- **Deterministic orthogonal benchmark** (DOB). Assume $D \geq n$, and fix orthonormal families $U = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n]$ and $V = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n]$. We then set $Z_i = \lambda \boldsymbol{u}_i \boldsymbol{v}_i^\top$. The shortcuts are now exactly orthogonal, contrast to shortcuts in DSM having small correlations.

Classifier and training. We consider the matrix linear classifier $f_W(X) = \langle W, X \rangle_F$ where $W \in \mathbb{R}^{d \times d}$, and minimize the empirical exponential loss

$$\mathcal{L}(W) = \frac{1}{n} \sum_{i=1}^{n} \exp(-\tilde{y}_i \langle W, X_i \rangle_F) .$$

Gradient descent (GD) and spectral gradient descent (SpecGD) iterates are defined by

$$W_{k+1}^{\mathrm{GD}} = W_k^{\mathrm{GD}} - \eta_k \nabla \mathcal{L}\big(W_k^{\mathrm{GD}}\big) , \tag{2.3}$$

$$W_{k+1}^{\mathrm{SpecGD}} = W_k^{\mathrm{SpecGD}} - \eta_k \, \mathrm{Polar}\Big(\nabla \mathcal{L}\big(W_k^{\mathrm{SpecGD}}\big)\Big) \tag{2.4}$$

for $k = 0, 1, 2, \ldots$, with $W_0 = 0$ and stepsizes $(\eta_k)_{k \geq 0}$. Here, $\mathrm{Polar}(A) = A[(A^\top A)^\dagger]^{1/2}$, where $\dagger$ denotes the Moore-Penrose pseudoinverse. SpecGD is the idealized full-batch counterpart of Muon, in which the gradient is replaced by its polar factor.

As in Fan et al. (2025), we impose the following stepsize condition.

Assumption A1. *The stepsizes $(\eta_k)_{k \geq 0}$ satisfy $\eta_k \downarrow 0$ and $\tau_k := \sum_{t<k} \eta_t \to \infty$.*

The following condition is used to establish the rate of directional convergence for W_k^{SpecGD}. It requires the stepsize to vary slowly across successive iterations, in the sense that $\eta_{k+1} = \eta_k + o(\eta_k^2)$. In particular, polynomially decaying stepsizes $\eta_k = (1+k)^{-a}$ with $a \in (0, 1)$ satisfy the condition.

Assumption A2. *The stepsizes $(\eta_k)_{k \geq 0}$ satisfy $\lim_{k \to \infty} \eta_k^{-2}(\eta_k - \eta_{k+1}) = 0$.*

Risk. We measure interpolation using the empirical 0-1 error with respect to the observed labels,

$$\widehat{\mathcal{R}}_n(W) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left\{\tilde{y}_i \langle W, X_i \rangle_F \leq 0\right\} .$$

We say that W interpolates the training data if $\widehat{\mathcal{R}}_n(W) = 0$. For an independent clean test sample (X, y) generated according to the same data model, we define the population 0-1 error as

$$\mathcal{R}(W) = \mathbb{P}\left(y \langle W, X \rangle_F \leq 0\right) .$$

Thus, interpolation is measured with respect to the corrupted training labels, whereas generalization is measured with respect to the underlying clean labels.

Remark 1 (Late-stage interpolation)**.** *For $D \geq n$, the training sample is linearly separable almost surely under the CSM and DSM (Lemma H.1). The max-margin results of Gunasekar et al. (2018) and Fan et al. (2025) therefore imply eventual interpolation for GD and SpecGD, respectively.*

3 Main Results

We begin with an informal presentation of our main statistical finding: shortcut geometry can determine which of GD and SpecGD generalizes better. We then give a precise characterization of the late-stage direction selected by SpecGD, before comparing its generalization with GD. Finally, we investigate the first SpecGD iterate and show how early training can differ from the late-stage limit.

3.1 Shortcut geometry can reverse the ordering of GD and SpecGD

We first illustrate the geometry-dependent reversal in Figure 2. We hold all model parameters fixed and change only the shortcut geometry. Both GD and SpecGD interpolate after the first update, so we focus on test error. With collapsed shortcuts, SpecGD generalizes nearly perfectly while GD remains close to chance; under random dispersed shortcuts, the ordering reverses, and GD generalizes substantially better. Thus, reorganizing the same sample-specific shortcuts across singular directions can switch which optimizer generalizes better after interpolation.

Within the dispersed setting, exact orthogonality and random near-orthogonality lead to sharply different late-stage behavior: SpecGD is close to chance in the former, while retaining nontrivial generalization in the latter.

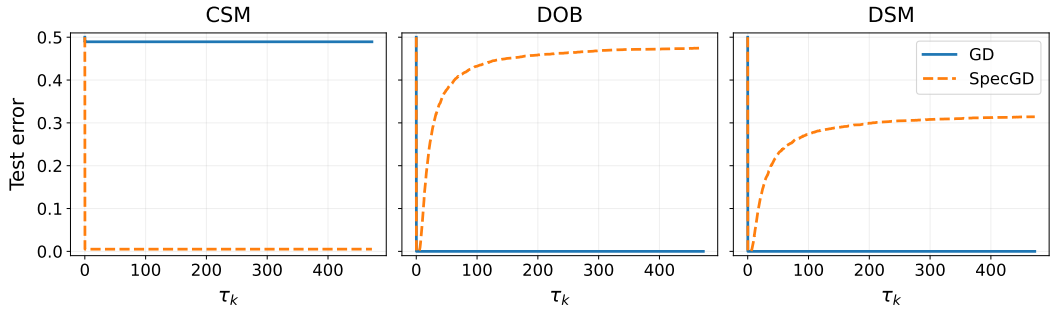

Figure 2: SpecGD generalizes for collapsed shortcuts; GD for dispersed shortcuts. Parameters: $n = 100$, $r = 1$, $D = 10^5$, $\rho = 0.45$, $\lambda = 125$, $\sigma_1 = 1$, $T = 3D$ and $\eta_k = 0.005(1+k)^{-0.1}$.

3.2 Directional convergence and Generalization of SpecGD

The two shortcut geometries lead to qualitatively different limiting directions as $k \to \infty$. We begin with the dispersed geometry (DSM and DOB), which exhibits a surprising phenomenon: exact orthogonality eliminates the signal from the convergent direction while tiny correlations under DSM generate a nonzero signal component. We then contrast this with the CSM.

Theorem 1 (Directional convergence under DSM)**.** *Assume the sample $\mathcal{D}_n$ is generated from the DSM and that Assumption A1 holds. Let $0 < \delta < e^{-1}$. There exist constants $c, C > 0$, depending only on ρ such that, if $(\log \frac{n}{\delta})^5 \le cn$, $D \ge Cn(\log \frac{n}{\delta})^4$, and $\frac{\lambda n}{sD} \le c$ then, with probability at least $1 - \delta$, the following statements hold.*

1. ***Directional convergence.*** *We have $W_k^{\mathrm{SpecGD}}/\tau_k \to W_{\mathrm{DSM}}^\star := a_{\mathrm{DSM}}^\star P_S + H_{\mathrm{DSM}}^\star$, where $\mathrm{rank}(H_{\mathrm{DSM}}^\star) = n$ and all nonzero singular values of $H_{\mathrm{DSM}}^\star$ are equal to 1.*
2. ***Signal strength.*** *Set $\Gamma_\rho := \frac{1}{8}(1-2\rho)\,[1 + 4\rho(1-\rho)]$. $a_{\mathrm{DSM}}^\star$ in 1. satisfies*

$$\left| a_{\mathrm{DSM}}^\star - \Gamma_\rho \frac{\lambda n}{sD} \right| \le C \frac{\lambda n}{sD} \left(\sqrt{\frac{\log(1/\delta)}{n}} + \frac{(\log \frac{n}{\delta})^3}{n} + \sqrt{\frac{n}{D}} \right). \tag{3.1}$$

Although $a_{\mathrm{DSM}}^\star$ is only of order $\lambda n/(sD)$, the corresponding test error is independent of D at leading order and decays exponentially with n as shown by the following corollary.

Corollary 1 (Generalization under DSM)**.** *Assume the hypotheses of Theorem 1. There exist constants $c, C > 0$, depending only on ρ such that, if $(\log \frac{n}{\delta})^5 \le cn$, $D \ge Cn(\log \frac{n}{\delta})^4$, and $\frac{\lambda n}{sD} \le c$, then, with probability at least $1 - \delta$ over the training sample $\mathcal{D}_n$, we have*

$$c_0 e^{-C_0 n} \le \mathcal{R}(W_{\mathrm{DSM}}^\star) \le 2e^{-C_1 n} \tag{3.2}$$

for a universal constant c_0 and constants $C_0, C_1 > 0$ which only depend on ρ. Moreover, $\mathcal{R}(W_k^{\mathrm{SpecGD}}) \to \mathcal{R}(W_{\mathrm{DSM}}^\star)$.

See Section 4 for a high-level description of the key proof techniques and Appendix C for details.

The absence of D from the test-error exponent is due to a matching scaling: while the signal contribution is of order n/D, the interaction of the rank-n memorization component with a fresh shortcut is of order $\sqrt{n}/D$, yielding a signal-to-noise ratio of order $\sqrt{n}$. The small signal in Theorem 1 is therefore statistically significant despite vanishing with D.

Residual shortcut correlations generate the small signal. Under DOB, exact orthogonality fixes the shortcut component of each update while the signal coefficient remains bounded. Indeed, we show

Proposition 1 (Directional convergence under DOB)**.** *Assume $|I_+| > |I_-| > 0$. Then, under DOB, W_k^{SpecGD} converges in direction to $\sum_i \tilde{y}_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$.*

The DSM and DOB spread shortcuts across n right singular directions. We now turn to the opposite extreme, the CSM, where all shortcuts share a single right singular direction.

Theorem 2 (Directional convergence under CSM)**.** *Assume that the sample $\mathcal{D}_n$ is generated from the CSM and that Assumption A1 holds. Let $0 < \delta < e^{-1}$. There exist constants $c, C > 0$, depending only on ρ, such that, if $\log \frac{n}{\delta} \le cn$ and $D \ge Cn \log \frac{n}{\delta}$ then, with probability at least $1 - \delta$ over $\mathcal{D}_n$, the following statements hold.*

1. ***Directional convergence.*** *We have $W_k^{\mathrm{SpecGD}}/\tau_k \to W_{\mathrm{CSM}}^\star := a_{\mathrm{CSM}}^\star P_S + H_{\mathrm{CSM}}^\star$, where $H_{\mathrm{CSM}}^\star = h^\star v_0^\top$ for some unit vector $h^\star$ explicitly characterized in equation* (D.11)*.*
2. ***Signal strength.*** *Set $\Gamma'_\rho := \frac{1-2\rho}{2\sqrt{\rho(1-\rho)}}$. $a_{\mathrm{CSM}}^\star$ in 1. satisfies*

$$a_{\mathrm{CSM}}^\star = \min\left\{1,\ \frac{\lambda}{s\sqrt{n}}\left(\Gamma'_\rho + \varepsilon_n\right)\right\}, \qquad |\varepsilon_n| \le C\left(\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{n}{D}}\right).$$

In contrast to the DSM, the memorization component now remains rank one and the signal coefficient does not vanish with D. For $\lambda \gg s\sqrt{n}$, $a_{\mathrm{CSM}}^\star$ saturates to one. The test error nevertheless remains small when $\lambda \ll s\sqrt{D}$, and approaches chance when $\lambda \gg s\sqrt{D}$.

Corollary 2 (Generalization under CSM)**.** *Assume the hypotheses of Theorem 2. There exist constants $c, C > 0$, depending only on ρ, such that, if $\log \frac{n}{\delta} \le cn$ and $D \ge Cn \log \frac{n}{\delta}$, then, with probability at least $1 - \delta$ over $\mathcal{D}_n$, we have*

$$c_o \exp\left(-c' D \min\left\{\frac{1}{n}, \frac{s^2}{\lambda^2}\right\}\right) \le \mathcal{R}(W_{\mathrm{CSM}}^\star) \le 2\exp\left(-cD \min\left\{\frac{1}{n}, \frac{s^2}{\lambda^2}\right\}\right)$$

where $c_0, c, c' > 0$ are constants. Moreover $\mathcal{R}\left(W_k^{\mathrm{SpecGD}}\right) \longrightarrow \mathcal{R}(W_{\mathrm{CSM}}^\star)$.

See Section D for a detailed proof.

Remark 2. *For both Theorem 1 and 2, under Assumption A2, the following convergence rate holds*

$$\left\| \frac{W_k^{\mathrm{SpecGD}}}{\tau_k} - (a^\star P_S + H^\star) \right\|_F = O\left(\frac{1 + \sum_{t<k} \eta_t^{3/2}}{\tau_k}\right). \tag{3.3}$$

The corollary reveals two regimes. For $\lambda^2 \lesssim ns^2$, the late-stage error decays as $e^{-\Theta(D/n)}$; for stronger shortcuts, it degrades as $e^{-\Theta(Ds^2/\lambda^2)}$, and approaches chance once $\lambda^2 \gg Ds^2$. Thus, unlike the DSM, late-stage generalization in the CSM depends also on D, s, λ.

We next compare with GD. The analysis is considerably simpler: the standard implicit-bias results (Soudry et al., 2018; Gunasekar et al., 2018) reduce it to the unique Frobenius max-margin classifier.

Comparison between GD and SpecGD in terms of generalization rates for representative regimes are summarized in Figure 1 and Table 1 when $\sigma_i = 1/\sqrt{r}$, i.e. $\|S\|_1 = \sqrt{r}$ while $\|S\|_F = 1$. Entry $1/2$ denotes convergence of the test error to chance, i.e. catastrophic overfitting. The most striking comparison arises in the parameter regime $n \ll \lambda^2 \ll D$, where the relative performance of SpecGD and GD is reversed by changing only the shortcut geometry.

For a more detailed analysis of GD and comparison, see Appendix E and J, respectively.

Table 1: Representataive test risks of GD and SpecGD. We set $\|S\|_F = 1$ and $\|S\|_1 = \sqrt{r}$. The displayed regimes are chosen so that one-step SpecGD improves over late-stage SpecGD.

	Parameter regime	GD (late)	SpecGD (late)	SpecGD (one step)
CSM	$n \ll \lambda^2 \ll r \wedge \sqrt{nD}$	$e^{-\Theta(Dn/\lambda^4)}$	$e^{-\Theta(D/n)}$	$e^{-\Omega(cD)}$
	$n \vee r \ll \lambda^2 \ll rn \wedge \sqrt{nD}$	$e^{-\Theta(Dn/\lambda^4)}$	$e^{-\Theta(D/n)}$	$e^{-\Omega(rD/\lambda^2)}$
	$r \vee \sqrt{nD} \ll \lambda^2 \ll rn$	$\frac{1}{2}$	$e^{-\Theta(D/n)}$	$e^{-\Omega(rD/\lambda^2)}$
DSM†	$n \ll \lambda^2 \ll D$	$e^{-\Theta(Dn/\lambda^2)}$	$e^{-\Theta(n)}$	$e^{-\Omega(\sqrt{r}\,D/\lambda)}$
	$D \ll \lambda^2 \ll D\sqrt{n}$	$e^{-\Theta(D^2n/\lambda^4)}$	$e^{-\Theta(n)}$	$e^{-\Omega(\sqrt{r}\,D/\lambda)}$
	$D\sqrt{n} \ll \lambda^2$	$\frac{1}{2}$	$e^{-\Theta(n)}$	$e^{-\Omega(\sqrt{r}\,D/\lambda)}$

† DSM rates assume $\lambda^2 \ll \frac{rD^2}{n^2}$, which satisfies a condition required for Theorem 1.

3.3 One step interpolation and generalization

We also evaluate SpecGD after a single iteration to complement our late-stage analysis. Since $W_0 = 0$, the gradient has the same form in both CSM and DSM and the signal contribution is identical.

$$-\operatorname{Polar}\left(\nabla \mathcal{L}(W_0)\right) = P_S + \operatorname{Polar}\left(\sum_{i=1}^{n} \tilde{y}_i Z_i\right).$$

The two models differ only in the aggregate shortcut component. In the collapsed model, the aggregate shortcut component has rank one, whereas in the dispersed model it typically has rank n. The following two results show that, despite arising from gradients of the same form, these different shortcut geometries lead to markedly different conditions for interpolation and generalization.

Theorem 3 (One-step interpolation and generalization under DSM)**.** *Assume the sample $\mathcal{D}_n$ is generated from the DSM. There exist universal constants $c, C > 0$ such that the following holds. Let $\delta \in (0,1)$ and suppose that $n \geq \frac{2}{(1-2\rho)^2} \log \frac{4}{\delta}$. Then, with probability at least $1-\delta$ over the training sample $\mathcal{D}_n$, the following statements hold.*

1. ***Generalization.*** *For an independent clean test sample (X, y),*

$$\mathcal{R}(W_{\mathrm{DSM},1}^{\mathrm{SpecGD}}) \leq 2\exp\left(-c\min\left\{\frac{D^2 s^2}{n\lambda^2}, \frac{Ds}{\lambda}\right\}\right).$$

2. ***Interpolation.*** *If, in addition, $\lambda > 2s$ and $D \geq Cn$, then $\widehat{\mathcal{R}}_n\left(W_1^{\mathrm{SpecGD}}\right) = 0$.*

By contrast, under CSM, the shortcut contribution is confined to a single singular direction. This changes the one-step benign-overfitting window substantially: interpolation now requires $\lambda \gtrsim s\sqrt{n}$, instead of $\lambda \gtrsim s$, while generalization holds only up to $\lambda \ll s\sqrt{D}$, instead of $\lambda \ll sD/\sqrt{n}$.

Theorem 4 (One-step interpolation and generalization under CSM)**.** *Assume the sample $\mathcal{D}_n$ is generated from the collapsed shortcut model. There exist universal constants $c, C > 0$ such that the following holds. Let $\delta \in (0,1)$ and suppose that $n \geq \frac{2}{(1-2\rho)^2} \log \frac{2}{\delta}$. Then, with probability at least $1-\delta$ over the training sample $\mathcal{D}_n$, the following statements hold.*

1. ***Generalization.*** *For an independent clean test sample (X, y),*

$$\mathcal{R}(W_{\mathrm{CSM},1}^{\mathrm{SpecGD}}) \leq 2\exp\left(-cD\min\left\{\frac{s^2}{\lambda^2}, 1\right\}\right).$$

2. ***Interpolation.*** *If, in addition, $\lambda > 2s\sqrt{n}$ and $D \geq Cn\log\frac{4n}{\delta}$, then $\widehat{\mathcal{R}}_n\left(W_1^{\mathrm{SpecGD}}\right) = 0$*

Under $\frac{\lambda n}{sD} \ll 1$, Theorem 3 and Corollary 1 imply $\mathcal{R}(W_1^{\mathrm{SpecGD}}) \ll \mathcal{R}(W^\star)$. In other words, the late-stage implicit bias of SpecGD provably fails to generalize better than early-stopped SpecGD.

The proofs of Theorem 3 and 4 are detailed in Section F and G, respectively.

4 TECHNICAL CONTRIBUTIONS: BEYOND STATIC MAX MARGIN

We focus on SpecGD under the DSM, the technically most challenging case. The CSM is simpler because all shortcuts share a common right singular direction, keeping their aggregate rank one. GD is likewise simpler, as its late-stage behavior reduces to the unique Frobenius max-margin classifier.

From margin maximization to directional selection. Directional convergence of SpecGD is essential for our generalization analysis in Section 3. While Fan et al. (2025) guarantees asymptotic margin maximization, it does not identify the direction selected by SpecGD since the primal max-margin solution may be non-unique in our overparameterized setting. Hence the margin maximization characterization does not determine the signal component governing generalization. Even uniqueness of the dual minimizer is insufficient on its own. We instead analyze the dual dynamics directly, bypassing the primal problem.

Dual-dynamical selection. Proposition B.1 gives the dual formulation

$$\gamma_\star := \max_{\|W\|_{\mathrm{op}}\le 1} \min_i \langle W, \tilde{y}_i X_i\rangle = \min_{\boldsymbol{\alpha}\in\Delta_n} F(\boldsymbol{\alpha}), \qquad F(\boldsymbol{\alpha}) := \left\| \sum_{i=1}^n \alpha_i \tilde{y}_i X_i \right\|_1 .$$

For convenience, set $Z(\boldsymbol{\alpha}) := \sum_{i=1}^n \alpha_i \tilde{y}_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, the dual shortcut aggregate. Then

$$\sum_{i=1}^n \alpha_i \tilde{y}_i X_i = \left(\sum_{i=1}^n \alpha_i \xi_i \right) S + \lambda Z(\boldsymbol{\alpha}).$$

The key observation is that the normalized exponential-loss weights

$$\alpha_{ik} := \frac{\exp\{-\tilde{y}_i \langle W_k^{\mathrm{SpecGD}}, X_i\rangle_F\}}{\sum_{j=1}^n \exp\{-\tilde{y}_j \langle W_k^{\mathrm{SpecGD}}, X_j\rangle_F\}}$$

are exactly exponentiated-gradient iterates for the dual objective F; see Lemma B.2. Hence the training dynamics induce optimization dynamics directly on the max-margin dual.

Theorem B.1 shows that if F admits a unique interior minimizer $\boldsymbol{\alpha}^\star$ satisfying the local relative growth condition

$$F(\boldsymbol{\alpha}) - F(\boldsymbol{\alpha}^\star) \ge \mu D_{\mathrm{KL}}(\boldsymbol{\alpha}^\star \| \boldsymbol{\alpha}),$$

then $\boldsymbol{\alpha}_k \to \boldsymbol{\alpha}^\star$. Convergence of the associated polar factors, together with a telescoping argument for the SpecGD iterates, then yields directional convergence. In particular, Proposition B.2 shows

$$W_k^{\mathrm{SpecGD}}/\tau_k \longrightarrow a^\star P_S + \mathrm{Polar}\big(Z(\boldsymbol{\alpha}^\star)\big).$$

The dual-dynamical formulation also gives quantitative information that is not provided by the static max-margin characterization. Under Assumption A2, we further obtain the convergence rate (3.3).

Thus, the dual dynamics identify both the direction selected by SpecGD and the convergence rate without requiring uniqueness of the primal max-margin problem.

The DSM dual problem. Proposition B.2 reduces the DSM analysis to verifying uniqueness, interiority, and local relative growth of the dual minimizer.

Using the orthogonality between the signal and shortcut blocks, the dual problem can be written as

$$\min_{\boldsymbol{\alpha}\in\Delta_n} F(\boldsymbol{\alpha}) = \min_{q\in[0,1]} \left\{ 2s \left| q - \frac{1}{2} \right| + \lambda \psi(q) \right\}, \qquad \psi(q) := \inf_{\substack{\boldsymbol{\alpha}\in\Delta_n \\ q(\boldsymbol{\alpha})=q}} \|Z(\boldsymbol{\alpha})\|_1, \tag{4.1}$$

where $q(\boldsymbol{\alpha}) := \sum_{i\in I_+} \alpha_i$ is the total dual mass assigned to uncorrupted observations. Since $\sum_{i=1}^n \alpha_i \xi_i = 2q(\boldsymbol{\alpha}) - 1$, $q(\boldsymbol{\alpha})$ directly measures the imbalance between the two groups and hence controls the signal component of the dual aggregate.

By (4.1), the dual optimization separates into a one-dimensional problem governing the signal balance q and an $(n-2)$-dimensional problem about the dual mass distribution among shortcuts.

Second-order geometry of the DSM. We need to address another difficulty for the DSM. The difficulty is that the inner optimization is degenerate under exact shortcut orthogonality; in fact, $\|Z(\boldsymbol{\alpha})\|_1 \equiv 1$, so there is no unique dual minimizer. In contrast, the random DSM shortcuts are only nearly orthogonal, and the difference appears at second order. Our expansion takes the form

$$\|Z(\boldsymbol{\alpha})\|_1 = 1 - \underbrace{\frac{1}{D}\sum_{i<j} h(\alpha_i, \alpha_j)}_{\text{deterministic curvature}} - \underbrace{\frac{1}{2}\sum_{i<j} h(\alpha_i, \alpha_j)\left(c_{ij}^2 - \frac{2}{D}\right)}_{\text{centered second-order fluctuation}} + R(\boldsymbol{\alpha}),$$

where $c_{ij} = \langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle - \langle \tilde{y}_i \boldsymbol{v}_i, \tilde{y}_j \boldsymbol{v}_j \rangle$ and $h(x, y) := \frac{xy}{x+y}$. The expansion makes clear why random near-orthogonality is fundamentally different from exact orthogonality. Under exact orthogonality, $c_{ij} = 0$, so the second-order fluctuation equals $\frac{1}{D}\sum_{i<j} h(\alpha_i, \alpha_j)$, and exactly cancels the deterministic curvature term, recovering the flat objective $\|Z(\boldsymbol{\alpha})\|_1 \equiv 1$. In the random DSM, however, c_{ij}^2 concentrates around its mean $2/D$. Hence the centered second-order fluctuation is small, and the deterministic D^{-1}-order curvature survives.

Thus, the crucial term is the deterministic curvature term. Proposition C.1 quantifies the curvature of $\|Z(\boldsymbol{\alpha})\|_1$, and Proposition C.2 uses it to identify a unique interior slice-wise optimizer $\boldsymbol{\alpha}^\star(q)$.

The main technical difficulty is to prove that this deterministic curvature survives in the random problem. This requires sharp *uniform* concentration of the centered second-order fluctuations and simultaneous control of the higher-order remainder $R(\boldsymbol{\alpha})$ over a neighborhood of the symmetric path of dual weights. Toward this end, we combine concentration inequalities for degenerate matrix U-statistics with uniform bounds on the higher-order remainder and its derivatives; see Subsection C.5.

Finally, Proposition C.3 controls the size of $\psi'(1/2)$. Combined with the cusp of $2s|q - 1/2|$, Proposition C.4 shows that the global minimizer satisfies $q(\boldsymbol{\alpha}^\star) = 1/2$. The same second-order curvature yields the local relative growth condition required by Theorem B.1; see Proposition C.5. These results verify the hypotheses of Proposition B.2 and complete the analysis for the DSM.

5 Discussion

Our results show that generalization is shaped jointly by shortcut geometry and optimizer geometry: reorganizing the same memorization features across singular directions can reverse the relative performance of GD and SpecGD. Figure 3 illustrates this transition continuously as the shortcut dimension K and strength λ vary, while additional experiments in Appendix I explore the effects of sample size, signal rank, and training time. In the dispersed setting, we further show that vanishing random correlations can qualitatively change the SpecGD limit relative to exact orthogonality. SpecGD may also interpolate and generalize after a single step while converging to a substantially worse late-stage classifier, making early stopping beneficial. On the technical side, our second-order dual analysis and dynamical selection argument may extend to other steepest-descent methods with analogous implicit-bias questions, such as SignGD. Natural extensions include higher-rank shortcuts, practical Muon variants, and nonlinear models.

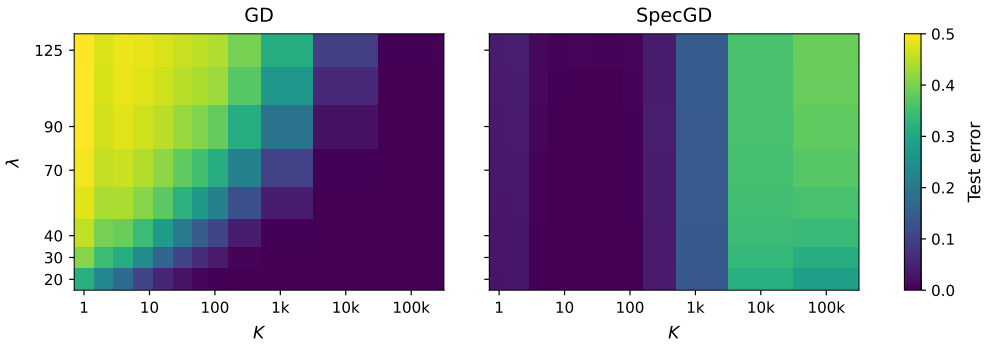

Figure 3: Test error vs. shortcut dimension K and strength λ for GD and SpecGD. Parameters: $n = 100$, $r = 1$, $D = 10^5$, $\rho = 0.45$, $||S||_F = ||S||_1 = 1$, $T = 10^4$, $\eta = 10^{-2}$.

AI USE STATEMENT

We used generative AI tools to assist with mathematical derivations and proof checking, numerical code development and debugging, manuscript editing, and exploratory discussions. All AI-assisted outputs, including mathematical claims and experiments, were independently verified by the authors, who take full responsibility for the final content of the paper.

REFERENCES

Essential AI, :, Ishaan Shah, Anthony M. Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Ashish Tanwer, Darsh J Shah, Khoi Nguyen, Kurt Smith, Michael Callahan, Michael Pust, Mohit Parmar, Peter Rushton, Platon Mazarakis, Ritvik Kapila, Saurabh Srivastava, Somanshu Singla, Tim Romanski, Yash Vanjani, and Ashish Vaswani. Practical efficiency of muon for pretraining, 2025.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 233–242. PMLR, 2017.

Simone Bombari and Marco Mondelli. How spurious features are memorized: Precise analysis for random and NTK features. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4267–4299. PMLR, 2024.

Guillaume Braun, Han Bao, Wei Huang, and Masaaki Imaizumi. Spectral gradient descent mitigates anisotropy-driven misalignment: A case study in phase retrieval. In *Forty-third International Conference on Machine Learning*, 2026.

Yuhang Cai, Kangjie Zhou, Jingfeng Wu, Song Mei, Michael Lindsey, and Peter Bartlett. Implicit bias of gradient descent for non-homogeneous deep networks. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=ruSU7xtH6v.

Yuan Cao, Quanquan Gu, and Mikhail Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8407–8418. Curran Associates, Inc., 2021.

George Casella and Roger L. Berger. *Statistical Inference*. Duxbury, 2 edition, 2002.

Niladri S. Chatterji and Philip M. Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research*, 22(129):1–30, 2021.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

Sara Dragutinović, Yedi Zhang, and Rajesh Ranganath. To use or not to use muon: How simplicity bias in optimizers matters, 2026.

Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and muon on multiclass separable data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Spencer Frei, Gal Vardi, Peter Bartlett, and Nathan Srebro. Benign overfitting in linear classifiers and leaky ReLU networks from KKT conditions for margin maximization. In *Proceedings of the Thirty-Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 3173–3228. PMLR, 2023.

Eitan Gronich and Gal Vardi. The implicit bias of adam and muon on smooth homogeneous neural networks. In *Forty-third International Conference on Machine Learning*, 2026.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1832–1841. PMLR, 10–15 Jul 2018.

Ichiro Hashimoto. Directional convergence, benign overfitting of gradient descent in leaky reLU two-layer neural networks. In *The Fourteenth International Conference on Learning Representations*, 2026.

Ichiro Hashimoto, Stanislav Volgushev, and Piotr Zwiernik. Universality of benign overfitting in binary linear classification, 2026.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17176–17186. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c76e4b2fa54f8506719a5c0dc14c2eb9-Paper.pdf.

Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato (eds.), *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pp. 772–804. PMLR, 16–19 Mar 2021. URL https://proceedings.mlr.press/v132/ji21a.html.

Jiarui Jiang, Wei Huang, Miao Zhang, Taiji Suzuki, and Liqiang Nie. Unveil benign overfitting for transformer in vision: Training dynamics, convergence, and generalization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

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

Changmin Kang, Jihun Yun, Baekrok Shin, Yeseul Cho, and Chulhee Yun. Uniform spectral growth and convergence of muon in lora-style matrix factorization, 2026.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. doi: 10.1214/aos/1015957395.

Eliahu Levy. Why do partitions occur in faa di bruno's chain rule for higher derivatives? *arXiv preprint math/0602183*, 2006.

Jichu Li, Xuan Tang, and Difan Zou. The implicit bias of steepest descent with mini-batch stochastic gradient. In *Forty-third International Conference on Machine Learning*, 2026.

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, 2025.

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, 2026.

Roey Magen, Shuning Shang, Zhiwei Xu, Spencer Frei, Wei Hu, and Gal Vardi. Benign overfitting in single-head attention. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/b889aef8f5204d22f80aa53d1a621bb4-Abstract-Conference.html.

Stanislav Minsker and Xiaohan Wei. Moment inequalities for matrix-valued U-statistics of order 2. *Electronic Journal of Probability*, 24:1–26, 2019.

Tianyu Ruan, Fengzhuo Zhang, Shuche Wang, and Shihua Zhang. Muon learns more robust and transferable features than adam, 2026.

Keitaro Sakamoto and Issei Sato. Benign overfitting in token selection of attention mechanism. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 52644–52727. PMLR, 13–19 Jul 2025.

Sholom Schechtman and Nicolas Schreuder. The late-stage training dynamics of (stochastic) subgradient descent on homogeneous neural networks. In Nika Haghtalab and Ankur Moitra (eds.), *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 5143–5172. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/schechtman25a.html.

Ohad Shamir. The implicit bias of benign overfitting. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178, pp. 448–478. PMLR, 02–05 Jul 2022.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *J. Mach. Learn. Res.*, 19(1):2822–2878, January 2018.

Gal Vardi. On the implicit bias in deep-learning algorithms, 2022. URL https://arxiv.org/abs/2208.12591.

Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How muon's spectral design benefits generalization: A study on imbalanced data. In *The Fourteenth International Conference on Learning Representations*, 2026.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2 edition, 2026.

Ke Wang and Christos Thrampoulidis. Binary classification of gaussian mixtures: Abundance of support vectors, benign overfitting, and regularization. *SIAM Journal on Mathematics of Data Science*, 4(1):260–284, 2022. doi: 10.1137/21M1415121. URL https://doi.org/10.1137/21M1415121.

Xingyu Xu and Yuantao Gu. Benign overfitting of non-smooth neural networks beyond lazy training. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 11094–11117. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/xu23k.html.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

A Additional related work

Additional implicit-bias literature. Implicit bias has been studied as a possible mechanism for understanding generalization in overparameterized models (Vardi, 2022), including margin-maximizing behavior in homogeneous and nearly homogeneous neural networks (Lyu & Li, 2020; Ji & Telgarsky, 2020; Cai et al., 2025; Schechtman & Schreuder, 2025). Recent work also studies implicit bias and optimization geometry for Muon and related spectral methods (Gronich & Vardi, 2026; Li et al., 2026; Chen et al., 2026).

Additional benign-overfitting literature. Further results on benign overfitting in classification include (Shamir, 2022; Xu & Gu, 2023), while recent work studies analogous phenomena in attention models and Transformers (Magen et al., 2025; Sakamoto & Sato, 2025; Jiang et al., 2024).

B Directional convergence via dual framework

In this section, we present our dual framework for establishing directional convergence, separating the general dual-dynamical mechanism from the model-specific arguments required to identify and control the limiting dual solution.

Throughout, set $N_i := \tilde{y}_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ and $A_i := \tilde{y}_i X_i = \xi_i S + \lambda N_i$.

B.1 Block geometry

Choose orthonormal complements $P_\perp$ and $Q_\perp$. Define $\mathcal{S} := \{PMQ^\top : M \in \mathbb{R}^{r\times r}\}$ and $\mathcal{N} := \{P_\perp N Q_\perp^\top : N \in \mathbb{R}^{D\times D}\}$. Then $S \in \mathcal{S}$, every $N_i \in \mathcal{N}$, and matrices in $\mathcal{S}$ and $\mathcal{N}$ have mutually orthogonal row and column spaces.

Lemma B.1 (Polar factor and Schatten 1-norm of orthogonal blocks)**.** *For every $c \in \mathbb{R}$ and $Z \in \mathcal{N}$, $\mathrm{Polar}(cS + \lambda Z) = \mathrm{sgn}(c) P_S + \mathrm{Polar}(Z)$, where the first term is zero when $c = 0$. Moreover, $\|cS + \lambda Z\|_1 = |c|s + \lambda \|Z\|_1$.*

Proof. Relative to the orthogonal row and column bases $[P\ P_\perp]$ and $[Q\ Q_\perp]$, the matrix is block diagonal:

$$\begin{pmatrix} c\Sigma & 0 \\ 0 & \lambda \overline{Z} \end{pmatrix}.$$

The singular values are the union of the singular values of the two blocks, which proves Schatten-1-norm additivity. Combining the compact singular value decompositions of the two blocks gives the polar-factor identity. □

B.2 Primal–dual reduction

The spectral max-margin value is $\gamma_\star := \max_{\|W\|_{\mathrm{op}}\le 1} \min_{i\in[n]} \langle W, A_i\rangle_F$.

Proposition B.1 (Exact dual and one-dimensional reduction)**.** *We have*

$$\gamma_\star = \min_{\boldsymbol{\alpha}\in\Delta_n} F(\boldsymbol{\alpha}), \qquad F(\boldsymbol{\alpha}) := \left\| \sum_{i=1}^n \alpha_i A_i \right\|_1.$$

If $q(\boldsymbol{\alpha}) := \sum_{i\in I_+} \alpha_i$, $\quad Z(\boldsymbol{\alpha}) := \sum_{i=1}^n \alpha_i N_i$, then $F(\boldsymbol{\alpha}) = 2s\left|q(\boldsymbol{\alpha}) - \frac{1}{2}\right| + \lambda \|Z(\boldsymbol{\alpha})\|_1$. Consequently,

$$\gamma_\star = \min_{q\in[0,1]} \left\{ 2s \left| q - \frac{1}{2} \right| + \lambda \psi(q) \right\}, \tag{B.1}$$

where

$$\psi(q) := \min_{\substack{\boldsymbol{\alpha}\in\Delta_n \\ q(\boldsymbol{\alpha})=q}} \|Z(\boldsymbol{\alpha})\|_1.$$

The function ψ is convex.

Proof. For every W, $\min_i \langle W, A_i\rangle_F = \min_{\boldsymbol{\alpha}\in\Delta_n} \sum_i \alpha_i \langle W, A_i\rangle_F$. Sion's minimax theorem gives $\gamma_\star = \min_{\boldsymbol{\alpha}\in\Delta_n} \max_{\|W\|_{\mathrm{op}}\le 1} \langle W, \sum_i \alpha_i A_i\rangle_F$. Spectral–Schatten-1 duality yields $\max_{\|W\|_{\mathrm{op}}\le 1} \langle W, M\rangle_F = \|M\|_1$. This proves the first identity.

Furthermore, $\sum_i \boldsymbol{\alpha}_i A_i = (2q(\boldsymbol{\alpha}) - 1)S + \lambda Z(\boldsymbol{\alpha})$, so Lemma B.1 proves the decomposition and the reduced problem.

To prove convexity of ψ, let $q_0, q_1 \in [0,1]$, choose feasible $\boldsymbol{\alpha}^{(0)}, \boldsymbol{\alpha}^{(1)}$, and let $t \in [0,1]$. Then $t\boldsymbol{\alpha}^{(1)} + (1-t)\boldsymbol{\alpha}^{(0)}$ is feasible for $tq_1 + (1-t)q_0$, and convexity of the Schatten 1-norm gives $\psi(tq_1 + (1-t)q_0) \le t\|Z(\boldsymbol{\alpha}^{(1)})\|_1 + (1-t)\|Z(\boldsymbol{\alpha}^{(0)})\|_1$. Taking infima proves convexity. □

B.3 Exact dual dynamics

Define

$$\alpha_{i,k} := \frac{\exp(-\left\langle W_k^{\mathrm{SpecGD}}, A_i \right\rangle_F)}{\sum_{j=1}^n \exp(-\left\langle W_k^{\mathrm{SpecGD}}, A_j \right\rangle_F)}. \tag{B.2}$$

Then $\boldsymbol{\alpha}_k \in \mathrm{int}(\Delta_n)$.

Set $M(\boldsymbol{\alpha}) := \sum_i \alpha_i A_i$. Since $-\nabla\mathcal{L}(W_k^{\mathrm{SpecGD}}) = \mathcal{L}(W_k^{\mathrm{SpecGD}})M(\boldsymbol{\alpha}_k)$, we have $G_k := \mathrm{Polar}(-\nabla\mathcal{L}(W_k^{\mathrm{SpecGD}})) = \mathrm{Polar}(M(\boldsymbol{\alpha}_k))$. Define $g_{i,k} := \langle G_k, A_i \rangle_F$.

Lemma B.2 (Exponentiated-gradient recursion)**.** *For every i, k, $\alpha_{i,k+1} = \frac{\alpha_{i,k}\exp(-\eta_k g_{i,k})}{\sum_j \alpha_{j,k}\exp(-\eta_k g_{j,k})}$. Moreover, $\boldsymbol{g}_k := (g_{1,k}, \ldots, g_{n,k}) \in \partial F(\boldsymbol{\alpha}_k), \quad \|g_k\|_\infty \le s + \lambda$.*

Proof. The recursion follows directly from $\exp(-\left\langle W_{k+1}^{\mathrm{SpecGD}}, A_i \right\rangle_F) = \exp(-\left\langle W_k^{\mathrm{SpecGD}}, A_i \right\rangle_F)e^{-\eta_k g_{i,k}}$.

Because $G_k = \mathrm{Polar}(M(\boldsymbol{\alpha}_k))$, we have $\|G_k\|_{\mathrm{op}} \le 1$, $\langle G_k, M(\boldsymbol{\alpha}_k)\rangle_F = \|M(\boldsymbol{\alpha}_k)\|_1$. Thus $G_k \in \partial\|\cdot\|_1(M(\boldsymbol{\alpha}_k))$. For $\boldsymbol{\beta} \in \Delta_n$, $F(\boldsymbol{\beta}) \ge F(\boldsymbol{\alpha}_k) + \sum_i(\beta_i - \alpha_{i,k})\langle G_k, A_i\rangle_F$, which proves $g_k \in \partial F(\boldsymbol{\alpha}_k)$.

Finally, $|g_{i,k}| \le \|G_k\|_{\mathrm{op}}\|A_i\|_1 \le s + \lambda$ by Lemma B.1. □

B.4 Dual convergence under local growth

For $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathrm{int}(\Delta_n)$, define $D_{\mathrm{KL}}(\boldsymbol{\alpha}\|\boldsymbol{\beta}) := \sum_i \alpha_i \log\frac{\alpha_i}{\beta_i}$.

Lemma B.3 (Local quadratic upper bound for relative entropy)**.** *Let $\boldsymbol{\alpha}^\star \in \mathrm{int}(\Delta_n)$. Then there exist constants $C_{\mathrm{KL}} > 0$ and $\varepsilon_{\mathrm{KL}} > 0$ such that $D_{\mathrm{KL}}(\boldsymbol{\alpha}^\star\|\boldsymbol{\alpha}) \le C_{\mathrm{KL}}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^\star\|_2^2$ for every $\boldsymbol{\alpha} \in \Delta_n$ satisfying $\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^\star\|_2 \le \varepsilon_{\mathrm{KL}}$.*

Proof. Let $a_\star := \min_i \alpha_i^\star > 0$, and choose $\varepsilon_{\mathrm{KL}} \le a_\star/2$. If $\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^\star\|_2 \le \varepsilon_{\mathrm{KL}}$, then, for every i,

$$\alpha_i \ge \alpha_i^\star - \|\boldsymbol{\alpha} - \boldsymbol{\alpha}^\star\|_\infty \ge \frac{\alpha_i^\star}{2}. \tag{B.3}$$

For $f_i(x) := \alpha_i^\star \log(\alpha_i^\star/x)$, we have $f_i(\alpha_i^\star) = 0$, $f_i'(\alpha_i^\star) = -1$, and $f_i''(x) = \alpha_i^\star/x^2$. By Taylor's theorem, for some $\widetilde{\alpha}_i$ between α_i and $\alpha_i^\star$,

$$f_i(\alpha_i) = -(\alpha_i - \alpha_i^\star) + \frac{1}{2}\frac{\alpha_i^\star}{\widetilde{\alpha}_i^2}(\alpha_i - \alpha_i^\star)^2.$$

By (B.3), we have $\widetilde{\alpha}_i \ge \alpha_i^\star/2$, and hence $\frac{\alpha_i^\star}{\widetilde{\alpha}_i^2} \le \frac{4}{\alpha_i^\star} \le \frac{4}{a_\star}$. Summing over i and using $\sum_i(\alpha_i - \alpha_i^\star) = 0$, we obtain

$$D_{\mathrm{KL}}(\boldsymbol{\alpha}^\star\|\boldsymbol{\alpha}) = \sum_i f_i(\alpha_i) \le \frac{2}{a_\star}\sum_i(\alpha_i - \alpha_i^\star)^2.$$

Thus the claim holds with $C_{\mathrm{KL}} = 2/a_\star$. □

Definition 1 (Local relative growth)**.** *A convex function F has local relative growth at an interior minimizer $\boldsymbol{\alpha}^\star$ if there exist $r, \mu > 0$ such that $F(\boldsymbol{\alpha}) - F(\boldsymbol{\alpha}^\star) \ge \mu D_{\mathrm{KL}}(\boldsymbol{\alpha}^\star\|\boldsymbol{\alpha})$ whenever $D_{\mathrm{KL}}(\boldsymbol{\alpha}^\star\|\boldsymbol{\alpha}) \le r$.*

Lemma B.4 (One-step relative-entropy inequality)**.** *If $\boldsymbol{\alpha}^\star$ minimizes F, then*

$$D_{\mathrm{KL}}(\boldsymbol{\alpha}^\star\|\boldsymbol{\alpha}_{k+1}) \le D_{\mathrm{KL}}(\boldsymbol{\alpha}^\star\|\boldsymbol{\alpha}_k) - \eta_k\big(F(\boldsymbol{\alpha}_k) - F(\boldsymbol{\alpha}^\star)\big) + \frac{(s+\lambda)^2}{2}\eta_k^2.$$

Proof. Let $Z_k^{\mathrm{EG}} := \sum_i \alpha_{i,k} e^{-\eta_k g_{i,k}}$. The exponentiated-gradient recursion gives

$$D_{\mathrm{KL}}(\boldsymbol{\alpha}^\star \| \boldsymbol{\alpha}_{k+1}) - D_{\mathrm{KL}}(\boldsymbol{\alpha}^\star \| \boldsymbol{\alpha}_k) = \eta_k \langle \boldsymbol{\alpha}^\star, \boldsymbol{g}_k \rangle + \log Z_k^{\mathrm{EG}}.$$

By Hoeffding's lemma, if a random variable X satisfies $X \in [a, b]$ almost surely, then $\log \mathbb{E} e^{\theta(X - \mathbb{E}X)} \leq \frac{\theta^2 (b-a)^2}{8}$. We apply this to the finite random variable X defined by $\mathbb{P}(X = g_{i,k}) = \alpha_{i,k}$. Since $|g_{i,k}| = |\langle G_k, A_i \rangle_F| \leq \|G_k\|_{\mathrm{op}} \|A_i\|_1 \leq s + \lambda$, we have $X \in [-B, B]$, where $B = s + \lambda$. Taking $\theta = -\eta_k$, Hoeffding's lemma gives $\log Z_k^{\mathrm{EG}} \leq -\eta_k \langle \boldsymbol{\alpha}_k, \boldsymbol{g}_k \rangle + \frac{(s+\lambda)^2}{2} \eta_k^2$. Thus,

$$D_{\mathrm{KL}}(\boldsymbol{\alpha}^\star \| \boldsymbol{\alpha}_{k+1}) - D_{\mathrm{KL}}(\boldsymbol{\alpha}^\star \| \boldsymbol{\alpha}_k) \leq \eta_k \langle \boldsymbol{\alpha}^\star - \boldsymbol{\alpha}_k, \boldsymbol{g}_k \rangle + \frac{(s+\lambda)^2}{2} \eta_k^2.$$

The subgradient inequality gives $F(\boldsymbol{\alpha}^\star) - F(\boldsymbol{\alpha}_k) \geq \langle \boldsymbol{g}_k, \boldsymbol{\alpha}^\star - \boldsymbol{\alpha}_k \rangle$, and the conclusion follows. □

Theorem B.1 (Dual convergence)**.** *Suppose F has a unique minimizer $\boldsymbol{\alpha}^\star \in \mathrm{int}(\Delta_n)$ and has local relative growth μ at $\boldsymbol{\alpha}^\star$. Then $\boldsymbol{\alpha}_k \to \boldsymbol{\alpha}^\star$.*

Moreover, if we further assume $\limsup_{k\to\infty} \frac{\eta_k - \eta_{k+1}}{\eta_k^2} < \mu$, then $\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}^\star\|_1 = O(\sqrt{\eta_k})$.

Proof. Write

$$d_k := D_{\mathrm{KL}}(\boldsymbol{\alpha}^\star \| \boldsymbol{\alpha}_k), \qquad \delta_k := F(\boldsymbol{\alpha}_k) - F(\boldsymbol{\alpha}^\star), \qquad C = \frac{(s+\lambda)^2}{2}.$$

Lemma B.4 gives $d_{k+1} \leq d_k - \eta_k \delta_k + C\eta_k^2$. Rearrangigng and summing from $k = 0$ to $K - 1$ gives

$$\sum_{k=0}^{K-1} \eta_k \delta_k \leq d_0 - d_K + C \sum_{k=0}^{K-1} \eta_k^2 \leq d_0 + C \sum_{k=0}^{K-1} \eta_k^2.$$

Let $\tau_K := \sum_{k=0}^{K-1} \eta_k$. Dividing by τ_K, we get

$$\frac{\sum_{k=0}^{K-1} \eta_k \delta_k}{\tau_K} \leq \frac{d_0}{\tau_K} + C \frac{\sum_{k=0}^{K-1} \eta_k^2}{\tau_K}.$$

Since $\sum_k \eta_k = \infty$, we have $\tau_K \to \infty$. Moreover, because $\eta_k \to 0$, $\frac{\sum_{k=0}^{K-1} \eta_k^2}{\sum_{k=0}^{K-1} \eta_k} \to 0$. Indeed, for any $\varepsilon > 0$, choose K_0 such that $\eta_k \leq \varepsilon$ for every $k \geq K_0$. Then $\sum_{k=0}^{K-1} \eta_k^2 \leq \sum_{k=0}^{K_0 - 1} \eta_k^2 + \varepsilon \sum_{k=K_0}^{K-1} \eta_k$. Dividing by τ_K and letting $K \to \infty$, we obtain $\limsup_{K\to\infty} \frac{\sum_{k=0}^{K-1} \eta_k^2}{\tau_K} \leq \varepsilon$. Since $\varepsilon > 0$ is arbitrary, the ratio converges to zero. Therefore, $\frac{\sum_{k=0}^{K-1} \eta_k \delta_k}{\tau_K} \to 0$. Since $\delta_k \geq 0$, there exists a subsequence $k_j \to \infty$ such that $\delta_{k_j} \to 0$. Compactness of Δ_n, uniqueness of the minimizer, and continuity of F yield $\boldsymbol{\alpha}_{k_j} \to \boldsymbol{\alpha}^\star$. Furthermore, continuity of $D_{\mathrm{KL}}(\boldsymbol{\alpha}^\star \| \cdot)$ at the interior point $\boldsymbol{\alpha}^\star$ yields $d_{k_j} \to 0$.

Let r, μ be local-growth constants. Choose a sufficiently late k_j so that $d_{k_j} < r/2$ and $C\eta_k \leq \frac{\mu r}{4}$ for every $k \geq k_j$. If $r/2 \leq d_k \leq r$, then $d_{k+1} \leq d_k - \mu \eta_k d_k + C\eta_k^2 \leq d_k - \frac{\mu r}{4} \eta_k < d_k$. If $d_k < r/2$, then sufficiently small η_k implies $d_{k+1} < r$. Thus the iterates eventually remain in the local region, where $d_{k+1} \leq (1 - \mu\eta_k) d_k + C\eta_k^2$.

For any $\varepsilon > 0$, there exists K such that $k \geq K$ implies $C\eta_k \leq \mu\varepsilon/2$ and $C\eta_k^2 \leq \varepsilon$ since $\eta_k \to 0$. If $d_k \geq \varepsilon$, then

$$\begin{aligned} d_{k+1} &\leq (1 - \mu\eta_k) d_k + C\eta_k^2 \\ &\leq d_k - \mu\eta_k \varepsilon + \mu\varepsilon\eta_k/2 \\ &= d_k - \mu\varepsilon\eta_k/2. \end{aligned}$$

Since $\sum_k \eta_k = \infty$, there exists $K_1 > K$ such that $d_{K_1} < \varepsilon$. Then, for $k \geq K_1$, $d_k < \varepsilon$ implies $d_{k+1} \leq 2\varepsilon$ due to $C\eta_k^2 \leq \varepsilon$ while $d_k \geq \varepsilon$ implies $d_{k+1} < d_k$, which by induction implies that $d_k \leq 2\varepsilon$ for $k \geq K_1$.

Hence $d_k \to 0$. Pinsker's inequality gives $\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}^\star\|_1^2 \leq 2 d_k \to 0$.

Suppose $\limsup_{k\to\infty} \frac{\eta_k-\eta_{k+1}}{\eta_k^2} < \mu$. Fix μ_0 so that $\limsup_{k\to\infty} \frac{\eta_k-\eta_{k+1}}{\eta_k^2} < \mu_0 < \mu$. Let K so that $k \geq K$ implies $d_{k+1} \leq (1-\mu\eta_k)d_k + C\eta_k^2$ and $\frac{\eta_k-\eta_{k+1}}{\eta_k^2} \leq \mu_0$. Set

$$B := \max\left\{\frac{d_K}{\eta_K}, \frac{C}{\mu-\mu_0}\right\}.$$

We now claim that $d_k \leq B\eta_k$ holds for all $k \geq K$. The claim holds at $k = K$ by the definition of B. Now suppose $d_k \leq B\eta_k$ holds. By noting that the definition of B implies that $B\mu - C \geq B\mu_0$, we have

$$\begin{aligned} d_{k+1} &\leq (1-\mu\eta_k)d_k + C\eta_k^2 \\ &\leq B(1-\mu\eta_k)\eta_k + C\eta_k^2 = B\eta_k - (B\mu - C)\eta_k^2 \\ &\leq B\eta_k - B\mu_0\eta_k^2 = B\eta_{k+1} + B(\eta_k - \eta_{k+1} - \mu_0\eta_k^2) \leq B\eta_{k+1}. \end{aligned}$$

Again, by Pinsker's inequality, we obtain $\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}^\star\|_1^2 = O(\eta_k)$. □

B.5 Which primal optimizer is selected?

Let $c_k := \sum_i \alpha_{i,k}\xi_i$, $Z_k := \sum_i \alpha_{i,k}N_i$. By Lemma B.1, $G_k = \sigma_k P_S + H_k$, $\sigma_k := \operatorname{sgn}(c_k)$, $H_k := \operatorname{Polar}(Z_k)$. Therefore,

$$W_k^{\mathrm{SpecGD}} = \left(\sum_{t<k}\eta_t\sigma_t\right)P_S + \sum_{t<k}\eta_t H_t. \tag{B.4}$$

Let $\boldsymbol{\alpha}^\star$ be the unique interior minimizer. Set $Z^\star := \sum_i \alpha_i^\star N_i$ and $H^\star := \operatorname{Polar}(Z^\star)$. Under DSM, since $D > n$, $U = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n]$ and $\widehat{V} = [\widehat{\boldsymbol{v}}_1, \ldots, \widehat{\boldsymbol{v}}_n]$ have full column rank almost surely. Since all $\alpha_i^\star > 0$, $Z^\star$ has rank n on its row and column spans. Lemma H.12 and H.11 implies that the polar factor is continuous at $Z^\star$. Thus, $\boldsymbol{\alpha}_k \to \boldsymbol{\alpha}^\star$ implies $H_k \to H^\star$. Under CSM, the convergence $H_k \to H^\star$ easily follows by noting $\operatorname{Polar}(Z(\boldsymbol{\alpha})) = \frac{\sum_i \alpha_i \boldsymbol{u_i}}{\|\sum_i \alpha_i \boldsymbol{u_i}\|_2}\boldsymbol{v}_0^\top$ and that $\sum_i \alpha_i \boldsymbol{u_i} \neq 0$ since $\boldsymbol{\alpha} \neq \mathbf{0}$.

Define $\Lambda_k := \frac{1}{n_+}\sum_{i\in I_+}\log\alpha_{i,k} - \frac{1}{n_-}\sum_{i\in I_-}\log\alpha_{i,k}$, and

$$\overline{h}_{+,k} := \frac{1}{n_+}\sum_{i\in I_+}\langle H_k, N_i\rangle_F, \qquad \overline{h}_{-,k} := \frac{1}{n_-}\sum_{i\in I_-}\langle H_k, N_i\rangle_F.$$

Lemma B.5 (Telescoping identity)**.** *For every k,*

$$\Lambda_{k+1} - \Lambda_k = -\eta_k\left[2s\sigma_k + \lambda(\overline{h}_{+,k} - \overline{h}_{-,k})\right].$$

Proof. From Lemma B.2, $\log\alpha_{i,k+1} = \log\alpha_{i,k} - \eta_k g_{i,k} - \log Z_k^{\mathrm{EG}}$. The normalizing term cancels between the two group averages. Moreover, $g_{i,k} = \langle\sigma_k P_S + H_k, \xi_i S + \lambda N_i\rangle_F = s\sigma_k\xi_i + \lambda\langle H_k, N_i\rangle_F$. Substitution proves the identity. □

Proposition B.2 (Dynamical selection formula)**.** *Suppose Assumption A1. If the unique dual minimizer is interior and F has local relative growth there, then $\frac{W_k^{\mathrm{SpecGD}}}{\tau_k} \longrightarrow a^\star P_S + H^\star$, where*

$$a^\star = -\frac{\lambda}{2s}(h_+^\star - h_-^\star), \tag{B.5}$$

with

$$h_+^\star := \frac{1}{n_+}\sum_{i\in I_+}\langle H^\star, N_i\rangle_F, \qquad h_-^\star := \frac{1}{n_-}\sum_{i\in I_-}\langle H^\star, N_i\rangle_F.$$

If we further assume Assumption A2, then we have

$$\left\|\frac{W_k^{\mathrm{SpecGD}}}{\tau_k} - (a^\star P_S + H^\star)\right\|_F = O\left(\frac{1+\sum_{t<k}\eta_t^{3/2}}{\tau_k}\right)$$

Proof. Theorem B.1 gives $\boldsymbol{\alpha}_k \to \boldsymbol{\alpha}^\star$. Hence Λ_k converges to a finite limit and $H_k \to H^\star$. Summing Lemma B.5, dividing by τ_k, and applying the weighted Cesàro theorem gives $\frac{\sum_{t<k}\eta_t\sigma_t}{\tau_k} \to -\frac{\lambda}{2s}(h_+^\star - h_-^\star)$. Dividing (B.4) by τ_k and applying weighted Cesàro convergence to H_t proves the result.

Suppose $\limsup_{k\to\infty} \frac{\eta_k - \eta_{k+1}}{\eta_k^2} = 0$. Then, by Theorem B.1 we have $\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}^\star\|_2 = O(\sqrt{\eta_k})$.

Since $\boldsymbol{\alpha} \to \mathrm{Polar}(Z(\boldsymbol{\alpha}))$ is Lipschitz continuous at $\boldsymbol{\alpha}^\star$, we have $\|H_k - H^\star\|_F = O(\sqrt{\eta_k})$. Then for the shortcut noise, we have

$$\left\| \frac{\sum_{t<k}\eta_t H_t}{\tau_k} - H^\star \right\|_F = O\left(\frac{\sum_{t<k}\eta_t^{3/2}}{\tau_k} \right).$$

Since $|\bar{h}_{+,k} - h_+^\star|, |\bar{h}_{+,k} - h_+^\star| = O(\|H_k - H^\star\|) = O(\sqrt{\eta_k})$, summing Lemma B.5 and dividing by τ_k implies that

$$\left| \frac{\sum_{t<k}\eta_t\sigma_t}{\tau_k} - \frac{\lambda}{2s}(h_-^\star - h_+^\star) \right| = O\left(\frac{1 + \sum_{t<k}\eta_t^{3/2}}{\tau_k} \right).$$

This gives the convergence rate for the signal direction. Combining these estimates proves the desired convergence rate. □

C Proof of Theorem 1

By Proposition B.2, the proof of Theorem 1 is reduced to the uniqueness, interiority of the dual minimizer $\boldsymbol{\alpha}^\star$ and relative local growth at $\boldsymbol{\alpha}^\star$. Under DSM, the key ingredient is the second-order expansion of $\|Z(\boldsymbol{\alpha})\|_1$.

Throughout, set $\hat{\boldsymbol{v}}_i := \tilde{y}_i \boldsymbol{v}_i$, so that $N_i = \boldsymbol{u}_i \hat{\boldsymbol{v}}_i^\top$.

C.1 Second-order expansion

This section derives the analytic expansion that turns the nuclear norm of the noise combination into a tractable quadratic function of the weights. Recall that $Z(\boldsymbol{\alpha}) = \sum_i \alpha_i \boldsymbol{u}_i \widehat{\boldsymbol{v}}_i^\top$. Since $\|\boldsymbol{u}_i \widehat{\boldsymbol{v}}_i^\top\|_1 = 1$, we have $\|Z(\boldsymbol{\alpha})\|_1 \le \sum_i \alpha_i = 1$, where the equality holds when the noise atoms are perfectly orthogonal.

In the random model the atoms are only approximately orthogonal. The goal of this section is to establish a second-order approximation of $\|Z(\boldsymbol{\alpha})\|_1$, which enables us to obtain a tight estimate of the noise norm gap $1 - \|Z(\boldsymbol{\alpha})\|_1$.

Write

$$U = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_n], \qquad \widehat{V} = [\widehat{\boldsymbol{v}}_1, \dots, \widehat{\boldsymbol{v}}_n], \qquad \Lambda_{\boldsymbol{\alpha}} := \mathrm{diag}(\alpha_1, \dots, \alpha_n). \tag{C.1}$$

Then $Z(\boldsymbol{\alpha}) = U\Lambda_{\boldsymbol{\alpha}}\widehat{V}^\top$. Also set

$$G_U := U^\top U = I + \Delta_U, \qquad G_V := \widehat{V}^\top \widehat{V} = I + \Delta_V, \qquad Y(\boldsymbol{\alpha}) := G_U^{1/2} \Lambda_{\boldsymbol{\alpha}} G_V^{1/2}. \tag{C.2}$$

The matrices Δ_U, Δ_V are symmetric with zero diagonal.

Throughout this section, we write

$$\mathsf{N}(Y) := \frac{1}{n}\|Y\|_1 \tag{C.3}$$

for the normalized Schatten 1-norm functional on $n \times n$ matrices. On nonsingular matrices we use the factorization

$$\mathsf{N} = \Psi \circ G, \qquad G(Y) := Y^\top Y, \qquad \Psi(M) := \frac{1}{n}\,\mathrm{tr}(M^{1/2}). \tag{C.4}$$

Lemma C.1 (Uniform derivatives of the matrix square-root map)**.** *Fix constants $0 < \ell < u < \infty$ and an integer $r \ge 1$. There exists a constant $C_r = C_r(\ell, u) < \infty$ such that the following holds in every dimension n.*

Let $A \in \mathbb{R}^{n\times n}$ be symmetric positive definite with $\ell I \preceq A \preceq uI$. Then the map $A \mapsto A^{1/2}$ is C^∞ in a neighborhood of A, and for every $1 \le q \le r$, $\left\|\mathrm{d}^q(A^{1/2})[H_1,\dots,H_q]\right\|_{\mathrm{op}} \le C_r \prod_{j=1}^q \|H_j\|_{\mathrm{op}}$. The constant C_r depends only on ℓ, u, r, and not on n.

Proof. Choose once and for all a positively oriented contour $\Gamma = \Gamma_{\ell,u}$ contained in the right half-plane, enclosing $[\ell, u]$, and satisfying $\operatorname{dist}(\Gamma, [\ell, u]) \ge \delta_{\ell,u} > 0$. The length of Γ, the quantity $\sup_{z\in\Gamma} |z^{1/2}|$, and $\delta_{\ell,u}^{-1}$ depend only on ℓ, u. By the holomorphic functional calculus, $A^{1/2} = \frac{1}{2\pi i}\int_\Gamma z^{1/2}(zI-A)^{-1}\,\mathrm{d}z$. Set $R_z(A) := (zI-A)^{-1}$. Since $z \in \Gamma$ stays at distance at least $\delta_{\ell,u}$ from the spectrum of A, $\|R_z(A)\|_{\mathrm{op}} \le \delta_{\ell,u}^{-1}$. The resolvent derivative is $\mathrm{d}R_z(A)[H] = R_z(A)HR_z(A)$. Iterating this identity gives

$$\mathrm{d}^q(A^{1/2})[H_1,\dots,H_q] = \frac{1}{2\pi i}\int_\Gamma z^{1/2} \sum_{\pi\in S_q} R_z(A)H_{\pi(1)}R_z(A)\cdots H_{\pi(q)}R_z(A)\,\mathrm{d}z.$$

Therefore,

$$\left\|\mathrm{d}^q(A^{1/2})[H_1,\dots,H_q]\right\|_{\mathrm{op}} \le \frac{\operatorname{length}(\Gamma)}{2\pi} \sup_{z\in\Gamma} |z^{1/2}|\, q!\, \delta_{\ell,u}^{-(q+1)} \prod_{j=1}^q \|H_j\|_{\mathrm{op}}.$$

The right-hand side depends only on ℓ, u, q, not on n. Taking the maximum over $1 \le q \le r$ proves the claim. □

The following is the standard partition form of Faà di Bruno's formula for Fréchet derivatives; see, for example, Levy (2006).

Lemma C.2 (Faà di Bruno formula for Fréchet derivatives)**.** *Let X, Y, Z be finite-dimensional normed vector spaces. Let $U \subset X$ and $V \subset Y$ be open sets, and let $G : U \to V, \quad F : V \to Z$ be C^q. Let $\mathcal{P}_q$ denote the set of all partitions of $\{1,\dots,q\}$. For a block $B \subset \{1,\dots,q\}$, define $G_B(x) := \mathrm{d}^{|B|}G(x)[h_i : i \in B]$, where the directions are taken in increasing order of their indices. Then*

$$\mathrm{d}^q(F\circ G)(x)[h_1,\dots,h_q] = \sum_{\pi\in\mathcal{P}_q} \mathrm{d}^{|\pi|}F(G(x))\left[G_{B_1}(x),\dots,G_{B_m}(x)\right],$$

where $\pi = \{B_1,\dots,B_m\}$ and $|\pi| = m$. Since the higher derivatives are symmetric, the order of the blocks and the order inside each block are immaterial.

Equivalently, the q-th derivative of a composition is obtained by summing over all ways of grouping the directions $\{h_1,\dots,h_q\}$: each block B is first inserted into the derivative $\mathrm{d}^{|B|}G(x)$, and the resulting outputs are then inserted into the derivative of F whose order is the number of blocks.

Lemma C.3 (Composition of uniformly bounded derivatives)**.** *Let X, Y, Z be normed vector spaces, and let $G : U \subset X \to Y, \quad F : V \subset Y \to Z$ be C^r maps with $G(U) \subset V$. Suppose that for every $1 \le m \le r$, there exist constants $A_m, B_m < \infty$ such that $\|\mathrm{d}^m F(y)[u_1,\dots,u_m]\|_Z \le A_m \prod_{j=1}^m \|u_j\|_Y$ for all $y \in V$, and $\|\mathrm{d}^m G(x)[h_1,\dots,h_m]\|_Y \le B_m \prod_{j=1}^m \|h_j\|_X$ for all $x \in U$. Then, for every $1 \le q \le r$, $\|\mathrm{d}^q(F\circ G)(x)[h_1,\dots,h_q]\|_Z \le C_q \prod_{j=1}^q \|h_j\|_X$, where one may take $C_q := \sum_{\pi\in\mathcal{P}_q} A_{|\pi|}\prod_{B\in\pi} B_{|B|}$. In particular, C_q depends only on $q, \quad A_1,\dots,A_q, \quad B_1,\dots,B_q$.*

Proof. By Faà di Bruno's formula,

$$\mathrm{d}^q(F\circ G)(x)[h_1,\dots,h_q] = \sum_{\pi\in\mathcal{P}_q} \mathrm{d}^{|\pi|}F(G(x))\left[G_{B_1}(x),\dots,G_{B_m}(x)\right],$$

where $\pi = \{B_1,\dots,B_m\}$, $|\pi| = m$, and $G_B(x) := \mathrm{d}^{|B|}G(x)[h_i : i \in B]$. For a fixed partition π, the assumptions give

$$\left\|\mathrm{d}^{|\pi|}F(G(x))\left[G_{B_1}(x),\dots,G_{B_m}(x)\right]\right\|_Z \le A_{|\pi|}\prod_{B\in\pi}\|G_B(x)\|_Y.$$

Moreover, $\|G_B(x)\|_Y \le B_{|B|}\prod_{i\in B}\|h_i\|_X$. Hence the contribution of π is bounded by $A_{|\pi|}\prod_{B\in\pi} B_{|B|}\prod_{j=1}^q \|h_j\|_X$. Summing over all partitions gives the claim. □

Lemma C.4 (Dimension-uniform derivatives of the normalized Schatten 1-norm)**.** *Fix constants $0 < a < b < \infty$ and an integer $r \geq 1$. There exists a constant $C_r < \infty$, depending only on a, b, r, such that the following holds for every dimension n.*

Let $Y \in \mathbb{R}^{n\times n}$ satisfy $a \leq \sigma_{\min}(Y) \leq \|Y\|_{\mathrm{op}} \leq b$. For N defined in (C.3)*, equivalently $\mathsf{N}(Y) = \frac{1}{n}\operatorname{tr}\big((Y^\top Y)^{1/2}\big)$, N is C^∞ in a neighborhood of Y, and for every $1 \leq q \leq r$, $|\mathrm{d}^q\mathsf{N}(Y)[E_1, \ldots, E_q]| \leq C_r \prod_{j=1}^q \|E_j\|_{\mathrm{op}}$. In particular, the constant is independent of n.*

Proof. Recall the factorization (C.4). On the set of symmetric positive-definite matrices whose spectra lie in $[a^2, b^2]$, Lemma C.1 implies that, for each $1 \leq m \leq r$, there is a constant $K_m(a, b)$, independent of n, such that $\big\|\mathrm{d}^m(M^{1/2})[H_1, \ldots, H_m]\big\|_{\mathrm{op}} \leq K_m(a, b) \prod_{j=1}^m \|H_j\|_{\mathrm{op}}$. Since $\frac{1}{n}|\operatorname{tr}(L)| \leq \|L\|_{\mathrm{op}}$, it follows that $|\mathrm{d}^m\Psi(M)[H_1, \ldots, H_m]| \leq K_m(a, b) \prod_{j=1}^m \|H_j\|_{\mathrm{op}}$.

By (C.4), it remains to control the derivatives of $G(Y) = Y^\top Y$. On the region $\|Y\|_{\mathrm{op}} \leq b$, the map G satisfies $\mathrm{d}G(Y)[E] = Y^\top E + E^\top Y, \quad \|\mathrm{d}G(Y)[E]\|_{\mathrm{op}} \leq 2b\|E\|_{\mathrm{op}}$, and

$$\mathrm{d}^2G(Y)[E_1, E_2] = E_1^\top E_2 + E_2^\top E_1, \qquad \|\mathrm{d}^2G(Y)[E_1, E_2]\|_{\mathrm{op}} \leq 2\|E_1\|_{\mathrm{op}}\|E_2\|_{\mathrm{op}}.$$

Moreover, $\mathrm{d}^m G = 0$ for $m \geq 3$. Therefore the derivatives of G up to order r are bounded by constants depending only on b and the derivative order. Applying Lemma C.3 to the factorization (C.4) gives the desired bound. All constants depend only on a, b, r, and not on n. □

Lemma C.5 (Rank-sensitive derivatives of the normalized Schatten 1-norm)**.** *Fix constants $0 < a < b < \infty$ and an integer $r \geq 1$. Consider N, defined in* (C.3)*, on the set of matrices satisfying $a \leq \sigma_{\min}(Y) \leq \|Y\|_{\mathrm{op}} \leq b$. For every $1 \leq q \leq r$, there is a constant $C_q = C_q(a, b) < \infty$, independent of n, such that the following bounds hold.*

First, if one direction, say E_1, is measured in Schatten 1-norm and the remaining directions are measured in operator norm, then

$$|\mathrm{d}^q\mathsf{N}(Y)[E_1, E_2, \ldots, E_q]| \leq \frac{C_q}{n}\|E_1\|_1 \prod_{j=2}^q \|E_j\|_{\mathrm{op}}.$$

Second, if two directions are measured in Frobenius norm and the remaining directions are measured in operator norm, then

$$|\mathrm{d}^q\mathsf{N}(Y)[E_1, E_2, E_3, \ldots, E_q]| \leq \frac{C_q}{n}\|E_1\|_F\|E_2\|_F \prod_{j=3}^q \|E_j\|_{\mathrm{op}}.$$

By symmetry of Fréchet derivatives, the distinguished directions may appear in any slots.

Proof. We use the factorization (C.4). For M whose spectrum lies in $[a^2, b^2]$, the holomorphic functional calculus gives $M^{1/2} = \frac{1}{2\pi i}\int_\Gamma z^{1/2}(zI - M)^{-1}\,\mathrm{d}z$, where the contour Γ and its distance from $[a^2, b^2]$ depend only on a, b. Hence every fixed-order derivative of Ψ is a finite sum of contour integrals of normalized traces of products of uniformly bounded resolvents and inserted directions. More precisely,

$$\mathrm{d}^m\Psi(M)[H_1, \ldots, H_m] = \frac{1}{2\pi i}\int_\Gamma z^{1/2} \sum_{\pi\in S_m} \frac{1}{n}\operatorname{tr}\left(R_z H_{\pi(1)} R_z \cdots H_{\pi(m)} R_z\right)\mathrm{d}z.$$

Recall that the resolvent factors $R_z = (zI - M)^{-1}$ are uniformly bounded in operator norm.

If one inserted direction, say H_1, is controlled in Schatten 1-norm, then cyclicity of trace and $|\operatorname{tr}(AB)| \leq \|A\|_1\|B\|_{\mathrm{op}}$ give $|\mathrm{d}^m\Psi(M)[H_1, H_2, \ldots, H_m]| \leq \frac{K_m}{n}\|H_1\|_1 \prod_{j=2}^m \|H_j\|_{\mathrm{op}}$. Similarly, if two inserted directions H_1, H_2 are controlled in Frobenius norm, then cyclicity of trace allows us to split each cyclic product at those two factors. More explicitly, after cyclically rotating the trace, each summand can be written as $\operatorname{tr}(AH_1BH_2)$, where the factors in A and B are resolvents and the remaining inserted directions. The resolvents are uniformly bounded in operator norm, and the remaining inserted directions are measured in operator norm. Hence

$$\|AH_1\|_F \leq K_m\|H_1\|_F \prod_{j\in J_1}\|H_j\|_{\mathrm{op}}, \qquad \|BH_2\|_F \leq K_m\|H_2\|_F \prod_{j\in J_2}\|H_j\|_{\mathrm{op}},$$

where $J_1 \cup J_2 = \{3, \dots, m\}$. Applying $|\operatorname{tr}(AB)| \le \|A\|_F \|B\|_F$ to the product $(AH_1)(BH_2)$ gives

$$|\mathrm{d}^m \Psi(M)[H_1, H_2, H_3, \dots, H_m]| \le \frac{K_m}{n} \|H_1\|_F \|H_2\|_F \prod_{j=3}^{m} \|H_j\|_{\mathrm{op}}.$$

It remains to pass these estimates from Ψ to N using the factorization (C.4). The derivatives of $G(Y) = Y^\top Y$ are $\mathrm{d}G(Y)[E] = Y^\top E + E^\top Y$, $\mathrm{d}^2 G(Y)[E, F] = E^\top F + F^\top E$, and all higher derivatives vanish. Multiplication by matrices with operator norm bounded by b preserves the above norm classes in the following sense: $\|AEB\|_1 \le \|A\|_{\mathrm{op}} \|E\|_1 \|B\|_{\mathrm{op}}$, $\|AEB\|_F \le \|A\|_{\mathrm{op}} \|E\|_F \|B\|_{\mathrm{op}}$, and

$$\|E^\top F\|_1 \le \|E\|_F \|F\|_F, \qquad \|E^\top F\|_1 \le \|E\|_1 \|F\|_{\mathrm{op}}, \qquad \|E^\top F\|_F \le \|E\|_{\mathrm{op}} \|F\|_F.$$

More precisely, if one direction E_1 is measured in Schatten 1-norm and the remaining directions are measured in operator norm, then the direction E_1 enters exactly one argument passed to a derivative of Ψ since each input direction belongs to exactly one block in the Faà di Bruno expansion. Then, every Faà di Bruno term can be bounded in the desired form by noticing $\|\mathrm{d}G(Y)[E_1]\|_1 \le 2b\|E_1\|_1$, $\|\mathrm{d}^2 G(Y)[E_1, E_i]\|_1 \le 2\|E_i\|_{\mathrm{op}} \|E_1\|_1$.

Similarly, if two directions E_1, E_2 are measured in Frobenius norm, and the remaining directions are measured in operator norm, every Faà di Bruno term can be bounded in the desired form by noticing

$$\|\mathrm{d}G(Y)[E_a]\|_F \le 2b\|E_a\|_F, \qquad \|\mathrm{d}^2 G(Y)[E_a, E_i]\|_F \le 2\|E_i\|_{\mathrm{op}} \|E_a\|_F, \qquad a \in \{1, 2\},$$

$\|\mathrm{d}^2 G(Y)[E_1, E_2]\|_1 \le 2\|E_1\|_F \|E_2\|_F$.

Thus, in every nonzero Faà di Bruno term for $\Psi \circ G$, a Schatten 1 controlled input direction produces an argument of Ψ with Schatten 1-norm control, and two Frobenius-controlled input directions either produce two Frobenius-controlled arguments of Ψ or one Schatten-1-controlled argument through $\mathrm{d}^2 G$. All other arguments are controlled in operator norm. Applying the preceding bounds for Ψ to each partition term and summing over the finitely many partitions proves the claim. □

Lemma C.6 (Diagonal perturbation estimates for the normalized remainder)**.** *Fix constants $0 < c < C < \infty$. There exist constants $\varrho_0 = \varrho_0(c, C) > 0$ and $K = K(c, C) < \infty$, independent of n, such that the following holds.*

Let $\widetilde{\Lambda}$ be a positive diagonal $n \times n$ matrix whose entries lie in $[c, C]$. Let $L_t = (I + t\Delta_U)^{1/2}$, $R_t = (I + t\Delta_V)^{1/2}$, $0 \le t \le 1$, where $\|\Delta_U\|_{\mathrm{op}} \le \varrho$, $\|\Delta_V\|_{\mathrm{op}} \le \varrho$, $0 < \varrho \le \varrho_0$. For diagonal matrices W_1, W_2, define $\varphi_{W_1,W_2}(t, s, r) := \mathsf{N}\big(L_t(\widetilde{\Lambda} + sW_1 + rW_2)R_t\big)$ and $\mathcal{R}_{W_1,W_2}(s, r) := \frac{1}{2}\int_0^1 (1-t)^2 \partial_t^3 \varphi_{W_1,W_2}(t, s, r)\,\mathrm{d}t$. Then $|\mathcal{R}_{0,0}(0, 0)| \le K\varrho^3$. Moreover, for every diagonal matrix W, $|\partial_s \mathcal{R}_{W,0}(0, 0)| \le K\varrho^3 \frac{\|W\|_1}{n}$, and, for every pair of diagonal matrices W_1, W_2, $|\partial_s \partial_r \mathcal{R}_{W_1,W_2}(0, 0)| \le K\varrho^3 \frac{\|W_1\|_F \|W_2\|_F}{n}$. Finally, the mixed Schatten-1/operator estimate

$$|\partial_s \partial_r \mathcal{R}_{W_1,W_2}(0, 0)| \le K\varrho^3 \frac{\|W_1\|_1 \|W_2\|_{\mathrm{op}}}{n} \tag{C.5}$$

also holds. By symmetry, the same estimate holds with W_1 and W_2 interchanged.

Proof. Choose $\varrho_0 \in (0, 1)$ sufficiently small. Since the entries of $\widetilde{\Lambda}$ lie in $[c, C]$, the singular values of $L_t \widetilde{\Lambda} R_t$ lie in a fixed interval $[a, b] \subset (0, \infty)$, depending only on c, C, uniformly over $0 \le t \le 1$. Thus the derivative estimates from Lemmas C.4 and C.5 apply uniformly along this path.

We first record the size of the matrix directions inserted into the Fréchet derivatives of N. By Lemma C.1, for every fixed $m \ge 0$, $\|\partial_t^m L_t\|_{\mathrm{op}} \le C_m \varrho^m$, $\|\partial_t^m R_t\|_{\mathrm{op}} \le C_m \varrho^m$. Hence, by the product rule,

$$\partial_t^m (L_t \widetilde{\Lambda} R_t) = \sum_{a+b=m} \binom{m}{a} (\partial_t^a L_t) \widetilde{\Lambda} (\partial_t^b R_t),$$

and therefore, for $m = 1, 2, 3$,

$$\|\partial_t^m (L_t \widetilde{\Lambda} R_t)\|_{\mathrm{op}} \le C_m \varrho^m. \tag{C.6}$$

Similarly, if $A_W(t,s) := L_t(\widetilde{\Lambda}+sW)R_t$, then $\partial_s A_W(t,s) = L_t W R_t$. Thus, for every fixed $a \geq 0$, $\partial_t^a \partial_s A_W(t,0) = \sum_{p=0}^{a} \binom{a}{p} (\partial_t^p L_t) W (\partial_t^{a-p} R_t)$. Consequently,

$$\|\partial_t^a \partial_s A_W(t,0)\|_1 \leq C_a \varrho^a \|W\|_1, \tag{C.7}$$

$$\|\partial_t^a \partial_s A_W(t,0)\|_{\mathrm{op}} \leq C_a \varrho^a \|W\|_{\mathrm{op}}, \tag{C.8}$$

and

$$\|\partial_t^a \partial_s A_W(t,0)\|_F \leq C_a \varrho^a \|W\|_F. \tag{C.9}$$

We now estimate the remainder itself. Set $A_0(t) := L_t \widetilde{\Lambda} R_t$. Then $\varphi_{0,0}(t,0,0) = \mathsf{N}(A_0(t))$. The one-variable chain rule gives

$$\begin{aligned}\partial_t^3 \varphi_{0,0}(t,0,0) = {} & \mathrm{d}\mathsf{N}(A_0(t))[A_0^{(3)}(t)] \\ & + 3\mathrm{d}^2\mathsf{N}(A_0(t))[A_0''(t), A_0'(t)] \\ & + \mathrm{d}^3\mathsf{N}(A_0(t))[A_0'(t), A_0'(t), A_0'(t)].\end{aligned}$$

By Lemma C.4 and (C.6), $|\partial_t^3 \varphi_{0,0}(t,0,0)| \leq C\varrho^3$. Therefore $|\mathcal{R}_{0,0}(0,0)| \leq \frac{1}{2}\int_0^1 (1-t)^2 C\varrho^3 \,\mathrm{d}t \leq K\varrho^3$.

Next we estimate the first derivative of the remainder in a diagonal matrix direction W. Let e_t, e_s denote the coordinate directions corresponding to t and s. By the coordinate-direction interpretation of scalar derivatives, $\partial_s \partial_t^3 \varphi_{W,0}(t,0,0) = \mathrm{d}^4(\mathsf{N}\circ A_W)(t,0)[e_t,e_t,e_t,e_s]$. We apply the Faà di Bruno formula from Lemma C.2 to $\mathsf{N}\circ A_W$. Each term is indexed by a partition $\pi \in \mathcal{P}_4$. Since exactly one inserted coordinate direction is e_s, exactly one block of π contains e_s. Suppose this block contains a copies of e_t, where $0 \leq a \leq 3$. The corresponding inserted direction is $\partial_t^a \partial_s A_W(t,0)$, and by (C.7) it is controlled in Schatten 1-norm by $C\varrho^a\|W\|_1$. All other blocks contain only copies of e_t. If their sizes are $b_1, \ldots, b_q$, then their inserted directions are pure t-derivatives and are controlled in operator norm, by (C.6), as $C\varrho^{b_1}, \ldots, C\varrho^{b_q}$. Since there are three copies of e_t in total, $a + b_1 + \cdots + b_q = 3$. Applying the one-Schatten- 1-input estimate from Lemma C.5, this partition term is bounded by $\frac{C}{n}\varrho^a\|W\|_1 \prod_{j=1}^{q} \varrho^{b_j} = \frac{C}{n}\varrho^3\|W\|_1$. There are only finitely many partitions in $\mathcal{P}_4$, so

$$|\partial_s \partial_t^3 \varphi_{W,0}(t,0,0)| \leq K\varrho^3 \frac{\|W\|_1}{n}. \tag{C.10}$$

Integrating in t gives $|\partial_s \mathcal{R}_{W,0}(0,0)| \leq K\varrho^3 \frac{\|W\|_1}{n}$.

Finally, we estimate the second derivative of the remainder in two diagonal matrix directions. Put $A_{W_1,W_2}(t,s,r) := L_t(\widetilde{\Lambda} + sW_1 + rW_2)R_t$. Let e_t, e_s, e_r denote the coordinate directions corresponding to t, s, r. Then

$$\partial_s \partial_r \partial_t^3 \varphi_{W_1,W_2}(t,0,0) = \mathrm{d}^5(\mathsf{N}\circ A_{W_1,W_2})(t,0,0)[e_t,e_t,e_t,e_s,e_r].$$

Again we use Lemma C.2. Since the dependence of A_{W_1,W_2} on s and r is affine and separate, $\partial_t^a \partial_s \partial_r A_{W_1,W_2}(t,s,r) = 0$ for every $a \geq 0$. Therefore every Faà di Bruno term whose partition has a block containing both e_s and e_r vanishes.

In every nonzero term, the block containing e_s and the block containing e_r are distinct. Suppose these two blocks contain a and b copies of e_t, respectively. Their inserted directions are $\partial_t^a \partial_s A_{W_1,W_2}(t,0,0)$ and $\partial_t^b \partial_r A_{W_1,W_2}(t,0,0)$. By (C.9), these are controlled in Frobenius norm by $C\varrho^a\|W_1\|_F$ and $C\varrho^b\|W_2\|_F$. The remaining blocks contain only copies of e_t. If their sizes are $c_1, \ldots, c_q$, then the corresponding inserted directions are pure t-derivatives and are controlled in operator norm by $C\varrho^{c_1}, \ldots, C\varrho^{c_q}$. Since $a + b + c_1 + \cdots + c_q = 3$, the two-Frobenius-input estimate from Lemma C.5 gives, for this partition term,

$$\frac{C}{n}\varrho^a\|W_1\|_F \varrho^b\|W_2\|_F \prod_{j=1}^{q} \varrho^{c_j} = \frac{C}{n}\varrho^3\|W_1\|_F\|W_2\|_F.$$

Summing over the finitely many partitions in $\mathcal{P}_5$, we obtain

$$\left|\partial_s \partial_r \partial_t^3 \varphi_{W_1,W_2}(t,0,0)\right| \leq K\varrho^3 \frac{\|W_1\|_F\|W_2\|_F}{n}. \tag{C.11}$$

Integrating in t gives $|\partial_s\partial_r\mathcal{R}_{W_1,W_2}(0,0)| \le K\varrho^3\frac{\|W_1\|_F\|W_2\|_F}{n}$.

We also prove the mixed Schatten-1/operator estimate. In every nonzero Faà di Bruno term, the block containing e_s and the block containing e_r are distinct. Suppose these two blocks contain a and b copies of e_t, respectively. The inserted direction corresponding to the e_s-block is $\partial_t^a\partial_s A_{W_1,W_2}(t,0,0)$, and by (C.7) it is bounded in Schatten 1-norm by $C\varrho^a\|W_1\|_1$. The inserted direction corresponding to the e_r-block is $\partial_t^b\partial_r A_{W_1,W_2}(t,0,0)$, and by (C.8) it is bounded in operator norm by $C\varrho^b\|W_2\|_{\mathrm{op}}$. All remaining blocks contain only copies of e_t, and their inserted directions are controlled in operator norm by $C\varrho^{c_1},\dots,C\varrho^{c_q}$. Since $a+b+c_1+\cdots+c_q=3$, the one-Schatten-1-input estimate from Lemma C.5 gives, for this partition term,

$$\frac{C}{n}\varrho^a\|W_1\|_1\varrho^b\|W_2\|_{\mathrm{op}}\prod_{\ell=1}^{q}\varrho^{c_\ell} = \frac{C}{n}\varrho^3\|W_1\|_1\|W_2\|_{\mathrm{op}}.$$

There are only finitely many partitions in $\mathcal{P}_5$, so

$$\left|\partial_s\partial_r\partial_t^3\varphi_{W_1,W_2}(t,0,0)\right| \le K\varrho^3\frac{\|W_1\|_1\|W_2\|_{\mathrm{op}}}{n}.$$

Integrating in t gives (C.5). Interchanging W_1 and W_2 gives the symmetric version. □

Throughout the sequel, we use the pairwise weight

$$h(x,y) := \frac{xy}{x+y}, \qquad x,y\ge 0,\ x+y>0, \tag{C.12}$$

with the convention $h(0,0)=0$.

Lemma C.7 (Dimension-uniform analytic remainder)**.** *Fix $0<c<C<\infty$. There exist constants $\varrho_0=\varrho_0(c,C)>0$ and $K=K(c,C)<\infty$, independent of n, such that the following holds. Suppose that $\boldsymbol{\alpha}\in\mathbb{R}^n$ satisfies $\sum_i\alpha_i=1$, $\frac{c}{n}\le\alpha_i\le\frac{C}{n}$ for all i, and that Δ_U,Δ_V are symmetric, have zero diagonal, and satisfy $\|\Delta_U\|_{\mathrm{op}}\le\varrho$, $\|\Delta_V\|_{\mathrm{op}}\le\varrho$, $0<\varrho\le\varrho_0$. Then, with h as in* (C.12)*,*

$$\|Z(\boldsymbol{\alpha})\|_1 = 1-\frac{1}{2}\sum_{i<j}h(\alpha_i,\alpha_j)\big((\Delta_U)_{ij}-(\Delta_V)_{ij}\big)^2 + R(\boldsymbol{\alpha}). \tag{C.13}$$

The remainder satisfies

$$|R(\boldsymbol{\alpha})|\le K\varrho^3,\qquad \|\nabla R(\boldsymbol{\alpha})\|_\infty\le K\varrho^3,\qquad \|\nabla^2R(\boldsymbol{\alpha})\|_{\mathrm{op}}\le Kn\varrho^3,\qquad \|\nabla^2R(\boldsymbol{\alpha})\|_{\infty\to\infty}\le Kn\varrho^3. \tag{C.14}$$

Moreover, if $q\mapsto\boldsymbol{\alpha}(q)$ is affine, $\xi:=\boldsymbol{\alpha}'(q)$, and $W_\xi:=n\,\mathrm{diag}(\xi)$, then

$$\left|\frac{d}{dq}R(\boldsymbol{\alpha}(q))\right| \le K\|W_\xi\|_{\mathrm{op}}\varrho^3. \tag{C.15}$$

Proof. By Lemma H.12, $\|Z(\boldsymbol{\alpha})\|_1=\|(I+\Delta_U)^{1/2}\Lambda_{\boldsymbol{\alpha}}(I+\Delta_V)^{1/2}\|_1$. Set $\widetilde{\Lambda}_{\boldsymbol{\alpha}}:=n\Lambda_{\boldsymbol{\alpha}}$. Then $\Lambda_{\boldsymbol{\alpha}}=n^{-1}\widetilde{\Lambda}_{\boldsymbol{\alpha}}$, and the entries of $\widetilde{\Lambda}_{\boldsymbol{\alpha}}$ lie in $[c,C]$. We use the normalized Schatten functional N from (C.3).

For $0\le t\le 1$, write

$$L_t:=(I+t\Delta_U)^{1/2},\qquad R_t:=(I+t\Delta_V)^{1/2},\qquad Y_{\boldsymbol{\alpha}}(t):=L_t\Lambda_{\boldsymbol{\alpha}}R_t,\qquad g_{\boldsymbol{\alpha}}(t):=\|Y_{\boldsymbol{\alpha}}(t)\|_1.$$

Then $g_{\boldsymbol{\alpha}}(1)=\|Z(\boldsymbol{\alpha})\|_1$ and $g_{\boldsymbol{\alpha}}(0)=\|\Lambda_{\boldsymbol{\alpha}}\|_1$. Taylor expansion of the square-root map gives

$$L_t = I+\frac{t}{2}\Delta_U-\frac{t^2}{8}\Delta_U^2+S_U(t),\qquad R_t = I+\frac{t}{2}\Delta_V-\frac{t^2}{8}\Delta_V^2+S_V(t),$$

with $\|S_U(t)\|_{\mathrm{op}},\|S_V(t)\|_{\mathrm{op}}\le Kt^3\varrho^3$. Since $\|\Lambda_{\boldsymbol{\alpha}}\|_{\mathrm{op}}\le C/n$, multiplication gives

$$Y_{\boldsymbol{\alpha}}(t)=\Lambda_{\boldsymbol{\alpha}}+tE_1+t^2E_2+\mathcal{R}_Y(t),\qquad \|\mathcal{R}_Y(t)\|_{\mathrm{op}}\le K\frac{t^3\varrho^3}{n},$$

where

$$E_1 = \frac{1}{2}(\Delta_U \Lambda_{\boldsymbol{\alpha}} + \Lambda_{\boldsymbol{\alpha}} \Delta_V), \qquad E_2 = -\frac{1}{8}\Delta_U^2 \Lambda_{\boldsymbol{\alpha}} - \frac{1}{8}\Lambda_{\boldsymbol{\alpha}} \Delta_V^2 + \frac{1}{4}\Delta_U \Lambda_{\boldsymbol{\alpha}} \Delta_V.$$

Thus $Y'_{\boldsymbol{\alpha}}(0) = E_1$ and $Y''_{\boldsymbol{\alpha}}(0) = 2E_2$.

Since $\Lambda_{\boldsymbol{\alpha}}$ is positive diagonal, Lemma H.10 gives $\mathrm{d}\|\cdot\|_1(\Lambda_{\boldsymbol{\alpha}})[H] = \langle I, H\rangle_F$. Hence $g'_{\boldsymbol{\alpha}}(0) = \langle I, E_1\rangle_F$, and the second-order chain rule gives $g''_{\boldsymbol{\alpha}}(0) = \mathrm{d}^2\|\cdot\|_1(\Lambda_{\boldsymbol{\alpha}})[E_1, E_1] + \mathrm{d}\|\cdot\|_1(\Lambda_{\boldsymbol{\alpha}})[2E_2]$. Therefore

$$\|Z(\boldsymbol{\alpha})\|_1 = \|\Lambda_{\boldsymbol{\alpha}}\|_1 + \langle I, E_1 + E_2\rangle_F + \frac{1}{2}\mathrm{d}^2\|\cdot\|_1(\Lambda_{\boldsymbol{\alpha}})[E_1, E_1] + R(\boldsymbol{\alpha}), \tag{C.16}$$

where $R(\boldsymbol{\alpha})$ is the third-order Taylor remainder. Since $\|\Lambda_{\boldsymbol{\alpha}}\|_1 = \sum_i \alpha_i = 1$, it remains to compute the two second-order contributions in (C.16).

First,

$$\langle I, E_1\rangle_F = \frac{1}{2}\operatorname{tr}(\Delta_U \Lambda_{\boldsymbol{\alpha}}) + \frac{1}{2}\operatorname{tr}(\Lambda_{\boldsymbol{\alpha}} \Delta_V) = 0$$

because Δ_U and Δ_V have zero diagonal. Also,

$$\operatorname{tr}(E_2) = -\frac{1}{8}\operatorname{tr}(\Delta_U^2 \Lambda_{\boldsymbol{\alpha}}) - \frac{1}{8}\operatorname{tr}(\Lambda_{\boldsymbol{\alpha}} \Delta_V^2) + \frac{1}{4}\operatorname{tr}(\Delta_U \Lambda_{\boldsymbol{\alpha}} \Delta_V),$$

so the contribution of each pair $i < j$ is

$$-\frac{\alpha_i + \alpha_j}{8}(\Delta_U)_{ij}^2 - \frac{\alpha_i + \alpha_j}{8}(\Delta_V)_{ij}^2 + \frac{\alpha_i + \alpha_j}{4}(\Delta_U)_{ij}(\Delta_V)_{ij}.$$

Thus

$$\langle I, E_2\rangle_F = -\sum_{i<j} \frac{\alpha_i + \alpha_j}{8}\big((\Delta_U)_{ij} - (\Delta_V)_{ij}\big)^2. \tag{C.17}$$

Next, for $i \neq j$, $(E_1)_{ij} - (E_1)_{ji} = \frac{1}{2}(\alpha_j - \alpha_i)\big((\Delta_U)_{ij} - (\Delta_V)_{ij}\big)$. Therefore Lemma H.14 gives

$$\frac{1}{2}\mathrm{d}^2\|\cdot\|_1(\Lambda_{\boldsymbol{\alpha}})[E_1, E_1] = \sum_{i<j} \frac{(\alpha_i - \alpha_j)^2}{8(\alpha_i + \alpha_j)}\big((\Delta_U)_{ij} - (\Delta_V)_{ij}\big)^2. \tag{C.18}$$

Combining (C.17) and (C.18), the coefficient of $\big((\Delta_U)_{ij} - (\Delta_V)_{ij}\big)^2$ is

$$-\frac{\alpha_i + \alpha_j}{8} + \frac{(\alpha_i - \alpha_j)^2}{8(\alpha_i + \alpha_j)} = -\frac{1}{2}\frac{\alpha_i \alpha_j}{\alpha_i + \alpha_j}.$$

Using (C.12), this proves (C.13).

It remains to prove (C.14) and (C.15). For diagonal W_1, W_2, define $\varphi_{W_1,W_2}(t,s,r) := \mathsf{N}\big(L_t(\widetilde{\Lambda}_{\boldsymbol{\alpha}} + sW_1 + rW_2)R_t\big)$, and $\mathcal{R}_{W_1,W_2}(s,r) := \frac{1}{2}\int_0^1 (1-t)^2 \partial_t^3 \varphi_{W_1,W_2}(t,s,r)\,\mathrm{d}t$. Then $R(\boldsymbol{\alpha}) = \mathcal{R}_{0,0}(0,0)$, so Lemma C.6 gives $|R(\boldsymbol{\alpha})| \le K\varrho^3$.

We now translate the diagonal-matrix derivative bounds to derivatives in $\boldsymbol{\alpha}$. A perturbation of $\boldsymbol{\alpha}$ in direction $\boldsymbol{v}$ corresponds to $W_{\boldsymbol{v}} := n\operatorname{diag}(\boldsymbol{v})$, since $\widetilde{\Lambda}_{\boldsymbol{\alpha}+s\boldsymbol{v}} = \widetilde{\Lambda}_{\boldsymbol{\alpha}} + sW_{\boldsymbol{v}}$. Taking $\boldsymbol{v} = \boldsymbol{e}_i$, $W_i := n\boldsymbol{e}_i\boldsymbol{e}_i^\top$, we have $\partial_{\alpha_i} R(\boldsymbol{\alpha}) = \partial_s \mathcal{R}_{W_i,0}(s,0)|_{s=0}$. Since $\|W_i\|_1 = n$, Lemma C.6 gives $|\partial_{\alpha_i} R(\boldsymbol{\alpha})| \le K\varrho^3$. Taking the maximum over i proves $\|\nabla R(\boldsymbol{\alpha})\|_\infty \le K\varrho^3$.

For the operator-norm Hessian bound, let $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^n$ and set $W_{\boldsymbol{v}} := n\operatorname{diag}(\boldsymbol{v})$, $W_{\boldsymbol{w}} := n\operatorname{diag}(\boldsymbol{w})$. Since $R(\boldsymbol{\alpha} + s\boldsymbol{v} + r\boldsymbol{w}) = \mathcal{R}_{W_{\boldsymbol{v}},W_{\boldsymbol{w}}}(s,r)$, Lemma C.6 gives $|\nabla^2 R(\boldsymbol{\alpha})[\boldsymbol{v},\boldsymbol{w}]| \le K\varrho^3 \frac{\|W_{\boldsymbol{v}}\|_F \|W_{\boldsymbol{w}}\|_F}{n} = Kn\varrho^3 \|\boldsymbol{v}\|_2 \|\boldsymbol{w}\|_2$. Taking the supremum over unit $\boldsymbol{v}, \boldsymbol{w}$ proves $\|\nabla^2 R(\boldsymbol{\alpha})\|_{\mathrm{op}} \le Kn\varrho^3$.

For the row-sum bound, write $H_R(\boldsymbol{\alpha}) := \nabla^2 R(\boldsymbol{\alpha})$. For each i, $\sum_{j=1}^n |(H_R(\boldsymbol{\alpha}))_{ij}| = \sup_{\|\boldsymbol{w}\|_\infty \le 1} |\boldsymbol{e}_i^\top H_R(\boldsymbol{\alpha})\boldsymbol{w}|$. Fix $\|\boldsymbol{w}\|_\infty \le 1$. With $W_i := n\boldsymbol{e}_i\boldsymbol{e}_i^\top$ and $W_{\boldsymbol{w}} := n\operatorname{diag}(\boldsymbol{w})$, we have $\|W_i\|_1 = n$ and $\|W_{\boldsymbol{w}}\|_{\mathrm{op}} \le n$. Therefore, by (C.5),

$$|\boldsymbol{e}_i^\top H_R(\boldsymbol{\alpha})\boldsymbol{v}w| = |\nabla^2 R(\boldsymbol{\alpha})[\boldsymbol{e}_i, \boldsymbol{w}]| \le K\varrho^3 \frac{\|W_i\|_1 \|W_{\boldsymbol{w}}\|_{\mathrm{op}}}{n} \le Kn\varrho^3.$$

Taking the supremum over w and then i proves the $\infty \to \infty$ bound in (C.14).

Finally, let $q \mapsto \boldsymbol{\alpha}(q)$ be affine, set $\boldsymbol{\beta} := \boldsymbol{\alpha}(q)$, $\boldsymbol{\xi} := \boldsymbol{\alpha}'(q)$, and $W_\xi := n\,\mathrm{diag}(\boldsymbol{\xi})$. Since $R(\boldsymbol{\beta} + s\boldsymbol{\xi}) = \mathcal{R}_{W_{\boldsymbol{\xi}},0}(s,0)$, $\frac{d}{dq}(R \circ \boldsymbol{\alpha})(q) = \partial_s \mathcal{R}_{W_{\boldsymbol{\xi}},0}(s,0)\big|_{s=0}$. Applying Lemma C.6 gives

$$\left|\frac{d}{dq}(R \circ \boldsymbol{\alpha})(q)\right| \le K\varrho^3 \frac{\|W_{\boldsymbol{\xi}}\|_1}{n} \le K\|W_{\boldsymbol{\xi}}\|_{\mathrm{op}}\varrho^3,$$

which proves (C.15). □

C.2 Deterministic averaged geometry

In this section we isolate the deterministic geometry of the averaged second-order term.

For $q \in [0,1]$, define the fixed group-mass affine set

$$\mathcal{A}_q := \left\{ \boldsymbol{\alpha} \in \mathbb{R}^n_+ : \sum_{i \in I_+} \alpha_i = q,\ \sum_{i \in I_-} \alpha_i = 1 - q \right\},$$

and the groupwise-uniform curve

$$\overline{\boldsymbol{\alpha}}_i(q) = \begin{cases} q/n_+, & i \in I_+, \\ (1-q)/n_-, & i \in I_-. \end{cases}$$

With h as in (C.12), define

$$\mathcal{H}(\boldsymbol{\alpha}) := \sum_{i<j} h(\alpha_i, \alpha_j), \qquad \mathcal{H}_{n_+,n_-}(q) := \mathcal{H}(\overline{\boldsymbol{\alpha}}(q)). \tag{C.19}$$

The main point of this section is not merely concavity of $\mathcal{H}$, but the quantitative fact that, near this groupwise-uniform curve, $\mathcal{H}$ has curvature of order n^2 in all within-group zero-sum directions. This curvature will later provide the deterministic concavity controlling within-group fluctuations.

The first lemma records the exact value of $\mathcal{H}$ along the groupwise-uniform curve and its derivatives with respect to q.

Lemma C.8 (Exact groupwise-uniform functional)**.** *For $q \in [0,1]$,*

$$\mathcal{H}_{n_+,n_-}(q) = \frac{n_+ - 1}{4} q + \frac{n_- - 1}{4}(1-q) + \frac{n_+ n_- q(1-q)}{n_+(1-q) + n_- q}. \tag{C.20}$$

Moreover,

$$\theta_{n_+,n_-} := \mathcal{H}'_{n_+,n_-}\left(\frac{1}{2}\right) = \frac{(n_+ - n_-)(n_+^2 + 6n_+n_- + n_-^2)}{4(n_+ + n_-)^2}. \tag{C.21}$$

Moreover,

$$\mathcal{H}''_{n_+,n_-}(q) = -\frac{2n_+^2 n_-^2}{(n_+(1-q) + n_- q)^3} < 0. \tag{C.22}$$

Proof. By direct computation. □

We next study the curvature of $\mathcal{H}$ in directions that keep the group masses fixed. These are the directions relevant after q has been chosen.

Define the tangent subspace

$$\mathcal{T} := \left\{ \boldsymbol{\beta} \in \mathbb{R}^n : \sum_{i \in I_+} \beta_i = 0,\ \sum_{i \in I_-} \beta_i = 0 \right\}.$$

Then $\mathcal{A}_q = \overline{\boldsymbol{\alpha}}(q) + \mathcal{T}$. Let $P_{\mathcal{T}}$ denote the Euclidean orthogonal projection onto $\mathcal{T}$. Equivalently, for $\boldsymbol{x} \in \mathbb{R}^n$, $(P_{\mathcal{T}}\boldsymbol{x})_i = x_i - \frac{1}{n_+}\sum_{\ell \in I_+} x_\ell$, $\quad i \in I_+$, and $(P_{\mathcal{T}}\boldsymbol{x})_i = x_i - \frac{1}{n_-}\sum_{\ell \in I_-} x_\ell$, $\quad i \in I_-$.

We write $I_{\mathcal{T}}$ for the identity operator on $\mathcal{T}$. Thus an operator inequality involving $I_{\mathcal{T}}$ is understood as an inequality after restricting both sides to the subspace $\mathcal{T}$.

Fix a small constant $r_0 \in (0, 1/4)$ and define the local q-window

$$\mathcal{I} := \left[\frac{1}{2} - r_0, \frac{1}{2} + r_0\right]. \tag{C.23}$$

The following lemma says that $\overline{\boldsymbol{\alpha}}(q)$ is a critical point of $\mathcal{H}$ after restricting to the fixed group-mass affine space $\mathcal{A}_q$, and that the restricted Hessian is negative definite with curvature scale n^2. Equivalently, the restricted gradient is $P_{\mathcal{T}}\nabla\mathcal{H}$, and the restricted Hessian is $P_{\mathcal{T}}\nabla^2\mathcal{H}P_{\mathcal{T}}$ acting on $\mathcal{T}$.

Lemma C.9 (Exact mean curvature)**.** *Assume $n_+/n, n_-/n \in [\kappa, 1-\kappa]$ for a fixed $\kappa > 0$. Uniformly over $q \in \mathcal{I}$, $P_{\mathcal{T}}\nabla\mathcal{H}(\overline{\boldsymbol{\alpha}}(q)) = 0$, and $cn^2 I_{\mathcal{T}} \preceq -P_{\mathcal{T}}\nabla^2\mathcal{H}(\overline{\boldsymbol{\alpha}}(q))P_{\mathcal{T}} \preceq Cn^2 I_{\mathcal{T}}$. Moreover, the restriction is a scalar multiple of the identity on each of the two within-group zero-sum subspaces. Consequently, for $L_q := -\frac{1}{D}P_{\mathcal{T}}\nabla^2\mathcal{H}(\overline{\boldsymbol{\alpha}}(q))P_{\mathcal{T}}$, we have $\|L_q^{-1}\|_{\mathrm{op}} \le C\frac{D}{n^2}$ and $\|L_q^{-1}\boldsymbol{x}\|_\infty \le C\frac{D}{n^2}\|\boldsymbol{x}\|_\infty$. Here and below, operator norms on $\mathcal{T}$ are taken with respect to the Euclidean norm inherited from $\mathbb{R}^n$.*

Proof. Fix $q \in \mathcal{I}$, and write $a := \frac{q}{n_+}, \quad b := \frac{1-q}{n_-}$. Since $q \in \mathcal{I} \subset (1/4, 3/4)$ and $n_+/n, n_-/n \in [\kappa, 1-\kappa]$, we have $a \asymp \frac{1}{n}, \quad b \asymp \frac{1}{n}, \quad a + b \asymp \frac{1}{n}$, uniformly over $q \in \mathcal{I}$.

We first prove the projected-gradient identity. For $x, y > 0$, $\partial_x h(x,y) = \frac{y^2}{(x+y)^2}, \quad \partial_y h(x,y) = \frac{x^2}{(x+y)^2}$. At $\overline{\boldsymbol{\alpha}}(q)$, all coordinates inside the same group are equal. Therefore the i-th coordinate of $\nabla\mathcal{H}(\overline{\boldsymbol{\alpha}}(q))$ is the same for all $i \in I_+$, and is also the same for all $i \in I_-$. More explicitly, for $i \in I_+$,

$$\begin{aligned}
\partial_{\alpha_i}\mathcal{H}(\overline{\boldsymbol{\alpha}}(q)) &= \sum_{j\in I_+} \partial_{\alpha_i} h(\alpha_i, \alpha_j)\big|_{\boldsymbol{\alpha}=\overline{\boldsymbol{\alpha}}(q)} + \sum_{j\in I_-} \partial_{\alpha_i} h(\alpha_i, \alpha_j)\big|_{\boldsymbol{\alpha}=\overline{\boldsymbol{\alpha}}(q)} \\
&= \sum_{j\in I_+} \frac{a^2}{(2a)^2} + \sum_{j\in I_-} \frac{b^2}{(a+b)^2} \\
&= \frac{n_+ - 1}{4} + n_- \frac{b^2}{(a+b)^2},
\end{aligned}$$

while for $i \in I_-$,

$$\begin{aligned}
\partial_{\alpha_i}\mathcal{H}(\overline{\boldsymbol{\alpha}}(q)) &= \sum_{j\in I_-} \partial_{\alpha_i} h(\alpha_i, \alpha_j)\big|_{\boldsymbol{\alpha}=\overline{\boldsymbol{\alpha}}(q)} + \sum_{j\in I_+} \partial_{\alpha_i} h(\alpha_i, \alpha_j)\big|_{\boldsymbol{\alpha}=\overline{\boldsymbol{\alpha}}(q)} \\
&= \sum_{j\in I_-} \frac{b^2}{(2b)^2} + \sum_{j\in I_+} \frac{a^2}{(b+a)^2} \\
&= \frac{n_- - 1}{4} + n_+ \frac{a^2}{(a+b)^2}.
\end{aligned}$$

Thus $\nabla\mathcal{H}(\overline{\boldsymbol{\alpha}}(q))$ is constant within each group. Since $P_{\mathcal{T}}$ subtracts the within-group average from each group, it annihilates every vector that is constant within each group. Hence $P_{\mathcal{T}}\nabla\mathcal{H}(\overline{\boldsymbol{\alpha}}(q)) = 0$.

We now turn to the Hessian. For one pair, the second directional derivative satisfies $-\nabla^2 h(x,y)[(u,v),(u,v)] = \frac{2(yu-xv)^2}{(x+y)^3}$. Therefore, for $\boldsymbol{\beta} \in \mathcal{T}$, $-\boldsymbol{\beta}^\top\nabla^2\mathcal{H}(\overline{\boldsymbol{\alpha}}(q))\boldsymbol{\beta} = S_{++}(\boldsymbol{\beta}) + S_{--}(\boldsymbol{\beta}) + S_{+-}(\boldsymbol{\beta})$, where the three terms correspond to pairs inside I_+, pairs inside I_-, and cross-group pairs.

For the I_+-within pairs, both coordinates equal a. Hence

$$S_{++}(\boldsymbol{\beta}) = \frac{1}{4a} \sum_{\substack{i<j \\ i,j\in I_+}} (\beta_i - \beta_j)^2.$$

Since $\boldsymbol{\beta} \in \mathcal{T}$, we have $\sum_{i \in I_+} \beta_i = 0$, and therefore

$$\sum_{\substack{i<j \\ i,j \in I_+}} (\beta_i - \beta_j)^2 = n_+ \sum_{i \in I_+} \beta_i^2.$$

Thus $S_{++}(\boldsymbol{\beta}) = \frac{n_+}{4a} \sum_{i \in I_+} \beta_i^2$. Similarly, $S_{--}(\boldsymbol{\beta}) = \frac{n_-}{4b} \sum_{i \in I_-} \beta_i^2$.

For the cross-group pairs, we have

$$S_{+-}(\boldsymbol{\beta}) = \frac{2}{(a+b)^3} \sum_{\substack{i \in I_+ \\ j \in I_-}} (b\beta_i - a\beta_j)^2.$$

Expanding the square gives

$$S_{+-}(\boldsymbol{\beta}) = \frac{2}{(a+b)^3} \left[n_- b^2 \sum_{i \in I_+} \beta_i^2 + n_+ a^2 \sum_{j \in I_-} \beta_j^2 - 2ab \left(\sum_{i \in I_+} \beta_i \right) \left(\sum_{j \in I_-} \beta_j \right) \right].$$

The last term vanishes because $\boldsymbol{\beta} \in \mathcal{T}$. Hence

$$S_{+-}(\boldsymbol{\beta}) = \frac{2n_- b^2}{(a+b)^3} \sum_{i \in I_+} \beta_i^2 + \frac{2n_+ a^2}{(a+b)^3} \sum_{j \in I_-} \beta_j^2.$$

Combining the three contributions, we obtain

$$-\boldsymbol{\beta}^\top \nabla^2 \mathcal{H}(\overline{\boldsymbol{\alpha}}(q)) \boldsymbol{\beta} = \lambda_+(q) \sum_{i \in I_+} \beta_i^2 + \lambda_-(q) \sum_{i \in I_-} \beta_i^2,$$

where $\lambda_+(q) := \frac{n_+}{4a} + \frac{2n_- b^2}{(a+b)^3}$, and $\lambda_-(q) := \frac{n_-}{4b} + \frac{2n_+ a^2}{(a+b)^3}$. Using

$$a \asymp \frac{1}{n}, \qquad b \asymp \frac{1}{n}, \qquad a+b \asymp \frac{1}{n}, \qquad n_+ \asymp n, \qquad n_- \asymp n,$$

we get $cn^2 \le \lambda_+(q), \lambda_-(q) \le Cn^2$ uniformly over $q \in \mathcal{I}$. Therefore, for every $\boldsymbol{\beta} \in \mathcal{T}$, $cn^2 \|\boldsymbol{\beta}\|_2^2 \le -\boldsymbol{\beta}^\top \nabla^2 \mathcal{H}(\overline{\boldsymbol{\alpha}}(q)) \boldsymbol{\beta} \le Cn^2 \|\boldsymbol{\beta}\|_2^2$. Equivalently, $cn^2 I_{\mathcal{T}} \preceq -P_{\mathcal{T}} \nabla^2 \mathcal{H}(\overline{\boldsymbol{\alpha}}(q)) P_{\mathcal{T}} \preceq Cn^2 I_{\mathcal{T}}$.

The displayed formula also shows the scalar-multiple structure. On the zero-sum subspace supported on I_+, the restricted operator $-P_{\mathcal{T}} \nabla^2 \mathcal{H}(\overline{\boldsymbol{\alpha}}(q)) P_{\mathcal{T}}$ acts as multiplication by $\lambda_+(q)$. On the zero-sum subspace supported on I_-, it acts as multiplication by $\lambda_-(q)$. Thus its inverse on $\mathcal{T}$ acts by multiplication by $\lambda_+(q)^{-1}$ on the first subspace and by $\lambda_-(q)^{-1}$ on the second. Since $\lambda_+(q), \lambda_-(q) \ge cn^2$, we obtain $\|L_q^{-1}\|_{\mathrm{op}} \le C \frac{D}{n^2}$. Moreover, if $\boldsymbol{x} \in \mathcal{T}$, then $L_q^{-1} \boldsymbol{x}$ is obtained by multiplying the I_+-component of $\boldsymbol{x}$ by $D/\lambda_+(q)$ and the I_--component of x by $D/\lambda_-(q)$. Therefore $\|L_q^{-1} \boldsymbol{x}\|_\infty \le C \frac{D}{n^2} \|\boldsymbol{x}\|_\infty$. □

C.3 Canonical spherical kernel

For

$$(\boldsymbol{u}, \widehat{\boldsymbol{v}}), (\boldsymbol{u}', \widehat{\boldsymbol{v}}') \in \mathbb{S}^{D-1} \times \mathbb{S}^{D-1},$$

define

$$\zeta\big((\boldsymbol{u}, \widehat{\boldsymbol{v}}), (\boldsymbol{u}', \widehat{\boldsymbol{v}}')\big) := \Big(\boldsymbol{u}^\top \boldsymbol{u}' - \widehat{\boldsymbol{v}}^\top \widehat{\boldsymbol{v}}' \Big)^2 - \frac{2}{D}.$$

Then, for $i \ne j$,

$$\zeta_{ij} = \zeta\big((\boldsymbol{u}_i, \widehat{\boldsymbol{v}}_i), (\boldsymbol{u}_j, \widehat{\boldsymbol{v}}_j)\big) = \big((\Delta_U)_{ij} - (\Delta_V)_{ij}\big)^2 - \frac{2}{D}. \tag{C.24}$$

For convenience, write $\omega_i := (\boldsymbol{u}_i, \widehat{\boldsymbol{v}}_i)$.

Lemma C.10 (Fourth moments on the sphere)**.** *Let* $\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}^{D-1})$. *Then* $\mathbf{E} u_1^4 = \frac{3}{D(D+2)}$, $\mathbf{E} u_1^2 u_2^2 = \frac{1}{D(D+2)}$, $\mathbf{E} u_1^3 u_2 = 0$*, More generally, for any deterministic unit vectors* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{D-1}$,

$$\mathbf{E}\left[(\boldsymbol{x}^\top \boldsymbol{u})^2 (\boldsymbol{y}^\top \boldsymbol{u})^2 \right] = \frac{1 + 2(\boldsymbol{x}^\top \boldsymbol{y})^2}{D(D+2)}.$$

Proof. Using the representation $\boldsymbol{u} = \boldsymbol{g}/\|\boldsymbol{g}\|_2$, with $\boldsymbol{g} \sim N(0, I_D)$, we have $u_1 = g_1/\|\boldsymbol{g}\|_2$ Recall that if $X \sim \Gamma(\alpha, \theta)$ and $Y \sim \Gamma(\beta, \theta)$ are independent gamma random variables with the same scale parameter, then $\frac{X}{X+Y} \sim \mathrm{Beta}(\alpha, \beta)$ (see (Casella & Berger, 2002, Exercise 4.24)). Now $g_1^2 \sim \chi_1^2 = \Gamma(1/2, 2)$, while $g_2^2 + \cdots + g_D^2 \sim \chi_{D-1}^2 = \Gamma\left(\frac{D-1}{2}, 2\right)$, and these two random variables are independent. Therefore $u_1^2 = \frac{g_1^2}{g_1^2 + \cdots + g_D^2} \sim \mathrm{Beta}\left(\frac{1}{2}, \frac{D-1}{2}\right)$. Therefore, we have $\mathbf{E}u_1^4 = \frac{3}{D(D+2)}$. Next, since $\sum_{k=1}^D u_k^2 = 1$, we have $1 = \left(\sum_{k=1}^D u_k^2\right)^2 = \sum_{k=1}^D u_k^4 + 2\sum_{1 \le k < \ell \le D} u_k^2 u_\ell^2$. Taking expectations and using symmetry gives $1 = D\mathbf{E}u_1^4 + D(D-1)\mathbf{E}u_1^2 u_2^2$. Using the formula for $\mathbf{E}u_1^4$, we obtain $\mathbf{E}u_1^2 u_2^2 = \frac{1}{D(D+2)}$.

Now let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{D-1}$, and set $a := \boldsymbol{x}^\top \boldsymbol{y}$. By rotational invariance, we may assume $\boldsymbol{x} = \boldsymbol{e}_1, \quad \boldsymbol{y} = a\boldsymbol{e}_1 + \sqrt{1-a^2}\boldsymbol{e}_2$. Then $\boldsymbol{x}^\top \boldsymbol{u} = u_1, \quad \boldsymbol{y}^\top \boldsymbol{u} = au_1 + \sqrt{1-a^2}u_2$. Therefore

$$\begin{aligned}\mathbf{E}\left[(\boldsymbol{x}^\top \boldsymbol{u})^2 (\boldsymbol{y}^\top \boldsymbol{u})^2\right] &= \mathbf{E}\left[u_1^2 \left(au_1 + \sqrt{1-a^2}u_2\right)^2\right] \\ &= a^2 \mathbf{E}u_1^4 + (1-a^2)\mathbf{E}u_1^2 u_2^2,\end{aligned}$$

because $\mathbf{E}u_1^3 u_2 = 0$ by symmetry. Substituting the two coordinate moment formulas gives

$$\mathbf{E}\left[(\boldsymbol{x}^\top \boldsymbol{u})^2 (\boldsymbol{y}^\top \boldsymbol{u})^2\right] = \frac{3a^2}{D(D+2)} + \frac{1-a^2}{D(D+2)} = \frac{1+2a^2}{D(D+2)}.$$

This proves the claim. □

Lemma C.11 (Conditional moments and degeneracy)**.** *Let $\omega_0 = (\boldsymbol{u}_0, \widehat{\boldsymbol{v}}_0) \in \mathbb{S}^{D-1} \times \mathbb{S}^{D-1}$ be fixed, and let $\omega = (\boldsymbol{u}, \widehat{\boldsymbol{v}})$ be random, where $\boldsymbol{u}$ and $\widehat{\boldsymbol{v}}$ are independent and uniformly distributed on $\mathbb{S}^{D-1}$. Then $\mathbf{E}_\omega \zeta(\omega_0, \omega) = 0$. For every $p \ge 1$, $\|\zeta(\omega_0, \omega)\|_{L_p(\omega)} \le C_\zeta \frac{p}{D}$ uniformly over ω_0. Moreover, $\mathbf{E}_\omega \zeta(\omega_0, \omega)^2 = \frac{6}{D(D+2)} + \frac{2}{D^2}$, and hence $\frac{4}{D^2} \le \mathbf{E}_\omega \zeta(\omega_0, \omega)^2 \le \frac{8}{D^2}$.*

Consequently, for every $i \ne j$,

$$\mathbf{E}[\zeta_{ij} \mid \omega_i] = 0, \qquad \|\zeta_{ij}\|_{L_p} \le C_\zeta \frac{p}{D}, \qquad \frac{4}{D^2} \le \mathbf{E}\zeta_{ij}^2 \le \frac{8}{D^2}.$$

Moreover, the conditional second-moment bound $\mathbf{E}[\zeta_{ij}^2 \mid \omega_i] \le \frac{8}{D^2}$ holds almost surely.

Proof. Write

$$\omega_0 = (\boldsymbol{u}_0, \widehat{\boldsymbol{v}}_0), \qquad \omega = (\boldsymbol{u}, \widehat{\boldsymbol{v}}),$$

and set $r := \boldsymbol{u}_0^\top \boldsymbol{u}, \quad t := \widehat{\boldsymbol{v}}_0^\top \widehat{\boldsymbol{v}}$. Since ω_0 is fixed, and by rotational invariance, the random variables r and t are independent, and each has the same distribution as $u_1 = \langle \boldsymbol{e}_1, \boldsymbol{u} \rangle, \quad \boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}^{D-1})$. Therefore $\mathbf{E}_\omega r = \mathbf{E}_\omega t = 0, \quad \mathbf{E}_\omega r^2 = \mathbf{E}_\omega t^2 = \frac{1}{D}$. Since $\zeta(\omega_0, \omega) = (r-t)^2 - \frac{2}{D}$, we obtain $\mathbf{E}_\omega \zeta(\omega_0, \omega) = \mathbf{E}_\omega (r-t)^2 - \frac{2}{D} = \frac{1}{D} + \frac{1}{D} - \frac{2}{D} = 0$.

Next, by the sub-Gaussianity of one-dimensional marginals of the uniform distribution on the sphere (Vershynin, 2026, Theorem 3.4.5), we have, for every $p \ge 1$, $\|r\|_{L_p(\omega)} = \|t\|_{L_p(\omega)} \le C_s\sqrt{\frac{p}{D}}$, where $C_s(> 1)$ is a universal constant. Hence

$$\begin{aligned}\|(r-t)^2\|_{L_p(\omega)} &= \|r-t\|_{L_{2p}(\omega)}^2 \\ &\le 2\|r\|_{L_{2p}(\omega)}^2 + 2\|t\|_{L_{2p}(\omega)}^2 \\ &\le 8C_s^2 \frac{p}{D}.\end{aligned}$$

Since $2/D \le 2p/D$, this gives $\|\zeta(\omega_0, \omega)\|_{L_p(\omega)} \le C_\zeta \frac{p}{D}$, where $C_\zeta = 8C_s^2 + 2$.

It remains to compute the second moment. By Lemma C.10, $\mathbf{E}_\omega r^4 = \mathbf{E}_\omega t^4 = \frac{3}{D(D+2)}$. Using independence of r and t,

$$\mathbf{E}_\omega (r-t)^4 = \mathbf{E}_\omega r^4 + 6\mathbf{E}_\omega r^2 \mathbf{E}_\omega t^2 + \mathbf{E}_\omega t^4 = \frac{6}{D(D+2)} + \frac{6}{D^2}.$$

Therefore

$$\begin{aligned}\mathbf{E}_\omega \zeta(\omega_0,\omega)^2 &= \mathbf{E}_\omega\left[\left((r-t)^2-\frac{2}{D}\right)^2\right]\\ &= \mathbf{E}_\omega(r-t)^4-\frac{4}{D}\mathbf{E}_\omega(r-t)^2+\frac{4}{D^2}\\ &= \frac{6}{D(D+2)}+\frac{6}{D^2}-\frac{8}{D^2}+\frac{4}{D^2}\\ &= \frac{6}{D(D+2)}+\frac{2}{D^2}.\end{aligned}$$

This lies between $4/D^2$ and $8/D^2$.

The final statements follow by applying the fixed-ω_0 identities with $\omega_0=\omega_i$ and $\omega=\omega_j$, and then conditioning on ω_i. □

Lemma C.12 (Spherical coherence moment bound)**.** *Let $\boldsymbol{u}_1,\dots,\boldsymbol{u}_n,\widehat{\boldsymbol{v}}_1,\dots,\widehat{\boldsymbol{v}}_n$ be independent random vectors, each uniformly distributed on $\mathbb{S}^{D-1}$. Let $L=\log n$. Then, for every $p\ge 1$,*

$$\left\|\max_{i\neq j}\left\{|\boldsymbol{u}_i^\top\boldsymbol{u}_j|,|\widehat{\boldsymbol{v}}_i^\top\widehat{\boldsymbol{v}}_j|\right\}\right\|_{L_{2p}}\le 2C_s n^{1/p}\sqrt{\frac{p}{D}}.$$

Proof. For each $i\neq j$, by rotational invariance, $\boldsymbol{u}_i^\top\boldsymbol{u}_j$ has the distribution of a spherical inner product. Hence, by the spherical inner-product moment bound used in Lemma C.11, for every $\ell\ge 1$, $\|\boldsymbol{u}_i^\top\boldsymbol{u}_j\|_{L_\ell}\le C_s\sqrt{\frac{\ell}{D}}$. The same bound holds for $\widehat{\boldsymbol{v}}_i^\top\widehat{\boldsymbol{v}}_j$.

Let

$$M:=\max_{i\neq j}\left\{|\boldsymbol{u}_i^\top\boldsymbol{u}_j|,|\widehat{\boldsymbol{v}}_i^\top\widehat{\boldsymbol{v}}_j|\right\}.$$

Using the elementary maximum inequality

$$\left\|\max_{j\in\mathcal{J}}|X_j|\right\|_{L_\ell}\le\left(\sum_{j\in\mathcal{J}}\mathbf{E}|X_j|^\ell\right)^{1/\ell},$$

with $\ell=2p$, we get

$$\begin{aligned}\|M\|_{L_{2p}} &\le \left(\sum_{i\neq j}\mathbf{E}|\boldsymbol{u}_i^\top\boldsymbol{u}_j|^{2p}+\sum_{i\neq j}\mathbf{E}|\widehat{\boldsymbol{v}}_i^\top\widehat{\boldsymbol{v}}_j|^{2p}\right)^{1/(2p)}\\ &\le (2n(n-1))^{1/(2p)}C_s\sqrt{\frac{2p}{D}}\\ &\le 2C_s n^{1/p}\sqrt{\frac{p}{D}}.\end{aligned}$$

□

The following shared-index covariance formula is useful for verifying the block-variance parameter in the matrix U-statistic inequality.

Lemma C.13 (Shared-index covariance)**.** *Let $(\boldsymbol{u},\widehat{\boldsymbol{v}})$ be an independent copy of $(\boldsymbol{u}_1,\widehat{\boldsymbol{v}}_1)$. Define*

$$\zeta_i(\boldsymbol{u},\widehat{\boldsymbol{v}}):=\zeta\big((\boldsymbol{u}_i,\widehat{\boldsymbol{v}}_i),(\boldsymbol{u},\widehat{\boldsymbol{v}})\big).$$

For $i\neq j$, put

$$a:=\boldsymbol{u}_i^\top\boldsymbol{u}_j,\qquad b:=\widehat{\boldsymbol{v}}_i^\top\widehat{\boldsymbol{v}}_j.$$

Then

$$\mathbf{E}_{\boldsymbol{u},\widehat{\boldsymbol{v}}}\left[\zeta_i(\boldsymbol{u},\widehat{\boldsymbol{v}})\zeta_j(\boldsymbol{u},\widehat{\boldsymbol{v}})\right]=\frac{2(Da^2-1)}{D^2(D+2)}+\frac{2(Db^2-1)}{D^2(D+2)}+\frac{4ab}{D^2}.$$

In particular, on the coherence event $\max_{i\neq j}\{|(\Delta_U)_{ij}|,|(\Delta_V)_{ij}|\}\le\mu$, the absolute value is at most $8\left(\frac{\mu^2}{D^2}+\frac{1}{D^3}\right)$.

Proof. Write

$$r_i = \boldsymbol{u}_i^\top \boldsymbol{u}, \qquad t_i = \widehat{\boldsymbol{v}}_i^\top \widehat{\boldsymbol{v}}.$$

Then $\zeta_i(\boldsymbol{u}, \widehat{\boldsymbol{v}}) = (r_i^2 - 1/D) + (t_i^2 - 1/D) - 2r_i t_i$. Since $\boldsymbol{u}$ and $\widehat{\boldsymbol{v}}$ are independent so as r_i and t_i conditioning on $\boldsymbol{u}_i, \widehat{\boldsymbol{v}}_i$.

By Lemma C.10, we obtain $\mathbf{E}_{\boldsymbol{u}} r_i^2 r_j^2 = \frac{1+2a^2}{D(D+2)}$.

Thus, $\mathbf{E}_{\boldsymbol{u}}[(r_i^2 - 1/D)(r_j^2 - 1/D)] = \frac{2(Da^2-1)}{D^2(D+2)}$. The corresponding t-term has the same form with b. The only mixed term that does not vanish is

$$\mathbf{E}_{\boldsymbol{u},\widehat{\boldsymbol{v}}}[(-2r_i t_i)(-2r_j t_j)] = 4\mathbf{E}_{\boldsymbol{u}}[r_i r_j]\, \mathbf{E}_{\widehat{\boldsymbol{v}}}[t_i t_j].$$

Since

$$\mathbf{E}_{\boldsymbol{u}}[r_i r_j] = \boldsymbol{u}_i^\top \mathbf{E}[\boldsymbol{u}\boldsymbol{u}^\top]\boldsymbol{u}_j = \frac{a}{D}, \qquad \mathbf{E}_{\widehat{\boldsymbol{v}}}[t_i t_j] = \frac{b}{D},$$

we obtain $4\mathbf{E}_{\boldsymbol{u}}[r_i r_j]\, \mathbf{E}_{\widehat{\boldsymbol{v}}}[t_i t_j] = \frac{4ab}{D^2}$. □

C.4 A SPECIALIZED MATRIX U-STATISTIC BOUND

We now record the only non-elementary concentration input. The result we need is a direct specialization of Theorem 3.1 of Minsker & Wei (2019) to a scalar completely degenerate kernel multiplied by deterministic sparse self-adjoint coefficient matrices.

We first state the Minsker–Wei inequality in the form used below. Let $\omega_1, \dots, \omega_n$ be i.i.d. random variables taking values in a measurable space $\mathcal{S}$, $\mathbb{S}^d \in \mathbb{R}^{d\times d}$ be the set of all symmetric matrices, and let $H_{ij} : \mathcal{S} \times \mathcal{S} \to \mathbb{S}^d, \quad i \neq j$ be matrix-valued kernels satisfying $H_{ij}(\omega, \omega') = H_{ji}(\omega', \omega)$. Assume that the kernels are completely degenerate, namely $\mathbf{E}_{\omega'} H_{ij}(\omega, \omega') = 0 \quad$ for every fixed ω. Let $\{\omega_i^{(1)}\}_{i=1}^n$ and $\{\omega_i^{(2)}\}_{i=1}^n$ be two independent copies of the original sample. Define the U-statistic $\mathsf{U}_H := \sum_{(i,j):\, i\neq j} H_{ij}(\omega_i, \omega_j)$, and let $\widetilde{G} \in \mathbb{H}^{nd}$ be the block matrix whose (i,j)-block is

$$\widetilde{G}_{ij} = \begin{cases} H_{ij}(\omega_i^{(1)}, \omega_j^{(2)}), & i \neq j, \\ 0, & i = j. \end{cases}$$

Theorem C.1 (Theorem 3.1, Minsker & Wei (2019))**.** *For every $p \geq 1$ with $r := \max\{p, \log(ed)\}$, we have*

$$\left(\mathbf{E}\|\mathsf{U}_H\|_{\mathrm{op}}^{2p}\right)^{1/(2p)} \leq C_{MW}\left[r^{3/2}A_p + rB + rC_p\right],$$

where

$$A_p := \left(\mathbf{E}\max_i \left\|\sum_{j:\, j\neq i} H_{ij}(\omega_i^{(1)}, \omega_j^{(2)})^2\right\|_{\mathrm{op}}^p\right)^{1/(2p)},$$

$$B := \left\|\sum_{(i,j):\, i\neq j} \mathbf{E}H_{ij}(\omega_i^{(1)}, \omega_j^{(2)})^2\right\|_{\mathrm{op}}^{1/2},$$

$$C_p := \left(\mathbf{E}\left\|\mathbf{E}_2 \widetilde{G}\widetilde{G}^\top\right\|_{\mathrm{op}}^p\right)^{1/(2p)},$$

and C_{MW} is a universal constant. Here $\mathbf{E}_2$ denotes expectation with respect to $\{\omega_i^{(2)}\}_{i=1}^n$, conditionally on $\{\omega_i^{(1)}\}_{i=1}^n$.

We apply this inequality to the canonical spherical kernel. For $\omega = (\boldsymbol{u}, \widehat{\boldsymbol{v}}), \quad \omega' = (\boldsymbol{u}', \widehat{\boldsymbol{v}}')$ in $\mathbb{S}^{D-1} \times \mathbb{S}^{D-1}$, recall that

$$\zeta(\omega, \omega') = \left(\boldsymbol{u}^\top \boldsymbol{u}' - \widehat{\boldsymbol{v}}^\top \widehat{\boldsymbol{v}}'\right)^2 - \frac{2}{D}.$$

We also use the following auxiliary lemma:

Lemma C.14 (Block row-sum bound)**.** *Let $\mathcal{A} = (A_{ij})_{i,j=1}^n$ be a symmetric block matrix, where each block A_{ij} is a square matrix of the same size. Then $\|\mathcal{A}\|_{\mathrm{op}} \le \max_i \sum_{j=1}^n \|A_{ij}\|_{\mathrm{op}}$. In particular, if $\max_i \|A_{ii}\|_{\mathrm{op}} \le D, \quad \max_{i\neq j} \|A_{ij}\|_{\mathrm{op}} \le R$, then $\|\mathcal{A}\|_{\mathrm{op}} \le D + (n-1)R$.*

Proof. Let $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$, where the $\boldsymbol{x}_i$'s are block vectors. Then

$$\begin{aligned}
\|\mathcal{A}\|_{\mathrm{op}} &= \sup_{\|\boldsymbol{x}\|_2=1} |\boldsymbol{x}^\top A \boldsymbol{x}| = \sup_{\|x\|_2=1} \left| \sum_{i,j=1}^n \boldsymbol{x}_i^\top A_{ij} \boldsymbol{x}_j \right| \\
&\le \sup_{\|\boldsymbol{x}\|_2=1} \sum_{i,j=1}^n \|A_{ij}\|_{\mathrm{op}} \|\boldsymbol{x}_i\|_2 \|\boldsymbol{x}_j\|_2 \\
&\le \sup_{\|\boldsymbol{x}\|_2=1} \frac{1}{2} \left(\sum_{i=1}^n \|\boldsymbol{x}_i\|_2^2 \sum_{j=1}^n \|A_{ij}\|_{\mathrm{op}} + \sum_{j=1}^n \|\boldsymbol{x}_j\|_2^2 \sum_{i=1}^n \|A_{ij}\|_{\mathrm{op}} \right) \\
&\le \sup_{\|\boldsymbol{x}\|_2=1} \frac{1}{2} \left(\sum_{i=1}^n \|\boldsymbol{x}_i\|_2^2 \sum_{j=1}^n \|A_{ij}\|_{\mathrm{op}} + \sum_{j=1}^n \|\boldsymbol{x}_j\|_2^2 \sum_{j=1}^n \|A_{ij}\|_{\mathrm{op}} \right) \\
&= \sup_{\|\boldsymbol{x}\|_2=1} \frac{1}{2} \max_i \sum_{j=1}^n \|A_{ij}\|_{\mathrm{op}} \left(\sum_{i=1}^n \|\boldsymbol{x}_i\|_2^2 + \sum_{i=j}^n \|\boldsymbol{x}_j\|_2^2 \right) \\
&\le \max_i \sum_{j=1}^n \|A_{ij}\|_{\mathrm{op}},
\end{aligned}$$

where the second inequality is due to $2\|\boldsymbol{x}_i\|_2 \|\boldsymbol{x}_j\|_2 \le \|\boldsymbol{x}_i\|_2^2 + \|\boldsymbol{x}_j\|_2^2$ and the third equality is due to the symmetry of $\mathcal{A}$.

□

Lemma C.15 (Sparse coefficient matrix U-statistic)**.** *Let $0 < \delta < e^{-1}$, and set $L_\delta := \log(n/\delta)$. Let $C_{ij} = C_{ji} \in \mathbb{R}^{n\times n}$ be deterministic symmetric matrices, each supported on coordinates $\{i, j\}$, and suppose $\|C_{ij}\|_{\mathrm{op}} \le Kn$ and $\left\|\sum_{i<j} C_{ij}^2\right\|_{\mathrm{op}} \le K^2 n^3$. Let $\mathsf{U} := \sum_{i<j} \zeta_{ij} C_{ij}$. If $D \ge nL_\delta$, then, with probability at least $1-\delta$, $\|\mathsf{U}\|_{\mathrm{op}} \le C_{\mathsf{U}} K \frac{n^{3/2} L_\delta^{5/2}}{D}$, where C_{U} is a universal constant.*

Moreover, suppose that $C_{ij}(q) = C_{ji}(q)$, $q \in \mathcal{I}$, is a continuously differentiable family of deterministic self-adjoint matrices supported on coordinates $\{i, j\}$. Assume that, for $\ell = 0, 1$, $\sup_{q\in\mathcal{I}} \sup_{i<j} \|\partial_q^\ell C_{ij}(q)\|_{\mathrm{op}} \le K_\ell n$ and

$$\sup_{q\in\mathcal{I}} \left\| \sum_{i<j} \left(\partial_q^\ell C_{ij}(q)\right)^2 \right\|_{\mathrm{op}} \le K_\ell^2 n^3.$$

Then, with probability at least $1-\delta$,

$$\sup_{q\in\mathcal{I}} \left\| \sum_{i<j} \zeta_{ij} C_{ij}(q) \right\|_{\mathrm{op}} \le C_{\mathsf{U}}(K_0 + K_1) \frac{n^{3/2} L_\delta^{5/2}}{D}.$$

Proof. We first prove the fixed-q statement. Set $p := \lceil L_\delta \rceil, \quad r := \max\{p, \log(en)\}$. Since $p \ge \log(n/\delta)$ and $\delta < e^{-1}$, we have $p \ge \log n, \quad p \ge \log(1/\delta), \quad p \ge \log(en)$. Thus, we have $1 \le n^{1/p} = \exp\left(\frac{\log n}{p}\right) \le e$, and $1 \le \delta^{-1/(2p)} = \exp\left(\frac{\log(1/\delta)}{2p}\right) \le \sqrt{e}$. Moreover $r = p$. Finally, since $p \le 2L_\delta$ for $L_\delta \ge 1$, the assumption $D \ge nL_\delta$ implies $\frac{np}{D} \le 2$.

We apply the Minsker–Wei inequality. For $\omega = (\boldsymbol{u}, \widehat{\boldsymbol{v}}), \quad \omega' = (\boldsymbol{u}', \widehat{\boldsymbol{v}}')$, define, for $i \neq j$, $H_{ij}(\omega, \omega') := \frac{1}{2}\zeta(\omega, \omega') C_{ij}$. The factor $1/2$ converts the full off-diagonal sum in the Minsker–Wei theorem into our upper-triangular sum. Indeed, since $C_{ij} = C_{ji}$ and ζ is symmetric,

$\sum_{i\neq j} H_{ij}(\omega_i,\omega_j) = \sum_{i<j}\zeta(\omega_i,\omega_j)C_{ij} = \mathsf{U}$. By Lemma C.11, the scalar kernel ζ is centered in each argument. Hence $\mathbf{E}_{\omega'} H_{ij}(\omega,\omega') = 0$, so the matrix-valued kernel H_{ij} is completely degenerate.

We now bound the three parameters in the Minsker–Wei inequality.

First, since Lemma C.11 gives $\mathbf{E}\zeta(\omega_i,\omega_j)^2 \le \frac{8}{D^2}$, we have $\mathbf{E}H_{ij}(\omega_i^{(1)},\omega_j^{(2)})^2 \preceq \frac{2}{D^2}C_{ij}^2$. Therefore

$$B := \left\|\sum_{i\neq j}\mathbf{E}H_{ij}(\omega_i^{(1)},\omega_j^{(2)})^2\right\|_{\mathrm{op}}^{1/2} \le \frac{2}{D}\left\|\sum_{i<j}C_{ij}^2\right\|_{\mathrm{op}}^{1/2} \le 2K\frac{n^{3/2}}{D}.$$

Next we estimate the row-square parameter A_p. Using $\|C_{ij}\|_{\mathrm{op}} \le Kn$, for each fixed i,

$$\begin{aligned}\left\|\sum_{j:\,j\neq i} H_{ij}(\omega_i^{(1)},\omega_j^{(2)})^2\right\|_{\mathrm{op}} &\le \sum_{j:\,j\neq i}\left\|H_{ij}(\omega_i^{(1)},\omega_j^{(2)})^2\right\|_{\mathrm{op}}\\ &\le \frac{1}{4}\sum_{j:\,j\neq i}|\zeta(\omega_i^{(1)},\omega_j^{(2)})|^2\|C_{ij}\|_{\mathrm{op}}^2\\ &\le \frac{K^2n^3}{4}\max_{j\neq i}|\zeta(\omega_i^{(1)},\omega_j^{(2)})|^2.\end{aligned}$$

Consequently,

$$A_p \le \frac{Kn^{3/2}}{2}\left\|\max_{i\neq j}|\zeta(\omega_i^{(1)},\omega_j^{(2)})|\right\|_{L_{2p}}.$$

By Lemma C.11, $\|\zeta(\omega_i,\omega_j)\|_{L_{2p}} \le C_\zeta\frac{p}{D}$. Therefore

$$\left\|\max_{i\neq j}|\zeta(\omega_i^{(1)},\omega_j^{(2)})|\right\|_{L_{2p}} \le \left(\sum_{i\neq j}\mathbf{E}|\zeta(\omega_i^{(1)},\omega_j^{(2)})|^{2p}\right)^{1/(2p)} \le n^{1/p}C_\zeta\frac{p}{D}.$$

Since $n^{1/p}\le e$, we obtain $A_p \le eC_\zeta K\frac{pn^{3/2}}{D}$.

It remains to control the block parameter C_p. Let $\widetilde{G}$ be the $n\times n$ block matrix whose (i,j)-block is $H_{ij}(\omega_i^{(1)},\omega_j^{(2)})$ for $i\neq j$, and zero on the diagonal. The diagonal (i,i)-block of $\mathbf{E}_2\widetilde{G}\widetilde{G}^\top$ is

$$\sum_{k:\,k\neq i}\mathbf{E}_\omega\left[H_{ik}(\omega_i^{(1)},\omega)^2\right] = \frac{1}{4}\sum_{k:\,k\neq i}\mathbf{E}_\omega\zeta(\omega_i^{(1)},\omega)^2C_{ik}^2.$$

By Lemma C.11, $\mathbf{E}_\omega\zeta(\omega_i^{(1)},\omega)^2 \le \frac{8}{D^2}$. Together with $\|C_{ik}\|_{\mathrm{op}} \le Kn$, this shows that each diagonal block has operator norm at most $2K^2\frac{n^3}{D^2}$.

For $i\neq j$, write

$$\Gamma_{ij} := \mathbf{E}_\omega\left[\zeta(\omega_i^{(1)},\omega)\zeta(\omega_j^{(1)},\omega)\right].$$

Then the off-diagonal (i,j)-block of $\mathbf{E}_2\widetilde{G}\widetilde{G}^\top$ is $\mathcal{B}_{ij} := \frac{1}{4}\sum_{k\notin\{i,j\}}\Gamma_{ij}C_{ik}C_{jk}$. By Lemma C.13,

$$|\Gamma_{ij}| \le 8\left(\frac{(\boldsymbol{u}_i^\top\boldsymbol{u}_j)^2 + (\widehat{\boldsymbol{v}}_i^\top\widehat{\boldsymbol{v}}_j)^2}{D^2} + \frac{1}{D^3}\right).$$

Moreover, since $\|C_{ik}C_{jk}\|_{\mathrm{op}} \le K^2n^2$, we have the pointwise bound

$$\|\mathcal{B}_{ij}\|_{\mathrm{op}} \le \frac{1}{4}\sum_{k\notin\{i,j\}}|\Gamma_{ij}|\,\|C_{ik}C_{jk}\|_{\mathrm{op}} \le \frac{1}{4}K^2n^3|\Gamma_{ij}|.$$

Consequently, $\max_{i\neq j}\|\mathcal{B}_{ij}\|_{\mathrm{op}} \le \frac{1}{4}K^2n^3\max_{i\neq j}|\Gamma_{ij}|$.

We now bound the right-hand side in L_p. Let

$$M := \max_{i \neq j} \left\{ |\boldsymbol{u}_i^\top \boldsymbol{u}_j|, |\widehat{\boldsymbol{v}}_i^\top \widehat{\boldsymbol{v}}_j| \right\}.$$

The preceding bound on Γ_{ij} gives $\max_{i \neq j} |\Gamma_{ij}| \leq 8 \left(\frac{M^2}{D^2} + \frac{1}{D^3} \right)$. Therefore

$$\left\| \max_{i \neq j} |\Gamma_{ij}| \right\|_{L_p} \leq 8 \left(\frac{\|M^2\|_{L_p}}{D^2} + \frac{1}{D^3} \right) = 8 \left(\frac{\|M\|_{L_{2p}}^2}{D^2} + \frac{1}{D^3} \right).$$

By Lemma C.12, $\|M\|_{L_{2p}} \leq 2C_s n^{1/p} \sqrt{\frac{p}{D}}$. Since $n^{1/p} \leq e$, this implies $\|M\|_{L_{2p}} \leq 2eC_s \sqrt{\frac{p}{D}}$. Since $C_\zeta = 8C_s^2 + 2$,

$$\left\| \max_{i \neq j} |\Gamma_{ij}| \right\|_{L_p} \leq 8 \left((2eC_s)^2 \frac{p}{D^3} + \frac{1}{D^3} \right) \leq (2e)^2 C_\zeta \frac{p}{D^3}.$$

Combining this with the pointwise block bound gives $\| \max_{i \neq j} \|\mathcal{B}_{ij}\|_{\mathrm{op}} \|_{L_p} \leq e^2 C_\zeta K^2 \frac{n^3 p}{D^3}$.

Combining the diagonal block bound with the off-diagonal block bound and using Lemma C.14, we obtain

$$\left\| \left\| \mathbf{E}_2 \widetilde{G} \widetilde{G}^* \right\|_{\mathrm{op}} \right\|_{L_p} \leq K^2 \left(2 \frac{n^3}{D^2} + e^2 C_\zeta \frac{n^4 p}{D^3} \right) \leq e^2 C_\zeta K^2 \left(\frac{n^3}{D^2} + \frac{n^4 p}{D^3} \right).$$

Since $2D \geq n 2L_\delta \geq np$, we have

$$\left\| \left\| \mathbf{E}_2 \widetilde{G} \widetilde{G}^* \right\|_{\mathrm{op}} \right\|_{L_p} \leq 3e^2 C_\zeta K^2 \frac{n^3}{D^2}.$$

Therefore

$$C_p := \left(\mathbf{E} \left\| \mathbf{E}_2 \widetilde{G} \widetilde{G}^* \right\|_{\mathrm{op}}^p \right)^{1/(2p)} \leq \sqrt{3} e C_\zeta^{1/2} K \frac{n^{3/2}}{D}.$$

Substituting the bounds

$$A_p \leq eC_\zeta K \frac{p n^{3/2}}{D}, \qquad B \leq 2K \frac{n^{3/2}}{D}, \qquad C_p \leq \sqrt{3} e C_\zeta^{1/2} K \frac{n^{3/2}}{D}$$

into the Minsker–Wei inequality gives

$$\begin{aligned} \left(\mathbf{E} \|\mathsf{U}\|_{\mathrm{op}}^{2p} \right)^{1/(2p)} &\leq C_{MW} \left(r^{3/2} A_p + rB + rC_p \right) \\ &\leq \sqrt{3} e C_\zeta C_{MW} K \frac{n^{3/2} p^{5/2}}{D}, \end{aligned} \tag{C.25}$$

where we used $r = p$. By Markov's inequality,

$$\mathbb{P} \left(\|\mathsf{U}\|_{\mathrm{op}} > \sqrt{3} e C_\zeta C_{MW} K \frac{n^{3/2} p^{5/2}}{D} \, \delta^{-1/(2p)} \right) \leq \delta.$$

Since $\delta^{-1/(2p)} \leq \sqrt{e}$, increasing C gives, with probability at least $1 - \delta$, $\|\mathsf{U}\|_{\mathrm{op}} \leq \sqrt{3} e^{3/2} C_\zeta C_{MW} K \frac{n^{3/2} p^{5/2}}{D}$. Finally, since $p = \lceil L_\delta \rceil \leq 2L_\delta$ for $L_\delta \geq 1$, we obtain $\|\mathsf{U}\|_{\mathrm{op}} \leq C_{\mathsf{U}} K \frac{n^{3/2} L_\delta^{5/2}}{D}$ with probability at least $1 - \delta$, where $C_{\mathsf{U}} = 2^{5/2} \sqrt{3} e^{3/2} C_\zeta C_{MW}$.

We now prove the uniform statement. Let $\mathsf{U}(q) := \sum_{i<j} \zeta_{ij} C_{ij}(q)$. Fix $q_\star \in \mathcal{I}$. By the fundamental theorem of calculus, for every $q \in \mathcal{I}$, $\mathsf{U}(q) = \mathsf{U}(q_\star) + \int_{q_\star}^q \mathsf{U}'(t) \, dt$, where $\mathsf{U}'(t) = \sum_{i<j} \zeta_{ij} \partial_q C_{ij}(t)$. Therefore

$$\sup_{q \in \mathcal{I}} \|\mathsf{U}(q)\|_{\mathrm{op}} \leq \|\mathsf{U}(q_\star)\|_{\mathrm{op}} + \int_{\mathcal{I}} \|\mathsf{U}'(t)\|_{\mathrm{op}} \, dt.$$

Taking L_{2p}-norms and using Minkowski's inequality gives

$$\left(\mathbf{E}\sup_{q\in\mathcal{I}}\|\mathsf{U}(q)\|_{\mathrm{op}}^{2p}\right)^{1/(2p)} \le \left(\mathbf{E}\|\mathsf{U}(q_\star)\|_{\mathrm{op}}^{2p}\right)^{1/(2p)} + \int_{\mathcal{I}}\left(\mathbf{E}\|\mathsf{U}'(t)\|_{\mathrm{op}}^{2p}\right)^{1/(2p)}\,dt.$$

The fixed-q moment estimate (C.25) just proved applies to $\mathsf{U}(q_\star)$ with coefficient parameter K_0, and to $\mathsf{U}'(t)$ with coefficient parameter K_1. Hence

$$\left(\mathbf{E}\sup_{q\in\mathcal{I}}\|\mathsf{U}(q)\|_{\mathrm{op}}^{2p}\right)^{1/(2p)} \le \sqrt{3}eC_\zeta C_{MW}(K_0+K_1)\frac{n^{3/2}p^{5/2}}{D}.$$

Applying Markov's inequality,

$$\mathbb{P}\left(\sup_{q\in\mathcal{I}}\|\mathsf{U}(q)\|_{\mathrm{op}} > \sqrt{3}eC_\zeta C_{MW}(K_0+K_1)\frac{n^{3/2}p^{5/2}}{D}\,\delta^{-1/(2p)}\right) \le \delta.$$

Since $\delta^{-1/(2p)} \le \sqrt{e}$, we have, with probability at least $1-\delta$,

$$\sup_{q\in\mathcal{I}}\|\mathsf{U}(q)\|_{\mathrm{op}} \le \sqrt{3}e^{3/2}C_\zeta C_{MW}(K_0+K_1)\frac{n^{3/2}p^{5/2}}{D}.$$

Using $p \le 2L_\delta$, we conclude that, with probability at least $1-\delta$, $\sup_{q\in\mathcal{I}}\|\mathsf{U}(q)\|_{\mathrm{op}} \le C_{\mathsf{U}}(K_0+K_1)\frac{n^{3/2}L_\delta^{5/2}}{D}$. □

Remark 3. *In the case when δ depends on n, e.g. $\delta = \frac{1}{n}$ the lemma still holds by taking $D \gtrsim \log n + \log\frac{1}{\delta}$*

C.5 Uniform local random geometry

The purpose of this section is to establish a high-probability local description of the random objective near the deterministic symmetric path $q \mapsto \overline{\boldsymbol{\alpha}}(q)$, $q \in \mathcal{I}$, where $\mathcal{I}$ is the local window defined in (C.23). The deterministic part of the objective has curvature of order n^2/D in the tangent directions, as shown in Lemma C.9. We show that, on a suitable local box around this path, all random and higher-order error terms are smaller than this curvature scale.

The argument has three components. First, we control the centered quadratic fluctuation $\mathcal{Q}$ and its derivatives along the symmetric path. Second, we show that these bounds remain stable when $\boldsymbol{\alpha} = \overline{\boldsymbol{\alpha}}(q)$ is perturbed by a small tangent vector β. Third, we combine these estimates with the deterministic mean curvature to obtain uniform strong concavity of the noise norm gap $\mathcal{G}$ on the local box.

Define the noise norm gap

$$\mathcal{G}(\boldsymbol{\alpha}) := 1 - \|Z(\boldsymbol{\alpha})\|_1. \tag{C.26}$$

With $\mathcal{H}$ as in (C.19), define

$$\overline{\mathcal{G}}(\boldsymbol{\alpha}) := \frac{1}{D}\mathcal{H}(\boldsymbol{\alpha}), \qquad \mathcal{Q}(\boldsymbol{\alpha}) := \frac{1}{2}\sum_{i<j}h(\alpha_i,\alpha_j)\zeta_{ij}. \tag{C.27}$$

Then Lemma C.7 and (C.24) give

$$\mathcal{G}(\boldsymbol{\alpha}) = \overline{\mathcal{G}}(\boldsymbol{\alpha}) + \mathcal{Q}(\boldsymbol{\alpha}) - R(\boldsymbol{\alpha}). \tag{C.28}$$

We first record deterministic coefficient bounds along the symmetric path. These bounds are used below to control $\mathcal{Q}$ and its derivatives by scalar and matrix U-statistic estimates.

Lemma C.16 (Geometry of the symmetric path)**.** *Assume that $n_+/n, n_-/n \in [\kappa, 1-\kappa]$, and let $\mathcal{I}$ be as in* (C.23)*. Define*

$$c_\alpha := \frac{1/2 - r_0}{1-\kappa}, \qquad C_\alpha := \frac{1/2+r_0}{\kappa}.$$

Then, uniformly over $q \in \mathcal{I}$ and $i \le n$,

$$\frac{c_\alpha}{n} \le \overline{\alpha}_i(q) \le \frac{C_\alpha}{n}, \qquad |\overline{\alpha}_i'(q)| \le \frac{1}{\kappa n}.$$

Moreover, for $h(x,y) = xy/(x+y)$, whenever $c_\alpha/n \le x, y \le C_\alpha/n$, we have $|h_x|, |h_y| \le M_1$, $|h_{xx}|, |h_{xy}|, |h_{yy}| \le M_2 n$, and $|h_{abc}| \le M_3 n^2$ for $a, b, c \in \{x, y\}$, where

$$M_1 := \frac{C_\alpha^2}{4c_\alpha^2}, \qquad M_2 := \frac{C_\alpha^2}{4c_\alpha^3}, \qquad M_3 := \frac{C_\alpha}{2c_\alpha^3} + \frac{3C_\alpha^2}{8c_\alpha^4}.$$

Proof. For $q \in \mathcal{I}$,

$$\overline{\alpha}_i(q) = \begin{cases} q/n_+, & i \in I_+, \\ (1-q)/n_-, & i \in I_-. \end{cases}$$

Since $q \in \mathcal{I}$ and $n_+/n, n_-/n \in [\kappa, 1-\kappa]$, the claimed upper and lower bounds on $\overline{\alpha}_i(q)$ follow immediately. Also $|\overline{\alpha}_i'(q)| \le (\kappa n)^{-1}$.

Next, $h_x(x,y) = \frac{y^2}{(x+y)^2}$, $\quad h_y(x,y) = \frac{x^2}{(x+y)^2}$, so $|h_x|, |h_y| \le M_1$ on the stated region. Also,

$$\nabla^2 h(x,y) = -\frac{2}{(x+y)^3} \begin{pmatrix} y^2 & -xy \\ -xy & x^2 \end{pmatrix},$$

which gives $|h_{xx}|, |h_{xy}|, |h_{yy}| \le M_2 n$.

For the third derivatives, for example, $|h_{xxx}(x,y)| = \frac{6y^2}{(x+y)^4} \le \frac{3C_\alpha^2}{8c_\alpha^4} n^2$, and

$$|h_{xxy}(x,y)| \le \frac{4y}{(x+y)^3} + \frac{6y^2}{(x+y)^4} \le \left(\frac{C_\alpha}{2c_\alpha^3} + \frac{3C_\alpha^2}{8c_\alpha^4} \right) n^2.$$

The other third derivatives are treated identically by symmetry. □

Lemma C.17 (First-derivative coefficients along the symmetric path)**.** *Let M_1, M_2 be as in Lemma C.16. For $i \ne j$, define $c_{ij}(q) := h_x(\overline{\alpha}_i(q), \overline{\alpha}_j(q))$. Then $\partial_i \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q)) = \frac{1}{2} \sum_{j:\, j \ne i} c_{ij}(q) \zeta_{ij}$, and*

$$\sup_{q \in \mathcal{I}} \sup_{i \ne j} |c_{ij}(q)| \le M_1, \qquad \sup_{q \in \mathcal{I}} \sup_{i \ne j} |\partial_q c_{ij}(q)| \le \frac{2M_2}{\kappa}.$$

For $i < j$, define $b_{ij}(q) := \frac{1}{2} \frac{d}{dq} h(\overline{\alpha}_i(q), \overline{\alpha}_j(q))$. Then

$$\partial_q \left[\mathcal{Q}(\overline{\boldsymbol{\alpha}}(q)) \right] = \sum_{i<j} b_{ij}(q) \zeta_{ij},$$

and

$$\sup_{q \in \mathcal{I}} \sup_{i<j} |b_{ij}(q)| \le \frac{M_1}{\kappa n}, \qquad \sup_{q \in \mathcal{I}} \sup_{i<j} |\partial_q b_{ij}(q)| \le \frac{2M_2}{\kappa^2 n}.$$

Proof. The representation of $\partial_i \mathcal{Q}$ follows by differentiating $\mathcal{Q}$ as defined in (C.27). The symmetry of h gives the same coefficient $c_{ij}(q)$ whether $j > i$ or $j < i$.

By Lemma C.16, $|c_{ij}(q)| \le M_1$. Moreover, by the chain rule,

$$|\partial_q c_{ij}(q)| \le M_2 n \left(|\overline{\alpha}_i'(q)| + |\overline{\alpha}_j'(q)| \right) \le \frac{2M_2}{\kappa}.$$

For b_{ij}, we have

$$b_{ij}(q) = \frac{1}{2} \left[h_x(\overline{\alpha}_i(q), \overline{\alpha}_j(q)) \overline{\alpha}_i'(q) + h_y(\overline{\alpha}_i(q), \overline{\alpha}_j(q)) \overline{\alpha}_j'(q) \right].$$

Thus $|b_{ij}(q)| \le M_1/(\kappa n)$. Differentiating once more in q, and using $\overline{\alpha}_i''(q) = 0$, gives $|\partial_q b_{ij}(q)| \le \frac{1}{2} \cdot 4M_2 n \frac{1}{\kappa^2 n^2} = \frac{2M_2}{\kappa^2 n}$. □

Lemma C.18 (Second-derivative coefficients along the symmetric path)**.** *Let M_2, M_3 be as in Lemma C.16. For $i < j$, define $C_{ij}^{(2)}(q) \in \mathbb{R}^{n\times n}$ by*

$$C_{ij}^{(2)}(q) := \frac{1}{2}\left[h_{xx}\boldsymbol{e}_i\boldsymbol{e}_i^\top + h_{xy}\left(\boldsymbol{e}_i\boldsymbol{e}_j^\top + \boldsymbol{e}_j\boldsymbol{e}_i^\top\right) + h_{yy}\boldsymbol{e}_j\boldsymbol{e}_j^\top\right],$$

where the derivatives of h are evaluated at $(\overline{\alpha}_i(q), \overline{\alpha}_j(q))$. Then $\nabla_\alpha^2 \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q)) = \sum_{i<j}\zeta_{ij}C_{ij}^{(2)}(q)$, and

$$\sup_{q\in\mathcal{I}}\sup_{i<j}\|C_{ij}^{(2)}(q)\|_{\mathrm{op}} \le M_2 n, \qquad \sup_{q\in\mathcal{I}}\sup_{i<j}\|\partial_q C_{ij}^{(2)}(q)\|_{\mathrm{op}} \le \frac{2M_3}{\kappa}n.$$

Moreover,

$$\sup_{q\in\mathcal{I}}\left\|\sum_{i<j}\left(C_{ij}^{(2)}(q)\right)^2\right\|_{\mathrm{op}} \le M_2^2 n^3, \qquad \sup_{q\in\mathcal{I}}\left\|\sum_{i<j}\left(\partial_q C_{ij}^{(2)}(q)\right)^2\right\|_{\mathrm{op}} \le \left(\frac{2M_3}{\kappa}\right)^2 n^3.$$

Finally, let $\boldsymbol{\xi}_0 := \frac{d}{dq}\overline{\boldsymbol{\alpha}}(q)$, $\quad \boldsymbol{d}_{ij}(q) := C_{ij}^{(2)}(q)\boldsymbol{\xi}_0$. Then

$$\nabla_\beta \partial_q\left[\mathcal{Q}(\overline{\boldsymbol{\alpha}}(q)+\boldsymbol{\beta})\right]\big|_{\boldsymbol{\beta}=0} = \sum_{i<j}\zeta_{ij}\boldsymbol{d}_{ij}(q).$$

Furthermore,

$$\sup_{q\in\mathcal{I}}\sup_{i<j}\|\boldsymbol{d}_{ij}(q)\|_2 \le \frac{\sqrt{2}M_2}{\kappa}, \qquad \sup_{q\in\mathcal{I}}\sup_{i<j}\|\partial_q\boldsymbol{d}_{ij}(q)\|_2 \le \frac{2\sqrt{2}M_3}{\kappa^2},$$

and

$$\sup_{q\in\mathcal{I}}\left\|\sum_{i<j}\boldsymbol{d}_{ij}(q)\boldsymbol{d}_{ij}(q)^\top\right\|_{\mathrm{op}} \le \frac{M_2^2}{\kappa^2}n^2, \qquad \sup_{q\in\mathcal{I}}\left\|\sum_{i<j}\partial_q\boldsymbol{d}_{ij}(q)\partial_q\boldsymbol{d}_{ij}(q)^\top\right\|_{\mathrm{op}} \le \frac{4M_3^2}{\kappa^4}n^2.$$

Proof. The identity for the Hessian follows by differentiating (C.27) twice in the α-coordinates.

By Lemma C.16, the entries of $C_{ij}^{(2)}(q)$ are bounded in absolute value by $(M_2/2)n$. Since $C_{ij}^{(2)}(q)$ is supported on the two coordinates $\{i, j\}$, its row sums are bounded by $M_2 n$. Hence $\|C_{ij}^{(2)}(q)\|_{\mathrm{op}} \le M_2 n$.

Similarly, the third-derivative bound from Lemma C.16 and the chain rule give, for $a, b \in \{x, y\}$,

$$\left|\frac{d}{dq}h_{ab}(\overline{\alpha}_i(q), \overline{\alpha}_j(q))\right| \le M_3 n^2\left(|\overline{\alpha}_i'(q)| + |\overline{\alpha}_j'(q)|\right) \le \frac{2M_3}{\kappa}n.$$

Since $C_{ij}^{(2)}(q)$ contains an additional factor $1/2$, each entry of $\partial_q C_{ij}^{(2)}(q)$ is bounded by $(M_3/\kappa)n$. Therefore its row sums are bounded by $\frac{2M_3}{\kappa}n$, and hence $\|\partial_q C_{ij}^{(2)}(q)\|_{\mathrm{op}} \le \frac{2M_3}{\kappa}n$.

We next prove the square-sum bounds. Let M_{ij}, $i < j$, be self-adjoint matrices supported on coordinates $\{i, j\}$, and suppose $\|M_{ij}\|_{\mathrm{op}} \le Kn$. Then, for every $\boldsymbol{x} \in \mathbb{R}^n$,

$$\begin{aligned}\boldsymbol{x}^\top\left(\sum_{i<j}M_{ij}^2\right)\boldsymbol{x} &= \sum_{i<j}\|M_{ij}\boldsymbol{x}\|_2^2 \\ &\le K^2n^2\sum_{i<j}(x_i^2 + x_j^2) \\ &= K^2n^2(n-1)\|\boldsymbol{x}\|_2^2 \\ &\le K^2n^3\|\boldsymbol{x}\|_2^2.\end{aligned}$$

Applying this with $M_{ij} = C_{ij}^{(2)}(q)$ and $K = M_2$, and then with $M_{ij} = \partial_q C_{ij}^{(2)}(q)$ and $K = 2M_3/\kappa$, proves the two square-sum bounds.

It remains to identify and bound the mixed-derivative coefficients. For fixed β, differentiating in q gives

$$\partial_q \left[\mathcal{Q}(\overline{\boldsymbol{\alpha}}(q)+\boldsymbol{\beta})\right] = \langle \nabla_\alpha \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q)+\boldsymbol{\beta}), \boldsymbol{\xi}_0 \rangle .$$

Taking the usual gradient with respect to β and setting $\beta = 0$, we obtain

$$\nabla_\beta \partial_q \left[\mathcal{Q}(\overline{\boldsymbol{\alpha}}(q)+\boldsymbol{\beta})\right]|_{\boldsymbol{\beta}=0} = \nabla^2_\alpha \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q))\boldsymbol{\xi}_0.$$

Using the Hessian representation already proved, this becomes

$$\nabla_\beta \partial_q \left[\mathcal{Q}(\overline{\boldsymbol{\alpha}}(q)+\boldsymbol{\beta})\right]|_{\boldsymbol{\beta}=0} = \sum_{i<j} \zeta_{ij} C^{(2)}_{ij}(q)\boldsymbol{\xi}_0 = \sum_{i<j} \zeta_{ij} \boldsymbol{d}_{ij}(q).$$

Since $\|(\xi_0)_{\{i,j\}}\|_2 \le \frac{\sqrt{2}}{\kappa n}$, we have

$$\|\boldsymbol{d}_{ij}(q)\|_2 \le \|C^{(2)}_{ij}(q)\|_{\mathrm{op}} \|(\boldsymbol{\xi}_0)_{\{i,j\}}\|_2 \le \frac{\sqrt{2}M_2}{\kappa}.$$

Because $\boldsymbol{\xi}_0$ is independent of q, $\partial_q \boldsymbol{d}_{ij}(q) = \partial_q C^{(2)}_{ij}(q)\boldsymbol{\xi}_0$, and therefore

$$\|\partial_q \boldsymbol{d}_{ij}(q)\|_2 \le \frac{2M_3}{\kappa} n \frac{\sqrt{2}}{\kappa n} = \frac{2\sqrt{2}M_3}{\kappa^2}.$$

Finally, using the preceding bounds and $\binom{n}{2} \le n^2/2$,

$$\begin{aligned}\left\| \sum_{i<j} \boldsymbol{d}_{ij}(q)\boldsymbol{d}_{ij}(q)^\top \right\|_{\mathrm{op}} &\le \sum_{i<j} \|\boldsymbol{d}_{ij}(q)\|_2^2 \\ &\le \binom{n}{2} \frac{2M_2^2}{\kappa^2} \le \frac{M_2^2}{\kappa^2} n^2.\end{aligned}$$

The same argument gives

$$\begin{aligned}\left\| \sum_{i<j} \partial_q \boldsymbol{d}_{ij}(q)\partial_q \boldsymbol{d}_{ij}(q)^\top \right\|_{\mathrm{op}} &\le \sum_{i<j} \|\partial_q \boldsymbol{d}_{ij}(q)\|_2^2 \\ &\le \binom{n}{2} \frac{8M_3^2}{\kappa^4} \le \frac{4M_3^2}{\kappa^4} n^2.\end{aligned}$$

This completes the proof. □

We next collect derivative bounds for the centered random fluctuation $\mathcal{Q}$ along the symmetric path $q \mapsto \overline{\boldsymbol{\alpha}}(q)$. By the second-order decomposition (C.28), the non-deterministic perturbation of the deterministic averaged objective is described by $\mathcal{Q}$ and the analytic remainder R. Moreover, (C.27) shows that $\mathcal{Q}$ is entirely encoded by the pairwise fluctuations ζ_{ij}. The deterministic curvature estimate in Lemma C.9 shows that the averaged objective has tangent-direction curvature of order n^2/D. Thus the relevant comparison is between this curvature scale and the derivatives of the perturbation terms. These estimates will later be combined with deterministic strong concavity to localize the maximizer near the symmetric path.

For a matrix $M \in \mathbb{R}^{n\times n}$, write $\|M\|_{2\to\infty} := \sup_{\|\boldsymbol{x}\|_2\le 1} \|M\boldsymbol{x}\|_\infty$, equivalently $\|M\|_{2\to\infty} = \max_{1\le i\le n} \|\boldsymbol{e}_i^\top M\|_2$.

Motivated by (C.28)–(C.27), we package the path derivative bounds together with a uniform bound on ζ_{ij} and the Gram operator-norm bounds needed to control R into a single event.

Definition 2 (Basic event)**.** *For $\boldsymbol{b} := \left(b_q,\, b_\nabla,\, b_{H,\mathrm{op}},\, b_{H,2\to\infty},\, b_{\nabla q}\right)$ and $\varepsilon, \nu > 0$, we define $\mathcal{E}(\boldsymbol{b}, \varepsilon, \nu)$ as the event on which, uniformly over $q \in \mathcal{I}$,*

$$|\partial_q[\mathcal{Q}(\overline{\boldsymbol{\alpha}}(q))]| \le b_q, \tag{C.29}$$

$$\|P_{\mathcal{T}} \nabla_\alpha \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q))\|_\infty \le b_\nabla, \qquad \|P_{\mathcal{T}} \nabla_\alpha \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q))\|_2 \le \sqrt{n} b_\nabla, \tag{C.30}$$

$$\|P_{\mathcal{T}} \nabla^2_\alpha \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q)) P_{\mathcal{T}}\|_{\mathrm{op}} \le b_{H,\mathrm{op}}, \qquad \|P_{\mathcal{T}} \nabla^2_\alpha \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q)) P_{\mathcal{T}}\|_{2\to\infty} \le b_{H,2\to\infty}, \tag{C.31}$$

$$\left\| P_{\mathcal{T}}\, \nabla_\beta \partial_q [\mathcal{Q}(\overline{\boldsymbol{\alpha}}(q)+\boldsymbol{\beta})]|_{\boldsymbol{\beta}=0} \right\|_2 \le b_{\nabla q}, \tag{C.32}$$

and

$$\|\Delta_U\|_{\mathrm{op}} \vee \|\Delta_V\|_{\mathrm{op}} \le \varepsilon, \tag{C.33}$$

$$\max_{i<j} |\zeta_{ij}| \le \nu. \tag{C.34}$$

Lemma C.19 (High-probability basic event)**.** *There exist constants $c_{\mathrm{path}}, K_q, K_\nabla, K_H, K_{H,\infty}, K_{\nabla q} > 0$, depending only on κ, r_0, and universal constants $K_G, K_\zeta > 0$, such that the following holds. Let $0 < \delta < e^{-1}$ and set $L := \log(n/\delta)$. Let $\boldsymbol{b}_{n,D,\delta}$ be the vector in Definition 2 whose components are*

$$\begin{aligned} b_q &:= K_q \frac{L^{5/2}}{D}, \qquad b_{\nabla q} := K_{\nabla q} \frac{nL^{5/2}}{D}, \qquad b_\nabla := K_\nabla \frac{\sqrt{nL} + L}{D}, \\ b_{H,\mathrm{op}} &:= K_H \frac{n^{3/2} L^{5/2}}{D}, \qquad b_{H,2\to\infty} := K_{H,\infty} \frac{n^{3/2} L}{D}. \end{aligned} \tag{C.35}$$

Define also

$$\varepsilon_{n,D,\delta} := K_G \sqrt{\frac{n+L}{D}}, \qquad \nu_{n,D,\delta} := K_\zeta \frac{L}{D}. \tag{C.36}$$

If $D \ge c_{\mathrm{path}} nL$, then

$$\mathbb{P}\left(\mathcal{E}(\boldsymbol{b}_{n,D,\delta}, \varepsilon_{n,D,\delta}, \nu_{n,D,\delta})\right) \ge 1 - \delta.$$

Proof. Let $\mathcal{N} \subset \mathcal{I}$ be a grid of mesh n^{-1}. Since $|\mathcal{I}| < 1$, we have $|\mathcal{N}| \le n$. We shall use six events, each with failure probability at most $\delta/6$. We repeatedly use the elementary bounds

$$\log \frac{6n}{\delta} = L + \log 6 \le 3L, \tag{C.37}$$

and

$$\log \frac{12n|\mathcal{N}|}{\delta} \le \log \frac{12n^2}{\delta} = L + \log 12 + \log n \le 5L. \tag{C.38}$$

Here we used $0 < \delta < e^{-1}$. We also use

$$\log \frac{24n^2}{\delta} = L + \log(24n) \le 6L. \tag{C.39}$$

Let c_G be the constant required in Lemma H.5 when the failure probability is $\delta/6$. Let $c_q, c_H, c_{\nabla q}$ be the constants required in the moving-coefficient estimates below, again with failure probability $\delta/6$. We choose

$$c_{\mathrm{path}} \ge \max\{6c_G, 3c_q, 3c_H, 3c_{\nabla q}\}. \tag{C.40}$$

Then $D \ge c_{\mathrm{path}} nL$, together with (C.37) and (C.39), allows all auxiliary estimates below to be applied.

First consider the gradient. By Lemma C.17,

$$\partial_i \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q)) = \frac{1}{2} \sum_{j:\, j\neq i} c_{ij}(q)\zeta_{ij}, \qquad |c_{ij}(q)| \le M_1, \qquad |\partial_q c_{ij}(q)| \le \frac{2M_2}{\kappa}.$$

By Lemma C.11, conditional on $(\boldsymbol{u}_i, \widehat{\boldsymbol{v}}_i)$, the variables ζ_{ij}, $j \neq i$, are independent and centered. The same lemma and Bernstein's inequality give

$$\mathbb{P}\left(|\partial_i \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q))| > C_1 \frac{\sqrt{ns} + s}{D} \right) \le 2e^{-s}, \tag{C.41}$$

where $C_1 \ge M_1 \max\{c_1^{-1}, c_1^{-1/2}\}$. Here $c_1 > 0$ is the absolute Bernstein constant.

Applying (C.41) with $s = \log \frac{12n|\mathcal{N}|}{\delta}$ and taking a union bound over $i \le n$ and $q \in \mathcal{N}$, we obtain an event $\mathcal{E}_{\nabla,\mathrm{grid}}$ satisfying

$$\mathbb{P}(\mathcal{E}^c_{\nabla,\mathrm{grid}}) \le \frac{\delta}{6}.$$

On this event, by (C.38),

$$\max_{q\in\mathcal{N}}\max_{i\le n}|\partial_i\mathcal{Q}(\overline{\boldsymbol{\alpha}}(q))|\le 5C_1\frac{\sqrt{nL}+L}{D}. \tag{C.42}$$

Next define the spherical-geometry event. By Lemma H.5, applied with failure probability $\delta/6$, there is an event $\mathcal{E}_G$ satisfying $\mathbb{P}(\mathcal{E}_G^{\mathrm{c}})\le\delta/6$ on which

$$\|\Delta_U\|_{\mathrm{op}}\vee\|\Delta_V\|_{\mathrm{op}}\le C_G\sqrt{\frac{n+\log(6n/\delta)}{D}}$$

and

$$\max_{i\neq j}\left\{|\boldsymbol{u}_i^\top\boldsymbol{u}_j|,|\widehat{\boldsymbol{v}}_i^\top\widehat{\boldsymbol{v}}_j|\right\}\le C_G\sqrt{\frac{\log(24n^2/\delta)}{D}},$$

where C_G is universal. By (C.37), (C.39), and $\sqrt{6}\le 3$, we may take

$$\|\Delta_U\|_{\mathrm{op}}\vee\|\Delta_V\|_{\mathrm{op}}\le K_G\sqrt{\frac{n+L}{D}},\qquad K_G:=3C_G, \tag{C.43}$$

and

$$\max_{i\neq j}\left\{|\boldsymbol{u}_i^\top\boldsymbol{u}_j|,|\widehat{\boldsymbol{v}}_i^\top\widehat{\boldsymbol{v}}_j|\right\}\le 3C_G\sqrt{\frac{L}{D}}. \tag{C.44}$$

Consequently, on $\mathcal{E}_G$,

$$\max_{i<j}|\zeta_{ij}|\le K_\zeta\frac{L}{D},\qquad K_\zeta:=36C_G^2+2. \tag{C.45}$$

We now pass the gradient estimate from the grid to the whole interval. Fix $i\le n$ and set $f_i(q):=\partial_i\mathcal{Q}(\overline{\boldsymbol{\alpha}}(q))$. Using Lemma C.17 and (C.45), on $\mathcal{E}_G$,

$$\sup_{q\in\mathcal{I}}|f_i'(q)|\le\frac{K_\zeta M_2}{\kappa}\frac{nL}{D}. \tag{C.46}$$

For arbitrary $q\in\mathcal{I}$, choose $q_0\in\mathcal{N}$ with $|q-q_0|\le n^{-1}$. Combining (C.42) and (C.46) gives, uniformly over $q\in\mathcal{I}$,

$$\|\nabla_\alpha\mathcal{Q}(\overline{\boldsymbol{\alpha}}(q))\|_\infty\le\left(5C_1+\frac{K_\zeta M_2}{\kappa}\right)\frac{\sqrt{nL}+L}{D}.$$

Since $\|P_{\mathcal{T}}\boldsymbol{x}\|_\infty\le 2\|\boldsymbol{x}\|_\infty$, we obtain

$$\|P_{\mathcal{T}}\nabla_\alpha\mathcal{Q}(\overline{\boldsymbol{\alpha}}(q))\|_\infty\le K_\nabla\frac{\sqrt{nL}+L}{D},\qquad K_\nabla:=2\left(5C_1+\frac{K_\zeta M_2}{\kappa}\right). \tag{C.47}$$

The Euclidean bound follows from (C.47) and $\|P_{\mathcal{T}}\boldsymbol{x}\|_2\le\|\boldsymbol{x}\|_2\le\sqrt{n}\|\boldsymbol{x}\|_\infty$:

$$\|P_{\mathcal{T}}\nabla_\alpha\mathcal{Q}(\overline{\boldsymbol{\alpha}}(q))\|_2\le\sqrt{n}K_\nabla\frac{\sqrt{nL}+L}{D}.$$

Next consider the scalar q-derivative. By Lemma C.17,

$$\partial_q[\mathcal{Q}(\overline{\boldsymbol{\alpha}}(q))]=\sum_{i<j}b_{ij}(q)\zeta_{ij},\qquad|b_{ij}(q)|\le\frac{B_0}{n},\qquad|\partial_q b_{ij}(q)|\le\frac{B_1}{n},$$

where

$$B_0:=\frac{M_1}{\kappa},\qquad B_1:=\frac{2M_2}{\kappa^2}.$$

Moreover,

$$\sup_{q\in\mathcal{I}}\sum_{i<j}b_{ij}(q)^2\le B_0^2,\qquad\sup_{q\in\mathcal{I}}\sum_{i<j}(\partial_q b_{ij}(q))^2\le B_1^2.$$

Applying the moving-coefficient estimate in Lemma C.15 with failure probability $\delta/6$ and coefficient scales

$$K_0 := \frac{B_0}{n^{3/2}}, \qquad K_1 := \frac{B_1}{n^{3/2}},$$

gives an event $\mathcal{E}_q$ satisfying $\mathbb{P}(\mathcal{E}_q^c) \le \delta/6$ on which

$$\sup_{q\in\mathcal{I}} |\partial_q[\mathcal{Q}(\overline{\boldsymbol{\alpha}}(q))]| \le C_U(B_0+B_1)\frac{\log^{5/2}(6n/\delta)}{D}.$$

By (C.37), this yields

$$\sup_{q\in\mathcal{I}} |\partial_q[\mathcal{Q}(\overline{\boldsymbol{\alpha}}(q))]| \le K_q\frac{L^{5/2}}{D}, \qquad K_q := 3^{5/2}C_U(B_0+B_1). \tag{C.48}$$

Next consider the Hessian operator norm. By Lemma C.18,

$$\nabla_\alpha^2\mathcal{Q}(\overline{\boldsymbol{\alpha}}(q)) = \sum_{i<j}\zeta_{ij}C_{ij}^{(2)}(q),$$

where

$$\sup_{q\in\mathcal{I}}\sup_{i<j}\|C_{ij}^{(2)}(q)\|_{\mathrm{op}} \le H_0 n, \qquad \sup_{q\in\mathcal{I}}\sup_{i<j}\|\partial_q C_{ij}^{(2)}(q)\|_{\mathrm{op}} \le H_1 n,$$

with

$$H_0 := M_2, \qquad H_1 := \frac{2M_3}{\kappa}.$$

Moreover,

$$\sup_{q\in\mathcal{I}}\left\|\sum_{i<j}(C_{ij}^{(2)}(q))^2\right\|_{\mathrm{op}} \le H_0^2 n^3, \qquad \sup_{q\in\mathcal{I}}\left\|\sum_{i<j}(\partial_q C_{ij}^{(2)}(q))^2\right\|_{\mathrm{op}} \le H_1^2 n^3.$$

Applying the moving-coefficient matrix estimate in Lemma C.15 with failure probability $\delta/6$ and coefficient scales

$$K_0 := H_0, \qquad K_1 := H_1,$$

gives an event $\mathcal{E}_{H,\mathrm{op}}$ satisfying $\mathbb{P}(\mathcal{E}_{H,\mathrm{op}}^c) \le \delta/6$ on which

$$\sup_{q\in\mathcal{I}}\left\|\nabla_\alpha^2\mathcal{Q}(\overline{\boldsymbol{\alpha}}(q))\right\|_{\mathrm{op}} \le C_U(H_0+H_1)\frac{n^{3/2}\log^{5/2}(6n/\delta)}{D}.$$

Since $P_{\mathcal{T}}$ is an orthogonal projection, the same bound holds after multiplication by $P_{\mathcal{T}}$ on both sides. Hence, by (C.37),

$$\sup_{q\in\mathcal{I}}\left\|P_{\mathcal{T}}\nabla_\alpha^2\mathcal{Q}(\overline{\boldsymbol{\alpha}}(q))P_{\mathcal{T}}\right\|_{\mathrm{op}} \le K_H\frac{n^{3/2}L^{5/2}}{D}, \qquad K_H := 3^{5/2}C_U(H_0+H_1). \tag{C.49}$$

We now prove the $2\to\infty$ Hessian bound. Write

$$H(q) := \nabla_\alpha^2\mathcal{Q}(\overline{\boldsymbol{\alpha}}(q)).$$

Recall that the off-diagonal entries $C_{ab}^{(2)}(q)$ is supported only on the coordinates $\{a,b\}$. Hence, for $i\neq j$,

$$H_{ij}(q) = \zeta_{ij}\big(C_{ij}^{(2)}(q)\big)_{ij}.$$

Combined with $\|C_{ij}^{(2)}(q)\|_{\mathrm{op}} \le H_0 n$ and (C.45), on $\mathcal{E}_G$,

$$|H_{ij}(q)| \le K_\zeta H_0\frac{nL}{D}.$$

Therefore, uniformly over $q\in\mathcal{I}$ and $i\le n$,

$$\left(\sum_{j\neq i}|H_{ij}(q)|^2\right)^{1/2} \le K_\zeta H_0\frac{n^{3/2}L}{D}. \tag{C.50}$$

It remains to control the diagonal entries. For fixed i, we can write

$$H_{ii}(q) = \sum_{j:\, j\neq i} a_{ij}(q)\zeta_{ij}, \qquad |a_{ij}(q)| \le H_0 n, \qquad |\partial_q a_{ij}(q)| \le H_1 n.$$

The same conditional Bernstein argument as in (C.41) gives, for every fixed i, q,

$$\mathbb{P}\left(|H_{ii}(q)| > C_3 \frac{n(\sqrt{ns}+s)}{D}\right) \le 2e^{-s}, \tag{C.51}$$

where $C_3 \ge H_0 \max\{c_1^{-1}, c_1^{-1/2}\}$. Applying (C.51) with $s = \log \frac{12n|\mathcal{N}|}{\delta}$ and taking a union bound over $i \le n$ and $q \in \mathcal{N}$, we obtain an event $\mathcal{E}_{\mathrm{diag,grid}}$ satisfying

$$\mathbb{P}(\mathcal{E}^c_{\mathrm{diag,grid}}) \le \frac{\delta}{6}.$$

On this event, by (C.38),

$$\max_{q\in\mathcal{N}} \max_{i\le n} |H_{ii}(q)| \le C_3 \frac{n(3\sqrt{nL}+5L)}{D}. \tag{C.52}$$

We next pass (C.52) from the grid to the whole interval. Set $g_i(q) := H_{ii}(q)$. By the derivative bound on a_{ij} and (C.45), on $\mathcal{E}_G$,

$$\sup_{q\in\mathcal{I}} |g_i'(q)| \le K_\zeta H_1 \frac{n^2 L}{D}. \tag{C.53}$$

For arbitrary $q \in \mathcal{I}$, choose $q_0 \in \mathcal{N}$ with $|q - q_0| \le n^{-1}$. Combining (C.52) and (C.53) gives

$$|H_{ii}(q)| \le C_3 \frac{n(3\sqrt{nL}+5L)}{D} + K_\zeta H_1 \frac{nL}{D}.$$

Since

$$3\sqrt{nL} + 5L \le 10\sqrt{n}\, L \qquad \text{and} \qquad nL \le n^{3/2}L,$$

we have, uniformly over $q \in \mathcal{I}$ and $i \le n$,

$$|H_{ii}(q)| \le (10C_3 + K_\zeta H_1)\frac{n^{3/2}L}{D}. \tag{C.54}$$

Combining (C.50) and (C.54), and using $\sqrt{a^2+b^2} \le a+b$, we obtain

$$\sup_{q\in\mathcal{I}} \|H(q)\|_{2\to\infty} \le K^{(0)}_{H,\infty} \frac{n^{3/2}L}{D}, \tag{C.55}$$

where

$$K^{(0)}_{H,\infty} := 10C_3 + K_\zeta H_1 + K_\zeta H_0.$$

Finally,

$$\|P_{\mathcal{T}}\boldsymbol{x}\|_\infty \le 2\|\boldsymbol{x}\|_\infty, \qquad \|P_{\mathcal{T}}\boldsymbol{x}\|_2 \le \|\boldsymbol{x}\|_2.$$

Therefore (C.55) implies

$$\sup_{q\in\mathcal{I}} \left\|P_{\mathcal{T}}\nabla^2_\alpha \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q))P_{\mathcal{T}}\right\|_{2\to\infty} \le K_{H,\infty}\frac{n^{3/2}L}{D}, \tag{C.56}$$

where

$$K_{H,\infty} := 2K^{(0)}_{H,\infty} = 2(10C_3 + K_\zeta H_1 + K_\zeta H_0).$$

Finally consider the mixed derivative. By Lemma C.18,

$$\nabla_\beta \partial_q \left[\mathcal{Q}(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta})\right]\big|_{\boldsymbol{\beta}=0} = \sum_{i<j} \zeta_{ij}\boldsymbol{d}_{ij}(q).$$

Set

$$D_0 := \frac{\sqrt{2}M_2}{\kappa}, \qquad D_1 := \frac{2\sqrt{2}M_3}{\kappa^2}.$$

Then

$$\sup_{q\in\mathcal{I}}\sup_{i<j}\|\boldsymbol{d}_{ij}(q)\|_2 \le D_0, \qquad \sup_{q\in\mathcal{I}}\sup_{i<j}\|\partial_q\boldsymbol{d}_{ij}(q)\|_2 \le D_1,$$

and

$$\sup_{q\in\mathcal{I}}\left\|\sum_{i<j}\boldsymbol{d}_{ij}(q)\boldsymbol{d}_{ij}(q)^\top\right\|_{\mathrm{op}} \le D_0^2n^2, \qquad \sup_{q\in\mathcal{I}}\left\|\sum_{i<j}\partial_q\boldsymbol{d}_{ij}(q)\partial_q\boldsymbol{d}_{ij}(q)^\top\right\|_{\mathrm{op}} \le D_1^2n^2.$$

For each $i<j$, introduce the self-adjoint dilation

$$\mathsf{M}_{ij}(q) := \begin{pmatrix} 0 & \boldsymbol{d}_{ij}(q)^\top \\ \boldsymbol{d}_{ij}(q) & 0 \end{pmatrix}.$$

Since

$$\|\mathsf{M}_{ij}(q)\|_{\mathrm{op}} = \|\boldsymbol{d}_{ij}(q)\|_2 \le D_0$$

and

$$\left\|\sum_{i<j}\mathsf{M}_{ij}(q)^2\right\|_{\mathrm{op}} \le D_0^2n^2,$$

with the analogous bounds for $\partial_q\mathsf{M}_{ij}(q)$ with D_1, the moving-coefficient estimate in Lemma C.15 applies with coefficient scales

$$K_0 := \frac{D_0}{\sqrt{n}}, \qquad K_1 := \frac{D_1}{\sqrt{n}}.$$

Therefore, with failure probability at most $\delta/6$,

$$\sup_{q\in\mathcal{I}}\left\|\sum_{i<j}\zeta_{ij}\mathsf{M}_{ij}(q)\right\|_{\mathrm{op}} \le C_U(D_0+D_1)\frac{n\log^{5/2}(6n/\delta)}{D}.$$

By (C.37), there is an event $\mathcal{E}_{\nabla q}$ satisfying $\mathbb{P}(\mathcal{E}_{\nabla q}^c)\le\delta/6$ on which

$$\sup_{q\in\mathcal{I}}\left\|\sum_{i<j}\zeta_{ij}\mathsf{M}_{ij}(q)\right\|_{\mathrm{op}} \le K_{\nabla q}\frac{nL^{5/2}}{D}, \qquad K_{\nabla q} := 3^{5/2}C_U(D_0+D_1).$$

By the definition of the dilation,

$$\left\|\sum_{i<j}\zeta_{ij}\mathsf{M}_{ij}(q)\right\|_{\mathrm{op}} = \left\|\sum_{i<j}\zeta_{ij}\boldsymbol{d}_{ij}(q)\right\|_2.$$

Since $P_{\mathcal{T}}$ is an orthogonal projection, we conclude that

$$\sup_{q\in\mathcal{I}}\left\|P_{\mathcal{T}}\,\nabla_\beta\partial_q\left[\mathcal{Q}(\overline{\boldsymbol{\alpha}}(q)+\boldsymbol{\beta})\right]\big|_{\boldsymbol{\beta}=0}\right\|_2 \le K_{\nabla q}\frac{nL^{5/2}}{D}. \tag{C.57}$$

Define

$$\mathcal{E}_0 := \mathcal{E}_{\nabla,\mathrm{grid}}\cap\mathcal{E}_G\cap\mathcal{E}_q\cap\mathcal{E}_{H,\mathrm{op}}\cap\mathcal{E}_{\mathrm{diag,grid}}\cap\mathcal{E}_{\nabla q}.$$

Each event has complement probability at most $\delta/6$, so

$$\mathbb{P}(\mathcal{E}_0^c)\le\delta.$$

On $\mathcal{E}_0$, (C.43) and (C.45) give the two geometric bounds in Definition 2, while (C.48), (C.47), (C.49), (C.56), and (C.57) give its path bounds. Hence

$$\mathcal{E}(\boldsymbol{b}_{n,D,\delta},\varepsilon_{n,D,\delta},\nu_{n,D,\delta})$$

holds on $\mathcal{E}_0$, which proves the lemma. □

C.6 Uniform strong concavity

Throughout the rest of the proof, we condition on a realization of the corruption signs satisfying

$$\kappa n \le n_- < n_+ \le (1-\kappa)n \tag{C.58}$$

for some fixed $0 < \kappa < 1/2$.

For deterministic event parameters $\boldsymbol{b}, \varepsilon, \nu$, we further work on the basic event $\mathcal{E}(\boldsymbol{b}, \varepsilon, \nu)$ from Definition 2. From this point until the final probability assembly, all arguments are deterministic under these conditions.

Define

$$\mathfrak{D}(r) := \frac{D}{n^2} b_{H,\mathrm{op}} + nr + Dn\nu r + \frac{D}{n}\varepsilon^3. \tag{C.59}$$

Let $\varrho_0 > 0$ and $K_R < \infty$ be the constants from Lemma C.7 corresponding to the coordinate range $[c_\alpha/(2n), 2C_\alpha/n]$. Throughout, $r_0 \in (0, 1/4)$ and the local window $\mathcal{I}$ are fixed as in (C.23).

Lemma C.20 (Uniformity over the local box)**.** *Let $r_\infty > 0$ and*

$$\mathcal{B}^\infty(r_\infty) := \{(q, \boldsymbol{\beta}) : q \in \mathcal{I},\ \boldsymbol{\beta} \in \mathcal{T},\ \|\boldsymbol{\beta}\|_\infty \le r_\infty\}.$$

There exists $C_{\mathrm{box}} \ge 1$, depending only on κ, r_0, such that, if $\mathcal{E}(\boldsymbol{b}, \varepsilon, \nu)$ holds, $\varepsilon \le \varrho_0$, and $nr_\infty \le c_\alpha/2$, then

$$\sup_{(q,\boldsymbol{\beta})\in\mathcal{B}^\infty(r_\infty)} \|P_{\mathcal{T}}[\nabla^2_\alpha \overline{\mathcal{G}}(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta}) - \nabla^2_\alpha \overline{\mathcal{G}}(\overline{\boldsymbol{\alpha}}(q))]P_{\mathcal{T}}\|_{\mathrm{op}} \le C_{\mathrm{box}} \frac{n^3}{D} r_\infty, \tag{C.60}$$

$$\sup_{(q,\boldsymbol{\beta})\in\mathcal{B}^\infty(r_\infty)} \|P_{\mathcal{T}}[\nabla^2_\alpha \overline{\mathcal{G}}(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta}) - \nabla^2_\alpha \overline{\mathcal{G}}(\overline{\boldsymbol{\alpha}}(q))]P_{\mathcal{T}}\|_{\infty\to\infty} \le C_{\mathrm{box}} \frac{n^3}{D} r_\infty, \tag{C.61}$$

$$\sup_{(q,\boldsymbol{\beta})\in\mathcal{B}^\infty(r_\infty)} \|\nabla^2_\alpha \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta}) - \nabla^2_\alpha \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q))\|_{\mathrm{op}} \le C_{\mathrm{box}} n^3 \nu r_\infty, \tag{C.62}$$

$$\sup_{(q,\boldsymbol{\beta})\in\mathcal{B}^\infty(r_\infty)} \|\nabla^2_\alpha \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta}) - \nabla^2_\alpha \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q))\|_{\infty\to\infty} \le C_{\mathrm{box}} n^3 \nu r_\infty, \tag{C.63}$$

and

$$\sup_{(q,\boldsymbol{\beta})\in\mathcal{B}^\infty(r_\infty)} \|\nabla^2_\alpha R(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta})\|_{\mathrm{op}} \le C_{\mathrm{box}} n\varepsilon^3, \tag{C.64}$$

$$\sup_{(q,\boldsymbol{\beta})\in\mathcal{B}^\infty(r_\infty)} \|\nabla^2_\alpha R(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta})\|_{\infty\to\infty} \le C_{\mathrm{box}} n\varepsilon^3. \tag{C.65}$$

Proof. By Lemma C.16 and $nr_\infty \le c_\alpha/2$, every coordinate in the local box satisfies

$$\frac{c_\alpha}{2n} \le \overline{\alpha}_i(q) + \beta_i \le \frac{2C_\alpha}{n}. \tag{C.66}$$

Hence every third derivative of h on this region is bounded in absolute value by $C_h n^2$, where C_h depends only on κ, r_0.

First we control the local variation of $\mathcal{Q}$. Let

$$A_Q(q, \boldsymbol{\beta}) := \nabla^2_\alpha \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta}) - \nabla^2_\alpha \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q)).$$

For $i < j$, let $C^{(2)}_{ij}(\boldsymbol{\alpha})$ denote the pairwise Hessian block, supported on $\{i, j\} \times \{i, j\}$. By the mean-value theorem and the preceding third-derivative bound, for $(k, \ell) \in \{i, j\} \times \{i, j\}$,

$$\left|[C^{(2)}_{ij}(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta}) - C^{(2)}_{ij}(\overline{\boldsymbol{\alpha}}(q))]_{k\ell}\right| \le C_h n^2 r_\infty. \tag{C.67}$$

By Definition 2, $\max_{i<j} |\zeta_{ij}| \le \nu$. For each row, only pairs containing that coordinate contribute, and each such pair contributes to at most two entries. Thus (C.67) gives, uniformly on the local box,

$$\|A_Q(q, \boldsymbol{\beta})\|_{\infty\to\infty} \le 2C_h n^3 \nu r_\infty. \tag{C.68}$$

Since $A_Q(q, \boldsymbol{\beta})$ is symmetric, its operator norm is bounded by the same row-sum bound. This proves (C.62) and (C.63) after enlarging C_{box}.

The corresponding bounds (C.60)–(C.61) follow from the same argument, replacing ζ_{ij} with deterministic $2/D$ and noting $P_{\mathcal{T}}$ is an orthogonal projection and $\|P_{\mathcal{T}} x\|_\infty \le 2\|x\|_\infty$, whence $\|P_{\mathcal{T}} A P_{\mathcal{T}}\|_{\infty\to\infty} \le 4\|A\|_{\infty\to\infty}$.

It remains to control the analytic remainder. By Definition 2, $\|\Delta_U\|_{\mathrm{op}} \vee \|\Delta_V\|_{\mathrm{op}} \le \varepsilon \le \varrho_0$. Therefore Lemma C.7 applies uniformly on the coordinate window (C.66), and therefore

$$\sup_{\mathcal{B}^\infty(r_\infty)} \|\nabla_\alpha^2 R\|_{\mathrm{op}} \vee \sup_{\mathcal{B}^\infty(r_\infty)} \|\nabla_\alpha^2 R\|_{\infty\to\infty} \le K_R n\varepsilon^3.$$

Choosing $C_{\mathrm{box}} \ge \max\{1, 16C_h, K_R\}$ proves (C.64) and (C.65). □

Proposition C.1 (Uniform strong concavity)**.** *Let $c_H > 0$ be the constant from Lemma C.9 such that $-P_{\mathcal{T}}\nabla_\alpha^2 \overline{\mathcal{G}}(\overline{\boldsymbol{\alpha}}(q))P_{\mathcal{T}} \succeq c_H(n^2/D)I_{\mathcal{T}}$ uniformly over $q \in \mathcal{I}$. There exists $c_{\mathrm{sc}} > 0$, depending only on κ, r_0, such that the following holds. Suppose $\mathcal{E}(\boldsymbol{b}, \varepsilon, \nu)$ holds, $\varepsilon \le \varrho_0$, and*

$$\mathfrak{D}(r_\infty) \le c_{\mathrm{sc}}. \tag{C.69}$$

Then, uniformly over $(q, \boldsymbol{\beta}) \in \mathcal{B}^\infty(r_\infty)$,

$$-P_{\mathcal{T}}\nabla_\alpha^2 \mathcal{G}(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta})P_{\mathcal{T}} \succeq \frac{c_H}{2}\frac{n^2}{D} I_{\mathcal{T}}.$$

Proof. Let $C_{\mathrm{box}} \ge 1$ be the constant from Lemma C.20. Choose $c_{\mathrm{sc}} > 0$ so small that

$$c_{\mathrm{sc}} \le \min\left\{\frac{c_\alpha}{2}, \frac{c_H}{8C_{\mathrm{box}}}\right\}.$$

Since $nr_\infty \le \mathfrak{D}(r_\infty) \le c_{\mathrm{sc}}$, the smallness assumption also guarantees the coordinate condition needed below.

First, Lemma C.9 and Lemma C.20 give

$$-P_{\mathcal{T}}\nabla_\alpha^2 \overline{\mathcal{G}}(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta})P_{\mathcal{T}} \succeq (c_H - C_{\mathrm{box}} n r_\infty)\frac{n^2}{D} I_{\mathcal{T}} \succeq \frac{7c_H}{8}\frac{n^2}{D} I_{\mathcal{T}}.$$

Next, by Definition 2,

$$\sup_{q\in\mathcal{I}} \|P_{\mathcal{T}}\nabla_\alpha^2 \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q))P_{\mathcal{T}}\|_{\mathrm{op}} \le b_{H,\mathrm{op}} \le \frac{c_H}{8}\frac{n^2}{D},$$

where the last inequality follows from $(D/n^2)b_{H,\mathrm{op}} \le \mathfrak{D}(r_\infty) \le c_{\mathrm{sc}}$ and $C_{\mathrm{box}} \ge 1$. Lemma C.20 similarly gives

$$\sup_{\mathcal{B}^\infty(r_\infty)} \|P_{\mathcal{T}}[\nabla_\alpha^2 \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta}) - \nabla_\alpha^2 \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q))]P_{\mathcal{T}}\|_{\mathrm{op}} \le \frac{c_H}{8}\frac{n^2}{D},$$

because $Dn\nu r_\infty \le \mathfrak{D}(r_\infty)$, and

$$\sup_{\mathcal{B}^\infty(r_\infty)} \|P_{\mathcal{T}}\nabla_\alpha^2 R(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta})P_{\mathcal{T}}\|_{\mathrm{op}} \le \frac{c_H}{8}\frac{n^2}{D},$$

because $(D/n)\varepsilon^3 \le \mathfrak{D}(r_\infty)$.

Finally, by (C.28),

$$-P_{\mathcal{T}}\nabla_\alpha^2 \mathcal{G} P_{\mathcal{T}} = -P_{\mathcal{T}}\nabla_\alpha^2 \overline{\mathcal{G}} P_{\mathcal{T}} - P_{\mathcal{T}}\nabla_\alpha^2 \mathcal{Q} P_{\mathcal{T}} + P_{\mathcal{T}}\nabla_\alpha^2 R P_{\mathcal{T}}.$$

Combining the preceding four bounds yields $-P_{\mathcal{T}}\nabla_\alpha^2 \mathcal{G} P_{\mathcal{T}} \succeq (c_H/2)(n^2/D)I_{\mathcal{T}}$, as claimed. □

C.7 THE UNIQUE INNER OPTIMIZER

For fixed $q \in \mathcal{I}$, we optimize $\boldsymbol{\beta} \mapsto \mathcal{G}(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta})$ over the tangent space $\mathcal{T}$. Its first-order condition is

$$F_q(\boldsymbol{\beta}) := P_{\mathcal{T}} \nabla \mathcal{G}(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta}) = 0.$$

We linearize this equation around $\boldsymbol{\beta} = 0$, using as the reference Hessian the deterministic mean curvature

$$L_q := -P_{\mathcal{T}} \nabla_\alpha^2 \overline{\mathcal{G}}(\overline{\boldsymbol{\alpha}}(q)) P_{\mathcal{T}}|_{\mathcal{T}}.$$

By Lemma C.9, L_q is uniformly positive definite and therefore invertible on $\mathcal{T}$. We collect the discrepancy between F_q and this linear approximation into

$$\mathfrak{r}_q(\boldsymbol{\beta}) := F_q(\boldsymbol{\beta}) - F_q(0) + L_q \boldsymbol{\beta}.$$

Thus

$$F_q(\boldsymbol{\beta}) = F_q(0) - L_q \boldsymbol{\beta} + \mathfrak{r}_q(\boldsymbol{\beta}),$$

where $F_q(0)$ is the forcing term at the reference path, $L_q \boldsymbol{\beta}$ is the leading linear response, and $\mathfrak{r}_q(\boldsymbol{\beta})$ contains the remaining local variation. Thus solving $F_q(\boldsymbol{\beta}) = 0$ is equivalent to finding a fixed point of

$$T_q(\boldsymbol{\beta}) := L_q^{-1} \big[F_q(0) + \mathfrak{r}_q(\boldsymbol{\beta}) \big].$$

We will prove existence and uniqueness of the inner optimizer by showing that T_q is a contraction on a suitable neighborhood of the origin.

Lemma C.21 (Mixed-norm control of the fixed-point remainder)**.** *There exists $C_{\mathrm{rem}} > 0$, depending only on κ, r_0, such that the following holds. Suppose $\mathcal{E}(\boldsymbol{b}, \varepsilon, \nu)$ holds, $\varepsilon \le \varrho_0$, and $\mathfrak{D}(r_\infty) \le c_{\mathrm{sc}}$. Then, uniformly over $q \in \mathcal{I}$ and $\boldsymbol{\beta}, \widetilde{\boldsymbol{\beta}} \in \mathcal{B}^\infty(r_\infty)$,*

$$\|T_q(\boldsymbol{\beta}) - T_q(\widetilde{\boldsymbol{\beta}})\|_2 \le C_{\mathrm{rem}} \mathfrak{D}(r_\infty) \|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|_2, \tag{C.70}$$

$$\|T_q(\boldsymbol{\beta}) - T_q(\widetilde{\boldsymbol{\beta}})\|_\infty \le C_{\mathrm{rem}} \frac{D}{n^2} b_{H,2\to\infty} \|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|_2 + C_{\mathrm{rem}} \mathfrak{D}(r_\infty) \|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|_\infty. \tag{C.71}$$

Proof. First note that $T_q(\boldsymbol{\beta}) - T_q(\widetilde{\boldsymbol{\beta}}) = L_q^{-1} \big[\mathfrak{r}_q(\boldsymbol{\beta}) - \mathfrak{r}_q(\widetilde{\boldsymbol{\beta}}) \big]$.

By Lemma C.9, uniformly over $q \in \mathcal{I}$,

$$\|L_q^{-1}\|_{2\to 2} \le C_L \frac{D}{n^2}, \qquad \|L_q^{-1} \boldsymbol{x}\|_\infty \le C_L \frac{D}{n^2} \|\boldsymbol{x}\|_\infty. \tag{C.72}$$

Fix $q \in \mathcal{I}$ and $\boldsymbol{\beta}, \widetilde{\boldsymbol{\beta}} \in \mathcal{B}^\infty(r_\infty)$. Set $\boldsymbol{h} := \boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}$ and $\boldsymbol{\gamma}_t := \widetilde{\boldsymbol{\beta}} + t\boldsymbol{h}$. Since $\mathcal{B}B^\infty(r_\infty)$ is convex, the fundamental theorem of calculus gives

$$\mathfrak{r}_q(\boldsymbol{\beta}) - \mathfrak{r}_q(\widetilde{\boldsymbol{\beta}}) = \int_0^1 P_{\mathcal{T}} [\nabla_\alpha^2 \mathcal{G}(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\gamma}_t) - \nabla_\alpha^2 \overline{\mathcal{G}}(\overline{\boldsymbol{\alpha}}(q))] P_{\mathcal{T}} \boldsymbol{h} \, dt. \tag{C.73}$$

Using (C.28), we estimate the four contributions in (C.73) separately.

For the deterministic mean variation, Lemma C.20 and (C.72) give, uniformly in t,

$$\|L_q^{-1} P_{\mathcal{T}} [\nabla^2 \overline{\mathcal{G}}(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\gamma}_t) - \nabla^2 \overline{\mathcal{G}}(\overline{\boldsymbol{\alpha}}(q))] P_{\mathcal{T}} \boldsymbol{h}\|_2 \le C_L C_{\mathrm{box}} n r_\infty \|\boldsymbol{h}\|_2,$$

and the same coefficient multiplies $\|\boldsymbol{h}\|_\infty$ in the ℓ_∞-bound.

For the frozen centered Hessian, the path bounds in Definition 2 and (C.72) yield

$$\|L_q^{-1} P_{\mathcal{T}} \nabla^2 \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q)) P_{\mathcal{T}} \boldsymbol{h}\|_2 \le C_L \frac{D}{n^2} b_{H,\mathrm{op}} \|\boldsymbol{h}\|_2,$$

and

$$\|L_q^{-1} P_{\mathcal{T}} \nabla^2 \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q)) P_{\mathcal{T}} \boldsymbol{h}\|_\infty \le C_L \frac{D}{n^2} b_{H,2\to\infty} \|\boldsymbol{h}\|_2.$$

For the local centered-Hessian variation, Lemma C.20, (C.72), and $\|P_{\mathcal{T}} A P_{\mathcal{T}}\|_{\infty\to\infty} \le 4\|A\|_{\infty\to\infty}$ give coefficients bounded by $C_L C_{\mathrm{box}} D n \nu r_\infty$ in ℓ_2 and by $4 C_L C_{\mathrm{box}} D n \nu r_\infty$ in ℓ_∞. Finally, the remainder-Hessian terms are bounded in the same way by $C_L C_{\mathrm{box}} (D/n) \varepsilon^3$ and $4 C_L C_{\mathrm{box}} (D/n) \varepsilon^3$, respectively.

Thus, with

$$C_{\mathrm{rem}} := 4 C_L C_{\mathrm{box}},$$

combining the four contributions in (C.73) gives (C.70)–(C.71) by the definition of $\mathfrak{D}$. □

For $r_2, r_\infty > 0$, write

$$\mathcal{B}(r_2, r_\infty) := \{\boldsymbol{\beta} \in \mathcal{T} : \|\boldsymbol{\beta}\|_2 \le r_2,\ \|\boldsymbol{\beta}\|_\infty \le r_\infty\}.$$

For deterministic event parameters $\boldsymbol{b}, \varepsilon, \nu$, define

$$\mathfrak{R} := \frac{D}{n^{3/2}}\big(b_\nabla + \varepsilon^3\big). \tag{C.74}$$

Proposition C.2 (Slice optimizer, regularity, and envelope formula)**.** *There exist constants $c_{\mathrm{fp}}, C_{\mathrm{in}} > 0$, depending only on κ, r_0, such that the following holds. Suppose $\mathcal{E}(\boldsymbol{b}, \varepsilon, \nu)$ holds, $\varepsilon \le \varrho_0$, and*

$$\mathfrak{D}\left(C_{\mathrm{in}}\left(\frac{1}{\sqrt{n}} + \frac{D}{n^2} b_{H,2\to\infty}\right)\mathfrak{R}\right) \le c_{\mathrm{fp}}. \tag{C.75}$$

Then, uniformly over $q \in \mathcal{I}$, Proposition C.1 applies on

$$\left\{\boldsymbol{\beta} \in \mathcal{T} : \|\boldsymbol{\beta}\|_\infty \le C_{\mathrm{in}}\left(\frac{1}{\sqrt{n}} + \frac{D}{n^2} b_{H,2\to\infty}\right)\mathfrak{R}\right\}.$$

Moreover, the slice $\mathcal{S}_q := \{\boldsymbol{\beta} \in \mathcal{T} : \overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta} \in \Delta_n\}$ admits a unique maximizer

$$\boldsymbol{\beta}^\star(q) := \arg\max_{\boldsymbol{\beta} \in \mathcal{S}_q} \mathcal{G}(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta}).$$

It is interior and satisfies

$$\|\boldsymbol{\beta}^\star(q)\|_2 \le \frac{C_{\mathrm{in}}}{2}\mathfrak{R}, \qquad \|\boldsymbol{\beta}^\star(q)\|_\infty \le \frac{C_{\mathrm{in}}}{2}\left(\frac{1}{\sqrt{n}} + \frac{D}{n^2} b_{H,2\to\infty}\right)\mathfrak{R}. \tag{C.76}$$

Moreover, $q \mapsto \boldsymbol{\beta}^\star(q)$ is continuously differentiable on $\mathcal{I}$, and, for $V(q) := \mathcal{G}(\boldsymbol{\alpha}^\star(q))$ with

$$\boldsymbol{\alpha}^\star(q) := \overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta}^\star(q),$$

we have

$$V'(q) = \langle \nabla\mathcal{G}(\boldsymbol{\alpha}^\star(q)), \overline{\boldsymbol{\alpha}}'(q)\rangle.$$

Proof. Let C_L be as in (C.72), let C_{rem} be the constant from Lemma C.21, and let c_{sc} be the constant from Proposition C.1. Set

$$C_0 := C_L \max\{1, 2K_R\},$$

choose

$$C_{\mathrm{in}} \ge 8C_0(1 + 8C_{\mathrm{rem}}), \qquad c_{\mathrm{fp}} \le \min\left\{c_{\mathrm{sc}}, \frac{1}{4C_{\mathrm{rem}}}\right\},$$

and, for the remainder of the proof, write

$$r_2 := 8C_0\mathfrak{R}, \qquad r_\infty := C_{\mathrm{in}}\left(\frac{1}{\sqrt{n}} + \frac{D}{n^2} b_{H,2\to\infty}\right)\mathfrak{R}.$$

Then (C.75) gives $\mathfrak{D}(r_\infty) \le c_{\mathrm{fp}} \le c_{\mathrm{sc}}$. Since $nr_\infty \le \mathfrak{D}(r_\infty)$ and $c_{\mathrm{sc}} \le c_\alpha/2$, we also have $nr_\infty \le c_\alpha/2$.

We first control $T_q(0) = L_q^{-1}F_q(0)$. Since Lemma C.9 implies $P_{\mathcal{T}}\nabla\overline{\mathcal{G}}(\overline{\boldsymbol{\alpha}}(q)) = 0$, (C.28) gives

$$F_q(0) = P_{\mathcal{T}}\nabla\mathcal{Q}(\overline{\boldsymbol{\alpha}}(q)) - P_{\mathcal{T}}\nabla R(\overline{\boldsymbol{\alpha}}(q)). \tag{C.77}$$

By Definition 2,

$$\|P_{\mathcal{T}}\nabla\mathcal{Q}(\overline{\boldsymbol{\alpha}}(q))\|_\infty \le b_\nabla, \qquad \|P_{\mathcal{T}}\nabla\mathcal{Q}(\overline{\boldsymbol{\alpha}}(q))\|_2 \le \sqrt{n}\, b_\nabla.$$

Also, Lemma C.7 gives $\|\nabla R\|_\infty \le K_R\varepsilon^3$. Since $\|P_{\mathcal{T}}x\|_\infty \le 2\|x\|_\infty$, $\|P_{\mathcal{T}}x\|_2 \le \sqrt{n}\|P_{\mathcal{T}}x\|_\infty$, and (C.72) holds, we obtain

$$\sup_{q\in\mathcal{I}}\|T_q(0)\|_2 \le C_0\mathfrak{R}, \qquad \sup_{q\in\mathcal{I}}\|T_q(0)\|_\infty \le \frac{C_0}{\sqrt{n}}\mathfrak{R}. \tag{C.78}$$

By the choice of c_{fp}, Proposition C.1 applies on the ℓ_∞-ball of radius r_∞, and Lemma C.21 applies to any $\boldsymbol{\beta}, \widetilde{\boldsymbol{\beta}} \in \mathcal{T}$ with $\|\boldsymbol{\beta}\|_\infty \vee \|\widetilde{\boldsymbol{\beta}}\|_\infty \le r_\infty$. Define

$$\mathcal{B}_{\mathrm{in}} := \mathcal{B}(r_2, r_\infty).$$

Since $nr_\infty \le c_\alpha/2$, Lemma C.16 implies $\mathcal{B}_{\mathrm{in}} \subset \mathcal{S}_q$.

Let $\boldsymbol{\beta} \in \mathcal{B}_{\mathrm{in}}$. By (C.78), Lemma C.21, and $\mathfrak{D}(r_\infty) \le c_{\mathrm{fp}}$,

$$\begin{aligned}\|T_q(\boldsymbol{\beta})\|_2 &\le \|T_q(0)\|_2 + \|T_q(\boldsymbol{\beta}) - T_q(0)\|_2 \\ &\le C_0\mathfrak{R} + C_{\mathrm{rem}} c_{\mathrm{fp}} r_2 \le \frac{1}{8} r_2 + \frac{1}{4} r_2 = \frac{3}{8} r_2.\end{aligned}$$

Similarly,

$$\begin{aligned}\|T_q(\boldsymbol{\beta})\|_\infty &\le \|T_q(0)\|_\infty + \|T_q(\boldsymbol{\beta}) - T_q(0)\|_\infty \\ &\le \frac{C_0}{\sqrt{n}}\mathfrak{R} + C_{\mathrm{rem}} \frac{D}{n^2} b_{H,2\to\infty} r_2 + C_{\mathrm{rem}} c_{\mathrm{fp}} r_\infty \\ &\le C_0(1 + 8C_{\mathrm{rem}}) \left(\frac{1}{\sqrt{n}} + \frac{D}{n^2} b_{H,2\to\infty} \right) \mathfrak{R} + \frac{1}{4} r_\infty \\ &\le \frac{1}{8} r_\infty + \frac{1}{4} r_\infty = \frac{3}{8} r_\infty.\end{aligned}$$

Thus T_q maps $\mathcal{B}_{\mathrm{in}}$ into itself. Moreover, for $\boldsymbol{\beta}, \widetilde{\boldsymbol{\beta}} \in \mathcal{B}_{\mathrm{in}}$, Lemma C.21 gives

$$\|T_q(\boldsymbol{\beta}) - T_q(\widetilde{\boldsymbol{\beta}})\|_2 \le C_{\mathrm{rem}} \mathfrak{D}(r_\infty) \|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|_2 \le \frac{1}{4} \|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|_2.$$

Banach's fixed-point theorem therefore gives a unique fixed point $\boldsymbol{\beta}^\star(q) \in \mathcal{B}_{\mathrm{in}}$, and hence

$$P_{\mathcal{T}} \nabla \mathcal{G}(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta}^\star(q)) = 0. \tag{C.79}$$

Since $r_2 = 8C_0\mathfrak{R} \le C_{\mathrm{in}}\mathfrak{R}$, the preceding bounds imply (C.76).

The set $\mathcal{S}_q$ is convex, and $\mathcal{G}(\overline{\boldsymbol{\alpha}}(q) + \cdot)$ is concave on it. Hence (C.79) shows that $\boldsymbol{\beta}^\star(q)$ is a maximizer on the full slice. To prove uniqueness, suppose $\boldsymbol{\beta}' \in \mathcal{S}_q$ is another maximizer. Concavity implies that

$$t \mapsto \mathcal{G}\big(\overline{\boldsymbol{\alpha}}(q) + (1-t)\boldsymbol{\beta}^\star(q) + t\boldsymbol{\beta}'\big)$$

is constant on $[0,1]$. Since the bounds above place $\boldsymbol{\beta}^\star(q)$ strictly inside $\mathcal{B}_{\mathrm{in}}$, for all sufficiently small $t > 0$ this segment remains in the region where Proposition C.1 gives strict concavity, a contradiction unless $\boldsymbol{\beta}' = \boldsymbol{\beta}^\star(q)$. Moreover, $\overline{\alpha}_i(q) + \beta_i^\star(q) \ge c_\alpha/(2n) > 0$, so the maximizer is interior.

Finally, define $\Psi(q, \boldsymbol{\beta}) := P_{\mathcal{T}} \nabla \mathcal{G}(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta})$. Then $\Psi(q, \boldsymbol{\beta}^\star(q)) = 0$, and $D_\beta \Psi = P_{\mathcal{T}} \nabla_\alpha^2 \mathcal{G} P_{\mathcal{T}}|_{\mathcal{T}}$ is invertible at the maximizer by Proposition C.1. The implicit-function theorem gives a local C^1 solution; uniqueness of the slice optimizer patches these local solutions into a C^1 map on $\mathcal{I}$. Differentiating $V(q) = \mathcal{G}(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta}^\star(q))$ gives

$$V'(q) = \langle \nabla \mathcal{G}(\boldsymbol{\alpha}^\star(q)), \overline{\boldsymbol{\alpha}}'(q) + (\boldsymbol{\beta}^\star)'(q) \rangle .$$

Since $(\boldsymbol{\beta}^\star)'(q) \in \mathcal{T}$, the second term vanishes by (C.79), yielding the stated envelope formula. □

C.8 Derivative of the optimized value

For a differentiable function F and vector $\boldsymbol{\xi}$, write $D_{\boldsymbol{\xi}} F(\boldsymbol{\alpha}) := \langle \nabla F(\boldsymbol{\alpha}), \boldsymbol{\xi} \rangle$.

Lemma C.22 (Local mixed derivative for the centered term)**.** *There exists $C_{\mathrm{mix}} > 0$, depending only on κ, r_0, such that, if $\mathcal{E}(\boldsymbol{b}, \varepsilon, \nu)$ holds, $\varepsilon \le \varrho_0$, and $nr_\infty \le c_\alpha/2$, then*

$$\sup_{\substack{q \in \mathcal{I},\ \boldsymbol{\beta} \in \mathcal{T} \\ \|\boldsymbol{\beta}\|_\infty \le r_\infty}} \|P_{\mathcal{T}} \nabla_\beta D_{\overline{\boldsymbol{\alpha}}'(q)} \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta})\|_2 \le C_{\mathrm{mix}} \big(b_{\nabla q} + n^{5/2} \nu r_\infty \big). \tag{C.80}$$

Proof. At $\boldsymbol{\beta} = 0$, Definition 2 gives

$$\sup_{q \in \mathcal{I}} \Big\| P_{\mathcal{T}}\ \nabla_\beta D_{\overline{\boldsymbol{\alpha}}'(q)} \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta})\big|_{\boldsymbol{\beta}=0} \Big\|_2 \le b_{\nabla q}.$$

It remains to control the variation in $\boldsymbol{\beta}$. For fixed q, the chain rule gives

$$\nabla_\beta D_{\overline{\boldsymbol{\alpha}}'(q)} \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta}) = \nabla_\alpha^2 \mathcal{Q}(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta}) \overline{\boldsymbol{\alpha}}'(q). \tag{C.81}$$

Writing $A_Q(q, \boldsymbol{\beta})$ as in the proof of Lemma C.20, the difference from $\boldsymbol{\beta} = 0$ is bounded by

$$\|A_Q(q, \boldsymbol{\beta})\|_{\mathrm{op}} \|\overline{\boldsymbol{\alpha}}'(q)\|_2.$$

By Lemma C.16, $\|\overline{\boldsymbol{\alpha}}'(q)\|_2 \le (\kappa\sqrt{n})^{-1}$, while (C.62) gives $\|A_Q(q, \boldsymbol{\beta})\|_{\mathrm{op}} \le C_{\mathrm{box}} n^3 \nu r_\infty$. Therefore the variation is at most $(C_{\mathrm{box}}/\kappa) n^{5/2} \nu r_\infty$. Taking $C_{\mathrm{mix}} \ge \max\{1, C_{\mathrm{box}}/\kappa\}$ proves (C.80). □

Given the deterministic event parameters, define

$$\mathfrak{E}_V := \frac{n^2}{D}\mathfrak{R}^2 + b_q + b_{\nabla q}\mathfrak{R} + n^{5/2}\nu \left(\frac{1}{\sqrt{n}} + \frac{D}{n^2} b_{H,2\to\infty} \right) \mathfrak{R}^2 + \varepsilon^3. \tag{C.82}$$

Proposition C.3 (Derivative approximation)**.** *There exists $C_V > 0$, depending only on κ, r_0, such that, under the hypotheses of Proposition C.2,*

$$\left| V'(1/2) - \frac{1}{D}\theta_{n_+,n_-} \right| \le C_V \mathfrak{E}_V. \tag{C.83}$$

Proof. Set

$$\boldsymbol{\xi}_0 := \overline{\boldsymbol{\alpha}}'(1/2), \qquad \overline{\boldsymbol{\alpha}} := \overline{\boldsymbol{\alpha}}(1/2), \qquad \boldsymbol{\beta}^\star := \boldsymbol{\beta}^\star(1/2).$$

Also set

$$C_\xi := \sup_{q \in \mathcal{I}} n \|\overline{\boldsymbol{\alpha}}'(q)\|_\infty \le \kappa^{-1},$$

where the last inequality follows from Lemma C.16. By Proposition C.2,

$$\|\boldsymbol{\beta}^\star\|_2 \le C_{\mathrm{in}}\mathfrak{R}, \qquad \|\boldsymbol{\beta}^\star\|_\infty \le C_{\mathrm{in}} \left(\frac{1}{\sqrt{n}} + \frac{D}{n^2} b_{H,2\to\infty} \right) \mathfrak{R}. \tag{C.84}$$

For brevity, write

$$r_\infty := C_{\mathrm{in}} \left(\frac{1}{\sqrt{n}} + \frac{D}{n^2} b_{H,2\to\infty} \right) \mathfrak{R}.$$

By the envelope formula in Proposition C.2,

$$V'(1/2) = D_{\boldsymbol{\xi}_0} \mathcal{G}(\overline{\boldsymbol{\alpha}} + \boldsymbol{\beta}^\star).$$

Using (C.28), we therefore have

$$V'(1/2) = D_{\boldsymbol{\xi}_0} \overline{\mathcal{G}}(\overline{\boldsymbol{\alpha}} + \boldsymbol{\beta}^\star) + D_{\boldsymbol{\xi}_0} \mathcal{Q}(\overline{\boldsymbol{\alpha}} + \boldsymbol{\beta}^\star) - D_{\boldsymbol{\xi}_0} R(\overline{\boldsymbol{\alpha}} + \boldsymbol{\beta}^\star). \tag{C.85}$$

We estimate the three terms separately.

First consider the deterministic mean term. Since $\overline{\mathcal{G}} = D^{-1}\mathcal{H}$, (C.21) gives

$$D_{\boldsymbol{\xi}_0} \overline{\mathcal{G}}(\overline{\boldsymbol{\alpha}}) = \frac{d}{dq} \overline{\mathcal{G}}(\overline{\boldsymbol{\alpha}}(q)) \Big|_{q=1/2} = \frac{1}{D}\theta_{n_+,n_-}. \tag{C.86}$$

Moreover, the first-order variation of $D_{\boldsymbol{\xi}_0}\overline{\mathcal{G}}$ in tangent directions vanishes. To see this, observe first that $\overline{\mathcal{G}}$ is invariant under every permutation of the coordinates that preserves the two groups I_+ and I_-. Both $\overline{\boldsymbol{\alpha}} = \overline{\boldsymbol{\alpha}}(1/2)$ and $\boldsymbol{\xi}_0 = \overline{\boldsymbol{\alpha}}'(1/2)$ are also constant on each of these two groups. Therefore the scalar function

$$\boldsymbol{\alpha} \longmapsto D_{\boldsymbol{\xi}_0} \overline{\mathcal{G}}(\boldsymbol{\alpha})$$

has the same within-group permutation invariance. Consequently, at the within-group symmetric point $\overline{\boldsymbol{\alpha}}$, its gradient cannot distinguish two coordinates belonging to the same group. Hence there exist scalars a_+ and a_- such that

$$\nabla_\alpha D_{\boldsymbol{\xi}_0} \overline{\mathcal{G}}(\overline{\boldsymbol{\alpha}}) = a_+ \mathbf{1}_{I_+} + a_- \mathbf{1}_{I_-}.$$

Since every $\boldsymbol{\beta} \in \mathcal{T}$ has zero sum on each group,

$$\left\langle \nabla_\alpha D_{\boldsymbol{\xi}_0} \overline{\mathcal{G}}(\overline{\boldsymbol{\alpha}}), \boldsymbol{\beta} \right\rangle = a_+ \sum_{i \in I_+} \beta_i + a_- \sum_{i \in I_-} \beta_i = 0.$$

Thus $\nabla_\alpha D_{\boldsymbol{\xi}_0} \overline{\mathcal{G}}(\overline{\boldsymbol{\alpha}}) \in \mathcal{T}^\perp$, and therefore

$$P_{\mathcal{T}} \nabla_\alpha D_{\boldsymbol{\xi}_0} \overline{\mathcal{G}}(\overline{\boldsymbol{\alpha}}) = 0. \tag{C.87}$$

We next bound the second tangent variation. On the coordinate window (C.66), the third derivatives of h are bounded by $C_h n^2$. Since $\|\boldsymbol{\xi}_0\|_\infty \le C_\xi / n$, each nonzero entry of a pairwise block contributing to $\nabla_\alpha^2 D_{\boldsymbol{\xi}_0} \overline{\mathcal{G}}$ is of order n/D. For a fixed row k, only the $n-1$ pairwise blocks indexed by pairs (i,j) with $k \in \{i,j\}$ can contribute. Consequently, the row-sum bound and symmetry give

$$\left\| P_{\mathcal{T}} \nabla_\alpha^2 D_{\boldsymbol{\xi}_0} \overline{\mathcal{G}}(\overline{\boldsymbol{\alpha}} + \boldsymbol{\beta}) P_{\mathcal{T}} \right\|_{\mathrm{op}} \le A_V \frac{n^2}{D}$$

uniformly over $\boldsymbol{\beta} \in \mathcal{T}$ with $\|\boldsymbol{\beta}\|_\infty \le r_\infty$, for a constant $A_V > 0$ depending only on κ, r_0. Since $\boldsymbol{\beta}^\star \in \mathcal{T}$, Taylor's theorem, (C.84), and (C.87) therefore imply

$$\left| D_{\boldsymbol{\xi}_0} \overline{\mathcal{G}}(\overline{\boldsymbol{\alpha}} + \boldsymbol{\beta}^\star) - D_{\boldsymbol{\xi}_0} \overline{\mathcal{G}}(\overline{\boldsymbol{\alpha}}) \right| \le A_V \frac{n^2}{D} \|\boldsymbol{\beta}^\star\|_2^2 \le A_V C_{\mathrm{in}}^2 \frac{n^2}{D} \mathfrak{R}^2. \tag{C.88}$$

Next consider the centered term. At the base point, Definition 2 gives

$$\left| D_{\boldsymbol{\xi}_0} \mathcal{Q}(\overline{\boldsymbol{\alpha}}) \right| \le b_q. \tag{C.89}$$

To control the displacement from $\overline{\boldsymbol{\alpha}}$ to $\overline{\boldsymbol{\alpha}} + \boldsymbol{\beta}^\star$, apply the fundamental theorem of calculus along the segment $\overline{\boldsymbol{\alpha}} + t\boldsymbol{\beta}^\star$. Since $\boldsymbol{\beta}^\star \in \mathcal{T}$,

$$\begin{aligned} &D_{\boldsymbol{\xi}_0} \mathcal{Q}(\overline{\boldsymbol{\alpha}} + \boldsymbol{\beta}^\star) - D_{\boldsymbol{\xi}_0} \mathcal{Q}(\overline{\boldsymbol{\alpha}}) \\ &\qquad = \int_0^1 \left\langle P_{\mathcal{T}} \nabla_\alpha D_{\boldsymbol{\xi}_0} \mathcal{Q}(\overline{\boldsymbol{\alpha}} + t\boldsymbol{\beta}^\star), \boldsymbol{\beta}^\star \right\rangle dt. \end{aligned} \tag{C.90}$$

Moreover, $\|t\boldsymbol{\beta}^\star\|_\infty \le r_\infty$ for $t \in [0,1]$. Hence Lemma C.22 and (C.84) give

$$\begin{aligned} &\left| D_{\boldsymbol{\xi}_0} \mathcal{Q}(\overline{\boldsymbol{\alpha}} + \boldsymbol{\beta}^\star) - D_{\boldsymbol{\xi}_0} \mathcal{Q}(\overline{\boldsymbol{\alpha}}) \right| \\ &\quad \le C_{\mathrm{mix}} \left(b_{\nabla q} + n^{5/2} \nu r_\infty \right) \|\boldsymbol{\beta}^\star\|_2 \\ &\quad \le C_{\mathrm{mix}} C_{\mathrm{in}} \left[b_{\nabla q} + C_{\mathrm{in}} n^{5/2} \nu \left(\frac{1}{\sqrt{n}} + \frac{D}{n^2} b_{H,2\to\infty} \right) \mathfrak{R} \right] \mathfrak{R}. \end{aligned} \tag{C.91}$$

Together with (C.89), this controls the centered contribution.

Finally consider the analytic remainder. Consider the affine path

$$\boldsymbol{\alpha}(t) := \overline{\boldsymbol{\alpha}} + \boldsymbol{\beta}^\star + t\boldsymbol{\xi}_0.$$

Then $\boldsymbol{\alpha}'(t) = \boldsymbol{\xi}_0$ and

$$\frac{d}{dt} R(\boldsymbol{\alpha}(t)) \Big|_{t=0} = D_{\boldsymbol{\xi}_0} R(\overline{\boldsymbol{\alpha}} + \boldsymbol{\beta}^\star).$$

Moreover,

$$\|n \operatorname{diag}(\boldsymbol{\xi}_0)\|_{\mathrm{op}} = n \|\boldsymbol{\xi}_0\|_\infty \le C_\xi.$$

Since the basic event gives $\varepsilon \le \varrho_0$, the affine-path derivative bound (C.15) therefore yields

$$\left| D_{\boldsymbol{\xi}_0} R(\overline{\boldsymbol{\alpha}} + \boldsymbol{\beta}^\star) \right| \le C_\xi K_R \varepsilon^3. \tag{C.92}$$

Combining (C.85), (C.86), (C.88), (C.89), (C.91), and (C.92), we obtain

$$\begin{aligned} \left| V'(1/2) - \frac{1}{D} \theta_{n_+,n_-} \right| \le C \Bigg[&\frac{n^2}{D} \mathfrak{R}^2 + b_q + b_{\nabla q} \mathfrak{R} \\ &+ n^{5/2} \nu \left(\frac{1}{\sqrt{n}} + \frac{D}{n^2} b_{H,2\to\infty} \right) \mathfrak{R}^2 + \varepsilon^3 \Bigg], \end{aligned}$$

where $C > 0$ depends only on κ, r_0. Choosing C_V sufficiently large and using (C.82) gives (C.83). □

C.9 Balanced minimizer, local growth, and positivity

Throughout this subsection, suppose that the hypotheses of Propositions C.2 are satisfied.

We also assume, whenever the balanced minimizer and positivity conclusions are invoked, that

$$C_V \mathfrak{E}_V \le \frac{1}{2D}\theta_{n_+,n_-}, \qquad \lambda\left(\frac{1}{D}\theta_{n_+,n_-} + C_V\mathfrak{E}_V\right) < 2s. \tag{C.93}$$

Recall $\psi(q) = 1 - V(q)$ and $\Phi(q) := 2s\left|q - \frac{1}{2}\right| + \lambda\psi(q)$.

We say that a dual vector $\boldsymbol{\alpha} \in \Delta_n$ is balanced if $q(\boldsymbol{\alpha}) = \frac{1}{2}$.

Proposition C.4 (Unique balanced minimizer)**.** *Under the standing assumptions of this subsection, suppose that* (C.93) *holds. Then $q_\star = \frac{1}{2}$ is the unique minimizer of Φ, and the full dual minimizer $\boldsymbol{\alpha}^\star := \boldsymbol{\alpha}^\star(1/2)$ is unique and interior.*

Proof. We first note that Φ is convex on $[0,1]$ since ψ is convex by Proposition B.1.

Since $n_+ > n_-$, we have $\theta_{n_+,n_-} > 0$. By Proposition C.3, $V'(1/2) \ge \frac{\theta_{n_+,n_-}}{D} - C_V\mathfrak{E}_V \ge \frac{1}{2D}\theta_{n_+,n_-} > 0$. The same proposition gives $V'(1/2) \le \frac{\theta_{n_+,n_-}}{D} + C_V\mathfrak{E}_V$. Thus $\lambda V'(1/2) < 2s$. The one-sided derivatives of the convex function Φ are $\Phi'_-(1/2) = -2s - \lambda V'(1/2) < 0$ and $\Phi'_+(1/2) = 2s - \lambda V'(1/2) > 0$. A convex function with strictly negative left derivative and strictly positive right derivative has a unique minimizer at the kink. Proposition C.2 gives a unique interior inner optimizer there. □

Proposition C.5 (Local relative growth)**.** *Under the standing assumptions of this subsection, suppose that* (C.93) *holds. Then F has local relative growth at $\boldsymbol{\alpha}^\star$.*

Proof. Set

$$r_{\mathrm{sc}} := C_{\mathrm{in}}\left(\frac{1}{\sqrt{n}} + \frac{D}{n^2}b_{H,2\to\infty}\right)\mathfrak{R}.$$

By Proposition C.2, Proposition C.1 applies uniformly over

$$q \in \mathcal{I}, \qquad \boldsymbol{\beta} \in \mathcal{T}, \qquad \|\boldsymbol{\beta}\|_\infty \le r_{\mathrm{sc}},$$

whereas

$$\|\boldsymbol{\beta}^\star(q)\|_\infty \le \frac{1}{2}r_{\mathrm{sc}} \tag{C.94}$$

uniformly over $q \in \mathcal{I}$.

Let $\boldsymbol{\alpha}$ be a feasible vector sufficiently close to $\boldsymbol{\alpha}^\star$. Then $q := q(\boldsymbol{\alpha})$ is sufficiently close to $1/2$, and we can write uniquely

$$\boldsymbol{\alpha} = \overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta}, \qquad \boldsymbol{\beta} \in \mathcal{T}.$$

Since $\boldsymbol{\beta}^\star(q)$ is continuous in q, by taking the neighborhood of $\boldsymbol{\alpha}^\star$ sufficiently small we may assume that

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^\star(q)\|_\infty \le \frac{1}{2}r_{\mathrm{sc}}.$$

Together with (C.94), this implies that the line segment joining $\boldsymbol{\beta}^\star(q)$ and $\boldsymbol{\beta}$ lies in the region on which Proposition C.1 applies.

By Proposition C.4, $q = 1/2$ is the unique minimizer of Φ, and the one-sided derivatives at $1/2$ are strictly negative on the left and strictly positive on the right. Hence there exist $c_1, \varepsilon_1 > 0$ such that

$$\Phi(q) - \Phi(1/2) \ge c_1|q - 1/2| \tag{C.95}$$

whenever $|q - 1/2| \le \varepsilon_1$.

Next, since $F(\boldsymbol{\alpha}) = 2s\left|q - \frac{1}{2}\right| + \lambda\|Z(\boldsymbol{\alpha})\|_1$ and $\Phi(q) = 2s\left|q - \frac{1}{2}\right| + \lambda\|Z(\boldsymbol{\alpha}^\star(q))\|_1$, we have

$$\begin{aligned} F(\boldsymbol{\alpha}) - F(\boldsymbol{\alpha}^\star) &= \Phi(q) - \Phi(1/2) \\ &\quad + \lambda\left(\|Z(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta})\|_1 - \|Z(\boldsymbol{\alpha}^\star(q))\|_1\right). \end{aligned} \tag{C.96}$$

On the line segment described above, Proposition C.1 implies that $\|Z(\overline{\boldsymbol{\alpha}}(q)+\cdot)\|_1 = 1-\mathcal{G}$ is strongly convex on $\mathcal{T}$, with Hessian bounded below by $(c_H/2)(n^2/D)I_{\mathcal{T}}$. Since $\boldsymbol{\alpha}^\star(q) = \overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta}^\star(q)$ is the tangent minimizer on the slice q, we have

$$\|Z(\overline{\boldsymbol{\alpha}}(q) + \boldsymbol{\beta})\|_1 - \|Z(\boldsymbol{\alpha}^\star(q))\|_1 \geq \frac{c_H}{4}\frac{n^2}{D}\|\boldsymbol{\beta} - \boldsymbol{\beta}^\star(q)\|_2^2. \tag{C.97}$$

Combining (C.95), (C.96), and (C.97) gives

$$F(\boldsymbol{\alpha}) - F(\boldsymbol{\alpha}^\star) \geq c_1|q - 1/2| + \frac{\lambda c_H}{4}\frac{n^2}{D}\|\boldsymbol{\beta} - \boldsymbol{\beta}^\star(q)\|_2^2. \tag{C.98}$$

On the other hand,

$$\boldsymbol{\alpha} - \boldsymbol{\alpha}^\star = \big(\overline{\boldsymbol{\alpha}}(q) - \overline{\boldsymbol{\alpha}}(1/2)\big) + \big(\boldsymbol{\beta} - \boldsymbol{\beta}^\star(q)\big) + \big(\boldsymbol{\beta}^\star(q) - \boldsymbol{\beta}^\star(1/2)\big).$$

The path $q \mapsto \overline{\boldsymbol{\alpha}}(q)$ is smooth, while Proposition C.2 gives that $q \mapsto \boldsymbol{\beta}^\star(q)$ is continuously differentiable. Hence there exists $C > 0$ such that, for q sufficiently close to $1/2$,

$$\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^\star\|_2^2 \leq C\left(|q - 1/2|^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}^\star(q)\|_2^2\right).$$

Since $|q - 1/2|^2 \leq |q - 1/2|$ locally, (C.98) implies that, in a sufficiently small neighborhood of $\boldsymbol{\alpha}^\star$, $F(\boldsymbol{\alpha}) - F(\boldsymbol{\alpha}^\star) \geq \widetilde{\mu}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^\star\|_2^2$ for some $\widetilde{\mu} > 0$.

Finally, by Lemma B.3, after possibly shrinking the neighborhood,

$$D_{\mathrm{KL}}(\boldsymbol{\alpha}^\star\|\boldsymbol{\alpha}) \leq C_{\mathrm{KL}}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^\star\|_2^2.$$

Thus

$$F(\boldsymbol{\alpha}) - F(\boldsymbol{\alpha}^\star) \geq \frac{\widetilde{\mu}}{C_{\mathrm{KL}}} D_{\mathrm{KL}}(\boldsymbol{\alpha}^\star\|\boldsymbol{\alpha})$$

throughout this neighborhood. Choosing $r > 0$ sufficiently small so that $D_{\mathrm{KL}}(\boldsymbol{\alpha}^\star\|\boldsymbol{\alpha}) \leq r$ implies that $\boldsymbol{\alpha}$ belongs to this neighborhood proves local relative growth. □

Lemma C.23 (Rank of positive noise combinations)**.** *Suppose $\mathcal{E}(\boldsymbol{b}, \varepsilon, \nu)$ holds with $\varepsilon < 1$. Then U and $\widehat{V}$ have full column rank, and, for every $\boldsymbol{\alpha} \in \mathrm{int}(\Delta_n)$,*

$$Z(\boldsymbol{\alpha}) = U\Lambda_{\boldsymbol{\alpha}}\widehat{V}^\top$$

has rank n. In particular, $\mathrm{rank}\, Z(\boldsymbol{\alpha}^\star) = n$ *whenever* $\boldsymbol{\alpha}^\star \in \mathrm{int}(\Delta_n)$.

Proof. By the geometric bounds in Definition 2, $\|G_U - I\|_{\mathrm{op}} \vee \|G_V - I\|_{\mathrm{op}} < 1$, so both Gram matrices are positive definite. Hence U and $\widehat{V}$ have full column rank. If $\boldsymbol{\alpha} \in \mathrm{int}(\Delta_n)$, then $\Lambda_{\boldsymbol{\alpha}}$ is invertible, and therefore $U\Lambda_{\boldsymbol{\alpha}}\widehat{V}^\top$ has rank n. □

Proposition C.6 (KKT and envelope identification)**.** *Under the standing assumptions of this subsection, suppose that* (C.93) *holds. Let $Z^\star := Z(\boldsymbol{\alpha}^\star)$, $H^\star := \mathrm{Polar}(Z^\star)$. Then* $\mathrm{rank}(Z^\star) = \mathrm{rank}(H^\star) = n$*, and $V'(1/2) = h_-^\star - h_+^\star$, where*

$$h_+^\star = \frac{1}{n_+}\sum_{i \in I_+}\langle H^\star, N_i\rangle_F, \qquad h_-^\star = \frac{1}{n_-}\sum_{i \in I_-}\langle H^\star, N_i\rangle_F.$$

The signal coefficient selected by the iterates is $a^\star = \frac{\lambda}{2s}V'(1/2)$. Moreover, $0 < a^\star < 1$.

Proof. By Proposition C.4, the full dual minimizer is $\boldsymbol{\alpha}^\star = \boldsymbol{\alpha}^\star(1/2)$, and it is interior. By Lemma C.23, $Z^\star = Z(\boldsymbol{\alpha}^\star) = U\Lambda_{\alpha^\star}\widehat{V}^\top$ has rank n. Since the polar factor has the same rank as the matrix to which it is applied, $\mathrm{rank}(H^\star) = n$. With the core notation from (C.2), Lemma H.12 gives $\|Z(\boldsymbol{\alpha})\|_1 = \|Y(\alpha)\|_1$. Since $G_U^{1/2}$ and $G_V^{1/2}$ are invertible and $\boldsymbol{\alpha}^\star \in \mathrm{int}(\Delta_n)$, the matrix $Y^\star := Y(\boldsymbol{\alpha}^\star) = G_U^{1/2}\Lambda_{\alpha^\star}G_V^{1/2}$ is invertible. Therefore Lemma H.10, applied to the full-rank $n \times n$ matrix $Y^\star$, implies that the coefficient map $\boldsymbol{\alpha} \mapsto \|Z(\boldsymbol{\alpha})\|_1$ is differentiable at $\boldsymbol{\alpha}^\star$. For the i-th

coordinate direction, $\frac{\partial}{\partial \alpha_i} Y(\boldsymbol{\alpha}^\star) = G_U^{1/2} E_{ii} G_V^{1/2}$, where E_{ii} is the diagonal matrix with a single 1 in the i-th diagonal entry. Hence

$$\frac{\partial}{\partial \alpha_i} \|Z(\boldsymbol{\alpha}^\star)\|_1 = \left\langle \mathrm{Polar}(Y^\star), G_U^{1/2} E_{ii} G_V^{1/2} \right\rangle_F .$$

On the other hand, Lemma H.12 gives $H^\star = \mathrm{Polar}(Z^\star) = \overline{U}\,\mathrm{Polar}(Y^\star)\,\overline{V}^\top$, where $\overline{U} := UG_U^{-1/2}, \overline{V} := \widehat{V} G_V^{-1/2}$. Since $N_i = \boldsymbol{u}_i \widehat{\boldsymbol{v}}_i^\top$, we have

$$\langle H^\star, N_i \rangle_F = \boldsymbol{u}_i^\top \overline{U}\,\mathrm{Polar}(Y^\star) \overline{V}^\top \widehat{\boldsymbol{v}}_i.$$

Using $U^\top \overline{U} = G_U^{1/2}, \quad \overline{V}^\top \widehat{V} = G_V^{1/2}$, this becomes

$$\langle H^\star, N_i \rangle_F = e_i^\top G_U^{1/2}\,\mathrm{Polar}(Y^\star) G_V^{1/2} e_i = \left\langle \mathrm{Polar}(Y^\star), G_U^{1/2} E_{ii} G_V^{1/2} \right\rangle_F .$$

Therefore $\frac{\partial}{\partial \alpha_i} \|Z(\boldsymbol{\alpha}^\star)\|_1 = \langle H^\star, N_i \rangle_F$. By Proposition C.2, $V'(1/2) = \langle \nabla \mathcal{G}(\boldsymbol{\alpha}^\star), \overline{\boldsymbol{\alpha}}'(1/2) \rangle$. Since $\mathcal{G}(\boldsymbol{\alpha}) = 1 - \|Z(\boldsymbol{\alpha})\|_1$, we have $\frac{\partial}{\partial \alpha_i} \mathcal{G}(\alpha^\star) = -\langle H^\star, N_i \rangle_F$. Using

$$\overline{\alpha}_i'(1/2) = \begin{cases} 1/n_+, & i \in I_+, \\ -1/n_-, & i \in I_-, \end{cases}$$

we get

$$V'(1/2) = -\frac{1}{n_+} \sum_{i \in I_+} \langle H^\star, N_i \rangle_F + \frac{1}{n_-} \sum_{i \in I_-} \langle H^\star, N_i \rangle_F = h_-^\star - h_+^\star.$$

Proposition B.2 now gives $a^\star = -\frac{\lambda}{2s}(h_+^\star - h_-^\star) = \frac{\lambda}{2s} V'(1/2)$. The inequalities $0 < a^\star < 1$ follow from Proposition C.4. Indeed, $V'(1/2) > 0$ gives $a^\star > 0$, and the strict right-derivative inequality $2s - \lambda V'(1/2) > 0$ is equivalent to $a^\star < 1$. □

C.10 Completion of the proof

We now return to the explicit high-probability scales. Let $0 < \delta < e^{-1}$, set $L := \log(n/\delta)$, and write

$$\boldsymbol{b} := \boldsymbol{b}_{n,D,\delta}, \qquad \varepsilon := \varepsilon_{n,D,\delta}, \qquad \nu := \nu_{n,D,\delta}.$$

Throughout this subsection, the deterministic quantities $\mathfrak{R}$, $\mathfrak{D}$, and $\mathfrak{E}_V$ are evaluated at these event thresholds. Define

$$R_{n,\delta} := \frac{\sqrt{L}}{n} + \frac{L}{n^{3/2}} + \frac{(n+L)^{3/2}}{n^{3/2}\sqrt{D}}. \tag{C.99}$$

By (C.35), (C.36), and (C.74), there exists $B_R \geq 1$, depending only on κ, r_0, such that

$$\mathfrak{R} \leq B_R R_{n,\delta}, \qquad \left(\frac{1}{\sqrt{n}} + \frac{D}{n^2} b_{H,2\to\infty} \right) \mathfrak{R} \leq B_R \frac{1+L}{\sqrt{n}} R_{n,\delta}. \tag{C.100}$$

Thus $R_{n,\delta}$ is an explicit upper envelope, up to a constant, for the deterministic localization scale $\mathfrak{R}$; the second inequality provides the corresponding ℓ_∞-scale.

Theorem C.2 (Conditional high-probability form)**.** *Fix $0 < \kappa < 1/2$, $s > 0$, and $\lambda > 0$. There exist constants $c_{\mathrm{dim}}, c_{\mathrm{loc}}, C_{\mathrm{sec}} > 0$, depending only on κ, r_0, such that the following holds. Condition on signs satisfying* (C.58)*. Let $0 < \delta < e^{-1}$, set $L = \log(n/\delta)$, and define $R_{n,\delta}$ by* (C.99)*. Assume*

$$D \geq c_{\mathrm{dim}} nL, \qquad \mathfrak{D}\left(C_{\mathrm{sec}} \frac{1+L}{\sqrt{n}} R_{n,\delta} \right) \leq c_{\mathrm{loc}}, \tag{C.101}$$

and

$$C_{\mathrm{sec}} \mathfrak{E}_V \leq \frac{1}{2D} \theta_{n_+,n_-}, \qquad \lambda \left(\frac{1}{D} \theta_{n_+,n_-} + C_{\mathrm{sec}} \mathfrak{E}_V \right) < 2s. \tag{C.102}$$

Then, with conditional probability at least $1 - \delta$ over the spherical noise variables, the dual spectral max-margin problem has a unique minimizer $\boldsymbol{\alpha}^\star \in \mathrm{int}(\Delta_n)$ with $q(\boldsymbol{\alpha}^\star) = 1/2$. Setting $Z^\star := \sum_i \alpha_i^\star N_i$ and $H^\star := \mathrm{Polar}(Z^\star)$, we have $\mathrm{rank}(Z^\star) = \mathrm{rank}(H^\star) = n$, $\boldsymbol{\alpha}_k \to \boldsymbol{\alpha}^\star$,

$$\frac{W_k^{\mathrm{SpecGD}}}{\tau_k} \longrightarrow a^\star P_S + H^\star, \qquad \left| a^\star - \frac{\lambda}{2sD} \theta_{n_+,n_-} \right| \leq \frac{\lambda}{2s} C_{\mathrm{sec}} \mathfrak{E}_V.$$

Moreover $0 < a^\star < 1$, and $W_k^{\mathrm{SpecGD}} / \|W_k^{\mathrm{SpecGD}}\|_{\mathrm{op}} \to a^\star P_S + H^\star$.

Proof. By increasing c_{dim} if necessary, (C.101) implies the dimensional hypothesis of Lemma C.19 and $\varepsilon \le \min\{\varrho_0, 1/2\}$. Hence, with conditional probability at least $1-\delta$, $\mathcal{E}(\boldsymbol{b}, \varepsilon, \nu)$ holds. We work on this event.

Let $c_{\mathrm{fp}}, C_{\mathrm{in}}$ be the constants from Proposition C.2 and C_V from Proposition C.3. Enlarge C_{sec} so that

$$C_{\mathrm{sec}} \ge \max\{B_R C_{\mathrm{in}}, C_V\},$$

and choose

$$c_{\mathrm{loc}} \le c_{\mathrm{fp}}.$$

By (C.100),

$$C_{\mathrm{in}} \left(\frac{1}{\sqrt{n}} + \frac{D}{n^2} b_{H,2\to\infty} \right) \mathfrak{R} \le C_{\mathrm{sec}} \frac{1+L}{\sqrt{n}} R_{n,\delta}.$$

Hence (C.101) and monotonicity of $\mathfrak{D}(r)$ imply the deterministic smallness hypothesis of Proposition C.2.

Proposition C.3 and $C_{\mathrm{sec}} \ge C_V$ give

$$\left| V'(1/2) - \frac{1}{D} \theta_{n_+, n_-} \right| \le C_{\mathrm{sec}} \mathfrak{E}_V. \tag{C.103}$$

Therefore (C.102) implies the deterministic assumptions (C.93). Proposition C.4 then yields $q_\star = 1/2$ and a unique interior full dual minimizer $\boldsymbol{\alpha}^\star = \boldsymbol{\alpha}^\star(1/2)$. Lemma C.23 gives $\mathrm{rank}(Z^\star) = n$, hence also $\mathrm{rank}(H^\star) = n$.

By Proposition C.5, F has local relative growth at $\boldsymbol{\alpha}^\star$, so Theorem B.1 gives $\boldsymbol{\alpha}_k \to \boldsymbol{\alpha}^\star$. Proposition C.6 gives $a^\star = \lambda V'(1/2)/(2s)$, while Proposition B.2 yields $W_k^{\mathrm{SpecGD}}/\tau_k \to a^\star P_S + H^\star$. Using (C.103) gives the claimed bound on $a^\star$, and Proposition C.6 gives $0 < a^\star < 1$. Finally, $\|a^\star P_S + H^\star\|_{\mathrm{op}} = 1$ by signal–noise block orthogonality, which yields the normalized directional convergence exactly as before. □

Proof of Theorem 1. The constants $c > 0$ and $C > 0$ in the theorem will be chosen successively below. Each time we shrink c or enlarge C, we do so only in terms of ρ and the constants from Theorem C.2; hence the final constants still have the allowed dependence.

Let

$$\mathcal{E}_{\mathrm{lab}}(\delta) := \left\{ \left| \frac{n_-}{n} - \rho \right| \le \sqrt{\frac{\log(4/\delta)}{2n}} \right\}.$$

By Hoeffding's inequality, $\mathbb{P}(\mathcal{E}_{\mathrm{lab}}(\delta)) \ge 1 - \frac{\delta}{2}$. By taking $c > 0$ sufficiently small, depending only on ρ, the assumption $L^5 \le cn$ implies that on $\mathcal{E}_{\mathrm{lab}}(\delta)$, $\frac{\rho}{2} n \le n_- \le n_+ \le \left(1 - \frac{\rho}{2}\right) n, \quad n_+ - n_- \ge \frac{1-2\rho}{2} n$. Thus (C.58) holds with $\kappa = \frac{\rho}{2}$.

Condition on a realization of the corruption signs in $\mathcal{E}_{\mathrm{lab}}(\delta)$. We apply Theorem C.2 with failure probability $\delta/2$. Let $L_2 := \log \frac{2n}{\delta}$. Then $L_2 \le 2L$. In the remainder of this proof, $\boldsymbol{b}, \varepsilon, \nu, \mathfrak{R}, \mathfrak{D}$, and $\mathfrak{E}_V$ are evaluated with failure probability $\delta/2$. Let $c_{\mathrm{dim}}, c_{\mathrm{loc}}, C_{\mathrm{sec}}$ be the constants in Theorem C.2 corresponding to $\kappa = \rho/2$. It remains to verify (C.101) and (C.102), with δ replaced by $\delta/2$.

First, by choosing $c \le 1$, $L^5 \le cn$ implies $L \le n$, and hence $n + L_2 \le 3n$. From (C.99), there is a universal constant $C_R > 0$ such that

$$R_{n,\delta/2} \le C_R \left(\frac{\sqrt{L}}{n} + \frac{1}{\sqrt{D}} \right). \tag{C.104}$$

Consequently,

$$\frac{1+L_2}{\sqrt{n}} R_{n,\delta/2} \le C_R \left(\frac{L^{3/2}}{n^{3/2}} + \frac{L}{\sqrt{nD}} \right). \tag{C.105}$$

By the definitions of $\mathfrak{D}$, the event thresholds, and (C.105), there exists $C_D > 0$, depending only on ρ, r_0, such that

$$\mathfrak{D}\left(C_{\mathrm{sec}} \frac{1+L_2}{\sqrt{n}} R_{n,\delta/2} \right) \le C_D \left(\frac{L^{5/2}}{\sqrt{n}} + L^2 \sqrt{\frac{n}{D}} + \sqrt{\frac{n}{D}} \right). \tag{C.106}$$

Again, $L^5 \le cn$ and $D \ge CnL^4$ make the right-hand side at most c_{loc} after decreasing c and increasing C. Also, increasing C if necessary gives $D \ge c_{\dim} n L_2$. Thus (C.101) holds with failure probability $\delta/2$.

Next, under the same assumptions, the deterministic derivative error satisfies

$$\mathfrak{E}_V \le C_E \frac{n}{D}\left(\frac{L^3}{n} + \sqrt{\frac{n}{D}}\right), \tag{C.107}$$

where $C_E > 0$ depends only on ρ, r_0. Indeed, this follows term by term from (C.82), (C.100), (C.104), and (C.105). The mixed terms are controlled using $L^5 \le cn$ and $D \ge CnL^4$.

On $\mathcal{E}_{\mathrm{lab}}(\delta)$, the conditional label-count bounds imply

$$c_\rho n \le \theta_{n_+,n_-} \le C_\rho n \tag{C.108}$$

for constants $c_\rho, C_\rho > 0$ depending only on ρ. Combining the lower bound in (C.108) with (C.107), and then taking $c > 0$ sufficiently small and $C > 0$ sufficiently large, gives $C_{\mathrm{sec}}\mathfrak{E}_V \le \frac{1}{2D}\theta_{n_+,n_-}$. For the second inequality in (C.102), use the first inequality just proved and the upper bound in (C.108). We obtain

$$\frac{1}{D}\theta_{n_+,n_-} + C_{\mathrm{sec}}\mathfrak{E}_V \le \frac{3}{2D}\theta_{n_+,n_-} \le \frac{3}{2}C_\rho \frac{n}{D},$$

where $C_\rho > 0$ depends only on ρ. Hence, by the assumption $\frac{\lambda n}{sD} \le c$, and by choosing $c > 0$ sufficiently small depending only on ρ, we get

$$\lambda\left(\frac{1}{D}\theta_{n_+,n_-} + C_{\mathrm{sec}}\mathfrak{E}_V\right) < 2s.$$

Thus all assumptions of Theorem C.2 hold with $\kappa = \rho/2$ and failure probability $\delta/2$. Conditional on $\mathcal{E}_{\mathrm{lab}}(\delta)$, the conclusions of Theorem C.2 hold with probability at least $1 - \delta/2$. A union bound gives probability at least $1 - \delta$ over the joint randomness.

It remains only to pass from the realized-count expression to the displayed main-theorem expression. Write $p_n := \frac{n_+}{n}$. Then

$$\frac{\theta_{n_+,n_-}}{n} = \frac{1}{4}(2p_n - 1)\left[1 + 4p_n(1 - p_n)\right]. \tag{C.109}$$

On $\mathcal{E}_{\mathrm{lab}}(\delta)$, $|p_n - (1-\rho)| = \left|\frac{n_-}{n} - \rho\right| \le \sqrt{\frac{\log(4/\delta)}{2n}}$. Since the function

$$p \longmapsto \frac{1}{4}(2p-1)\left[1 + 4p(1-p)\right]$$

has bounded derivative on $[0,1]$, there exists a universal constant C_θ such that, by (C.109),

$$\left|\frac{\theta_{n_+,n_-}}{n} - \frac{1}{4}(1-2\rho)\left[1 + 4\rho(1-\rho)\right]\right| \le C_\theta\sqrt{\frac{\log(4/\delta)}{n}}. \tag{C.110}$$

By Theorem C.2,

$$\left|a^\star - \frac{\lambda}{2sD}\theta_{n_+,n_-}\right| \le \frac{\lambda}{2s}C_{\mathrm{sec}}\mathfrak{E}_V.$$

Combining this with (C.107) and (C.110) and increasing C if necessary yields

$$\left|a^\star - \frac{\lambda n}{8sD}(1-2\rho)\left[1 + 4\rho(1-\rho)\right]\right| \le C\frac{\lambda n}{sD}\left(\sqrt{\frac{\log(1/\delta)}{n}} + \frac{L^3}{n} + \sqrt{\frac{n}{D}}\right).$$

This is (3.1). The remaining claims are exactly the conclusions of Theorem C.2. □

C.11 Completion of the proof of Corollary 1

Proof of Corollary 1. First choose $c, C > 0$ so that Theorem 1 holds. On the event from Theorem 1, $W^\star = a^\star P_S + H^\star$, $\mathrm{rank}(H^\star) = n$, $0 < a^\star < 1$. For a fresh test point,

$$y\langle W^\star, X\rangle_F = y\left\langle a^\star P_S + H^\star, yS + \lambda \boldsymbol{u}\boldsymbol{v}^\top \right\rangle_F.$$

Using the signal-noise block orthogonality, $\langle P_S, \boldsymbol{u}\boldsymbol{v}^\top\rangle_F = 0, \quad \langle H^\star, S\rangle_F = 0$. Therefore

$$y\langle W^\star, X\rangle_F = a^\star\langle P_S, S\rangle_F + \lambda y\langle H^\star, \boldsymbol{u}\boldsymbol{v}^\top\rangle_F = a^\star s + \lambda y\langle H^\star, \boldsymbol{u}\boldsymbol{v}^\top\rangle_F.$$

Let

$$T := \langle H^\star, \boldsymbol{u}\boldsymbol{v}^\top\rangle_F, \qquad r := \frac{a^\star s}{\lambda}.$$

Conditioning further on y, the error event is $\{T \le -r\}$ if $y = 1$ and $\{T \ge r\}$ if $y = -1$. Since, conditionally on the training data, T is symmetric about zero, these two probabilities are equal. Hence

$$\mathcal{R}(W^\star) = \frac{1}{2}\mathbb{P}\left(\left|\langle H^\star, \boldsymbol{u}\boldsymbol{v}^\top\rangle_F\right| \ge \frac{a^\star s}{\lambda} \,\middle|\, \mathcal{D}_n\right).$$

Applying Lemma H.7 with $H = H^\star$ and $r_H = n$, we obtain

$$\mathcal{R}(W^\star) \le 2\exp\left[-C' \min\left\{\frac{D^2}{n}\left(\frac{a^\star s}{\lambda}\right)^2, D\frac{a^\star s}{\lambda}\right\}\right]. \tag{C.111}$$

It remains to derive the simplified bound. By Theorem 1,

$$\left|a^\star - \frac{\lambda n}{sD}\Gamma_\rho\right| \le C\frac{\lambda n}{sD}\left(\sqrt{\frac{\log(1/\delta)}{n}} + \frac{L^3}{n} + \sqrt{\frac{n}{D}}\right).$$

Thus, by replacing c with a sufficiently small constant and C with a sufficiently large constant, we have $\frac{\lambda n}{sD}\frac{\Gamma_\rho}{2} \le a^\star \le \frac{\lambda n}{sD}\frac{3\Gamma_\rho}{2}$. Equivalently, $\frac{\Gamma_\rho}{2}\frac{n}{D} \le \frac{a^\star s}{\lambda} \le \frac{3\Gamma_\rho}{2}\frac{n}{D}$. Substituting this lower bound into (C.111) gives

$$\mathcal{R}(W^\star) \le 2\exp\left[-C' \min\left\{\Gamma_\rho^2 n, \Gamma_\rho n\right\}\right] = 2\exp(-C'\Gamma_\rho^2 n)$$

after renaming C', where the equality is due to $\Gamma_\rho < 1/4 < 1$ on $\rho \in (0, 1/2)$. Setting $C_0 := C'\Gamma_\rho^2$, we obtain the desired upper bound.

For the lower bound, let K_0 be the universal constant from Lemma H.7. By replacing C with a larger constant if necessary so that $C \ge K_0$, we have $D \ge Cn \ge K_0 n$.

Since $\frac{a^\star s}{\lambda} \le \frac{3\Gamma_\rho}{2}\frac{n}{D}$ and $\Gamma_\rho < 1/4$, we have

$$\frac{a^\star s}{\lambda} < \frac{n}{D}.$$

Therefore the lower-tail estimate (H.10) applies with $H = H^\star$, $r_H = n$, and $t = a^\star s/\lambda$. Combining this with the tail identity gives

$$\mathcal{R}(W^\star) \ge \frac{c_0}{2}\exp\left[-C_0'\frac{D^2}{n}\left(\frac{a^\star s}{\lambda}\right)^2\right].$$

Using again

$$\frac{a^\star s}{\lambda} \le \frac{3\Gamma_\rho}{2}\frac{n}{D},$$

we obtain

$$\mathcal{R}(W^\star) \ge \frac{c_0}{2}\exp\left[-\frac{9}{4}C_0'\Gamma_\rho^2 n\right].$$

Renaming the universal constant c_0 and setting $C_0 = \frac{9}{4}C_0'\Gamma_\rho^2$ proves $\mathcal{R}(W^\star) \ge c_0 e^{-C_0 n}$.

Finally, Theorem 1 gives $\frac{W_k^{\mathrm{SpecGD}}}{\|W_k^{\mathrm{SpecGD}}\|_{\mathrm{op}}} \to W^\star$. The classifier is invariant under positive rescaling of W_k^{SpecGD}. Moreover, for fixed training data, the fresh-noise distribution is continuous, so $\mathbb{P}\left(y\langle W^\star, X\rangle_F = 0 \,|\, \mathcal{D}_n\right) = 0$. Therefore the indicators $\mathbf{1}\{y\langle W_k^{\mathrm{SpecGD}}, X\rangle_F \le 0\}$ converge almost surely to $\mathbf{1}\{y\langle W^\star, X\rangle_F \le 0\}$, and dominated convergence gives $\mathcal{R}(W_k^{\mathrm{SpecGD}}) \to \mathcal{R}(W^\star)$. □

C.12 Proof of Proposition 1

Proof. Set $a_{\star\star} := (2\|S\|_1)^{-1}\log(n_+/n_-)$. For $W \in \mathbb{R}^{d\times d}$, let $m_i(W) := \langle W, \widetilde{y}_i X_i\rangle_F = \langle W, \xi_i S + \lambda N_i\rangle_F$ and $q_i(W) := e^{-m_i(W)}$. At iteration k, write $q_{i,k} := q_i(W_k)$ and $s_k := \sum_i \xi_i q_{i,k}$. Then

$$-\nabla L(W_k) = \frac{1}{n}\left(s_k S + \lambda \sum_{i=1}^n q_{i,k} N_i\right).$$

The factor $1/n$ is irrelevant for the polar factor. Since the signal block is orthogonal to the noise block, and the N_i's have mutually orthogonal left and right singular directions, $q_{i,k} > 0$ implies

$$\mathrm{Polar}(-\nabla L(W_k)) = \mathrm{sgn}(s_k) P_S + \sum_{i=1}^n N_i,$$

where $\mathrm{sgn}(0) = 0$.

Thus, starting from $W_0 = 0$, the iterates have the exact form

$$W_k = a_k P_S + \tau_k \sum_{i=1}^n N_i, \qquad \tau_k := \sum_{\ell<k} \eta_\ell,$$

where $a_0 = 0$ and $a_{k+1} = a_k + \eta_k \,\mathrm{sgn}(s_k)$. Using this representation and orthogonality,

$$m_i(W_k) = \xi_i \|S\|_1 a_k + \lambda \tau_k,$$

so

$$s_k = e^{-\lambda\tau_k}\left(n_+ e^{-\|S\|_1 a_k} - n_- e^{\|S\|_1 a_k}\right).$$

Hence

$$\mathrm{sgn}(s_k) = \mathrm{sgn}(a_{\star\star} - a_k),$$

and therefore the signal coefficient obeys the scalar recursion

$$a_{k+1} = a_k + \eta_k \,\mathrm{sgn}(a_{\star\star} - a_k).$$

Let $K := \sup\{k : \tau_k < a_{\star\star}\}$. Since $\sum_k \eta_k = \infty$, this K is finite. Before crossing $a_{\star\star}$, the recursion gives $a_k = \tau_k$, and hence $|a_K - a_{\star\star}| \le \eta_K$. Since the stepsizes are non-increasing, the recursion moves toward $a_{\star\star}$ by steps of size at most η_K for all $k \ge K$. Consequently

$$|a_k - a_{\star\star}| \le \eta_K, \qquad k \ge K.$$

Thus $|a_k| \le a_{\star\star} + \eta_K$ for $k \ge K$. Dividing the exact representation of W_k by τ_k gives

$$\left\| \frac{W_k^{\mathrm{SpecGD}}}{\tau_k} - \sum_{i=1}^n N_i \right\|_F = \frac{|a_k|}{\tau_k}\|P_S\|_F \le \frac{(a_{\star\star} + \eta_K)\sqrt{r}}{\tau_k}, \qquad k \ge K.$$

Since $\tau_k \to \infty$, the claimed directional convergence follows. □

D Proof of Theorem 2

We will use the strategy developed in Section C for the DSM.

Proof of Theorem 2. Absorbing $\tilde{y}_i$ o $\boldsymbol{u}_i$, let $U := [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n] \in \mathbb{R}^{D\times n}$, $G := U^\top U$, and $\boldsymbol{\xi} := (\xi_1, \ldots, \xi_n)^\top$.

Step 1: exact dual objective. By Proposition B.1, the dual objective is

$$F(\boldsymbol{\alpha}) = \left\| \sum_{i=1}^{n} \alpha_i \widetilde{y}_i X_i \right\|_1, \qquad \boldsymbol{\alpha} \in \Delta_n.$$

Under the CSM,

$$\sum_{i=1}^{n} \alpha_i \widetilde{y}_i X_i = (\boldsymbol{\xi}^\top \boldsymbol{\alpha}) S + \lambda (U\boldsymbol{\alpha}) \boldsymbol{v}_0^\top.$$

The two terms have mutually orthogonal row and column spaces, so their nuclear norms add. Moreover, $(U\boldsymbol{\alpha})\boldsymbol{v}_0^\top$ has rank one and, since $\|\boldsymbol{v}_0\|_2 = 1$,

$$\|(U\boldsymbol{\alpha})\boldsymbol{v}_0^\top\|_1 = \|U\boldsymbol{\alpha}\|_2 = \sqrt{\boldsymbol{\alpha}^\top G \boldsymbol{\alpha}}.$$

By Lemma B.1 and $\|S\|_1 = s$,

$$F(\boldsymbol{\alpha}) = s|\boldsymbol{\xi}^\top \boldsymbol{\alpha}| + \lambda \sqrt{\boldsymbol{\alpha}^\top G \boldsymbol{\alpha}}. \tag{D.1}$$

Step 2: high-probability geometry. Let $\widehat{\rho} := n_-/n$. By Hoeffding's inequality, with probability at least $1 - \delta/3$,

$$|\widehat{\rho} - \rho| \le \sqrt{\frac{\log(6/\delta)}{2n}}. \tag{D.2}$$

By Lemma H.4, applied with failure probability $\delta/3$ and enlarging the universal constant C_G, there exists a universal constant $K > 0$ such that

$$\|G - I_n\|_{\mathrm{op}} \le K \sqrt{\frac{n + \log(n/\delta)}{D}}. \tag{D.3}$$

Increasing the dimensional constant if necessary, the right-hand side is at most $1/2$, and therefore

$$\|G^{-1} - I_n\|_{\mathrm{op}} \le K \sqrt{\frac{n + \log(n/\delta)}{D}}, \tag{D.4}$$

after enlarging the same universal constant K if necessary.

By Lemma H.6, applied with failure probability $\delta/3$, and enlarging K once more if necessary,

$$\max_{\boldsymbol{x} \in \{\mathbf{1}, \boldsymbol{\xi}\}} \|(G^{-1} - I_n)\boldsymbol{x}\|_\infty \le K \sqrt{\frac{n \log(n/\delta)}{D}}. \tag{D.5}$$

Let $\mathcal{E}$ be the intersection of these three events. Then $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$, and we work on $\mathcal{E}$ throughout. Define

$$\varepsilon_{\mathrm{op}} := K \sqrt{\frac{n + \log(n/\delta)}{D}}, \qquad \varepsilon_\infty := K \sqrt{\frac{n \log(n/\delta)}{D}}.$$

By increasing the constants in the assumptions if necessary, $\widehat{\rho}$ lies in a fixed compact subinterval of $(0, 1/2)$ depending only on ρ.

Step 3: solution of the dual problem. For $\boldsymbol{\alpha} \in \Delta_n$, let $q(\boldsymbol{\alpha}) := \sum_{i \in I_+} \alpha_i$, so that $\boldsymbol{\xi}^\top \boldsymbol{\alpha} = 2q(\boldsymbol{\alpha}) - 1$, and define

$$\psi(q) := \min_{\substack{\boldsymbol{\alpha} \in \Delta_n \\ q(\boldsymbol{\alpha}) = q}} \sqrt{\boldsymbol{\alpha}^\top G \boldsymbol{\alpha}}.$$

Set

$$A := \mathbf{1}^\top G^{-1} \mathbf{1}, \qquad B := \mathbf{1}^\top G^{-1} \boldsymbol{\xi}, \qquad C := \boldsymbol{\xi}^\top G^{-1} \boldsymbol{\xi}, \qquad \Delta := AC - B^2.$$

The operator-norm bound in Step 2 immediately gives

$$\left| \frac{A}{n} - 1 \right| \vee \left| \frac{B}{n} - (1 - 2\widehat{\rho}) \right| \vee \left| \frac{C}{n} - 1 \right| \le \varepsilon_{\mathrm{op}}. \tag{D.6}$$

We first solve the inner problem.

Lemma D.1. *On $\mathcal{E}$, for every $q \in [1/2,\, 1/2 + B/(2A)]$, the minimizer defining $\psi(q)$ is unique, belongs to $\mathrm{int}(\Delta_n)$, and is given by*

$$\boldsymbol{\alpha}^\star(q) = \frac{G^{-1}\left[\{C - (2q-1)B\}\mathbf{1} + \{(2q-1)A - B\}\boldsymbol{\xi}\right]}{\Delta}. \tag{D.7}$$

Moreover,

$$\psi(q)^2 = \frac{C - 2(2q-1)B + (2q-1)^2 A}{\Delta}. \tag{D.8}$$

Proof. Since $G \succ 0$, consider first the equality-constrained problem

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \boldsymbol{\alpha}^\top G \boldsymbol{\alpha} \quad \text{subject to} \quad \mathbf{1}^\top\boldsymbol{\alpha} = 1, \qquad \boldsymbol{\xi}^\top\boldsymbol{\alpha} = 2q - 1.$$

Its objective is strictly convex. Introducing Lagrange multipliers $a, b \in \mathbb{R}$, the stationarity condition gives

$$G\boldsymbol{\alpha} = a\mathbf{1} + b\boldsymbol{\xi},$$

and hence $\boldsymbol{\alpha} = G^{-1}(a\mathbf{1} + b\boldsymbol{\xi})$. The two constraints therefore imply

$$\begin{pmatrix} A & B \\ B & C \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 1 \\ 2q-1 \end{pmatrix}.$$

On $\mathcal{E}$, both I_+ and I_- are nonempty, so $\mathbf{1}$ and $\boldsymbol{\xi}$ are linearly independent. Since $G^{-1} \succ 0$, the matrix above is their Gram matrix in the G^{-1}-inner product and hence $\Delta = AC - B^2 > 0$. Thus

$$a = \frac{C - (2q-1)B}{\Delta}, \qquad b = \frac{(2q-1)A - B}{\Delta},$$

which gives (D.7). Moreover, using $G\boldsymbol{\alpha}^\star = a\mathbf{1} + b\boldsymbol{\xi}$ and the two constraints,

$$\boldsymbol{\alpha}^{\star\top} G \boldsymbol{\alpha}^\star = a + b(2q-1) = \frac{C - 2(2q-1)B + (2q-1)^2 A}{\Delta},$$

proving (D.8).

It remains to verify that $\boldsymbol{\alpha}^\star(q)$ is strictly positive. By (D.6),

$$\frac{1}{2} + \frac{B}{2A} = 1 - \widehat{\rho} + O_\rho(\varepsilon_{\mathrm{op}}).$$

Hence, after increasing the dimensional constant if necessary, $1 - q \geq \rho/2$ throughout the stated interval.

When $G = I_n$, the corresponding equality-constrained minimizer is

$$\overline{\alpha}_{0,i}(q) = \begin{cases} q/n_+, & i \in I_+, \\ (1-q)/n_-, & i \in I_-, \end{cases}$$

and therefore

$$\min_{i\in[n]} \overline{\alpha}_{0,i}(q) \geq \frac{\rho}{2n}.$$

Let $a_0(q), b_0(q)$ denote the corresponding Lagrange multipliers when $G = I_n$, so that $\overline{\boldsymbol{\alpha}}_0(q) = a_0(q)\mathbf{1} + b_0(q)\boldsymbol{\xi}$. Since $\widehat{\rho}$ stays in a fixed compact subset of $(0, 1/2)$, (D.6) and the explicit formulas for a, b give, uniformly over the stated interval,

$$|a - a_0(q)| + |b - b_0(q)| \leq K_\rho \frac{\varepsilon_{\mathrm{op}}}{n}, \qquad |a| + |b| \leq \frac{K_\rho}{n},$$

for a constant $K_\rho > 0$ depending only on ρ. Therefore

$$\begin{aligned} \boldsymbol{\alpha}^\star(q) - \overline{\boldsymbol{\alpha}}_0(q) = {} & (G^{-1} - I_n)(a\mathbf{1} + b\boldsymbol{\xi}) \\ & + (a - a_0(q))\mathbf{1} + (b - b_0(q))\boldsymbol{\xi}. \end{aligned}$$

Using the inverse-Gram coordinate bound from Step 2,

$$\|\boldsymbol{\alpha}^\star(q) - \overline{\boldsymbol{\alpha}}_0(q)\|_\infty \leq K_\rho \frac{\varepsilon_\infty + \varepsilon_{\mathrm{op}}}{n} \leq 2K_\rho \frac{\varepsilon_\infty}{n}.$$

Increasing the dimensional constant once more if necessary, the right-hand side is at most $\rho/(4n)$. Hence

$$\min_{i\in[n]} \alpha_i^\star(q) \geq \frac{\rho}{4n} > 0.$$

Thus the equality-constrained minimizer belongs to $\mathrm{int}(\Delta_n)$ and is therefore the unique minimizer defining $\psi(q)$. □

By (D.1), the dual problem reduces to

$$\min_{q\in[0,1]} \left\{ 2s \left| q - \frac{1}{2} \right| + \lambda\psi(q) \right\}.$$

From (D.8),

$$\psi'(q) = \frac{2\{A(2q-1) - B\}}{\Delta\psi(q)},$$

and hence ψ is uniquely minimized at $q_{\mathrm{sh}} := \frac{1}{2} + \frac{B}{2A}$. By (D.6), $B > 0$, so $q_{\mathrm{sh}} > 1/2$.

We next localize the minimizer of the reduced dual problem. If $q < 1/2$, increasing q toward $1/2$ decreases both $|q - 1/2|$ and $\psi(q)$, since $q < 1/2 < q_{\mathrm{sh}}$. Similarly, if $q > q_{\mathrm{sh}}$, decreasing q toward q_{sh} decreases both terms. Therefore every minimizer belongs to $[1/2, q_{\mathrm{sh}}]$.

On this interval, differentiating once more gives

$$\psi''(q) = \frac{4}{\Delta\psi(q)^3} > 0.$$

Thus the reduced objective is strictly convex on $[1/2, q_{\mathrm{sh}}]$ and admits a unique minimizer, denoted by $q^\star$. Set $\boldsymbol{\alpha}^\star := \boldsymbol{\alpha}^\star(q^\star)$. By Lemma D.1, $\boldsymbol{\alpha}^\star \in \mathrm{int}(\Delta_n)$ and is the unique minimizer of F.

Finally,

$$\psi'\left(\frac{1}{2}\right) = -\frac{2B}{\sqrt{C\Delta}}.$$

Since the right derivative of the reduced objective at $q = 1/2$ is $2s - \frac{2\lambda B}{\sqrt{C\Delta}}$, we obtain

$$q^\star = \frac{1}{2} \quad \Longleftrightarrow \quad \frac{\lambda B}{s\sqrt{C\Delta}} \leq 1; \tag{D.9}$$

otherwise $q^\star \in (1/2, q_{\mathrm{sh}})$.

Step 4: convergence of the dual weights. By Step 3, F has a unique minimizer $\boldsymbol{\alpha}^\star \in \mathrm{int}(\Delta_n)$. By Theorem B.1, it remains only to verify local relative growth at $\boldsymbol{\alpha}^\star$.

Set $r(\boldsymbol{\alpha}) := \sqrt{\boldsymbol{\alpha}^\top G\boldsymbol{\alpha}}$. For every $\boldsymbol{\alpha} \in \Delta_n$ and every $\boldsymbol{v}$ satisfying $\mathbf{1}^\top \boldsymbol{v} = 0$,

$$\boldsymbol{v}^\top \nabla^2 r(\boldsymbol{\alpha}) \boldsymbol{v} = \frac{(\boldsymbol{v}^\top G\boldsymbol{v})(\boldsymbol{\alpha}^\top G\boldsymbol{\alpha}) - (\boldsymbol{v}^\top G\boldsymbol{\alpha})^2}{(\boldsymbol{\alpha}^\top G\boldsymbol{\alpha})^{3/2}}.$$

By Cauchy–Schwarz in the G-inner product, this quantity is nonnegative, with equality only if $\boldsymbol{v}$ and $\boldsymbol{\alpha}$ are collinear. Since $\mathbf{1}^\top \boldsymbol{v} = 0$ while $\mathbf{1}^\top \boldsymbol{\alpha} = 1$, equality is impossible for $\boldsymbol{v} \neq 0$. Hence, by continuity, there exist a neighborhood $\mathcal{U}$ of $\boldsymbol{\alpha}^\star$ and $\mu_0 > 0$ such that

$$\boldsymbol{v}^\top \nabla^2 r(\boldsymbol{\alpha}) \boldsymbol{v} \geq \mu_0 \|\boldsymbol{v}\|_2^2$$

for every $\boldsymbol{\alpha} \in \mathcal{U}$ and every tangent vector $\boldsymbol{v}$. In other words, $r(\boldsymbol{\alpha})$ is μ_0-strongly convex on $\mathcal{U}$ relative to Δ_n.

Since $\boldsymbol{\alpha} \mapsto s|\boldsymbol{\xi}^\top \boldsymbol{\alpha}|$ is convex, it follows that

$$F(\boldsymbol{\alpha}) - F(\boldsymbol{\alpha}^\star) \geq \frac{\lambda\mu_0}{2} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}^\star\|_2^2, \qquad \boldsymbol{\alpha} \in \mathcal{U}.$$

Because $\boldsymbol{\alpha}^\star \in \mathrm{int}(\Delta_n)$, Lemma B.3 gives, after shrinking $\mathcal{U}$ if necessary,

$$D_{\mathrm{KL}}(\boldsymbol{\alpha}^\star \| \boldsymbol{\alpha}) \leq C_{\mathrm{KL}} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}^\star\|_2^2.$$

Thus F has local relative growth at $\boldsymbol{\alpha}^\star$. Let $\boldsymbol{\alpha}_k$ be the normalized loss weights as in (B.2). Theorem B.1 yields

$$\boldsymbol{\alpha}_k \longrightarrow \boldsymbol{\alpha}^\star. \tag{D.10}$$

Step 5: characterization of the primal direction. Define

$$H(\boldsymbol{\alpha}) := \frac{U\boldsymbol{\alpha}}{\sqrt{\boldsymbol{\alpha}^\top G\boldsymbol{\alpha}}}\boldsymbol{v}_0^\top.$$

By Step 4, $\boldsymbol{\alpha}_k \to \boldsymbol{\alpha}^\star = \boldsymbol{\alpha}^\star(q^\star)$, and hence

$$H_k \longrightarrow H^\star := \frac{U\boldsymbol{\alpha}^\star}{\sqrt{\boldsymbol{\alpha}^{\star\top} G\boldsymbol{\alpha}^\star}}\boldsymbol{v}_0^\top. \tag{D.11}$$

In particular, $H^\star$ has rank one and its unique nonzero singular value is 1.

Since the dual minimizer is unique and interior and F has local relative growth, Proposition B.2 gives

$$\frac{W_k^{\mathrm{SpecGD}}}{\tau_k} \longrightarrow a^\star P_S + H^\star, \qquad a^\star = -\frac{\lambda}{2s}(h_+^\star - h_-^\star).$$

For $t := 2q-1$, (D.7) gives

$$G\boldsymbol{\alpha}^\star(q) = \frac{\{C - tB\}\mathbf{1} + \{tA - B\}\boldsymbol{\xi}}{\Delta}.$$

Since $\boldsymbol{\alpha}^\star(q)^\top G\boldsymbol{\alpha}^\star(q) = \psi(q)^2$, we obtain

$$h_+(q) - h_-(q) = \frac{2\{(2q-1)A - B\}}{\Delta\psi(q)} = \psi'(q).$$

Consequently,

$$a^\star = -\frac{\lambda}{2s}\psi'(q^\star).$$

Since $q^\star$ minimizes $2s(q-1/2) + \lambda\psi(q)$ over $[1/2, q_{\mathrm{sh}}]$ and ψ' is strictly increasing,

$$-\frac{\lambda}{2s}\psi'(q^\star) = \min\left\{1,\ -\frac{\lambda}{2s}\psi'(1/2)\right\}.$$

Using

$$\psi'(1/2) = -\frac{2B}{\sqrt{C\Delta}},$$

we conclude that

$$a^\star = \min\left\{1,\ \frac{\lambda B}{s\sqrt{C\Delta}}\right\}. \tag{D.12}$$

Step 6: approximation of the signal coefficient. Define

$$\Phi(a,b,c) := \frac{b}{\sqrt{c(ac-b^2)}}.$$

Since $\widehat{\rho}$ remains in a fixed compact subset of $(0,1/2)$, Φ is Lipschitz on a fixed neighborhood of $(1, 1-2\widehat{\rho}, 1)$. Noting that

$$\sqrt{n}\,\frac{B}{\sqrt{C\Delta}} = \Phi\left(\frac{A}{n}, \frac{B}{n}, \frac{C}{n}\right),$$

we obtain, by (D.6),

$$\left|\sqrt{n}\,\frac{B}{\sqrt{C\Delta}} - \frac{1-2\widehat{\rho}}{2\sqrt{\widehat{\rho}(1-\widehat{\rho})}}\right| \le K_\rho\varepsilon_{\mathrm{op}} \le K_\rho K\sqrt{\frac{n+\log(n/\delta)}{D}}.$$

By (D.2) and the Lipschitz continuity of $q \mapsto (1-2q)/(2\sqrt{q(1-q)})$ on a fixed neighborhood of ρ,

$$\left|\sqrt{n}\,\frac{B}{\sqrt{C\Delta}} - \Gamma_\rho'\right| \le K_\rho\left(\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{n+\log(n/\delta)}{D}}\right).$$

Finally, since $x \mapsto \min\{1,x\}$ is 1-Lipschitz, (D.12) yields

$$\left|a^\star - \min\left\{1,\ \Gamma_\rho'\frac{\lambda}{s\sqrt{n}}\right\}\right| \le K_\rho\frac{\lambda}{s\sqrt{n}}\left(\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{n+\log(n/\delta)}{D}}\right).$$

Replacing K_ρ with a larger constant, this completes the proof. □

Proof of Corollary 2. Let (X, y) be an independent clean sample. We may write

$$yX = S + \lambda \boldsymbol{u}\boldsymbol{v}_0^\top, \qquad \boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}^{D-1}).$$

Since the signal and shortcut blocks are orthogonal,

$$y\langle W^\star, X\rangle_F = a^\star s + \lambda \langle \boldsymbol{h}^\star, \boldsymbol{u}\rangle,$$

where $\|\boldsymbol{h}^\star\|_2 = 1$. Hence, conditionally on the training sample,

$$\mathcal{R}(W^\star) = \mathbb{P}\left(\langle \boldsymbol{h}^\star, \boldsymbol{u}\rangle \le -\frac{a^\star s}{\lambda} \,\middle|\, \mathcal{D}_n \right).$$

Set $t := \dfrac{a^\star s}{\lambda}$. By Theorem 2,

$$t = \min\left\{ \frac{s}{\lambda}, \frac{\Gamma'_\rho + \varepsilon_n}{\sqrt{n}} \right\}, \qquad |\varepsilon_n| \le C\left(\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{n}{D}} \right).$$

By choosing the constants in the assumptions of Theorem 2 appropriately, we may ensure that

$$c_\rho \min\left\{ \frac{s}{\lambda}, \frac{1}{\sqrt{n}} \right\} \le t \le C_\rho \min\left\{ \frac{s}{\lambda}, \frac{1}{\sqrt{n}} \right\},$$

for constants $c_\rho, C_\rho > 0$ depending only on ρ. Moreover, since $\log(n/\delta) \ge 1$, by decreasing c further if necessary, we may also ensure that $t \le 1/2$.

Applying Lemma H.2 conditionally on $\mathcal{D}_n$ gives

$$c_0 \exp(-C_0 D t^2) \le \mathcal{R}(W^\star) \le 2\exp(-cDt^2).$$

Combining this with the preceding bounds on t yields

$$c_0 \exp\left(-c_1 D \min\left\{ \frac{1}{n}, \frac{s^2}{\lambda^2} \right\} \right) \le \mathcal{R}(W^\star) \le 2\exp\left(-c_2 D \min\left\{ \frac{1}{n}, \frac{s^2}{\lambda^2} \right\} \right),$$

where $c_0, c_1, c_2 > 0$ depend only on ρ.

Finally, the convergence of the risks follows by the same argument as in Corollary 1. □

E Analysis of GD

In this section, we study the late-stage implicit bias and generalization of GD under both DSM and CSM.

By Lemma H.1, the training sample is linearly separable when $D \ge n$. Therefore, the standard implicit bias result (Gunasekar et al., 2018) implies that W_k^{GD} converges in direction toward the unique Euclidean max-margin classifier $W_{\mathrm{GD}}^\star$, where uniqueness follows from strict convexity.

The Euclidean hard-margin problem for both DSM and CSM has the form

$$\min_W \frac{1}{2}\|W\|_F^2 \qquad \text{subject to} \qquad \langle W, \widetilde{y}_i X_i\rangle_F \ge 1, \quad i \in [n]. \tag{E.1}$$

Its Lagrangian is

$$\mathcal{L}(W, \boldsymbol{\alpha}) = \frac{1}{2}\|W\|_F^2 + \sum_{i=1}^n \alpha_i\big(1 - \langle W, \widetilde{y}_i X_i\rangle_F\big), \qquad \boldsymbol{\alpha} \ge 0.$$

Minimization over W gives $W = \sum_{i=1}^n \alpha_i \widetilde{y}_i X_i$, and hence the dual problem is

$$\max_{\boldsymbol{\alpha}\ge 0}\left\{ \mathbf{1}^\top \boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^\top K \boldsymbol{\alpha} \right\}, \tag{E.2}$$

where $K = (K_{ij})_{i,j=1}^n$ is the Gram matrix of the signed observations, with $K_{ij} := \langle \widetilde{y}_i X_i, \widetilde{y}_j X_j\rangle_F$. Set $s_F := \|S\|_F$ and $G := (\langle N_i, N_j\rangle_F)_{i,j=1}^n = (U^\top U) \circ (\widehat{V}^\top \widehat{V})$. Since $\widetilde{y}_i X_i = \xi_i S + \lambda N_i$, signal–shortcut orthogonality gives

$$K = \lambda^2 G + s_F^2 \boldsymbol{\xi}\boldsymbol{\xi}^\top. \tag{E.3}$$

The following lemma isolates the common GD dual algebra; the DSM and CSM analyses differ only in the control of G^{-1} and the positivity of the resulting optimizer.

Lemma E.1 (Unconstrained GD dual solution)**.** *Suppose $G \succ 0$. Dropping the constraint $\boldsymbol{\alpha} \geq 0$ in* (E.2)*, the unique maximizer is*

$$\boldsymbol{\alpha}^\star = \frac{1}{\lambda^2} G^{-1} \left(\mathbf{1} - a^\star_{\mathrm{GD}} s_F^2 \boldsymbol{\xi} \right), \qquad a^\star_{\mathrm{GD}} = \frac{\boldsymbol{\xi}^\top G^{-1} \mathbf{1}}{\lambda^2 + s_F^2 \boldsymbol{\xi}^\top G^{-1} \boldsymbol{\xi}} = \boldsymbol{\xi}^\top \boldsymbol{\alpha}^\star.$$

If $\boldsymbol{\alpha}^\star > 0$, then it is also the optimizer of (E.2)*, and*

$$W^\star_{\mathrm{GD}} = a^\star_{\mathrm{GD}} S + \lambda \sum_{i=1}^n \alpha_i^\star N_i.$$

Proof. The unconstrained first-order condition gives $\boldsymbol{\alpha}^\star = K^{-1}\mathbf{1}$. Applying the Sherman–Morrison formula to (E.3) gives the displayed formula and $a^\star_{\mathrm{GD}} = \boldsymbol{\xi}^\top \boldsymbol{\alpha}^\star$. If $\boldsymbol{\alpha}^\star > 0$, the inequality constraints are inactive, and the final identity follows from $W^\star_{\mathrm{GD}} = \sum_i \alpha_i^\star \widetilde{y}_i X_i$. □

We now study DSM and CSM separately.

Proposition E.1 (Directional convergence of GD under DSM)**.** *Assume the sample $\mathcal{D}_n$ is generated from the DSM and that assumption A1 holds. Let $0 < \delta < e^{-1}$. There exists a constant $C_{\dim} > 0$, depending only on ρ, such that, if*

$$n \geq C_{\dim} \log(4/\delta), \qquad D \geq C_{\dim} n \log(8n^2/\delta),$$

then, with probability at least $1 - \delta$ over the training sample, GD converges in direction toward the unique Euclidean max-margin classifier

$$W^\star_{\mathrm{GD}} = a^\star_{\mathrm{GD}} S + H^\star_{\mathrm{GD}}.$$

Moreover, $\operatorname{rank}(H^\star_{\mathrm{GD}}) = n$, and there exists a universal constant $C > 0$ such that

$$\left| a^\star_{\mathrm{GD}} - \frac{(1-2\rho)n}{\lambda^2 + n\|S\|_F^2} \right| \leq C \frac{n}{\lambda^2 + n\|S\|_F^2} \left(\sqrt{\frac{\log(4/\delta)}{n}} + \frac{n \log(8n^2/\delta)}{D} \right).$$

There also exist constants $c_H, C_H > 0$, depending only on ρ, such that

$$\frac{c_H}{\lambda} \leq \sigma_n(H^\star_{\mathrm{GD}}) \leq \sigma_1(H^\star_{\mathrm{GD}}) \leq \frac{C_H}{\lambda}.$$

Proof. Since multiplication by $\widetilde{y}_i \in \{\pm 1\}$ preserves the uniform distribution on the sphere, the columns of $\widehat{V}$ satisfy the assumptions of Lemma H.5.

Concentration of the shortcut Gram matrix. Apply Lemma H.5 with failure probability $\delta/2$. On the resulting event,

$$\max_{i \neq j} \{ |\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle|, |\langle \widehat{\boldsymbol{v}}_i, \widehat{\boldsymbol{v}}_j \rangle| \} \leq C_G \sqrt{\frac{\log(8n^2/\delta)}{D}}.$$

Since $G_{ii} = 1$ and $|G_{ij}| = |\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle| |\langle \widehat{\boldsymbol{v}}_i, \widehat{\boldsymbol{v}}_j \rangle|$ for $i \neq j$, this gives $\|G - I_n\|_{\infty \to \infty} \leq C_G^2 n \log(8n^2/\delta)/D =: \varepsilon_G$. Choosing $C_{\dim}$ sufficiently large, depending only on ρ, we may assume $\varepsilon_G \leq 1/2$. Thus G is invertible and

$$\|G^{-1} - I_n\|_{\infty \to \infty} \leq \frac{\varepsilon_G}{1 - \varepsilon_G} \leq 2\varepsilon_G. \tag{E.4}$$

After increasing $C_{\dim}$ if necessary, the same application of Lemma H.5 also gives

$$\frac{1}{2} I_n \preceq U^\top U \preceq \frac{3}{2} I_n, \qquad \frac{1}{2} I_n \preceq \widehat{V}^\top \widehat{V} \preceq \frac{3}{2} I_n. \tag{E.5}$$

By Lemma E.1, the unconstrained dual optimizer is explicit in terms of G^{-1}. We next approximate its signal coefficient and then verify positivity.

Approximation of the signal coefficient. Set $m := 1 - 2\rho$. By Hoeffding's inequality, with probability at least $1 - \delta/2$,

$$\left|\mathbf{1}^\top \boldsymbol{\xi} - mn\right| \le \sqrt{2n \log(4/\delta)}. \tag{E.6}$$

We henceforth work on the intersection of this event with the Gram concentration event, which has probability at least $1 - \delta$.

Set $b := \boldsymbol{\xi}^\top G^{-1}\mathbf{1}$ and $c := \boldsymbol{\xi}^\top G^{-1}\boldsymbol{\xi}$. By (E.4), $|b - \boldsymbol{\xi}^\top \mathbf{1}| \le 2n\varepsilon_G$ and $|c - n| \le 2n\varepsilon_G$. Hence, by (E.6),

$$|b - mn| \le \sqrt{2n \log(4/\delta)} + 2n\varepsilon_G, \qquad |c - n| \le 2n\varepsilon_G. \tag{E.7}$$

By Lemma E.1, $a^\star_{\mathrm{GD}} = b/(\lambda^2 + s_F^2 c)$. Increasing C_{dim} if necessary, we may assume $\varepsilon_G \le 1/4$, so $n/2 \le c \le 3n/2$ and hence $\lambda^2 + s_F^2 c \asymp \lambda^2 + n s_F^2$ with universal constants. Moreover,

$$a^\star_{\mathrm{GD}} - \frac{mn}{\lambda^2 + n s_F^2} = \frac{b - mn}{\lambda^2 + s_F^2 c} + \frac{mn s_F^2 (n - c)}{(\lambda^2 + s_F^2 c)(\lambda^2 + n s_F^2)}.$$

Using (E.7) and the definition of ε_G therefore gives

$$\left| a^\star_{\mathrm{GD}} - \frac{mn}{\lambda^2 + n s_F^2} \right| \le C \frac{n}{\lambda^2 + n s_F^2} \left(\sqrt{\frac{\log(4/\delta)}{n}} + \frac{n \log(8n^2/\delta)}{D} \right). \tag{E.8}$$

Positivity of the dual weights. By choosing C_{dim} sufficiently large depending only on ρ, the quantity in parentheses in (E.8), multiplied by C, is at most $\min\{m/2, \rho\}$. Hence $a^\star_{\mathrm{GD}} \ge \frac{m}{2}\, n/(\lambda^2 + n s_F^2) > 0$ and $0 < a^\star_{\mathrm{GD}} s_F^2 \le m + \rho = 1 - \rho$.

Set $\boldsymbol{z} := \mathbf{1} - a^\star_{\mathrm{GD}} s_F^2 \boldsymbol{\xi}$. Then $\rho \le z_i \le 2$ for every i, and by (E.4), $\|G^{-1}\boldsymbol{z} - \boldsymbol{z}\|_\infty \le \|G^{-1} - I_n\|_{\infty\to\infty} \|\boldsymbol{z}\|_\infty \le 4\varepsilon_G$. Increasing C_{dim} once more so that $4\varepsilon_G \le \rho/2$, we obtain $\rho/2 \le (G^{-1}\boldsymbol{z})_i \le 3$. Thus Lemma E.1 yields

$$\frac{\rho}{2\lambda^2} \le \alpha_i^\star \le \frac{3}{\lambda^2}, \qquad i \in [n]. \tag{E.9}$$

In particular, $\boldsymbol{\alpha}^\star > 0$, so the same lemma gives $W^\star_{\mathrm{GD}} = a^\star_{\mathrm{GD}} S + H^\star_{\mathrm{GD}}$, where $H^\star_{\mathrm{GD}} := \lambda \sum_{i=1}^n \alpha_i^\star N_i$.

Singular values of the shortcut component. Let $\Lambda^\star := \mathrm{diag}(\alpha_1^\star, \dots, \alpha_n^\star)$, so that $H^\star_{\mathrm{GD}} = \lambda U \Lambda^\star \widehat{V}^\top$. By Lemma H.12, its nonzero singular values are λ times those of $(U^\top U)^{1/2} \Lambda^\star (\widehat{V}^\top \widehat{V})^{1/2}$. Thus, by (E.5) and (E.9),

$$\sigma_n(H^\star_{\mathrm{GD}}) \ge \lambda \sqrt{\lambda_{\min}(U^\top U)}\, \min_i \alpha_i^\star \sqrt{\lambda_{\min}(\widehat{V}^\top \widehat{V})} \ge \frac{\rho}{4\lambda},$$
$$\sigma_1(H^\star_{\mathrm{GD}}) \le \lambda \sqrt{\lambda_{\max}(U^\top U)}\, \max_i \alpha_i^\star \sqrt{\lambda_{\max}(\widehat{V}^\top \widehat{V})} \le \frac{9}{2\lambda}.$$

Hence the claimed singular-value bounds hold with $c_H = \rho/4$ and $C_H = 9/2$, and $\mathrm{rank}(H^\star_{\mathrm{GD}}) = n$. The directional convergence of W_k^{GD} toward $W^\star_{\mathrm{GD}}$ follows from the common implicit-bias argument at the beginning of this section. □

Proposition E.2 (Directional convergence of GD under CSM)**.** *Assume the sample $\mathcal{D}_n$ is generated from the CSM and that Assumption A1 holds. Let $0 < \delta < e^{-1}$. There exists a constant $C_{\mathrm{dim}} > 0$, depending only on ρ, such that, if*

$$n \ge C_{\mathrm{dim}} \log(4/\delta), \qquad D \ge C_{\mathrm{dim}} n \log(8n/\delta),$$

then, with probability at least $1 - \delta$ over the training sample, GD converges in direction toward the unique Euclidean max-margin classifier

$$W^\star_{\mathrm{GD}} = a^\star_{\mathrm{GD}} S + H^\star_{\mathrm{GD}},$$

where $\mathrm{rank}(H^\star_{\mathrm{GD}}) = 1$.

Moreover, there exists a universal constant $C > 0$ such that

$$\left| a^\star_{\mathrm{GD}} - \frac{(1-2\rho)n}{\lambda^2 + n\|S\|_F^2} \right| \le C \frac{n}{\lambda^2 + n\|S\|_F^2} \left(\sqrt{\frac{\log(4/\delta)}{n}} + \sqrt{\frac{n\log(8n/\delta)}{D}} \right).$$

There also exists a constant $C_H > 0$, depending only on ρ, such that

$$\left| \frac{\lambda}{\sqrt{n}} \|H^\star_{\mathrm{GD}}\|_F - \sqrt{4\rho(1-\rho) + (1-2\rho)^2 \frac{\lambda^4}{(\lambda^2 + n\|S\|_F^2)^2}} \right| \le C_H \left(\sqrt{\frac{\log(4/\delta)}{n}} + \sqrt{\frac{n\log(8n/\delta)}{D}} \right).$$

Proof. Set $m := 1 - 2\rho$ and $\varepsilon_G := C\sqrt{n\log(8n/\delta)/D}$.

Concentration. Under the CSM, after absorbing $\widetilde{y}_i$ into $\boldsymbol{u}_i$, $G = U^\top U$. Applying Lemma H.6 with failure probability $\delta/2$ gives

$$\frac{1}{2} I_n \preceq G \preceq \frac{3}{2} I_n, \qquad \max_{\boldsymbol{x} \in \{\mathbf{1}, \boldsymbol{\xi}\}} \|(G^{-1} - I_n)\boldsymbol{x}\|_\infty \le \varepsilon_G. \tag{E.10}$$

Here and below, universal constants may change from line to line. By Hoeffding's inequality, with failure probability at most $\delta/2$, $|\mathbf{1}^\top \boldsymbol{\xi} - mn| \le \sqrt{2n\log(4/\delta)}$. We work on the intersection of these events, which has probability at least $1 - \delta$.

Approximation of $a^\star_{\mathrm{GD}}$. Set $b := \boldsymbol{\xi}^\top G^{-1}\mathbf{1}$ and $c := \boldsymbol{\xi}^\top G^{-1}\boldsymbol{\xi}$. By (E.10), $|b - \boldsymbol{\xi}^\top \mathbf{1}| \le n\varepsilon_G$ and $|c - n| \le n\varepsilon_G$. Hence $|b - mn| \le \sqrt{2n\log(4/\delta)} + n\varepsilon_G$ and $|c - n| \le n\varepsilon_G$. Moreover, $2n/3 \le c \le 2n$, so $\lambda^2 + s_F^2 c \asymp \lambda^2 + n s_F^2$.

By Lemma E.1, $a^\star_{\mathrm{GD}} = b/(\lambda^2 + s_F^2 c)$, and

$$a^\star_{\mathrm{GD}} - \frac{mn}{\lambda^2 + n s_F^2} = \frac{b - mn}{\lambda^2 + s_F^2 c} + \frac{mn s_F^2 (n - c)}{(\lambda^2 + s_F^2 c)(\lambda^2 + n s_F^2)}.$$

Therefore,

$$\left| a^\star_{\mathrm{GD}} - \frac{mn}{\lambda^2 + n s_F^2} \right| \le C \frac{n}{\lambda^2 + n s_F^2} \left(\sqrt{\frac{\log(4/\delta)}{n}} + \sqrt{\frac{n\log(8n/\delta)}{D}} \right). \tag{E.11}$$

Positivity. Set $\varepsilon := \sqrt{\log(4/\delta)/n} + \sqrt{n\log(8n/\delta)/D}$. By choosing C_{dim} sufficiently large depending only on ρ, the normalized right-hand side of (E.11) is at most $\min\{m/2, \rho\}$. Consequently, $a^\star_{\mathrm{GD}} \ge \frac{m}{2}\, n/(\lambda^2 + n s_F^2) > 0$ and $0 \le a^\star_{\mathrm{GD}} s_F^2 \le m + \rho = 1 - \rho$.

Set $\boldsymbol{z} := \mathbf{1} - a^\star_{\mathrm{GD}} s_F^2 \boldsymbol{\xi}$. Then $\rho \le z_i \le 2$ for every i, while

$$\|(G^{-1} - I_n)\boldsymbol{z}\|_\infty \le \|(G^{-1} - I_n)\mathbf{1}\|_\infty + a^\star_{\mathrm{GD}} s_F^2 \|(G^{-1} - I_n)\boldsymbol{\xi}\|_\infty \le 2\varepsilon_G.$$

Increasing C_{dim} if necessary so that $2\varepsilon_G \le \rho/2$, we obtain $(G^{-1}\boldsymbol{z})_i \ge \rho/2$. Hence Lemma E.1 gives $\alpha_i^\star \ge \rho/(2\lambda^2) > 0$ for every i.

Since $N_i = \boldsymbol{u}_i \boldsymbol{v}_0^\top$ under CSM, the same lemma gives $W^\star_{\mathrm{GD}} = a^\star_{\mathrm{GD}} S + H^\star_{\mathrm{GD}}$, where $H^\star_{\mathrm{GD}} := \lambda U \boldsymbol{\alpha}^\star \boldsymbol{v}_0^\top$. Since $G = U^\top U \succ 0$ and $\boldsymbol{\alpha}^\star \ne 0$, $\mathrm{rank}(H^\star_{\mathrm{GD}}) = 1$.

Shortcut norm. By Lemma E.1, $\boldsymbol{\alpha}^\star = \lambda^{-2} G^{-1}\boldsymbol{z}$, and hence

$$\frac{\lambda^2}{n} \|H^\star_{\mathrm{GD}}\|_F^2 = \frac{\lambda^4}{n} (\boldsymbol{\alpha}^\star)^\top G \boldsymbol{\alpha}^\star = \frac{1}{n} \boldsymbol{z}^\top G^{-1} \boldsymbol{z}. \tag{E.12}$$

Since $\|\boldsymbol{z}\|_1 \le 2n$ and $\|(G^{-1} - I_n)\boldsymbol{z}\|_\infty \le 2\varepsilon_G$,

$$\left| \frac{1}{n} \boldsymbol{z}^\top G^{-1} \boldsymbol{z} - \frac{1}{n} \|\boldsymbol{z}\|_2^2 \right| \le C\varepsilon_G.$$

Set $\overline{\xi} := n^{-1}\mathbf{1}^\top \boldsymbol{\xi}$, $\theta := a^\star_{\mathrm{GD}} s_F^2$, and $\theta_0 := mns_F^2/(\lambda^2 + ns_F^2)$. Then $n^{-1}\|\boldsymbol{z}\|_2^2 = 1 - 2\theta\overline{\xi} + \theta^2$, while (E.11) and Hoeffding's bound give $|\theta - \theta_0| \le C\varepsilon$ and $|\overline{\xi} - m| \le C\sqrt{\log(4/\delta)/n}$. Since θ and θ_0 are uniformly bounded,

$$\left|\frac{1}{n}\|\boldsymbol{z}\|_2^2 - (1 - 2m\theta_0 + \theta_0^2)\right| \le C\varepsilon.$$

Moreover,

$$1 - 2m\theta_0 + \theta_0^2 = 4\rho(1-\rho) + m^2\frac{\lambda^4}{(\lambda^2 + ns_F^2)^2}.$$

Combining these estimates with (E.12) yields

$$\left|\frac{\lambda^2}{n}\|H^\star_{\mathrm{GD}}\|_F^2 - \left[4\rho(1-\rho) + m^2\frac{\lambda^4}{(\lambda^2 + ns_F^2)^2}\right]\right| \le C\varepsilon.$$

The quantity in brackets is at least $4\rho(1-\rho) > 0$; hence $|\sqrt{x} - \sqrt{y}| = |x-y|/(\sqrt{x} + \sqrt{y})$ gives

$$\left|\frac{\lambda}{\sqrt{n}}\|H^\star_{\mathrm{GD}}\|_F - \sqrt{4\rho(1-\rho) + (1-2\rho)^2\frac{\lambda^4}{(\lambda^2 + n\|S\|_F^2)^2}}\right| \le C_H\varepsilon,$$

where $C_H > 0$ depends only on ρ. □

Corollary E.1. *Assume the sample $\mathcal{D}_n$ is generated from the DSM and that Assumption A1 holds. Let $0 < \delta < e^{-1}$. There exist constants $C_{\dim}, c, C > 0$, depending only on ρ, such that, if*

$$n \ge C_{\dim}\log(4/\delta), \qquad D \ge C_{\dim} n\log(8n^2/\delta),$$

then, with probability at least $1 - \delta$ over the training sample,

$$\mathcal{R}(W^\star_{\mathrm{GD}}) \le 4\exp\left[-c\min\left\{\frac{D^2 n\|S\|_F^4}{(\lambda^2 + n\|S\|_F^2)^2}, \frac{Dn\|S\|_F^2}{\lambda^2 + n\|S\|_F^2}\right\}\right], \tag{E.13}$$

$$\mathcal{R}(W^\star_{\mathrm{GD}}) \ge \frac{1}{2} - Ce^{-cn} - C\frac{D\sqrt{n}\,\|S\|_F^2}{\lambda^2 + n\|S\|_F^2}. \tag{E.14}$$

Consequently,

$$\lambda^2 \ll D\sqrt{n}\,\|S\|_F^2 \quad\Longrightarrow\quad \mathcal{R}(W^\star_{\mathrm{GD}}) \to 0,$$

whereas

$$\lambda^2 \gg D\sqrt{n}\,\|S\|_F^2 \quad\Longrightarrow\quad \mathcal{R}(W^\star_{\mathrm{GD}}) \to \frac{1}{2}.$$

Proof. Increase $C_{\dim}$, if necessary, so that the conclusions of Proposition E.1 hold and its approximation of $a^\star_{\mathrm{GD}}$ implies

$$a^\star_{\mathrm{GD}}\|S\|_F^2 \asymp \frac{n\|S\|_F^2}{\lambda^2 + n\|S\|_F^2}. \tag{E.15}$$

We work on the corresponding event, which has probability at least $1 - \delta$.

Set $B^\star_{\mathrm{GD}} := \lambda H^\star_{\mathrm{GD}}$. Proposition E.1 gives $\mathrm{rank}(B^\star_{\mathrm{GD}}) = n$ and $c \le \sigma_n(B^\star_{\mathrm{GD}}) \le \sigma_1(B^\star_{\mathrm{GD}}) \le C$, and therefore

$$\|B^\star_{\mathrm{GD}}\|_F \asymp \sqrt{n}, \qquad \|B^\star_{\mathrm{GD}}\|_{\mathrm{op}} \asymp 1. \tag{E.16}$$

For an independent clean test sample, $yX = S + \lambda\boldsymbol{u}\boldsymbol{v}^\top$, where $\boldsymbol{u}, \boldsymbol{v}$ are independent and uniformly distributed on $\mathbb{S}^{D-1}$. Hence

$$y\langle W^\star_{\mathrm{GD}}, X\rangle_F = a^\star_{\mathrm{GD}}\|S\|_F^2 + \boldsymbol{u}^\top B^\star_{\mathrm{GD}}\boldsymbol{v}.$$

Upper bound. Set $t := a^\star_{\mathrm{GD}}\|S\|_F^2$. By Lemma H.7,

$$\mathcal{R}(W^\star_{\mathrm{GD}}) \le 4\exp\left[-c\min\left\{\frac{D^2t^2}{\|B^\star_{\mathrm{GD}}\|_F^2}, \frac{Dt}{\|B^\star_{\mathrm{GD}}\|_{\mathrm{op}}}\right\}\right].$$

Using (E.15) and (E.16) gives (E.13).

Lower bound. Let $B^\star_{\mathrm{GD}} = U_B \Sigma_B V_B^\top$ be its compact SVD. By Proposition E.1, $\mathrm{rank}(B^\star_{\mathrm{GD}}) = n$ and $\sigma_n(B^\star_{\mathrm{GD}}) \geq c_H$, where $c_H > 0$ depends only on ρ.

Let $P := U_B U_B^\top$ be the orthogonal projection onto the left singular space of $B^\star_{\mathrm{GD}}$. Then

$$\begin{aligned} \|(B^\star_{\mathrm{GD}})^\top \boldsymbol{u}\|_2 &= \|V_B \Sigma_B U_B^\top \boldsymbol{u}\|_2 = \|\Sigma_B U_B^\top \boldsymbol{u}\|_2 \\ &\geq \sigma_n(B^\star_{\mathrm{GD}}) \|U_B^\top \boldsymbol{u}\|_2 = \sigma_n(B^\star_{\mathrm{GD}}) \|P\boldsymbol{u}\|_2 \geq c_H \|P\boldsymbol{u}\|_2. \end{aligned} \tag{E.17}$$

Here we used $V_B^\top V_B = I_n$.

Since P has rank n, Lemma H.3 gives

$$\mathbb{P}\left(\|P\boldsymbol{u}\|_2 \leq \frac{1}{2}\sqrt{\frac{n}{D}} \,\middle|\, \mathcal{D}_n \right) \leq 2e^{-n/16}. \tag{E.18}$$

Together with (E.17), this implies

$$\|(B^\star_{\mathrm{GD}})^\top \boldsymbol{u}\|_2 \geq \frac{c_H}{2}\sqrt{\frac{n}{D}} \tag{E.19}$$

except on an event of probability at most $2e^{-n/16}$.

Now condition on such a realization of $\boldsymbol{u}$. Since $\boldsymbol{v} \sim \mathrm{Unif}(\mathbb{S}^{D-1})$ is independent of $\boldsymbol{u}$, Lemma H.2 gives, for every $t > 0$,

$$\mathbb{P}\left(\left| \boldsymbol{u}^\top B^\star_{\mathrm{GD}} \boldsymbol{v} \right| \leq t \,\middle|\, \boldsymbol{u}, \mathcal{D}_n \right) \leq C\sqrt{D}\, \frac{t}{\|(B^\star_{\mathrm{GD}})^\top \boldsymbol{u}\|_2} \leq C_\rho \frac{Dt}{\sqrt{n}},$$

where the last inequality follows from (E.19). Integrating over $\boldsymbol{u}$ gives

$$\mathbb{P}\left(\left| \boldsymbol{u}^\top B^\star_{\mathrm{GD}} \boldsymbol{v} \right| \leq t \,\middle|\, \mathcal{D}_n \right) \leq Ce^{-cn} + C_\rho \frac{Dt}{\sqrt{n}}.$$

Finally, $Z := \boldsymbol{u}^\top B^\star_{\mathrm{GD}} \boldsymbol{v}$ has a continuous distribution symmetric about zero. Since

$$\mathcal{R}(W^\star_{\mathrm{GD}}) = \mathbb{P}(Z \leq -t \mid \mathcal{D}_n) = \frac{1}{2} - \frac{1}{2}\mathbb{P}(|Z| < t \mid \mathcal{D}_n),$$

the preceding bound together with (E.15) yields

$$\mathcal{R}(W^\star_{\mathrm{GD}}) \geq \frac{1}{2} - Ce^{-cn} - C\frac{D\sqrt{n}\,\|S\|_F^2}{\lambda^2 + n\|S\|_F^2},$$

which proves (E.14). □

Corollary E.2. *Assume the sample $\mathcal{D}_n$ is generated from the CSM and that Assumption A1 holds. Let $0 < \delta < e^{-1}$. There exist constants $C_{\mathrm{dim}}, c, C > 0$, depending only on ρ, such that, if*

$$n \geq C_{\mathrm{dim}} \log(4/\delta), \qquad D \geq C_{\mathrm{dim}} n \log(8n/\delta),$$

then, with probability at least $1 - \delta$ over the training sample,

$$\mathcal{R}(W^\star_{\mathrm{GD}}) \leq 2\exp\left[-c\frac{Dn\|S\|_F^4}{(\lambda^2 + n\|S\|_F^2)^2} \right], \tag{E.20}$$

$$\mathcal{R}(W^\star_{\mathrm{GD}}) \geq \frac{1}{2} - C\frac{\sqrt{nD}\,\|S\|_F^2}{\lambda^2 + n\|S\|_F^2}. \tag{E.21}$$

Consequently,

$$\lambda^2 \ll \|S\|_F^2 \sqrt{nD} \quad \Longrightarrow \quad \mathcal{R}(W^\star_{\mathrm{GD}}) \to 0,$$

whereas

$$\lambda^2 \gg \|S\|_F^2 \sqrt{nD} \quad \Longrightarrow \quad \mathcal{R}(W^\star_{\mathrm{GD}}) \to \frac{1}{2}.$$

Proof. Increase $C_{\dim}$, if necessary, so that the conclusions of Proposition E.2 hold and its estimates imply

$$a^\star_{\mathrm{GD}}\|S\|_F^2 \asymp_\rho \frac{n\|S\|_F^2}{\lambda^2 + n\|S\|_F^2}, \qquad \lambda\|H^\star_{\mathrm{GD}}\|_F \asymp_\rho \sqrt{n}. \tag{E.22}$$

We work on the corresponding event, which has probability at least $1-\delta$.

Since $\mathrm{rank}(H^\star_{\mathrm{GD}}) = 1$, write $H^\star_{\mathrm{GD}} = \boldsymbol{h}^\star_{\mathrm{GD}}\boldsymbol{v}_0^\top$ with $\|\boldsymbol{v}_0\|_2 = 1$, so that $\|H^\star_{\mathrm{GD}}\|_F = \|\boldsymbol{h}^\star_{\mathrm{GD}}\|_2$. For an independent clean test sample, by spherical symmetry we may write $yX = S + \lambda\boldsymbol{u}\boldsymbol{v}_0^\top$, where $\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}^{D-1})$ is independent of the training sample. Hence

$$y\langle W^\star_{\mathrm{GD}}, X\rangle_F = a^\star_{\mathrm{GD}}\|S\|_F^2 + \lambda\langle \boldsymbol{h}^\star_{\mathrm{GD}}, \boldsymbol{u}\rangle.$$

Therefore, conditionally on $\mathcal{D}_n$,

$$\mathcal{R}(W^\star_{\mathrm{GD}}) = \mathbb{P}(Z \le -t \mid \mathcal{D}_n), \qquad Z := \left\langle \frac{\boldsymbol{h}^\star_{\mathrm{GD}}}{\|\boldsymbol{h}^\star_{\mathrm{GD}}\|_2}, \boldsymbol{u}\right\rangle, \qquad t := \frac{a^\star_{\mathrm{GD}}\|S\|_F^2}{\lambda\|\boldsymbol{h}^\star_{\mathrm{GD}}\|_2}.$$

By (E.22),

$$t \asymp_\rho \frac{\sqrt{n}\,\|S\|_F^2}{\lambda^2 + n\|S\|_F^2}. \tag{E.23}$$

By symmetry and Lemma H.2,

$$\frac{1}{2} - C\sqrt{D}\,t \le \mathcal{R}(W^\star_{\mathrm{GD}}) \le 2e^{-cDt^2}.$$

Together with (E.23), this gives (E.20) and (E.21).

Finally, if $\lambda^2 \ll \|S\|_F^2\sqrt{nD}$, then, since $D \gtrsim n\log n$, $\sqrt{nD}\,\|S\|_F^2/(\lambda^2 + n\|S\|_F^2) \to \infty$. Thus (E.20) gives $\mathcal{R}(W^\star_{\mathrm{GD}}) \to 0$.

Conversely, if $\lambda^2 \gg \|S\|_F^2\sqrt{nD}$, then $\sqrt{nD}\,\|S\|_F^2/(\lambda^2 + n\|S\|_F^2) \to 0$. Hence (E.21), together with $\mathcal{R}(W^\star_{\mathrm{GD}}) \le 1/2$ by symmetry and the positive signal coefficient, gives $\mathcal{R}(W^\star_{\mathrm{GD}}) \to \frac{1}{2}$. □

F Proof of Theorem 3

Lemma F.1 (One-step spherical polar alignment)**.** *Let $U = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_n]$ and $\widehat{V} = [\widehat{\boldsymbol{v}}_1, \dots, \widehat{\boldsymbol{v}}_n]$, where all columns are independent and uniformly distributed on $\mathbb{S}^{D-1}$. Set $N := U\widehat{V}^\top = \sum_{i=1}^n \boldsymbol{u}_i\widehat{\boldsymbol{v}}_i^\top$ and $H := \mathrm{Polar}(N)$. There exist universal constants $c, C > 0$ such that, if*

$$D \ge c\max\{n, \log(1/\delta)\},$$

then, with probability at least $1-\delta$,

$$\max_{1\le i\le n}\left|\left\langle H, \boldsymbol{u}_i\widehat{\boldsymbol{v}}_i^\top\right\rangle_F - 1\right| \le C\sqrt{\frac{\max\{n, \log(1/\delta)\}}{D}}.$$

Proof. Let $c_G, C_G > 0$ be the universal constants in Lemma H.5, and set $G_U := U^\top U$ and $G_V := \widehat{V}^\top\widehat{V}$. Since $\log n \le n$, we have $n + \log(n/\delta) \le 3\max\{n, \log(1/\delta)\}$. Thus, by taking $c > 0$ sufficiently large, Lemma H.5 implies that, with probability at least $1-\delta$,

$$\|G_U - I\|_{\mathrm{op}} \vee \|G_V - I\|_{\mathrm{op}} \le C_G\sqrt{\frac{n + \log(n/\delta)}{D}} \le C_0\sqrt{\frac{\max\{n, \log(1/\delta)\}}{D}} =: \eta,$$

where $C_0 > 0$ is a universal constant. Increasing c, we have $\eta \le 1/4$. We work on this event.

Set $\overline{U} := UG_U^{-1/2}$, $\overline{V} := \widehat{V}G_V^{-1/2}$, and $Y := G_U^{1/2}G_V^{1/2}$. Then $\overline{U}^\top\overline{U} = \overline{V}^\top\overline{V} = I_n$ and $N = \overline{U}Y\overline{V}^\top$, so Lemma H.12 gives $H = \overline{U}\,\mathrm{Polar}(Y)\overline{V}^\top$.

For $N_i := \boldsymbol{u}_i\widehat{\boldsymbol{v}}_i^\top$,

$$\langle H, N_i\rangle_F = e_i^\top G_U^{1/2}\,\mathrm{Polar}(Y)G_V^{1/2}e_i,$$

where we used $U^\top \overline{U} = G_U^{1/2}$ and $\overline{V}^\top \widehat{V} = G_V^{1/2}$.

Since $\eta \le 1/4$, $\|G_U - I\|_{\mathrm{op}} \vee \|G_V - I\|_{\mathrm{op}} \le \eta < 1$. The square-root map is uniformly Lipschitz on a fixed compact subset of the positive definite cone, so, for a universal constant $C_1 > 0$,

$$\|G_U^{1/2} - I\|_{\mathrm{op}} \vee \|G_V^{1/2} - I\|_{\mathrm{op}} \le C_1 \eta.$$

Since $Y - I = (G_U^{1/2} - I)G_V^{1/2} + (G_V^{1/2} - I)$, it follows that $\|Y - I\|_{\mathrm{op}} \le C_2\eta$ for a universal constant $C_2 > 0$.

Moreover, $Y^\top Y - I = G_V^{1/2}(G_U - I)G_V^{1/2} + (G_V - I)$, and hence

$$\|Y^\top Y - I\|_{\mathrm{op}} \le (2 + \eta)\eta \le \frac{9}{16}.$$

Thus $Y^\top Y$ lies in a fixed compact subset of the positive definite cone. The inverse square-root map is therefore uniformly Lipschitz there, so, for a universal constant $C_3 > 0$,

$$\|(Y^\top Y)^{-1/2} - I\|_{\mathrm{op}} \le C_3 \eta.$$

Since $\mathrm{Polar}(Y) - I = (Y - I)(Y^\top Y)^{-1/2} + \{(Y^\top Y)^{-1/2} - I\}$, and $\|(Y^\top Y)^{-1/2}\|_{\mathrm{op}}$ is uniformly bounded, the triangle inequality gives

$$\|\,\mathrm{Polar}(Y) - I\|_{\mathrm{op}} \le C_4 \eta$$

for a universal constant $C_4 > 0$. Finally, $G_U^{1/2}\,\mathrm{Polar}(Y)G_V^{1/2} - I = (G_U^{1/2} - I)\,\mathrm{Polar}(Y)G_V^{1/2} + (\mathrm{Polar}(Y) - I)G_V^{1/2} + (G_V^{1/2} - I)$, so another application of the triangle inequality yields

$$\left\| G_U^{1/2}\,\mathrm{Polar}(Y)G_V^{1/2} - I \right\|_{\mathrm{op}} \le C_5 \eta.$$

Thus

$$\max_{i \in [n]} |\langle H, N_i \rangle_F - 1| \le C_0 C_5 \sqrt{\frac{\max\{n, \log(1/\delta)\}}{D}}.$$

Renaming $C_0 C_5$ as C proves the claim. □

We are now ready to show Theorem 3. Actually, we show slightly sharper result:

Theorem F.1 (One-step interpolation and generalization under dispersed shortcuts)**.** *Assume the sample $\mathcal{D}_n$ is generated from the DSM. There exist universal constants $c, C > 0$ such that the following holds. Let $\delta \in (0,1)$ and suppose that $n \ge \frac{2}{(1-2\rho)^2} \log \frac{2}{\delta}$. Then, with probability at least $1 - \delta$ over the training sample $\mathcal{D}_n$, the following statements hold.*

1. ***Generalization.*** *For an independent test sample (X, y),*

$$\mathbb{P}\left(y \langle W_1^{\mathrm{SpecGD}}, X \rangle_F \le 0 \,\middle|\, \mathcal{D}_n \right) \le 2 \exp\left(-c \min\left\{ \frac{D^2 s^2}{n\lambda^2}, \frac{Ds}{\lambda} \right\} \right).$$

2. ***Interpolation.*** *If, in addition, $\lambda > s$ and $D \ge C\left(1 - \frac{s}{\lambda}\right)^{-2} \max\left\{n, \log \frac{2}{\delta}\right\}$, then*

$$\widehat{\mathcal{R}}_n\left(W_1^{\mathrm{SpecGD}} \right) = 0.$$

Proof. Let $\mathcal{E}_{\mathrm{sign}} := \{n_+ > n_-\}$. Since $n_- = \sum_{i=1}^n \mathbf{1}_{\{\xi_i = -1\}} \sim \mathrm{Bin}(n, \rho)$, Hoeffding's inequality gives

$$\mathbb{P}(\mathcal{E}_{\mathrm{sign}}^c) \le \exp\left\{ -\frac{1}{2} n (1 - 2\rho)^2 \right\} \le \frac{\delta}{2},$$

where the last inequality follows from the lower bound on n. We work on $\mathcal{E}_{\mathrm{sign}}$.

Set $s := \|S\|_1$ and $N := \sum_{i=1}^n N_i$. Since $W_0^{\mathrm{SpecGD}} = 0$,

$$-\nabla \mathcal{L}(W_0^{\mathrm{SpecGD}}) = \frac{1}{n} \sum_{i=1}^n A_i = \frac{1}{n} \left\{ (n_+ - n_-) S + \lambda N \right\}.$$

Because $n_+ > n_-$ and the signal and shortcut blocks are orthogonal,

$$\mathrm{Polar}\left(\sum_{i=1}^{n} A_i\right) = P_S + \mathrm{Polar}(N).$$

Writing $H := \mathrm{Polar}(N)$, we therefore have $W_1^{\mathrm{SpecGD}} = \eta_0(P_S + H)$.

We first prove the conditional generalization bound. Conditional on the training data, H is deterministic, with $\mathrm{rank}(H) = \mathrm{rank}(N) \le n$, $\|H\|_{\mathrm{op}} = 1$, and $\|H\|_F^2 = \mathrm{rank}(H) \le n$ almost surely. For a fresh test point $X = yS + \lambda \boldsymbol{u}\boldsymbol{v}^\top$, signal–noise block orthogonality gives

$$y\langle W_1^{\mathrm{SpecGD}}, X\rangle_F = \eta_0 \left[s + \lambda y \langle H, \boldsymbol{u}\boldsymbol{v}^\top\rangle_F\right].$$

Set $T := \langle H, \boldsymbol{u}\boldsymbol{v}^\top\rangle_F$. Conditional on the training data, T is symmetric, and multiplication by the independent sign y preserves its distribution. Hence

$$\mathbb{P}\left(y\langle W_1^{\mathrm{SpecGD}}, X\rangle_F \le 0 \,\middle|\, \mathcal{D}_n\right) = \frac{1}{2}\mathbb{P}\left(|T| \ge \frac{s}{\lambda} \,\middle|\, \mathcal{D}_n\right).$$

Applying Lemma H.7 and using $\|H\|_F^2 \le n$ and $\|H\|_{\mathrm{op}} = 1$ gives

$$\mathbb{P}\left(y\langle W_1^{\mathrm{SpecGD}}, X\rangle_F \le 0 \,\middle|\, \mathcal{D}_n\right) \le 2\exp\left[-c\min\left\{\frac{D^2 s^2}{n\lambda^2}, \frac{Ds}{\lambda}\right\}\right].$$

It remains to prove interpolation. Assume $\lambda > s$. Applying Lemma F.1 with failure probability $\delta/2$, the dimensional assumption and a sufficiently large universal constant C imply that, with probability at least $1 - \delta/2$,

$$\max_{i\in[n]} |\langle H, N_i\rangle_F - 1| \le C_0\sqrt{\frac{\max\{n, \log(2/\delta)\}}{D}} \le \frac{1}{2}\left(1 - \frac{s}{\lambda}\right),$$

where $C_0 > 0$ is a universal constant. Therefore

$$\min_{i\in[n]} \langle H, N_i\rangle_F \ge \frac{1}{2}\left(1 + \frac{s}{\lambda}\right) > \frac{s}{\lambda}.$$

On the intersection of this event with $\mathcal{E}_{\mathrm{sign}}$, for every $i \in [n]$,

$$\eta_0^{-1}\langle W_1^{\mathrm{SpecGD}}, \widetilde{y}_i X_i\rangle_F = \langle P_S + H, \xi_i S + \lambda N_i\rangle_F = \xi_i s + \lambda\langle H, N_i\rangle_F.$$

If $i \in I_+$, this is positive, while if $i \in I_-$, it is larger than $-s + \lambda(s/\lambda) = 0$. Thus

$$\min_{i\in[n]} \langle W_1^{\mathrm{SpecGD}}, \widetilde{y}_i X_i\rangle_F > 0,$$

and hence $\widehat{R}_n(W_1^{\mathrm{SpecGD}}) = 0$.

Finally, $\mathcal{E}_{\mathrm{sign}}$ fails with probability at most $\delta/2$ and the alignment event with probability at most $\delta/2$. Thus their intersection has probability at least $1 - \delta$, and both conclusions hold there. □

G Proof of Theorem 4

In this section, we prove Theorem 4. Actually, We show the following slightly sharper result.

Theorem G.1 (One-step interpolation and generalization under collapsed shortcuts)**.** *Assume the sample $\mathcal{D}_n$ is generated from the CSM. There exist universal constants $c, C > 0$ such that the following holds. Let $\delta \in (0, 1)$ and suppose that*

$$n \ge \frac{2}{(1-2\rho)^2}\log\frac{2}{\delta}.$$

Then, with probability at least $1 - \delta$ over the training sample $\mathcal{D}_n$, the following statements hold.

1. ***Generalization.*** *For an independent clean test sample* (X, y)*,*

$$\mathbb{P}\left(y\langle W_1^{\mathrm{SpecGD}}, X\rangle_F \le 0 \,\middle|\, \mathcal{D}_n\right) \le 2\exp\left(-cD\min\left\{\frac{s^2}{\lambda^2}, 1\right\}\right).$$

2. ***Interpolation.*** *If, in addition,* $\lambda > s\sqrt{n}$ *and* $D \ge C\left(1 - \frac{s\sqrt{n}}{\lambda}\right)^{-2} n\log\frac{4n}{\delta}$*, then*

$$\widehat{\mathcal{R}}_n\left(W_1^{\mathrm{SpecGD}}\right) = 0.$$

Proof. Let $\mathcal{E}_{\mathrm{sign}} := \{n_+ > n_-\}$, where $n_\pm := |\{i : \xi_i = \pm 1\}|$. By Hoeffding's inequality and the assumption on n,

$$\mathbb{P}(\mathcal{E}_{\mathrm{sign}}^c) \le \exp\left\{-\frac{1}{2}n(1-2\rho)^2\right\} \le \frac{\delta}{2}.$$

We work on $\mathcal{E}_{\mathrm{sign}}$. By signal–shortcut block orthogonality,

$$W_1^{\mathrm{SpecGD}} = \eta_0\left(P_S + \boldsymbol{h}\boldsymbol{v}_0^\top\right), \qquad \boldsymbol{h} := \frac{\sum_{i=1}^n \tilde{y}_i\boldsymbol{u}_i}{\|\sum_{i=1}^n \tilde{y}_i\boldsymbol{u}_i\|_2}.$$

We first prove the conditional generalization bound. Let (X, y) be an independent clean test sample, so that $yX = S + \lambda y\boldsymbol{u}\boldsymbol{v}_0^\top$. Then $\eta_0^{-1}y\langle W_1^{\mathrm{SpecGD}}, X\rangle_F = s + \lambda\langle \boldsymbol{h}, y\boldsymbol{u}\rangle$. By spherical symmetry, $y\boldsymbol{u}$ is uniformly distributed on $\mathbb{S}^{D-1}$. Conditional on $\mathcal{D}_n$, $\boldsymbol{h}$ is a fixed unit vector, so Lemma H.2 gives

$$\mathbb{P}\left(y\langle W_1^{\mathrm{SpecGD}}, X\rangle_F \le 0 \,\middle|\, \mathcal{D}_n\right) \le 2\exp\left(-cD\min\left\{\frac{s^2}{\lambda^2}, 1\right\}\right).$$

It remains to prove interpolation. Assume $\lambda > s\sqrt{n}$ and set $\kappa := 1 - s\sqrt{n}/\lambda \in (0, 1)$. We show that, with probability at least $1 - \delta/2$, $\min_{i\in[n]}\langle \boldsymbol{h}, \tilde{y}_i\boldsymbol{u}_i\rangle > s/\lambda$. For each $i \in [n]$, set $r_i := \sum_{j\neq i}\langle \tilde{y}_i\boldsymbol{u}_i, \tilde{y}_j\boldsymbol{u}_j\rangle$. Conditionally on $\tilde{y}_i\boldsymbol{u}_i$, the summands are independent, centered, and sub-Gaussian, so, for a universal constant $c > 0$ and $t \in (0, 1)$, $\mathbb{P}(|r_i| > t \mid \tilde{y}_i\boldsymbol{u}_i) \le 2\exp(-cDt^2/n)$. A union bound gives

$$\mathbb{P}\left(\max_{i\in[n]}|r_i| > t\right) \le 2n\exp\left(-c\frac{Dt^2}{n}\right).$$

On the event $\max_i |r_i| \le t$, since $\|\sum_j \tilde{y}_j\boldsymbol{u}_j\|_2^2 = n + \sum_j r_j \le n(1+t)$, for $t \le 1/2$,

$$\langle \boldsymbol{h}, \tilde{y}_i\boldsymbol{u}_i\rangle = \frac{1 + r_i}{\|\sum_j \tilde{y}_j\boldsymbol{u}_j\|_2} \ge \frac{1-t}{\sqrt{n(1+t)}} \ge \frac{1-2t}{\sqrt{n}}.$$

Taking $t = \kappa/4$, the dimensional assumption implies, for C sufficiently large, that with probability at least $1 - \delta/2$,

$$\min_{i\in[n]}\langle \boldsymbol{h}, \tilde{y}_i\boldsymbol{u}_i\rangle \ge \frac{1 - \kappa/2}{\sqrt{n}} > \frac{1-\kappa}{\sqrt{n}} = \frac{s}{\lambda}.$$

On the intersection of this event with $\mathcal{E}_{\mathrm{sign}}$, for every $i \in [n]$,

$$\eta_0^{-1}\left\langle W_1^{\mathrm{SpecGD}}, \widetilde{y}_i X_i\right\rangle_F = \xi_i s + \lambda\langle \boldsymbol{h}, \tilde{y}_i\boldsymbol{u}_i\rangle.$$

If $i \in I_+$, this is positive, while if $i \in I_-$, it is larger than $-s + \lambda(s/\lambda) = 0$. Hence $\widehat{\mathcal{R}}_n(W_1^{\mathrm{SpecGD}}) = 0$.

Finally, $\mathcal{E}_{\mathrm{sign}}$ and the interpolation event fail with probability at most $\delta/2$ each, so their intersection has probability at least $1 - \delta$, and both conclusions hold there. □

H AUXILIARY FACTS

H.1 LINEAR SEPARABILITY

Lemma H.1 (Linear separability of the training sample)**.** *Suppose that $D \geq n$. Then, under the K-dimensional shortcut model for any $1 \leq K \leq D$, the training sample $\{(X_i, \widetilde{y}_i)\}_{i=1}^n$ is linearly separable with respect to the observed labels almost surely. More precisely, almost surely there exists $W_{\mathrm{sep}} \in \mathbb{R}^{d\times d}$ such that*

$$\widetilde{y}_i \langle W_{\mathrm{sep}}, X_i\rangle_F = 1, \qquad i \in [n].$$

Consequently, both the spectral and Frobenius max-margin values are strictly positive.

Proof. Recall that $X_i = y_i S + \lambda \boldsymbol{u}_i \boldsymbol{v}_i^\top$, where $\boldsymbol{u}_i$ are drawn independently and uniformly from the unit sphere of the D-dimensional shortcut subspace $\mathcal{U}$. Since $D \geq n$, the vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ are linearly independent almost surely.

On this event, let $\boldsymbol{u}_1^\sharp, \ldots, \boldsymbol{u}_n^\sharp \in \mathrm{span}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ denote the dual family satisfying $\langle \boldsymbol{u}_i^\sharp, \boldsymbol{u}_j\rangle = \delta_{ij}$, and define

$$W_{\mathrm{sep}} := \frac{1}{\lambda} \sum_{i=1}^n \widetilde{y}_i \boldsymbol{u}_i^\sharp \boldsymbol{v}_i^\top.$$

Since $\boldsymbol{u}_i^\sharp$ and $\boldsymbol{v}_i$ belong to the left and right shortcut subspaces, respectively, which are orthogonal to the corresponding signal subspaces of S, we have $\langle W_{\mathrm{sep}}, S\rangle_F = 0$. Moreover, for every $j \in [n]$,

$$\langle W_{\mathrm{sep}}, Z_j\rangle_F = \sum_{i=1}^n \widetilde{y}_i \langle \boldsymbol{u}_i^\sharp, \boldsymbol{u}_j\rangle \langle \boldsymbol{v}_i, \boldsymbol{v}_j\rangle = \widetilde{y}_j \|\boldsymbol{v}_j\|_2^2 = \widetilde{y}_j.$$

Hence $\widetilde{y}_j \langle W_{\mathrm{sep}}, X_j\rangle_F = \widetilde{y}_j \langle W_{\mathrm{sep}}, y_j S + Z_j\rangle_F = \widetilde{y}_j^2 = 1$, proving strict linear separability.

In particular, if $\gamma_{\mathrm{op}} := \max_{\|W\|_{\mathrm{op}}\leq 1} \min_{i\in[n]} \widetilde{y}_i \langle W, X_i\rangle_F$, then normalization of W_{sep} gives $\gamma_{\mathrm{op}} \geq \|W_{\mathrm{sep}}\|_{\mathrm{op}}^{-1} > 0$. The same argument with Frobenius normalization gives a strictly positive Frobenius max-margin value. □

H.2 CONCENTRATION INEQUALITIES

Lemma H.2 (Spherical linear tails and small balls)**.** *Let $\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}^{D-1})$ and let $\boldsymbol{w} \in \mathbb{R}^D \setminus \{0\}$ be fixed. There exist universal constants $c, C, c_0, C_0 > 0$ such that*

$$\mathbb{P}(|\langle \boldsymbol{u}, \boldsymbol{w}\rangle| \geq t\|\boldsymbol{w}\|_2) \leq 2e^{-cDt^2}, \qquad t > 0,$$

and

$$\mathbb{P}(|\langle \boldsymbol{u}, \boldsymbol{w}\rangle| \leq t\|\boldsymbol{w}\|_2) \leq C\sqrt{D}\, t, \qquad t > 0.$$

Moreover, for $0 \leq t \leq 1/2$,

$$\mathbb{P}(\langle \boldsymbol{u}, \boldsymbol{w}\rangle \leq -t\|\boldsymbol{w}\|_2) \geq c_0 e^{-C_0 Dt^2}.$$

Proof. By rotational invariance, $Z := \langle \boldsymbol{u}, \boldsymbol{w}\rangle / \|\boldsymbol{w}\|_2$ has the same distribution as the first coordinate of a uniform vector on $\mathbb{S}^{D-1}$. The upper-tail bound follows directly from the standard spherical concentration inequality (Vershynin, 2026, Theorem 3.4.5).

For the remaining claims, Z has density

$$f_D(z) = b_D(1-z^2)^{(D-3)/2}, \qquad b_D = \frac{\Gamma(D/2)}{\sqrt{\pi}\Gamma((D-1)/2)} \asymp \sqrt{D}, \qquad |z| < 1.$$

Hence $\mathbb{P}(|Z| \leq t) \leq 2t\|f_D\|_\infty \leq C\sqrt{D}\, t$. For $0 \leq t \leq 1/2$, using $\log(1-z^2) \geq -Cz^2$ for $0 \leq z \leq 3/4$ gives $f_D(z) \geq c\sqrt{D}\, e^{-CDz^2}$. Therefore,

$$\mathbb{P}(Z \geq t) \geq \int_t^{t+(4\sqrt{D})^{-1}} f_D(z)\, dz \geq c_0 e^{-C_0 Dt^2}.$$

By symmetry, the same bound holds for $\mathbb{P}(Z \leq -t)$. □

Lemma H.3 (Spherical projection norm)**.** *Let $\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}^{D-1})$, and let P be an orthogonal projection of rank $r \le D$. Then*

$$\mathbb{P}\left(\|P\boldsymbol{u}\|_2 \le \frac{1}{2}\sqrt{\frac{r}{D}}\right) \le 2e^{-r/16}.$$

Proof. Write $\boldsymbol{u} = \boldsymbol{g}/\|\boldsymbol{g}\|_2$ with $\boldsymbol{g} \sim N(0, I_D)$. Then $\|P\boldsymbol{g}\|_2^2 \sim \chi_r^2$ and $\|\boldsymbol{g}\|_2^2 \sim \chi_D^2$. By the Laurent–Massart inequalities,

$$\mathbb{P}\left(\|P\boldsymbol{g}\|_2^2 < \frac{r}{2} \text{ or } \|\boldsymbol{g}\|_2^2 > 2D\right) \le 2e^{-r/16}.$$

On the complementary event, $\|P\boldsymbol{u}\|_2 = \|P\boldsymbol{g}\|_2/\|\boldsymbol{g}\|_2 \ge \frac{1}{2}\sqrt{r/D}$. □

Lemma H.4 (Spherical Gram concentration)**.** *Let $U = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_n]$, where $\boldsymbol{u}_1, \dots, \boldsymbol{u}_n$ are independent and uniformly distributed on $\mathbb{S}^{D-1}$. Then there exist universal constants $c_G, C_G > 0$ such that, if $D \ge c_G\{n + \log(n/\delta)\}$, then, with probability at least $1 - \delta$,*

$$\|U^\top U - I_n\|_{\mathrm{op}} \le C_G\sqrt{\frac{n + \log(n/\delta)}{D}},$$

and

$$\max_{i \ne j} |\boldsymbol{u}_i^\top \boldsymbol{u}_j| \le C_G\sqrt{\frac{\log(2n^2/\delta)}{D}}.$$

Proof. Write $\boldsymbol{u}_i = \boldsymbol{g}_i/\|\boldsymbol{g}_i\|_2$, where $\boldsymbol{g}_i \sim N(0, I_D)$ independently, and let $G = [\boldsymbol{g}_1, \dots, \boldsymbol{g}_n] \in \mathbb{R}^{D\times n}$. The rows of G are independent, mean-zero, isotropic, sub-Gaussian random vectors in $\mathbb{R}^n$. Hence, by the standard singular-value bound (Vershynin, 2026, Theorem 4.6.1), for every $t \ge 0$,

$$\mathbb{P}\left(\left\|\frac{1}{D}G^\top G - I\right\|_{\mathrm{op}} > C\left[\sqrt{\frac{n + t^2}{D}} + \frac{n + t^2}{D}\right]\right) \le 2e^{-t^2}.$$

Taking $t^2 = \log(8/\delta)$ gives, with probability at least $1 - \delta/4$,

$$\left\|\frac{1}{D}G^\top G - I\right\|_{\mathrm{op}} \le C\left[\sqrt{\frac{n + \log(8/\delta)}{D}} + \frac{n + \log(8/\delta)}{D}\right].$$

By increasing c_G if necessary, on the same event

$$\left\|\frac{1}{D}G^\top G - I\right\|_{\mathrm{op}} \le 2C\sqrt{\frac{n + \log(8n/\delta)}{D}}.$$

We next control the normalization factors. By the Laurent–Massart chi-square concentration inequality (Laurent & Massart, 2000, Lemma 1), if $X \sim \chi_D^2$, then for every $s > 0$, $\mathbb{P}(X - D \ge 2\sqrt{Ds} + 2s) \le e^{-s}$ and $\mathbb{P}(D - X \ge 2\sqrt{Ds}) \le e^{-s}$. Consequently,

$$\mathbb{P}\left(\left|\frac{X}{D} - 1\right| > 2\left[\sqrt{\frac{s}{D}} + \frac{s}{D}\right]\right) \le 2e^{-s}.$$

Applying this with $X = \|\boldsymbol{g}_i\|_2^2$ and $s = \log(8n/\delta)$, and taking a union bound over $i \in [n]$, gives, with probability at least $1 - \delta/4$,

$$\max_{i\le n}\left|\frac{\|\boldsymbol{g}_i\|_2^2}{D} - 1\right| \le 2\left[\sqrt{\frac{\log(8n/\delta)}{D}} + \frac{\log(8n/\delta)}{D}\right].$$

Increasing c_G if necessary, we may further assume

$$\max_{i\le n}\left|\frac{\|\boldsymbol{g}_i\|_2^2}{D} - 1\right| \le 4\sqrt{\frac{\log(8n/\delta)}{D}} \le \frac{1}{2}.$$

Define

$$\Lambda := \operatorname{diag}\left(\frac{\sqrt{D}}{\|\boldsymbol{g}_1\|_2}, \dots, \frac{\sqrt{D}}{\|\boldsymbol{g}_n\|_2}\right), \qquad X := \frac{1}{D} G^\top G.$$

Then $U = D^{-1/2} G\Lambda$, so $U^\top U = \Lambda X \Lambda$ and $U^\top U - I_n = \Lambda(X - I_n)\Lambda + (\Lambda^2 - I_n)$. The preceding normalization bound gives $\|\Lambda\|_{\mathrm{op}} \le \sqrt{2}$. Moreover,

$$\left|\frac{D}{\|\boldsymbol{g}_i\|_2^2} - 1\right| = \frac{\left|\frac{\|\boldsymbol{g}_i\|_2^2}{D} - 1\right|}{\frac{\|\boldsymbol{g}_i\|_2^2}{D}} \le 2\left|\frac{\|\boldsymbol{g}_i\|_2^2}{D} - 1\right|,$$

and therefore $\|\Lambda^2 - I_n\|_{\mathrm{op}} \le 2\max_{i\le n} |\|\boldsymbol{g}_i\|_2^2/D - 1|$. Combining these bounds yields

$$\begin{aligned} \|U^\top U - I_n\|_{\mathrm{op}} &\le \|\Lambda\|_{\mathrm{op}}^2 \|X - I_n\|_{\mathrm{op}} + \|\Lambda^2 - I_n\|_{\mathrm{op}} \\ &\le (4C + 8)\sqrt{\frac{n + \log(8n/\delta)}{D}}. \end{aligned} \tag{H.1}$$

Thus the desired Gram bound holds with probability at least $1 - \delta/2$.

It remains to prove the coherence bound. By rotational invariance and standard spherical concentration (Vershynin, 2026, Theorem 3.4.5), for every $i \neq j$ and $x > 0$, $\mathbb{P}(|\boldsymbol{u}_i^\top \boldsymbol{u}_j| > x) \le 2e^{-cDx^2}$. Setting $x := \sqrt{\log(2n^2/\delta)/(cD)}$ and taking a union bound over the at most $n(n-1)/2$ relevant inner products gives

$$\mathbb{P}\left(\max_{i\neq j} |\boldsymbol{u}_i^\top \boldsymbol{u}_j| > \sqrt{\frac{\log(2n^2/\delta)}{cD}}\right) \le n^2 \exp\left(-cD\frac{\log(2n^2/\delta)}{cD}\right) = \frac{\delta}{2}. \tag{H.2}$$

Combining (H.1) and (H.2) and choosing C_G sufficiently large proves the result. □

Lemma H.5 (Simultaneous spherical Gram concentration)**.** *Let*

$$U = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_n], \qquad \widehat{V} = [\widehat{\boldsymbol{v}}_1, \dots, \widehat{\boldsymbol{v}}_n],$$

where the columns of both matrices are independent and uniformly distributed on $\mathbb{S}^{D-1}$. There exist universal constants $c_G, C_G > 0$ such that, if $D \ge c_G(n + \log(n/\delta))$, then, with probability at least $1 - \delta$,

$$\max\left\{\|U^\top U - I_n\|_{\mathrm{op}}, \|\widehat{V}^\top \widehat{V} - I_n\|_{\mathrm{op}}\right\} \le C_G\sqrt{\frac{n + \log(n/\delta)}{D}},$$

and

$$\max_{i\neq j}\left\{|\boldsymbol{u}_i^\top \boldsymbol{u}_j|, |\widehat{\boldsymbol{v}}_i^\top \widehat{\boldsymbol{v}}_j|\right\} \le C_G\sqrt{\frac{\log(4n^2/\delta)}{D}}.$$

Proof. Apply Lemma H.4 separately to U and $\widehat{V}$, each with failure probability $\delta/2$, take a union bound, and enlarging C_G. □

Lemma H.6 (Inverse spherical Gram concentration)**.** *Let*

$$U = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_n], \qquad G_U := U^\top U,$$

where $\boldsymbol{u}_1, \dots, \boldsymbol{u}_n$ are independent and uniformly distributed on $\mathbb{S}^{D-1}$. Let $\boldsymbol{\xi} \in \{-1, 1\}^n$ be independent of U. There exist universal constants $c_I, C_I > 0$ such that, if $D \ge c_I n \log(8n/\delta)$, then, with probability at least $1 - \delta$,

$$\frac{1}{2} I_n \preceq G_U \preceq \frac{3}{2} I_n, \tag{H.3}$$

and

$$\max_{\boldsymbol{x}\in\{\mathbf{1},\boldsymbol{\xi}\}} \left\|G_U^{-1}\boldsymbol{x} - \boldsymbol{x}\right\|_\infty \le C_I\sqrt{\frac{n\log(8n/\delta)}{D}}. \tag{H.4}$$

Proof. Apply Lemma H.4 with failure probability $\delta/2$. There exists a universal constant $C_G > 0$ such that, on an event $\mathcal{E}$ with $\mathbb{P}(\mathcal{E}) \geq 1 - \delta/2$,

$$\|G_U - I_n\|_{\mathrm{op}} \leq C_G \sqrt{\frac{n + \log(2n/\delta)}{D}} =: \varepsilon. \tag{H.5}$$

By increasing c_I if necessary, we may assume $\varepsilon \leq 1/4$. Hence, on $\mathcal{E}$, $\frac{3}{4} I_n \preceq G_U \preceq \frac{5}{4} I_n$, which in particular implies (H.3).

It remains to establish (H.4). Fix $\boldsymbol{x} \in \{\mathbf{1}, \boldsymbol{\xi}\}$ and $i \in [n]$. After moving the ith coordinate to the first position, write

$$G_U = \begin{pmatrix} 1 & \boldsymbol{g}_i^\top \\ \boldsymbol{g}_i & G_{U,-i} \end{pmatrix}, \qquad \boldsymbol{x} = \begin{pmatrix} x_i \\ \boldsymbol{x}_{-i} \end{pmatrix},$$

where $G_{U,-i} := U_{-i}^\top U_{-i}$ and $\boldsymbol{g}_i := U_{-i}^\top \boldsymbol{u}_i$. The Schur-complement formula gives

$$(G_U^{-1}\boldsymbol{x})_i = \frac{x_i - \boldsymbol{g}_i^\top G_{U,-i}^{-1} \boldsymbol{x}_{-i}}{1 - \boldsymbol{g}_i^\top G_{U,-i}^{-1} \boldsymbol{g}_i},$$

and therefore

$$(G_U^{-1}\boldsymbol{x})_i - x_i = \frac{x_i \boldsymbol{g}_i^\top G_{U,-i}^{-1} \boldsymbol{g}_i - \boldsymbol{g}_i^\top G_{U,-i}^{-1} \boldsymbol{x}_{-i}}{1 - \boldsymbol{g}_i^\top G_{U,-i}^{-1} \boldsymbol{g}_i}. \tag{H.6}$$

We first control the denominator. Since $G_{U,-i}$ is a principal submatrix of G_U, (H.5) implies, on $\mathcal{E}$, $(1-\varepsilon) I_{n-1} \preceq G_{U,-i} \preceq (1+\varepsilon) I_{n-1}$, and hence $\|G_{U,-i}^{-1}\|_{\mathrm{op}} \leq (1-\varepsilon)^{-1} \leq 2$. Moreover, $\|\boldsymbol{g}_i\|_2 \leq \|G_U - I_n\|_{\mathrm{op}} \leq \varepsilon$. Thus $0 \leq \boldsymbol{g}_i^\top G_{U,-i}^{-1} \boldsymbol{g}_i \leq 2\varepsilon^2 \leq 1/8$, and therefore

$$1 - \boldsymbol{g}_i^\top G_{U,-i}^{-1} \boldsymbol{g}_i \geq \frac{1}{2}. \tag{H.7}$$

It remains to control $\boldsymbol{g}_i^\top G_{U,-i}^{-1} \boldsymbol{x}_{-i}$. For each $i \in [n]$, define $\mathcal{E}_{-i} := \{\|G_{U,-i} - I_{n-1}\|_{\mathrm{op}} \leq \varepsilon\}$. Since $G_{U,-i} - I_{n-1}$ is a principal submatrix of $G_U - I_n$, we have $\mathcal{E} \subseteq \mathcal{E}_{-i}$.

On $\mathcal{E}_{-i}$,

$$\begin{aligned} \left\| U_{-i} G_{U,-i}^{-1} \boldsymbol{x}_{-i} \right\|_2^2 &= \boldsymbol{x}_{-i}^\top G_{U,-i}^{-1} U_{-i}^\top U_{-i} G_{U,-i}^{-1} \boldsymbol{x}_{-i} \\ &= \boldsymbol{x}_{-i}^\top G_{U,-i}^{-1} \boldsymbol{x}_{-i} \leq \frac{\|\boldsymbol{x}_{-i}\|_2^2}{1-\varepsilon} \leq 2n. \end{aligned}$$

Furthermore,

$$\boldsymbol{g}_i^\top G_{U,-i}^{-1} \boldsymbol{x}_{-i} = \left\langle \boldsymbol{u}_i,\, U_{-i} G_{U,-i}^{-1} \boldsymbol{x}_{-i} \right\rangle.$$

Conditionally on U_{-i} and $\boldsymbol{\xi}$, the vector $U_{-i} G_{U,-i}^{-1} \boldsymbol{x}_{-i}$ is fixed, whereas $\boldsymbol{u}_i$ remains uniformly distributed on $\mathbb{S}^{D-1}$. Standard spherical concentration therefore gives, for every $t > 0$,

$$\mathbb{P}\left(\mathcal{E}_{-i} \cap \left\{ \left| \boldsymbol{g}_i^\top G_{U,-i}^{-1} \boldsymbol{x}_{-i} \right| > t \right\} \,\middle|\, U_{-i}, \boldsymbol{\xi} \right) \leq 2 \exp\left(-c \frac{D t^2}{n} \right),$$

where we used (H.2).

Set $L := \log(8n/\delta)$ and $t := C_0 \sqrt{nL/D}$, where $C_0 > 0$ is a sufficiently large universal constant. Since $\mathcal{E} \subseteq \mathcal{E}_{-i}$, a union bound over $i \in [n]$ and $\boldsymbol{x} \in \{\mathbf{1}, \boldsymbol{\xi}\}$ gives

$$\mathbb{P}\left(\mathcal{E} \cap \left\{ \max_{\substack{i \in [n] \\ \boldsymbol{x} \in \{\mathbf{1}, \boldsymbol{\xi}\}}} \left| \boldsymbol{g}_i^\top G_{U,-i}^{-1} \boldsymbol{x}_{-i} \right| > t \right\} \right) \leq 4n \exp(-c C_0^2 L) \leq \frac{\delta}{2}.$$

Thus, outside an event of probability at most $\delta/2$, simultaneously for every $i \in [n]$ and $\boldsymbol{x} \in \{\mathbf{1}, \boldsymbol{\xi}\}$,

$$\left| \boldsymbol{g}_i^\top G_{U,-i}^{-1} \boldsymbol{x}_{-i} \right| \leq C_0 \sqrt{\frac{n \log(8n/\delta)}{D}}. \tag{H.8}$$

Finally, by (H.5), $\boldsymbol{g}_i^\top G_{U,-i}^{-1}\boldsymbol{g}_i \le 2\varepsilon^2 \le C\{n+\log(2n/\delta)\}/D$. Under the dimensional assumption $D \ge c_I n\log(8n/\delta)$,

$$\frac{n+\log(2n/\delta)}{D} \le C\sqrt{\frac{n\log(8n/\delta)}{D}}.$$

Combining this estimate with (H.7), (H.8), and (H.6), we obtain

$$\left|(G_U^{-1}\boldsymbol{x})_i - x_i\right| \le C_I\sqrt{\frac{n\log(8n/\delta)}{D}}$$

simultaneously for all $i \in [n]$ and $\boldsymbol{x} \in \{\mathbf{1}, \boldsymbol{\xi}\}$. Together with $\mathbb{P}(\mathcal{E}) \ge 1-\delta/2$ and (H.2), this event has probability at least $1-\delta$, proving (H.4). □

Lemma H.7 (Spherical bilinear tails)**.** *Let $H \in \mathbb{R}^{D\times D}$ be deterministic with $\operatorname{rank}(H) = r_H \ge 1$. Let $\boldsymbol{u}, \boldsymbol{v} \sim \operatorname{Unif}(\mathbb{S}^{D-1})$ be independent. Then there is a universal constant $c > 0$ such that, for all $t > 0$,*

$$\mathbb{P}\left(\left|\left\langle H, \boldsymbol{u}\boldsymbol{v}^\top\right\rangle_F\right| \ge t\right) \le 4\exp\left[-c\min\left\{\frac{D^2t^2}{\|H\|_F^2}, \frac{Dt}{\|H\|_{\mathrm{op}}}\right\}\right]. \tag{H.9}$$

Moreover, there exist universal constants $K_0, c_0, C_0 > 0$ such that, if $D \ge K_0\|H\|_F^2/\|H\|_{\mathrm{op}}^2$, then, for all $0 < t \le \|H\|_F^2/(D\|H\|_{\mathrm{op}})$,

$$\mathbb{P}\left(\left|\left\langle H, \boldsymbol{u}\boldsymbol{v}^\top\right\rangle_F\right| \ge t\right) \ge c_0\exp\left[-C_0\frac{D^2t^2}{\|H\|_F^2}\right]. \tag{H.10}$$

Proof. Let $H = U\Sigma V^\top$ be the compact SVD of H, where $\Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_{r_H})$ and $\sigma_1 \ge \cdots \ge \sigma_{r_H} > 0$. By rotational invariance, we may assume without loss of generality that $U = V = (I_{r_H}, 0)^\top$. Thus, with $T := \langle H, \boldsymbol{u}\boldsymbol{v}^\top\rangle_F$, we have $T = \boldsymbol{u}^\top H\boldsymbol{v} = \sum_{j=1}^{r_H}\sigma_j u_j v_j$.

Let $\boldsymbol{g}, \boldsymbol{h} \sim N(0, I_D)$ be independent. Then $\boldsymbol{u} = \boldsymbol{g}/\|\boldsymbol{g}\|_2$ and $\boldsymbol{v} = \boldsymbol{h}/\|\boldsymbol{h}\|_2$, and hence

$$T = \frac{S_H}{\|\boldsymbol{g}\|_2\|\boldsymbol{h}\|_2}, \qquad S_H := \sum_{j=1}^{r_H}\sigma_j g_j h_j. \tag{H.11}$$

We first prove (H.9). Let $\mathcal{E}_{\mathrm{den}}^- := \{\|\boldsymbol{g}\|_2^2 \ge D/2,\ \|\boldsymbol{h}\|_2^2 \ge D/2\}$. By the standard lower-tail bound for chi-square random variables, $\mathbb{P}((\mathcal{E}_{\mathrm{den}}^-)^c) \le 2e^{-c_1 D}$ for a universal constant $c_1 > 0$. On $\mathcal{E}_{\mathrm{den}}^-$, (H.11) implies $|T| \ge t \Rightarrow |S_H| \ge Dt/2$. Therefore

$$\mathbb{P}(|T| \ge t) \le 2e^{-c_1 D} + \mathbb{P}\left(|S_H| \ge \frac{Dt}{2}\right). \tag{H.12}$$

Since $S_H = \sum_{j=1}^{r_H}\sigma_j g_j h_j$ is a weighted sum of independent centered sub-exponential random variables, Bernstein's inequality gives a universal constant $c_2 > 0$ such that, for all $x > 0$,

$$\mathbb{P}(|S_H| \ge x) \le 2\exp\left[-c_2\min\left\{\frac{x^2}{\sum_{j=1}^{r_H}\sigma_j^2}, \frac{x}{\sigma_1}\right\}\right].$$

Since $\sum_j \sigma_j^2 = \|H\|_F^2$ and $\sigma_1 = \|H\|_{\mathrm{op}}$, taking $x = Dt/2$ gives, after changing the universal constant,

$$\mathbb{P}\left(|S_H| \ge \frac{Dt}{2}\right) \le 2\exp\left[-c_3\min\left\{\frac{D^2t^2}{\|H\|_F^2}, \frac{Dt}{\|H\|_{\mathrm{op}}}\right\}\right].$$

If $0 < t \le \|H\|_{\mathrm{op}}$, then $\min\{D^2t^2/\|H\|_F^2, Dt/\|H\|_{\mathrm{op}}\} \le D$, so the denominator error in (H.12) is absorbed into the same exponential. This proves (H.9) for $0 < t \le \|H\|_{\mathrm{op}}$. If $t > \|H\|_{\mathrm{op}}$, its left-hand side is zero because $|T| \le \|H\|_{\mathrm{op}}$. Thus (H.9) holds for all $t > 0$.

We now prove (H.10). Assume $D \ge K_0\|H\|_F^2/\|H\|_{\mathrm{op}}^2$, where $K_0 \ge 1$ will be chosen sufficiently large, and define $\sigma_g^2 := \sum_{j=1}^{r_H}\sigma_j^2 g_j^2$ and

$$\mathcal{E}_g := \left\{\sigma_g^2 \ge \frac{\|H\|_F^2}{2}, \quad \|\boldsymbol{g}\|_2^2 \le 2D\right\}.$$

We first show that $\mathcal{E}_g$ has probability bounded below by a universal constant. We have $\mathbb{E}\sigma_g^2 = \|H\|_F^2$, while

$$\mathbb{E}(\sigma_g^2)^2 = \sum_{j=1}^{r_H} \sigma_j^4 \mathbb{E} g_j^4 + \sum_{i \neq j} \sigma_i^2 \sigma_j^2 \mathbb{E} g_i^2 \mathbb{E} g_j^2 = \|H\|_F^4 + 2\sum_{j=1}^{r_H} \sigma_j^4 \leq 3\|H\|_F^4.$$

Hence, by Paley–Zygmund, $\mathbb{P}(\sigma_g^2 \geq \|H\|_F^2/2) \geq 1/12$. On the other hand, the standard chi-square upper-tail bound gives $\mathbb{P}(\|\boldsymbol{g}\|_2^2 > 2D) \leq e^{-cD}$. Since $D \geq K_0 \|H\|_F^2/\|H\|_{\mathrm{op}}^2 \geq K_0$, choosing K_0 sufficiently large gives $e^{-cD} \leq 1/24$. Therefore

$$\mathbb{P}(\mathcal{E}_g) \geq \frac{1}{12} - \frac{1}{24} = \frac{1}{24}. \tag{H.13}$$

Fix $\boldsymbol{g} \in \mathcal{E}_g$, and let $\boldsymbol{e}_g \in \mathbb{R}^D$ be the unit vector whose first r_H coordinates are $(\sigma_1 g_1, \ldots, \sigma_{r_H} g_{r_H})/\sigma_g$ and whose remaining coordinates are zero. Decompose $\boldsymbol{h} = Z\boldsymbol{e}_g + \boldsymbol{h}_\perp$, where $Z \sim N(0,1)$, $\boldsymbol{h}_\perp \sim N(0, I_{D-1})$, and Z is independent of $\boldsymbol{h}_\perp$. Then $S_H = \sigma_g Z$ and $\|\boldsymbol{h}\|_2^2 = Z^2 + \|\boldsymbol{h}_\perp\|_2^2$.

We use the following elementary Gaussian consequence: there exist universal constants $b_1, c_1', C_1' > 0$ such that, for all $D \geq 1$ and $0 \leq a \leq b_1\sqrt{D}$,

$$\mathbb{P}\left(|Z| \geq a, \quad Z^2 \leq \frac{D}{2}\right) \geq c_1' e^{-C_1' a^2}.$$

Also, by a chi-square bound, there is a universal constant $p > 0$ such that $\mathbb{P}(\|\boldsymbol{h}_\perp\|_2^2 \leq 3D/2) \geq p$. Since Z and $\boldsymbol{h}_\perp$ are independent, for $a := x/\sigma_g \leq b_1\sqrt{D}$,

$$\mathbb{P}\left(|S_H| \geq x, \quad \|\boldsymbol{h}\|_2^2 \leq 2D \,\middle|\, \boldsymbol{g}\right) \geq p c_1' e^{-C_1' a^2}.$$

On $\mathcal{E}_g$, $\sigma_g^2 \geq \|H\|_F^2/2$, and hence $a^2 \leq 2x^2/\|H\|_F^2$. Thus, after changing universal constants,

$$\mathbb{P}\left(|S_H| \geq x, \quad \|\boldsymbol{h}\|_2^2 \leq 2D \,\middle|\, \boldsymbol{g}\right) \geq c_2' e^{-C_2' x^2/\|H\|_F^2}, \tag{H.14}$$

whenever $0 < x \leq 2\|H\|_F^2/\|H\|_{\mathrm{op}}$, provided K_0 is sufficiently large. Indeed, in this range, on $\mathcal{E}_g$, $a = x/\sigma_g \leq \sqrt{2}\, x/\|H\|_F \leq 2\sqrt{2}\, \|H\|_F/\|H\|_{\mathrm{op}}$. Since $D \geq K_0 \|H\|_F^2/\|H\|_{\mathrm{op}}^2$, this gives $a \leq 2\sqrt{2}\,\sqrt{D}/\sqrt{K_0}$. Choosing $K_0 \geq 8/b_1^2$ ensures $a \leq b_1\sqrt{D}$.

Integrating (H.14) over $\mathcal{E}_g$ and using (H.13), we obtain

$$\mathbb{P}\left(|S_H| \geq x, \quad \|\boldsymbol{g}\|_2^2 \leq 2D, \quad \|\boldsymbol{h}\|_2^2 \leq 2D\right) \geq \frac{c_2'}{24} e^{-C_2' x^2/\|H\|_F^2}, \qquad 0 < x \leq \frac{2\|H\|_F^2}{\|H\|_{\mathrm{op}}}. \tag{H.15}$$

Absorbing $c_2'/24$ into a universal constant $c_0' > 0$ gives

$$\mathbb{P}\left(|S_H| \geq x, \quad \|\boldsymbol{g}\|_2^2 \leq 2D, \quad \|\boldsymbol{h}\|_2^2 \leq 2D\right) \geq c_0' e^{-C_2' x^2/\|H\|_F^2}, \qquad 0 < x \leq \frac{2\|H\|_F^2}{\|H\|_{\mathrm{op}}}.$$

Finally, take $x := 2Dt$. If $0 < t \leq \|H\|_F^2/(D\|H\|_{\mathrm{op}})$, then $0 < x \leq 2\|H\|_F^2/\|H\|_{\mathrm{op}}$. On the event in the preceding display, $\|\boldsymbol{g}\|_2\|\boldsymbol{h}\|_2 \leq 2D$, so (H.11) gives $|T| \geq |S_H|/(2D) \geq x/(2D) = t$. Therefore

$$\mathbb{P}(|T| \geq t) \geq c_0' \exp\left[-4C_2' \frac{D^2 t^2}{\|H\|_F^2}\right].$$

Renaming constants proves (H.10). □

H.3 Polar factor

Throughout this section, for a matrix X, write a compact singular value decomposition as

$$X = U\Sigma V^\top, \qquad \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r), \qquad \sigma_i > 0.$$

Lemma H.8 (Basic properties of the compact polar factor)**.** *Let $X \neq 0$. Then*

$$\|\operatorname{Polar}(X)\|_{\mathrm{op}} = 1, \qquad \langle \operatorname{Polar}(X), X\rangle_F = \|X\|_1, \qquad \operatorname{Polar}(cX) = \operatorname{sgn}(c)\operatorname{Polar}(X)$$

for every $c \in \mathbb{R} \setminus \{0\}$.

Proof. Let $X = U\Sigma V^\top$ be a compact singular value decomposition. Then $\operatorname{Polar}(X) = UV^\top$, so $\|\operatorname{Polar}(X)\|_{\mathrm{op}} = \|UV^\top\|_{\mathrm{op}} = 1$. Moreover,

$$\langle \operatorname{Polar}(X), X\rangle_F = \operatorname{tr}\big((UV^\top)^\top U\Sigma V^\top\big) = \operatorname{tr}(\Sigma) = \|X\|_1.$$

If $c > 0$, then $cX = U(c\Sigma)V^\top$, so $\operatorname{Polar}(cX) = \operatorname{Polar}(X)$. If $c < 0$, then $cX = U((-c)\Sigma)(-V)^\top$, so $\operatorname{Polar}(cX) = -\operatorname{Polar}(X)$. □

Lemma H.9 (Spectral–Schatten-1 duality and subgradient)**.** *For every matrix X,*

$$\|X\|_1 = \sup_{\|G\|_{\mathrm{op}} \le 1} \langle G, X\rangle_F.$$

The supremum is attained at $G = \operatorname{Polar}(X)$. Moreover, $\operatorname{Polar}(X) \in \partial\|X\|_1$; that is, for every matrix Y,

$$\|Y\|_1 \ge \|X\|_1 + \langle \operatorname{Polar}(X), Y - X\rangle_F.$$

Proof. If $X = 0$, the claims are immediate. Suppose $X \neq 0$, and write $X = U\Sigma V^\top$. For any G satisfying $\|G\|_{\mathrm{op}} \le 1$,

$$\langle G, X\rangle_F = \operatorname{tr}\big((U^\top GV)^\top \Sigma\big) \le \sum_{i=1}^{r} \sigma_i = \|X\|_1,$$

since $\|U^\top GV\|_{\mathrm{op}} \le 1$ implies that each diagonal entry of $U^\top GV$ has absolute value at most one. Taking $G = \operatorname{Polar}(X) = UV^\top$ gives equality by Lemma H.8, proving the duality.

Consequently, for every matrix Y, $\|Y\|_1 \ge \langle \operatorname{Polar}(X), Y\rangle_F = \|X\|_1 + \langle \operatorname{Polar}(X), Y - X\rangle_F$, where the equality follows from Lemma H.8. This is the subgradient inequality. □

Lemma H.10 (Gradient of the Schatten 1-norm at a full-rank matrix)**.** *Let $X \in \mathbb{R}^{m\times n}$ have full column rank. Then $\nabla\|X\|_1 = \operatorname{Polar}(X)$. Equivalently, for every perturbation $E \in \mathbb{R}^{m\times n}$,*

$$\mathrm{d}\|\cdot\|_1(X)[E] = \frac{d}{dt}\Big|_{t=0} \|X + tE\|_1 = \langle \operatorname{Polar}(X), E\rangle_F.$$

The analogous statement holds for full row rank matrices by applying the result to $X^\top$.

Proof. Assume X has full column rank and set $H := (X^\top X)^{1/2}$. Then $H \succ 0$, $\|X\|_1 = \operatorname{tr}(H)$, and $\operatorname{Polar}(X) = XH^{-1}$.

For an arbitrary perturbation E, let $H(t) := ((X + tE)^\top (X + tE))^{1/2}$. Since X has full column rank, so does $X + tE$ for all t sufficiently close to 0, and hence $\|X + tE\|_1 = \operatorname{tr}(H(t))$ for such t. The square-root map is smooth on the positive definite cone, for instance by the holomorphic functional calculus representation

$$A^{1/2} = \frac{1}{2\pi i}\int_\Gamma z^{1/2}(zI - A)^{-1}\,dz,$$

with Γ enclosing the spectrum of A in the right half-plane. Thus $H(t)$ is differentiable near $t = 0$.

Writing $\dot H := H'(0)$ and differentiating $H(t)^2 = (X + tE)^\top(X + tE)$ gives $H\dot H + \dot H H = X^\top E + E^\top X$. Multiplying by H^{-1}, taking traces, and using cyclicity yields

$$2\operatorname{tr}(\dot H) = 2\operatorname{tr}\big((XH^{-1})^\top E\big) = 2\langle \operatorname{Polar}(X), E\rangle_F.$$

Therefore $\frac{d}{dt}\big|_{t=0}\|X + tE\|_1 = \operatorname{tr}(\dot H) = \langle \operatorname{Polar}(X), E\rangle_F$. Hence $\nabla\|X\|_1 = \operatorname{Polar}(X)$. The full row rank case follows by applying the result to $X^\top$. □

Lemma H.11 (Local regularity of the polar factor)**.** *Let $X \in \mathbb{R}^{m\times n}$ have full column rank. Then the map $Y \mapsto \operatorname{Polar}(Y)$ is locally Lipschitz in a neighborhood of X. In particular, if $X_j \to X$, then $\operatorname{Polar}(X_j) \to \operatorname{Polar}(X)$. The same statements hold for full row rank matrices.*

Proof. The set of full-column-rank matrices is open, and on this set $\mathrm{Polar}(Y) = Y(Y^\top Y)^{-1/2}$. Since $A \mapsto A^{-1/2}$ is smooth on the positive definite cone (for instance by the holomorphic functional calculus), the polar map is smooth, and hence locally Lipschitz, in a neighborhood of X. The continuity assertion follows immediately. The full row rank case follows by applying the result to $X^\top$. □

Lemma H.12 (Orthonormalization identity)**.** *Let $U \in \mathbb{R}^{m_1 \times n}$ and $V \in \mathbb{R}^{m_2 \times n}$ have full column rank. Let $G_U := U^\top U$ and $G_V := V^\top V$. Then, for every $M \in \mathbb{R}^{n \times n}$, the nonzero singular values of $UMV^\top$ coincide with those of $G_U^{1/2} M G_V^{1/2}$:*

$$\sigma_j(UMV^\top) = \sigma_j(G_U^{1/2} M G_V^{1/2}), \qquad j \in [n].$$

Consequently,

$$\mathrm{rank}(UMV^\top) = \mathrm{rank}(M), \qquad \|UMV^\top\|_1 = \|G_U^{1/2} M G_V^{1/2}\|_1.$$

Moreover,

$$\mathrm{Polar}(UMV^\top) = \overline{U}\ \mathrm{Polar}(G_U^{1/2} M G_V^{1/2})\, \overline{V}^\top, \qquad \overline{U} := U G_U^{-1/2}, \quad \overline{V} := V G_V^{-1/2}.$$

Proof. By definition, $\overline{U}^\top \overline{U} = G_U^{-1/2} U^\top U G_U^{-1/2} = I$, and similarly $\overline{V}^\top \overline{V} = I$. Also $U = \overline{U} G_U^{1/2}$ and $V = \overline{V} G_V^{1/2}$, so

$$UMV^\top = \overline{U} \left(G_U^{1/2} M G_V^{1/2} \right) \overline{V}^\top.$$

Multiplication on the left and right by matrices with orthonormal columns preserves all nonzero singular values. Hence $\|UMV^\top\|_1 = \|G_U^{1/2} M G_V^{1/2}\|_1$, and the same singular value decomposition shows that the compact polar factor is $\overline{U}\ \mathrm{Polar}(G_U^{1/2} M G_V^{1/2})\, \overline{V}^\top$. □

Lemma H.13 (Differential of the polar factor at a positive diagonal matrix)**.** *Let $\Lambda = \mathrm{diag}(d_1, \dots, d_n)$, where $d_i > 0$. For a perturbation $E \in \mathbb{R}^{n \times n}$, define $X(t) := \Lambda + tE$. Then*

$$\mathrm{Polar}(X(t)) = I + tK + o(t),$$

where K is skew-symmetric and is determined by $K\Lambda + \Lambda K = E - E^\top$. Equivalently,

$$K_{ij} = \frac{E_{ij} - E_{ji}}{d_i + d_j} \quad (i \neq j), \qquad K_{ii} = 0.$$

Proof. For sufficiently small t, the matrix $X(t)$ is invertible. Write its polar decomposition as $X(t) = Q(t)H(t)$, where $Q(t)$ is orthogonal and $H(t)$ is symmetric positive definite. Since Λ is positive diagonal, $Q(0) = I$ and $H(0) = \Lambda$. Let $K := Q'(0)$. Differentiating $Q(t)^\top Q(t) = I$ at $t = 0$ gives $K^\top + K = 0$, so K is skew-symmetric.

Differentiating $\Lambda + tE = Q(t)H(t)$ at $t = 0$ gives $E = K\Lambda + H'(0)$. Since $H'(0)$ is symmetric, taking transposes gives $E^\top = -\Lambda K + H'(0)$. Subtracting yields $E - E^\top = K\Lambda + \Lambda K$. Since Λ is diagonal, $(K\Lambda + \Lambda K)_{ij} = (d_i + d_j)K_{ij}$, and hence $K_{ij} = (E_{ij} - E_{ji})/(d_i + d_j)$ for $i \neq j$. For $i = j$, skew-symmetry gives $K_{ii} = 0$. □

Lemma H.14 (Hessian of the Schatten 1-norm at a positive diagonal matrix)**.** *Let $\Lambda = \mathrm{diag}(d_1, \dots, d_n)$, where $d_i > 0$. Then, for every perturbation $E \in \mathbb{R}^{n \times n}$,*

$$\nabla^2 \|\cdot\|_1(\Lambda)[E, E] = \sum_{i<j} \frac{(E_{ij} - E_{ji})^2}{d_i + d_j}.$$

Consequently,

$$\|\Lambda + tE\|_1 = \mathrm{tr}(\Lambda) + t\,\mathrm{tr}(E) + \frac{t^2}{2} \sum_{i<j} \frac{(E_{ij} - E_{ji})^2}{d_i + d_j} + o(t^2).$$

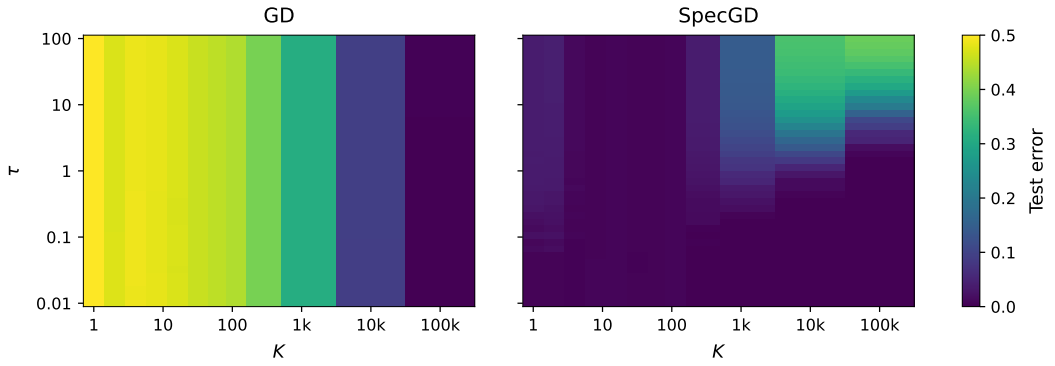

Figure 4: Test error as a function of shortcut dimension K and cumulative optimization time $\tau := \sum_{\ell=0}^{T} \eta_\ell$ for GD and SpecGD. Parameters: $n = 100$, $r = 1$, $D = d - r = 10^5$, $\rho = 0.45$, $\lambda = 125$, $\eta = 10^{-2}$, and $T = 10^4$

Proof. Since Λ is invertible, the Schatten 1-norm is differentiable in a neighborhood of Λ, and by Lemma H.10, $\nabla\|X\|_1 = \mathrm{Polar}(X)$ for X near Λ. Thus, for $X(t) = \Lambda + tE$,

$$\frac{d}{dt}\|X(t)\|_1 = \langle \mathrm{Polar}(X(t)), E\rangle_F.$$

At $t = 0$, $\mathrm{Polar}(\Lambda) = I$, so the first derivative is $\langle I, E\rangle_F = \mathrm{tr}(E)$. By Lemma H.13, $\mathrm{Polar}(X(t)) = I + tK + o(t)$, where $K_{ij} = (E_{ij} - E_{ji})/(d_i + d_j)$ for $i \neq j$ and $K_{ii} = 0$. Therefore,

$$\nabla^2\|\cdot\|_1(\Lambda)[E, E] = \frac{d}{dt}\Big|_{t=0} \langle \mathrm{Polar}(\Lambda + tE), E\rangle_F = \langle K, E\rangle_F.$$

Since K is skew-symmetric,

$$\langle K, E\rangle_F = \sum_{i<j} K_{ij}(E_{ij} - E_{ji}) = \sum_{i<j} \frac{(E_{ij} - E_{ji})^2}{d_i + d_j}.$$

The Taylor expansion follows from differentiability of the gradient in a neighborhood of Λ. □

I Additional numerical experiments

We provide additional numerical experiments.

Early stopping mitigates overfitting for SpecGD, but not GD in the dispersed shortcut model. Figure 4 shows how generalization evolves along training as the shortcut dimension K and the cumulative optimization time $\tau := \sum_{\ell=0}^{T} \eta_\ell$ vary. For SpecGD, increasing the training time can substantially degrade test performance, especially as the shortcuts become more dispersed, while an intermediate stopping time preserves good generalization over a much broader range of K. In contrast, GD exhibits little comparable benefit from early stopping: once interpolation is reached, its generalization behavior is largely determined by the shortcut geometry. Thus, training time provides an additional control over the implicit bias of SpecGD, whereas it has a much weaker effect on GD.

Influence of the sample size n. As illustrated by Figure 5, under the dispersed model, increasing the sample size improves the performance of SpecGD.

J Comparison between SpecGD and GD

The following table summarizes our theoretical findings.

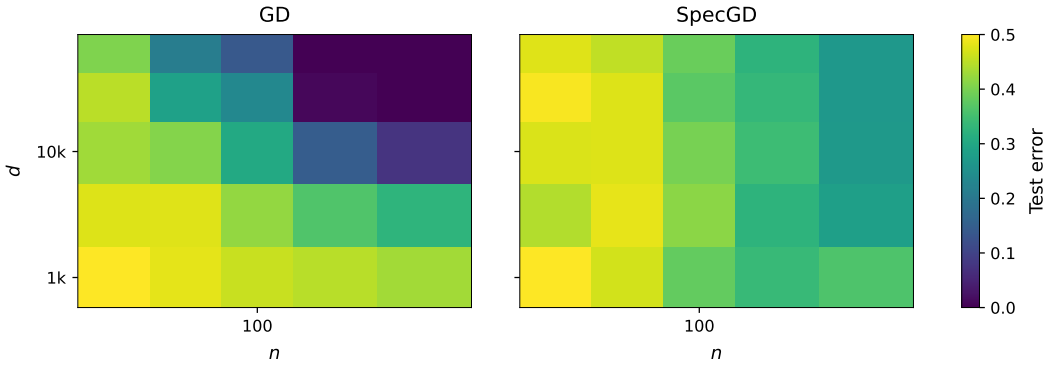

Figure 5: Test error as a function of the sample size n and ambient dimension d for GD and SpecGD in the random dispersed model, with $r = 1$ and $K = D = d - r$. Parameters: $\rho = 0.45$, $\lambda = 125$, and $\eta = 10^{-2}$. We compare all configurations at fixed normalized optimization time $n\tau/D = 1.5$. Results are averaged over 10 independent runs.

Table 2: Comparison of the late-stage generalization of GD and SpecGD.

	Parameter regime	$R(W^\star_{\mathrm{GD}})$		$R(W^\star)$
CSM	$\lambda^2 \ll n\|S\|_F^2$	$e^{-\Theta\left(\frac{D}{n}\right)}$		$e^{-\Theta\left(\frac{D}{n}\right)}$
	$n\|S\|_F^2 \ll \lambda^2 \ll \sqrt{nD}\|S\|_F^2 \wedge n\|S\|_1^2$	$e^{-\Theta\left(\frac{Dn\|S\|_F^4}{\lambda^4}\right)}$	$\gg$	$e^{-\Theta\left(\frac{D}{n}\right)}$
	$n\|S\|_1^2 \ll \lambda^2 \ll \sqrt{nD}\|S\|_F^2$	$e^{-\Theta\left(\frac{Dn\|S\|_F^4}{\lambda^4}\right)}$	$\gg$	$e^{-\Theta\left(\frac{D\|S\|_1^2}{\lambda^2}\right)}$
	$\sqrt{nD}\|S\|_F^2 \ll \lambda^2 \ll D\|S\|_1^2$	$\frac{1}{2}$	$\gg$	$e^{-\Theta\left(\frac{D}{n} \wedge \frac{D\|S\|_1^2}{\lambda^2}\right)}$
	$\lambda^2 \gg D\|S\|_1^2$	$\frac{1}{2}$		$\frac{1}{2}$
DSM[1]	$\lambda^2 \ll n\|S\|_F^2 \wedge \frac{D^2\|S\|_1^2}{n^2}$	$e^{-\Theta(D)}$	$\ll$	$e^{-\Theta(n)}$
	$n\|S\|_F^2 \ll \lambda^2 \ll D\|S\|_F^2 \wedge \frac{D^2\|S\|_1^2}{n^2}$	$e^{-\Theta\left(\frac{Dn\|S\|_F^2}{\lambda^2}\right)}$	$\ll$	$e^{-\Theta(n)}$
	$D\|S\|_F^2 \ll \lambda^2 \ll D\sqrt{n}\,\|S\|_F^2 \wedge \frac{D^2\|S\|_1^2}{n^2}$	$e^{-\Theta\left(\frac{D^2n\|S\|_F^4}{\lambda^4}\right)}$	$\gg$	$e^{-\Theta(n)}$
	$D\sqrt{n}\,\|S\|_F^2 \ll \lambda^2 \ll \frac{D^2\|S\|_1^2}{n^2}$	$\frac{1}{2}$	$\gg$	$e^{-\Theta(n)}$
	$\frac{D^2\|S\|_1^2}{n^2} \ll \lambda^2 \ll n\|S\|_F^2$	$e^{-\Theta(D)}$		*not characterized*
	$n\|S\|_F^2 \vee \frac{D^2\|S\|_1^2}{n^2} \ll \lambda^2 \ll D\|S\|_F^2$	$e^{-\Theta\left(\frac{Dn\|S\|_F^2}{\lambda^2}\right)}$		*not characterized*
	$D\|S\|_F^2 \vee \frac{D^2\|S\|_1^2}{n^2} \ll \lambda^2 \ll D\sqrt{n}\,\|S\|_F^2$	$e^{-\Theta\left(\frac{D^2n\|S\|_F^4}{\lambda^4}\right)}$		*not characterized*
	$\lambda^2 \gg D\sqrt{n}\,\|S\|_F^2 \vee \frac{D^2\|S\|_1^2}{n^2}$	$\frac{1}{2}$		*not characterized*

[1] The relative magnitude of $\frac{D^2\|S\|_1^2}{n^2}$ and the GD transition scales $n\|S\|_F^2$, $D\|S\|_F^2$, and $D\sqrt{n}\,\|S\|_F^2$ depends on D, n, S; consequently, some displayed parameter regimes may be empty.