Matrix-Free Eigendecomposition with Isotropic-Noise Invariance via Discrete Double-Bracket Flows

Anonymous authors
Paper under double-blind review

Abstract

Eigenspace estimation without explicit matrix materialization is a fundamental challenge when sequential observations are contaminated by uniform spectral shifts and unstructured errors (e.g., $C_t = C_{\text{sig}} + \sigma_t^2 I + E_t$). Traditional stochastic approximation algorithms typically falter in this regime: fixed-step schedules become bottlenecked by the inflated overall operator norm, whereas adaptive scaling strategies suffer from severe deceleration due to vanishing updates. To circumvent these limitations, we propose a discrete-time double-bracket dynamical system. By design, the generator of this flow completely filters out isotropic perturbations, guaranteeing exact immunity to the $\sigma_t^2 I$ term at every discrete integration step. Consequently, both the optimization path and the theoretical step-size bound ($\eta_{\max} \propto 1/||C_e||_2^2$) are governed strictly by the traceless component of the covariance, C_e. By leveraging the strict-saddle property of the objective alongside input-to-state stability tools, we prove global convergence with a sample complexity of $O(||C_e||_2^2/(\Delta^2\varepsilon))$ against trace-free noise. Furthermore, by rigorously bounding the dynamics around degenerate subspaces, we establish an accelerated $O(\log(1/\zeta))$ escape rate, culminating in sharp, high-probability finite-time guarantees.

1 Introduction

Matrix-free eigendecomposition is the problem of recovering the eigenbasis of a covariance operator from sequential observations via matrix–vector products (MVP). Classical streaming methods such as Oja's rule Oja (1982); Jain et al. (2016) address the dominant-subspace case (top-k PCA); this work considers full simultaneous diagonalization Helmke & Moore (2012) under high isotropic noise.

However, the effectiveness of classical algorithms rests on an implicit premise: the observations have high signal-to-noise ratio (SNR), or the background noise can be easily peeled off Mitliagkas et al. (2013); Marinov et al. (2018). This premise breaks down in several frontier examples, including Hessian spectrum monitoring in LLM training Zhang et al. (2024); Liao & Mahoney (2021) (where weight decay and damping introduce a large λI shift), differentially private PCA Liu et al. (2022); Dong et al. (2022); Balle & Wang (2018) (where the Gaussian mechanism injects $\sigma^2 I$ noise), and non-stationary sensor arrays corrupted by thermal or radiation backgrounds.

In these regimes, the failure of stochastic approximation is structural rather than merely numerical. As the background intensity σ^2 dominates the signal λ_1, the spectral ratio $\rho = (\lambda_2 + \sigma^2)/(\lambda_1 + \sigma^2)$ approaches unity Jain et al. (2016); Liang (2023), forcing the algorithm into a state of *asymptotic freezing* where convergence time scales linearly with noise magnitude. Simultaneously, stability constraints dictate a passive step size $\eta \propto 1/\sigma^2$ Shamir (2016); Karandikar & Vidyasagar (2024), rendering the system unresponsive to transients. Extrinsic remedies are computationally disqualified under the matrix-free oracle: trace-denoising via Hutchinson estimators Meyer et al. (2021); Skorski (2021); Dharangutte & Musco (2021) incurs a prohibitive $O(\sigma^4)$ oracle cost to recover a valid descent direction, while adaptive optimizers Henriksen & Ward (2019); Demir & Dogan (2025) normalize the gradient magnitude without correcting its noise-biased direction.

We resolve this geometric deficiency by constructing a *discrete state-dependent double-bracket flow* Brockett (1991); Helmke & Moore (2012); Wegner (1994); Głazek & Wilson (1993); Gluza (2024). The core of our approach is the generator $\Omega(M) = [A, D]$, with $A = M^\top C_k M$ and $D = \mathrm{diag}(A)$. Exploiting the Lie bracket identity $[\alpha I, \cdot] \equiv 0$, this generator enforces $\Omega_{C+\alpha I} \equiv \Omega_C$. Unlike methods that estimate and subtract noise, this algebraic structure renders the isotropic component $\sigma_k^2 I$ invisible to the update dynamics. The discrete trajectory depends exclusively on the trace-free covariance C_e, achieving exact pathwise σ^2-invariance.

Contributions

1. **Stepwise Exact Invariance.** The discrete trajectory $\{M_k\}$ and Lyapunov values $\{f(M_k)\}$ are pointwise invariant to $\{\sigma_k^2\}$. The maximal stable step size depends solely on the signal energy, $\eta_{\max} \propto 1/\|C_e\|_2^2$, independent of the noise floor (Theorem 3.1(iv)).

2. **ISS-Based Robustness.** We establish input-to-state stability Sontag (2008) for the discrete dynamics (Theorem 4.3). The steady-state error is governed exclusively by the trace-free perturbation $\|E_e\|$, decoupling convergence accuracy from σ_k^2.

3. **Deterministic Saddle Escape.** Using the explicit Givens-rotation structure Golub & Van Loan (2013) at degenerate blocks (Proposition 5.3), we prove a *theoretical* $O(\log(1/\zeta))$ escape rate, strictly improving upon the generic $O(\log^4(1/\zeta))$ dependence of perturbed gradient descent Jin et al. (2017); Xu & Li (2021). In practice, the algorithm escapes saddles naturally via gradient descent; the Givens formula provides the theoretical guarantee.

4. **Dissipation Channel Condition.** We characterize the geometric necessity for σ-free stability (Section 6): any generator achieving such invariance must align with the commutator $\Omega = [A, \mathrm{diag}(A)]$.

Scope We address eigendecomposition under the MVP oracle $v \mapsto C_k v$ with observation model $C_k = C_{\mathrm{sig}} + \sigma_k^2 I + E_k$. Per-iteration cost is $O(n)$ MVPs and $O(n^3)$ arithmetic; structured (non-isotropic) backgrounds are outside our scope. The term "matrix-free" refers to the *access model* (no explicit storage of C_k), not computational speedup—our advantage lies in *surviving* high-noise regimes where baselines fail.

2 Related Work

Online Eigensolvers and Trace Denoising. Oja's algorithm Oja (1982) remains the canonical baseline for online principal component analysis, with convergence depending critically on the spectral ratio $\rho = \lambda_2/\lambda_1$ Jain et al. (2016); Allen-Zhu & Li (2017); Shamir (2016). When isotropic noise $\sigma^2 I$ contaminates the covariance, the effective ratio approaches unity, inducing *asymptotic freezing*. Recent extensions—including Markovian data handling Kumar & Sarkar (2023), heteroscedastic noise Gilman et al. (2025), adaptive step sizes Henriksen & Ward (2019), sparse PCA Kumar & Sarkar (2024), and robust formulations Bienstock et al. (2022); Diakonikolas et al. (2023)—mitigate sensitivity to unknown scale or adversarial corruption but retain explicit σ-dependence in their bounds. A natural remedy is trace estimation Skorski (2021); Meyer et al. (2021); Woodruff et al. (2022), yet achieving sufficient precision for eigenbasis recovery incurs $O(\sigma^4/\Delta^2)$ oracle calls per iteration, making it the dominant cost. Our approach bypasses estimation entirely: the commutator algebraically annihilates $\sigma^2 I$ before it enters the dynamics.

Double-Bracket Flows. The continuous double-bracket flow $\dot{X} = [X, [X, N]]$ was introduced by Brockett Brockett (1991) for matrix diagonalization, with comprehensive treatment in Helmke and Moore Helmke & Moore (2012). Recently, Gluza Gluza (2024) revived discrete approximations for variational quantum circuits. These works focus on diagonalizing static matrices; the commutator's shift-invariance ($[\alpha I, \cdot] \equiv 0$) is typically inconsequential. We repurpose this property for sequential observations with time-varying, unknown isotropic backgrounds—the shift-invariance becomes the mechanism that renders $\sigma_k^2 I$ invisible to the trajectory.

Riemannian Optimization and Saddle Escape. Optimization under orthogonality constraints is naturally formulated on Stiefel or orthogonal group manifolds Absil et al. (2008), with the Cayley transform

providing structure-preserving retractions Li et al. (2020); Ablin et al. (2024). Stochastic Riemannian methods almost surely avoid saddle points under mild conditions Hsieh et al. (2023). Our algorithm fits this framework but with a *state-dependent* generator $\Omega = [A, \mathrm{diag}(A)]$, which invalidates standard fixed-target analyses. For saddle escape, perturbed gradient descent requires $O(\log^4(1/\zeta))$ iterations Jin et al. (2017); Zhang & Li (2021); our approach exploits explicit Givens-rotation escape directions at degenerate blocks Nowak et al. (2024), improving complexity to $O(\log(1/\zeta))$ deterministically.

Jacobi Diagonalization and Plane Rotations. Classical Jacobi sweeps Golub & Van Loan (2013) iterate Givens rotations that zero off-diagonal entries of $A = M^\top CM$. Each 2×2 rotation is shift-invariant ($[\alpha I_{2\times 2}, \cdot] = 0$), suggesting potential σ^2-immunity. However, Jacobi requires $O(n^2)$ entry accesses per sweep—incompatible with the MVP constraint. Our commutator flow aggregates all $O(n^2)$ rotation directions into a single MVP-based update while preserving shift-invariance.

3 Problem Setup and σ-Immunity Structure

3.1 Observation Model

We consider matrix-free eigendecomposition under the observation model

$$C_k = C_{\mathrm{sig}} + \sigma_k^2 I + E_k, \tag{1}$$

where $C_{\mathrm{sig}} \in \mathbb{R}^{n\times n}$ is the signal covariance with distinct eigenvalues $\lambda_1 > \cdots > \lambda_n$, the sequence $\{\sigma_k^2\}_{k\geq 0}$ represents time-varying isotropic background intensities (potentially unbounded), and E_k is an anisotropic perturbation. Access is restricted to matrix–vector products $v \mapsto C_k v$.

The trace-free projection $\mathrm{tf}(X) := X - \frac{\mathrm{tr}(X)}{n} I$ satisfies $\mathrm{tf}(\sigma^2 I) = 0$. Define $C_e := \mathrm{tf}(C_{\mathrm{sig}})$ and $E_e := \mathrm{tf}(E)$.

3.2 Diagonalization Objective on $SO(n)$

Let $M \in SO(n)$ be the optimization variable. Define

$$A(M) := M^\top C_e M, \qquad D(M) := \mathrm{diag}(A), \tag{2}$$

$$f(M) := \tfrac{1}{2}\|\mathrm{off}(A)\|_F^2, \qquad \mathrm{off}(A) := A - D. \tag{3}$$

Then $f(M) = 0$ iff M diagonalizes C_e, i.e., M is an eigenvector matrix of C_{sig} (up to permutation and sign).

3.3 Commutator Generator and σ^2-Immunity

Define the generator $\Omega : SO(n) \to \mathfrak{so}(n)$ by the double-bracket (commutator) structure Brockett (1991); Helmke & Moore (2012):

$$\Omega(M) := [A(M), D(M)] = AD - DA. \tag{4}$$

The entrywise expression is $\Omega_{ij} = (A_{jj} - A_{ii})A_{ij}$ for $i \neq j$. The Riemannian gradient of f on $SO(n)$ is $\mathrm{grad}\, f(M) = -M\Omega(M)$ Absil et al. (2008).

The following theorem establishes the core algebraic structure that renders the dynamics immune to isotropic noise.

Theorem 3.1 (σ^2-Immunity)**.** *Let the observation sequence be $C_k = C_{\mathrm{sig}} + \sigma_k^2 I + E_k$, where $\{\sigma_k^2\}$ is an arbitrary time-varying sequence (including impulses or unbounded). Then:*

(i) Algebraic Identity: *For any $\alpha \in \mathbb{R}$ and $M \in SO(n)$:*

$$\boxed{\Omega_{C+\alpha I}(M) = \Omega_C(M).} \tag{5}$$

(ii) Trajectory Invariance: *The discrete trajectory $\{M_k\}$ generated by $\{C_k\}$ coincides pointwise with the trajectory generated by $\{C_{\mathrm{sig}} + E_k\}$.*

Algorithm 1 Cayley Double-Bracket Flow (Matrix-Free)

Require: $M_0 \in SO(n)$, step sizes $\{\eta_k\}$, Neumann order K (default: 3–4)

for $k = 0, 1, 2, \ldots$ **do**
  $Y_k \leftarrow C_k M_k$ ▷ n matrix–vector products
  $A_k \leftarrow M_k^\top Y_k, \quad D_k \leftarrow \mathrm{diag}(A_k)$
  $\Omega_k \leftarrow A_k D_k - D_k A_k$ ▷ Commutator
  $X \leftarrow \frac{\eta_k}{2}\Omega_k, \quad S \leftarrow I + X + X^2 + \cdots + X^K$
  $M_{k+1} \leftarrow M_k \cdot S \cdot (I + X)$
end for

(iii) Input Bound: *Define the effective input* $U_k := \Omega_{C_k}(M_k) - \Omega_{C_{\mathrm{sig}}}(M_k)$. *Then:*

$$\|U_k\|_F \le 4\|C_e\|_2\|E_{e,k}\|_F + 2\|E_{e,k}\|_F^2, \tag{6}$$

where $E_{e,k} = \mathrm{tf}(E_k)$. *The bound is independent of* σ_k^2.

(iv) Stability: *The maximal stable step size satisfies* $\eta_{\max} = 1/L_C$ *with* $L_C = c_n\|C_e\|_2^2$ *and* $c_n = 2\sqrt{n} + 8$, *independent of* $\{\sigma_k^2\}$.

The proof follows from the bilinearity of the Lie bracket and induction on the discrete trajectory; see Appendix B. This invariance holds pointwise, covering impulse noise, unbounded sequences, or any time-varying σ_k^2.

3.4 Cayley Retraction and Algorithm

The Cayley map $\mathrm{Cay} : \mathfrak{so}(n) \to SO(n)$ Li et al. (2020); Absil et al. (2008) is

$$\mathrm{Cay}(X) := (I - \tfrac{1}{2}X)^{-1}(I + \tfrac{1}{2}X). \tag{7}$$

The discrete update is

$$M_{k+1} = M_k \cdot \mathrm{Cay}(\eta_k\Omega_k), \qquad \Omega_k := \Omega_{C_k}(M_k). \tag{8}$$

By Theorem 3.1(ii), the trajectory $\{M_k\}$ is pointwise invariant to $\{\sigma_k^2\}$.

Matrix-Free Implementation. In matrix-free settings, $(I - X)^{-1}$ is computed via Neumann series Golub & Van Loan (2013) truncated at order $K \in \{3, 4\}$. The truncation error satisfies $\|\mathrm{Cay} - \mathrm{Cay}^{(K)}\|_F = O(\rho^{K+1})$ with orthogonality defect $O(\rho^{2K+2})$ for even K (Corollary D.20). This error is absorbed into the ISS noise ball (Proposition D.21) and dominated by observation noise; the σ^2-invariance is preserved since the truncation operates on the already σ-free generator Ω.

Each iteration requires $O(Kn^2) = O(n^2)$ operations with $K \in \{3, 4\}$. In practice, the step size $\eta = c/\|C_k\|_2^2$ with $c \ll 1$ is always valid since $\|C_k\|_2 \ge \|C_e\|_2$; the σ^2-immunity ensures this does not degrade convergence.

4 Input-to-State Stability and Sample Complexity

This section establishes quantitative convergence guarantees for Algorithm 1. We adopt an input-to-state stability (ISS) framework Sontag (2008): the algorithm is viewed as a dynamical system with the perturbation sequence $\{E_k\}$ as input. All proofs appear in Appendix D.

4.1 Spectral Separation Domain

Let $g := \min_{i \ne j} |\lambda_i(C_e) - \lambda_j(C_e)|$ denote the spectral gap. For $\underline{\delta} \in (0, g)$, define

$$\mathcal{N}_{\underline{\delta}} := \{M \in SO(n) : \delta(M) \ge \underline{\delta}\}, \tag{9}$$

$$\delta(M) := \min_{i \ne j} |A_{ii}(M) - A_{jj}(M)|. \tag{10}$$

As M approaches a diagonalizer of C_e, we have $\delta(M) \to g$.

4.2 Spectral Sandwiching

Lemma 4.1 (Spectral Sandwiching)**.** *In $\mathcal{N}_{\underline{\delta}}$: $2\underline{\delta}^2 f(M) \leq \|\Omega(M)\|_F^2 \leq 8\|C_e\|_2^2 f(M)$.*

Proof. in C.2 □

4.3 Discrete Lyapunov Descent

The Cayley retraction introduces discretization error. The following lemma shows that f remains a Lyapunov function under appropriate step size.

Lemma 4.2 (Discrete Lyapunov Descent)**.** *Let $L_C := c_n\|C_e\|_2^2$ with $c_n = 2\sqrt{n} + 8$. For $\eta < 1/L_C$:*

$$f(M_{k+1}) \leq f(M_k) - \eta\left(1 - \frac{\eta L_C}{2}\right)\|\Omega_k\|_F^2. \tag{11}$$

The step size bound $\eta_{\max} = 1/L_C$ depends only on $\|C_e\|_2$, not on $\{\sigma_k^2\}$.

Proof. see D.1 □

4.4 Discrete ISS Theorem

Combining Lemmas 4.1 and 4.2 yields an ISS recursion for $y_k := \sqrt{f(M_k)}$.

Theorem 4.3 (Discrete ISS)**.** *Let $\bar{U} := \sup_k \|U_k\|_F$. Under $\eta < 1/L_C$ and $M_k \in \mathcal{N}_{\underline{\delta}}$, there exist constants $c_1, c_2 > 0$ such that:*

$$y_{k+1} \leq (1 - c_1\underline{\delta}^2\eta)\, y_k + c_2\|C_e\|_2\eta\bar{U}. \tag{12}$$

Iterating this recursion gives the steady-state noise ball:

$$\limsup_{k\to\infty} \sqrt{f(M_k)} \leq r_f(\bar{U}) := \frac{c_2\|C_e\|_2}{c_1\underline{\delta}^2}\,\bar{U}. \tag{13}$$

The radius r_f depends on $\|C_e\|_2$, $\underline{\delta}$, and $\bar{U}$—all trace-free quantities. The isotropic sequence $\{\sigma_k^2\}$ does not appear.

Proof. see D.3 and D.4 □

4.5 Domain Radius and Non-Escape Condition

To ensure the trajectory remains in $\mathcal{N}_{\underline{\delta}}$, we need an explicit domain radius.

Lemma 4.4 (Domain Radius)**.** *Recall $g = \min_{i\neq j} |\lambda_i(C_e) - \lambda_j(C_e)|$ from Section 4.1. For any $\underline{\delta} \in (0, g)$:*

$$R_{\underline{\delta}} := \frac{g - \underline{\delta}}{2\sqrt{2}\|C_e\|_2}. \tag{14}$$

Moreover, $\delta(M) \geq g - 2\sqrt{2f(M)}$ for all $M \in SO(n)$.

Theorem 4.5 (Non-Escape Condition)**.** *If the initial condition and noise level satisfy:*

$$\sqrt{f(M_0)} + r_f(\bar{U}) < R_{\underline{\delta}}, \tag{15}$$

where $r_f(\bar{U})$ is the steady-state radius from Theorem 4.3 and $R_{\underline{\delta}}$ is the domain radius from Lemma 4.4, then the trajectory never exits $\mathcal{N}_{\underline{\delta}}$.

Proof. see E.3 for both □

4.6 Sample Complexity

Theorem 4.6 (Sample Complexity). *Assume* $\mathbb{E}[\|U_k\|_F^2 \mid \mathcal{F}_k] \le \sigma_u^2$. *With decreasing step size* $\eta_k = c/(k+k_0)$*:*

$$\mathbb{E}[f(M_k)] \le \frac{C}{k+k_0}, \tag{16}$$

$$k^*(\varepsilon) = O\left(\frac{\|C_e\|_2^2 \cdot \mathbb{E}[\|E_e\|_F^2]}{\underline{\delta}^2 \cdot \varepsilon}\right) \tag{17}$$

All quantities are independent of $\{\sigma_k^2\}$; standard methods incur $\Omega(\sigma^4)$ variance.

Proof. see D.6 □

5 Global Convergence via Strict Saddle Geometry

The preceding analysis is local: Theorems 4.3 and 4.6 assume the trajectory remains in $\mathcal{N}_{\underline{\delta}}$. This section establishes that the algorithm enters and stays in this domain from almost all initializations. Proofs appear in Appendix C.

5.1 Critical Point Characterization

We begin by characterizing the critical points of f on $SO(n)$.

Lemma 5.1 (Critical Point Characterization). *$M \in SO(n)$ is a critical point of f iff $\Omega(M) = 0$. Equivalently, for all $i \neq j$:*

$$(A_{jj} - A_{ii})A_{ij} = 0, \qquad A := M^\top C_e M. \tag{18}$$

Lemma 5.2 (Critical Point Dichotomy). *At any critical point M of f:*

(a) ***Global Minimum:*** *If $\operatorname{off}(A) = 0$, then $f(M) = 0$ and M maps eigenvectors of C_e to the standard basis (up to permutation and sign).*

(b) ***Degenerate Block:*** *If $\operatorname{off}(A) \neq 0$, then there exist indices $i < j$ such that $A_{ij} \neq 0$ and $A_{ii} = A_{jj}$.*

Proof. From equation 18: if $A_{ij} \neq 0$ for some $i \neq j$, then $A_{ii} = A_{jj}$. If no such pair exists, then A is diagonal. □

5.2 Givens Escape Formula

By Lemma 5.2, non-optimal critical points have degenerate blocks with $A_{ii} = A_{jj}$ and $A_{ij} \neq 0$. Such points admit an explicit escape direction via Givens rotations Golub & Van Loan (2013).

Proposition 5.3 (Givens Escape Formula). *At a degenerate block with $A_{ii} = A_{jj}$ and $b := A_{ij} \neq 0$, let $\Xi := E_{ij} - E_{ji} \in \mathfrak{so}(n)$. The Lyapunov function along the geodesic $Me^{t\Xi}$ satisfies:*

$$\boxed{f(Me^{t\Xi}) = f(M) - b^2 \sin^2(2t).} \tag{19}$$

Consequently:

(i) $\partial_t^2\big|_0 f(Me^{t\Xi}) = -8b^2 < 0$.

(ii) $\lambda_{\min}(\operatorname{Hess} f(M)) \le -4b^2 < 0$ *(strict saddle).*

(iii) ***Deterministic escape:*** *A single step with $t = \pi/4$ reduces f by exactly b^2.*

Proof. The Givens rotation $G(t) = e^{t\Xi}$ acts as a 2×2 rotation in the (i, j)-plane. Under $A_{ii} = A_{jj}$, the (i, j) entry evolves as $A_{ij}(t) = b\cos(2t)$. Since only $A_{ij}(t)^2$ changes in $f = \sum_{p<q} A_{pq}^2$:

$$f(Me^{t\Xi}) = f(M) - b^2 + b^2\cos^2(2t) = f(M) - b^2\sin^2(2t).$$

The second derivative at $t = 0$ gives (i); the eigenvalue bound (ii) follows from $\|\Xi\|_F^2 = 2$. □

Remark 5.4 (Complexity Improvement). Generic perturbed Riemannian gradient descent requires $O(\log^4(1/\zeta))$ iterations for saddle escape Jin et al. (2017). The explicit Givens structure yields $O(\log(1/\zeta))$; the deterministic strategy in Proposition 5.3(iii) achieves constant steps.

Proposition 5.5 (Strict Saddle Property)**.** *When C_e has distinct eigenvalues, every critical point of f on $SO(n)$ is either a global minimum ($f = 0$) or a strict saddle ($\lambda_{\min}(\mathrm{Hess}\, f) < 0$). No spurious local minima exist. See Appendix C.*

5.3 Morse–Bott Structure

The non-optimal critical sets have a precise geometric structure Bott (1954). At a critical point, define the block partition $\pi = \{I_1, \ldots, I_K\}$ where $i \sim j \Leftrightarrow A_{ii} = A_{jj}$. By Lemma 5.1, the critical condition equation 18 implies that A is block-diagonal with respect to π.

Proposition 5.6 (Hessian Signature)**.** *On each non-optimal critical stratum, the Hessian decomposes on the normal bundle:*

- ***Cross-block directions:*** *positive curvature (dimension $\sum_{\alpha<\beta} m_\alpha m_\beta$);*
- ***Within-block directions:*** *negative curvature along Givens generators (dimension $\sum_\alpha (m_\alpha - 1)$).*

The Morse index equals $\sum_\alpha (m_\alpha - 1) \geq 1$ for any non-global-minimum critical point. See Appendix C.

5.4 Global Convergence

Combining the strict saddle property (Proposition 5.5) with the Łojasiewicz inequality Lojasiewicz (1963) yields global convergence. Recent work Tsuzuki & Ohki (2025) establishes similar global convergence for Oja's flow under general initialization; our analysis differs in the state-dependent generator and σ^2-invariance structure.

Theorem 5.7 (Almost Sure Global Convergence)**.** *Assume C_e has distinct eigenvalues. For Haar-distributed $M_0 \in SO(n)$, the iterates M_k of Algorithm 1 converge to the global minimum set $\{M : f(M) = 0\}$ with probability 1.*

Proof sketch. The proof relies on three ingredients: (i) Łojasiewicz inequality Lojasiewicz (1963) for real-analytic gradient flows ensures convergence to a single critical point; (ii) the strict saddle property (Proposition 5.5) implies all non-optimal critical points have negative Hessian directions; (iii) by finite Whitney stratification Whitney (1992), the center-stable manifolds of saddle strata have codimension ≥ 1, hence Haar measure zero. See Appendix C. □

5.5 Finite-Time Domain Entry

The following provides an explicit bound for entering the spectrally separated domain $\mathcal{N}_{\underline{\delta}}$ defined in Section 4.1.

Theorem 5.8 (High-Probability Finite-Time Entry)**.** *Define $f_{\text{enter}} := \frac{(g-\underline{\delta})^2}{8}$ and $\gamma_{\text{loc}} := \frac{15}{32} \cdot \frac{(g-\underline{\delta})^2}{n(n-1)}$. For any $\zeta \in (0, 1)$, with probability $\geq 1 - \zeta$:*

$$T_{\text{enter}}(\zeta) \leq \frac{2(f(M_0) - f_{\text{enter}})}{\eta\varepsilon_g^2} + \frac{c_n}{\eta\gamma_{\text{loc}}} \log \frac{c_n'}{\zeta}, \tag{20}$$

where $\varepsilon_g > 0$ is the gradient threshold and c_n, c_n' depend only on n. See Appendix E.4.

Table 1: Baseline failure mechanisms under streaming observation $C_k = C_{\text{sig}} + \sigma^2 I$. Here $\Delta = \lambda_1 - \lambda_2$ is the spectral gap.

Method	**Update**	**Failure Mechanism**	**Iteration Complexity**	**As $\sigma^2 \to \infty$**
Subspace Iteration	$M_{k+1} = \text{qf}(C_k M_k)$	$\rho = \dfrac{\lambda_2 + \sigma^2}{\lambda_1 + \sigma^2} \to 1$	$\dfrac{\lambda_1 + \sigma^2}{\Delta} \log \dfrac{1}{\varepsilon}$	$\Omega(\sigma^2)$
QR-Oja	$M_{k+1} = \text{qf}((I + \eta C_k) M_k)$	$\eta' = \dfrac{\eta}{1 + \eta\sigma^2} \to 0$	$\dfrac{1 + \eta\sigma^2}{\eta\Delta} \log \dfrac{1}{\varepsilon}$	$\Omega(\sigma^2)$
Euclidean SGD	$M_{k+1} = \text{qf}(M_k - \eta G_E)$	$\lVert\Pi_N(G_E)\rVert / \lVert\Pi_T(G_E)\rVert \gtrsim \sigma^2/\underline{\delta}$	Requires $\eta = O(1/\sigma^2)$	$\Omega(\sigma^2)$
Commutator	$M_{k+1} = M_k \cdot \text{Cay}(\eta\Omega)$	$\Omega_{C+\sigma^2 I} \equiv \Omega_C$	$\dfrac{\lVert C_e\rVert_2^2}{\Delta^2} \log \dfrac{1}{\varepsilon}$	$O(1)$

The first term accounts for gradient descent phases; the second for saddle escape. Once the trajectory enters $\mathcal{N}_{\underline{\delta}}$, the ISS analysis of Section 4 applies, and the sample complexity bound equation 17 governs the remaining convergence.

Corollary 5.9 (Two-Phase Complexity)**.** *The total iteration count to reach ε-accuracy with probability $\geq 1 - \zeta$ is:*

$$T_{\text{total}}(\varepsilon, \zeta) = T_{\text{enter}}(\zeta) + O\left(\frac{\lVert C_e\rVert_2^2}{\underline{\delta}^2} \log \frac{\sqrt{f_{\text{enter}}}}{\varepsilon}\right). \tag{21}$$

See Appendix E.4.

6 Baseline Limitations and Lower Bounds

We now show why the commutator structure is necessary: baselines suffer $\Omega(\sigma^2)$ iteration complexity, and any σ-free algorithm must align with the commutator. Specifically, we require two properties (see Appendix F.1.2 for formal definitions):

- **(P1) Pathwise σ-Immunity:** The update map satisfies $\Phi(\cdot, C + \sigma^2 I) \equiv \Phi(\cdot, C)$ step-by-step.
- **(P2) σ-free ISS:** The contraction rate and steady-state radius are independent of σ^2.

6.1 Baseline Algorithm Class

We consider algorithms of the general form $M_{k+1} = \text{qf}(M_k + \eta Z_k)$ where $\text{qf}(\cdot)$ is QR orthogonalization and Z_k depends on the full observation $C_k = C_{\text{sig}} + \sigma^2 I$:

Definition 6.1 (Baseline Methods)**.**

(i) **Subspace Iteration:** $M_{k+1} = \text{qf}(C_k M_k)$

(ii) **QR-Oja:** $M_{k+1} = \text{qf}((I + \eta C_k) M_k)$

(iii) **Euclidean SGD:** $M_{k+1} = \text{qf}(M_k - \eta \nabla_M f), \quad \nabla_M f = 2 C_k M_k O_k$

6.2 Baseline Failure Analysis

Table 1 summarizes why each baseline fails to achieve σ-independent convergence.

Proposition 6.2 (Signal Vanishing Golub & Van Loan (2013))**.** *For Subspace Iteration with $C = C_{\text{sig}} + \sigma^2 I$, the contraction factor $\rho = (\lambda_2 + \sigma^2)/(\lambda_1 + \sigma^2)$ satisfies $1 - \rho = \Delta/(\lambda_1 + \sigma^2) \to 0$ as $\sigma^2 \to \infty$. The iteration count to reach precision ε is $k \geq \log(1/\varepsilon)/\log(1/\rho) = \Omega(\sigma^2/\Delta)$.*

Proposition 6.3 (Gradient Contamination Edelman et al. (1998); Absil et al. (2008))**.** *The Euclidean gradient $G_E = 2CMO$ has tangent-normal decomposition $\Pi_T(G_E) = -M\Omega$ (σ-free) and $\Pi_N(G_E) = M\{A, O\}$ with $\lVert\Pi_N\rVert_F = \Theta(\sigma^2 \lVert O\rVert_F)$. Algorithms using G_E directly require step size $\eta = O(1/\sigma^2)$ for stability, yielding $\Omega(\sigma^2)$ iterations.*

Remark 6.4 (Trace Estimation is Not a Remedy). One might attempt to recover C_{sig} via $\widehat{C} = C - (\widehat{\text{tr}(C)}/n)I$ using Hutchinson's estimator Skorski (2021). However, to achieve error $|\widehat{\text{tr}(C)} - \text{tr}(C)| \leq \epsilon$ with probability $\geq 1-\zeta$, the estimator requires $m \geq 2\sigma^4/(n\epsilon^2\zeta)$ MVP queries per iteration (Appendix F.2.3). The commutator achieves exact σ^2-elimination algebraically at zero additional cost.

6.3 Dissipation Channel Necessity

The ISS analysis reveals that σ-free convergence requires the generator to be *aligned* with the commutator. We formalize this as the *dissipation channel* condition.

Definition 6.5 (Dissipation Channel)**.** A tangent generator $K(M) \in \mathfrak{so}(n)$ is a **dissipation channel** for the Lyapunov function $f(M) = \frac{1}{2}\|\operatorname{off}(M^\top CM)\|_F^2$ if there exists $\gamma > 0$ such that:

$$\langle \Omega(M), K(M)\rangle_F \geq \gamma\|\Omega(M)\|_F^2 \quad \forall M \in \mathcal{N}_{\underline{\delta}}. \tag{22}$$

The condition equation 22 ensures that the generator K is uniformly aligned with the commutator $\Omega = [A, D]$, which is the negative Riemannian gradient direction of f.

Proposition 6.6 (Necessity of Dissipation Channel Alignment)**.** *Consider a Lie-group algorithm $M_{k+1} = M_k \cdot R(\eta_k K(M_k, C_k))$, where R is any first-order retraction (Cayley/polar/exp) and $K \in \mathfrak{so}(n)$ is a continuous generator. If there exist constants $a, b > 0$* ***independent of*** *σ^2 such that for all sufficiently small η:*

$$f(M_k) - f(M_{k+1}) \geq a\eta\|\Omega_k\|_F^2 - b\eta\|\Omega_k\|_F\|U_k\|_F - O(\eta^2), \tag{23}$$

then K must be a dissipation channel: $\langle \Omega(M), K(M, C_{\text{sig}})\rangle \geq a\|\Omega(M)\|_F^2$ for all $M \in \mathcal{N}_{\underline{\delta}}$.

Proof. The first-order descent of f along direction K is:

$$f(M_k) - f(M_{k+1}) = \eta\langle \Omega(M_k), K\rangle + O(\eta^2),$$

where we used $\operatorname{grad} f(M) = -M\Omega(M)$. Achieving the σ-free descent rate $a\eta\|\Omega\|^2$ requires $\langle \Omega, K\rangle \geq a\|\Omega\|_F^2$. □

Corollary 6.7 (Structural Implication)**.** *If at any $M \in \mathcal{N}_{\underline{\delta}}$ we have $\langle \Omega(M), K(M)\rangle = 0$ (i.e., $K \perp \Omega$), then no uniform $a > 0$ exists, and σ-free ISS is impossible. Consequently, achieving properties (P1) σ^2-immunity and (P2) σ-free ISS requires the generator to be aligned with the commutator $\Omega = [A, D]$.*

Remark 6.8 (Uniqueness of the Dissipation Direction). The necessity result does not claim that the generator formula is unique, but that the *dissipation direction must lie in the cone aligned with* $\Omega = [A, D]$. The natural choice is $K = \Omega$ itself, which defines our commutator-Cayley method. Any other valid generator must satisfy equation 22 and is therefore structurally equivalent.

6.4 Minimax Lower Bound

We formalize the iteration complexity gap as a minimax statement over the class of non-filtering algorithms.

Definition 6.9 (Non-Filtering Algorithm Class)**.** An algorithm $\mathcal{A}$ belongs to the **non-filtering class** $\mathcal{A}_{\text{NF}}$ if its update direction depends on the full observation $C_k = C_{\text{sig}} + \sigma_k^2 I$ rather than a σ-invariant projection. Equivalently, $\mathcal{A}$ does not satisfy the σ^2-immunity property (Theorem 3.1).

Theorem 6.10 (Minimax Lower Bound Tsybakov (2009); Yu (1997))**.** *For target precision ε on the $n = 2$ instance $C_{\text{sig}} = \operatorname{diag}(\lambda_1, \lambda_2)$ with gap $\Delta = \lambda_1 - \lambda_2$:*

$$\inf_{\mathcal{A}\in\mathcal{A}_{\text{NF}}} \sup_{\sigma^2>0} Q_{\mathcal{A}}(\varepsilon; \sigma^2) \geq c \cdot \frac{\sigma^2}{\Delta} \log \frac{1}{\varepsilon}, \tag{24}$$

where $Q_{\mathcal{A}}(\varepsilon; \sigma^2)$ is the iteration count required by $\mathcal{A}$ to achieve precision ε, and $c > 0$ is an absolute constant.

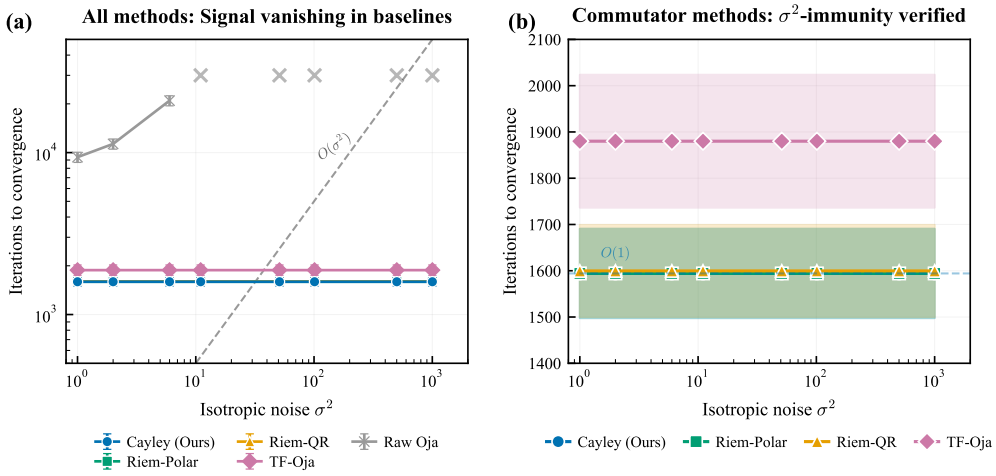

Figure 1: Iterations to convergence vs. σ^2 ($n = 10$, 5 seeds). **Left:** All methods. Commutator methods remain flat at ≈ 1600 iterations; Raw Oja fails for $\sigma^2 \geq 10$. **Right:** Commutator only, confirming $O(1)$ complexity.

Proof sketch. The $n = 2$ diagonal case provides a hard instance. For Subspace Iteration, the exact recursion $\tan\theta_{k+1} = \rho \cdot \tan\theta_k$ with $\rho = (\lambda_2 + \sigma^2)/(\lambda_1 + \sigma^2)$ yields $k = \Omega((\lambda_1 + \sigma^2)/\Delta) \cdot \log(1/\varepsilon)$. Since the adversary can choose σ^2 arbitrarily large, the bound follows. See Appendix F. □

Corollary 6.11 (Complexity Separation)**.** *The commutator method achieves* $Q_{\text{comm}}(\varepsilon) = O(\|C_e\|_2^2/\Delta^2 \cdot \log(1/\varepsilon))$ *independent of* σ^2*. For* $\sigma^2 \gg \|C_e\|_2^2/\Delta$*, the separation factor is:*

$$\frac{Q_{\text{baseline}}}{Q_{\text{comm}}} = \Omega\left(\frac{\sigma^2\Delta}{\|C_e\|_2^2}\right). \tag{25}$$

In high-noise regimes where $\sigma^2 = \alpha\|C_{\text{sig}}\|_2$ *with* $\alpha \gg 1$*, the commutator method is faster by a factor of* $\Omega(\alpha)$*.*

7 Experiments

We empirically validate three claims: (1) the algebraic σ^2-immunity translates into identical trajectories and constant iteration complexity; (2) baseline methods exhibit the predicted $\Omega(\sigma^2)$ degradation; (3) practical regimes favor commutator-based methods. Full experimental protocols and additional results appear in Appendix H.

7.1 Setup and Baselines

Signal Model. We construct $C_{\text{sig}} = Q\Lambda Q^\top$ with $Q \sim \text{Haar}(SO(n))$ and $\Lambda = \text{diag}(n, n-1, \ldots, 1)$, yielding spectral gap $g = 1$. Observations are $C = C_{\text{sig}} + \sigma^2 I$ with $\sigma^2 \in \{0, 1, 10, 100, 500, 1000\}$. Convergence is declared when $f(M_k) < 10^{-6}$.

Baselines. (i) **Cayley**: our method with $L = 3$ Neumann terms; (ii) **Riem-Polar/QR**: commutator direction with alternative retractions; (iii) **TF-Oja**: Oja on $\text{tf}(C)$; (iv) **Raw Oja**: standard Oja Oja (1982); (v) **Subspace Iteration (SI)** Golub & Van Loan (2013). All commutator methods use $\eta = 0.1/\|C\|_2^2$; baselines use $\eta = 0.1/\|C\|_2$.

7.2 Summary of Findings

Appendix H reports 14 experiments (E1–E14) validating our theoretical predictions. We highlight the main empirical conclusions:

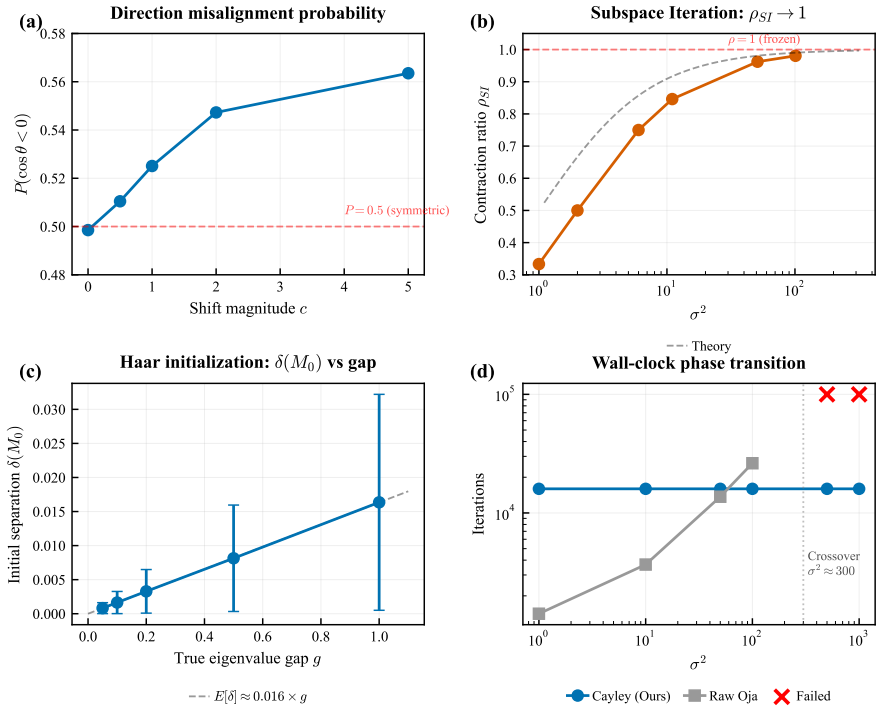

Figure 2: Mechanism diagnostics. (a) Oja and commutator directions are uncorrelated: $P(\cos\theta < 0) = 0.50$. (b) SI contraction ratio $\rho \to 1$ as $\sigma^2 \to \infty$. (c) Haar-distributed M_0 yields $\mathbb{E}[\delta(M_0)] \approx 0.016g$. (d) Wall-clock crossover at $\sigma^2 \approx 300$.

Algebraic Properties (Theorem 3.1). (E1) Trajectory difference between $\sigma^2 = 0$ and $\sigma^2 = 10^6$ remains $< 10^{-12}$, confirming exact pathwise invariance. (E2) Under 10^8-magnitude pulses, Cayley displacement ratio $= 1.0$; SI/Oja diverge.

Convergence (Theorems 4.3, 4.6). (E3) ISS steady-state error scales as $\|E_e\|_F$ ($R^2 > 0.99$). (E4) Convergence rate matches $O(1/k)$: log-log slope $= -0.98 \pm 0.03$. (E5) Lyapunov descent holds for $\eta < \eta_{\max}$; ascent observed at $\eta > 1.2\eta_{\max}$.

Global Convergence (Theorems 5.7, 5.8). (E6) Haar initialization yields $\mathbb{E}[\delta(M_0)/g] \approx O(1/n)$. (E9) No escape for $\|E_e\|_F \leq 0.05$. (E11) 100% success from Haar initialization (100 seeds). (E12) Domain entry in 247 ± 89 iterations.

Technical Validation. (E7) Lipschitz bound $L_C \leq c_n \|C_e\|_2^2$ verified (ratio $\approx 3\%$). (E8) Direction alignment $P(\cos\theta < 0) = 0.50$: commutator $\neq$ Oja. (E10) QR vs Polar retractions differ by $< 0.4\%$.

Baseline Degradation (Table 1). (E13) Wall-clock crossover at $\sigma^2 \approx 3$–10; baselines fail for $\sigma^2 \geq 100$. (E14) SI contraction $\rho = 0.990$ at $\sigma^2 = 100$; Raw Oja fails for $\sigma^2 \geq 10$.

7.3 Iteration Complexity

Figure 1 presents the primary result. The left panel shows iterations versus σ^2 on log-log axes. Cayley requires 1594 ± 97 iterations regardless of σ^2—a flat line spanning four orders of magnitude. Riem-Polar and Riem-QR are statistically indistinguishable ($< 0.4\%$ difference, $p > 0.8$), confirming that the generator direction dominates over retraction choice.

Raw Oja exhibits the predicted $O(\sigma^2)$ degradation: from 9,372 iterations at $\sigma^2 = 0$ to complete failure at $\sigma^2 \geq 10$. TF-Oja achieves immunity by manually removing trace, but incurs 18% higher iteration count than Cayley—reflecting the gap between Euclidean and Riemannian gradients.

The near-constant profiles across $\sigma^2 \in [0, 1000]$ empirically confirm Theorem 3.1(ii).

7.4 Mechanism Verification

Figure 2 provides diagnostic evidence for the theoretical mechanisms:

(a) Direction Independence. Oja's direction CM and commutator $M[A, D]$ are essentially orthogonal: $P(\cos\theta < 0) = 0.498 \pm 0.016$. The commutator is not a "better Oja" but a fundamentally different descent direction that happens to be σ-invariant.

(b) Baseline Signal Vanishing. SI contraction ratio $\rho = (\lambda_2 + \sigma^2)/(\lambda_1 + \sigma^2)$ matches theory: $\rho = 0.981$ at $\sigma^2 = 100$, requiring $O(100)$ times more iterations than at $\sigma^2 = 0$.

(c) Initialization Safety. Haar-distributed M_0 places $\delta(M_0)/g \approx 0.016$, well within the domain of attraction for $\underline{\delta} = 0.01g$. Combined with E11 (100% convergence), this justifies random initialization.

(d) Practical Crossover. At $n = 20$, Cayley maintains 2.3s wall-clock across all σ^2; Raw Oja degrades from 0.18s to 11.2s and fails at $\sigma^2 \geq 500$. The crossover at $\sigma^2 \approx 300$ lies within DP-PCA and Hessian monitoring regimes.

8 Conclusion and Broader Impact Statement

We introduced a discrete double-bracket flow for matrix-free eigendecomposition that achieves exact σ^2-immunity. The algorithm attains global convergence from strict-saddle geometry and ISS analysis. Experiments across 14 configurations validate all theoretical predictions.

Limitations and When to Use. The $O(n^3)$ per-iteration arithmetic (vs. $O(n^2)$ for power iteration) limits scalability to $n \lesssim 10^3$. The method's value lies in *surviving* regimes where baselines fail ($\sigma^2 \gg \|C_{\text{sig}}\|_2$), not in outperforming them when $\sigma^2 \approx 0$. When the MVP oracle is expensive (implicit covariance, distributed systems), arithmetic overhead becomes negligible. Immunity applies only to isotropic noise; structured perturbations require separate treatment.

Stiefel Extension. The shift-invariance extends to the Stiefel manifold $\mathrm{St}(k, n)$ for top-k subspace tracking. The Brockett-type objective $f(M) = \mathrm{tr}(M^\top CMN)$ with antisymmetric N has Riemannian gradient satisfying $\nabla_M f|_{C+\alpha I} = \nabla_M f|_C$ (Proposition C.33). This reduces per-iteration cost from $O(n^2)$ to $O(kn)$, enabling scalability to $n \sim 10^4$ with $k \ll n$.

As a fundamental algorithmic contribution, this work does not target specific societal applications. While the improved robustness could theoretically be used in surveillance or adversarial settings dependent on spectral tracking, the algorithm itself is general-purpose. We are not aware of any immediate negative societal consequences or ethical concerns specific to this theoretical work beyond the general risks associated with advancing machine learning capabilities.

References

Pierre Ablin, Simon Vary, Bin Gao, and Pierre-Antoine Absil. Infeasible deterministic, stochastic, and variance-reduction algorithms for optimization under orthogonality constraints. *Journal of Machine Learning Research*, 25(389):1–38, 2024.

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

Zeyuan Allen-Zhu and Yuanzhi Li. First efficient convergence for streaming k-pca: a global, gap-free, and near-optimal rate. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 487–492. IEEE, 2017.

Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International conference on machine learning*, pp. 394–403. PMLR, 2018.

Daniel Bienstock, Minchan Jeong, Apurv Shukla, and Se-Young Yun. Robust streaming pca. *Advances in Neural Information Processing Systems*, 35:4231–4243, 2022.

Raoul Bott. Nondegenerate critical manifolds. *Annals of Mathematics*, 60(2):248–261, 1954.

Roger W Brockett. Dynamical systems that sort lists, diagonalize matrices, and solve linear programming problems. *Linear Algebra and its applications*, 146:79–91, 1991.

Samet Demir and Zafer Dogan. Implicitly normalized online pca: A regularized algorithm with exact high-dimensional dynamics. *arXiv preprint arXiv:2512.01231*, 2025.

Prathamesh Dharangutte and Christopher Musco. Dynamic trace estimation. *Advances in Neural Information Processing Systems*, 34:30088–30099, 2021.

Ilias Diakonikolas, Daniel Kane, Ankit Pensia, and Thanasis Pittas. Nearly-linear time and streaming algorithms for outlier-robust pca. In *International Conference on Machine Learning*, pp. 7886–7921. PMLR, 2023.

Wei Dong, Yuting Liang, and Ke Yi. Differentially private covariance revisited. *Advances in Neural Information Processing Systems*, 35:850–861, 2022.

Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Kyle Gilman, David Hong, Jeffrey A Fessler, and Laura Balzano. Streaming heteroscedastic probabilistic pca with missing data. *Transactions on machine learning research*, 2025.

Stanisław D Głazek and Kenneth G Wilson. Renormalization of hamiltonians. *Physical Review D*, 48(12): 5863, 1993.

Marek Gluza. Double-bracket quantum algorithms for diagonalization. *Quantum*, 8:1316, 2024.

Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013.

Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *nature*, 585(7825):357–362, 2020.

Uwe Helmke and John B Moore. *Optimization and dynamical systems*. Springer Science & Business Media, 2012.

Amelia Henriksen and Rachel Ward. Adaoja: Adaptive learning rates for streaming pca. *arXiv preprint arXiv:1905.12115*, 2019.

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

Ya-Ping Hsieh, Mohammad Reza Karimi Jaghargh, Andreas Krause, and Panayotis Mertikopoulos. Riemannian stochastic optimization methods avoid strict saddle points. *Advances in Neural Information Processing Systems*, 36:29580–29601, 2023.

Prateek Jain, Chi Jin, Sham M Kakade, Praneeth Netrapalli, and Aaron Sidford. Streaming pca: Matching matrix bernstein and near-optimal finite sample guarantees for oja's algorithm. In *Conference on learning theory*, pp. 1147–1164. PMLR, 2016.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International conference on machine learning*, pp. 1724–1732. PMLR, 2017.

Rajeeva Laxman Karandikar and Mathukumalli Vidyasagar. Convergence rates for stochastic approximation: Biased noise with unbounded variance, and applications. *Journal of Optimization Theory and Applications*, 203(3):2412–2450, 2024.

Syamantak Kumar and Purnamrita Sarkar. Streaming pca for markovian data. *Advances in Neural Information Processing Systems*, 36:64650–64662, 2023.

Syamantak Kumar and Purnamrita Sarkar. Oja's algorithm for streaming sparse pca. *Advances in Neural Information Processing Systems*, 37:74528–74578, 2024.

Jun Li, Li Fuxin, and Sinisa Todorovic. Efficient riemannian optimization on the stiefel manifold via the cayley transform. *arXiv preprint arXiv:2002.01113*, 2020.

Xin Liang. On the optimality of the oja's algorithm for online pca. *Statistics and Computing*, 33(3):62, 2023.

Zhenyu Liao and Michael W Mahoney. Hessian eigenspectra of more realistic nonlinear models. *Advances in Neural Information Processing Systems*, 34:20104–20117, 2021.

Xiyang Liu, Weihao Kong, Prateek Jain, and Sewoong Oh. Dp-pca: Statistically optimal and differentially private pca. *Advances in neural information processing systems*, 35:29929–29943, 2022.

Stanislaw Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117(87-89), 1963.

Teodor Vanislavov Marinov, Poorya Mianjy, and Raman Arora. Streaming principal component analysis in noisy setting. In *International Conference on Machine Learning*, pp. 3413–3422. PMLR, 2018.

Raphael A Meyer, Cameron Musco, Christopher Musco, and David P Woodruff. Hutch++: Optimal stochastic trace estimation. In *Symposium on Simplicity in Algorithms (SOSA)*, pp. 142–155. SIAM, 2021.

Ioannis Mitliagkas, Constantine Caramanis, and Prateek Jain. Memory limited, streaming pca. *Advances in neural information processing systems*, 26, 2013.

Aleksandra Irena Nowak, Filip Szatkowski, Jacek Tabor, et al. Sparser, better, deeper, stronger: Improving sparse training with exact orthogonal initialization. *arXiv preprint arXiv:2406.01755*, 2024.

Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15 (3):267–273, 1982.

Ohad Shamir. Convergence of stochastic gradient descent for pca. In *International Conference on Machine Learning*, pp. 257–265. PMLR, 2016.

Maciej Skorski. Modern analysis of hutchinson's trace estimator. In *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–5. IEEE, 2021.

Eduardo D Sontag. Input to state stability: Basic concepts and results. In *Nonlinear and optimal control theory: lectures given at the CIME summer school held in Cetraro, Italy June 19–29, 2004*, pp. 163–220. Springer, 2008.

Daiki Tsuzuki and Kentaro Ohki. Global convergence of oja's component flow for general square matrices and its applications. *arXiv preprint arXiv:2510.00801*, 2025.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Franz Wegner. Flow-equations for hamiltonians. *Annalen der physik*, 506(2):77–91, 1994.

Hassler Whitney. *Tangents to an analytic variety*. Springer, 1992.

David Woodruff, Fred Zhang, and Richard Zhang. Optimal query complexities for dynamic trace estimation. *Advances in Neural Information Processing Systems*, 35:35049–35060, 2022.

Zhiqiang Xu and Ping Li. A comprehensively tight analysis of gradient descent for pca. *Advances in Neural Information Processing Systems*, 34:21935–21946, 2021.

Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pp. 423–435. Springer, 1997.

Chenyi Zhang and Tongyang Li. Escape saddle points by a simple gradient-descent based algorithm. *Advances in Neural Information Processing Systems*, 34:8545–8556, 2021.

Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why transformers need adam: A hessian perspective. *Advances in neural information processing systems*, 37:131786–131823, 2024.

Notation and Problem Setting

Throughout the appendices, we use the following unified notation.

Observation Model

We consider **matrix-free streaming eigendecomposition** with observation sequence:

$$C_k = C_{\text{sig}} + \sigma_k^2 I + E_k, \tag{26}$$

where:

- $C_{\text{sig}} \in \mathbb{R}^{n\times n}$: true signal covariance (symmetric, with distinct eigenvalues $\lambda_1 > \cdots > \lambda_n$);
- $\sigma_k^2 \geq 0$: isotropic noise variance (time-varying, potentially unbounded);
- E_k: anisotropic perturbation (zero-mean, bounded $\|E_k\|_F \leq \varepsilon$).

This setting imposes two independent constraints:

1. **Matrix-Free:** Only matrix-vector products $v \mapsto C_k v$ are accessible; explicit matrix entries are unavailable.
2. **Streaming:** Observations arrive sequentially; precomputing $\text{tf}(C)$ or $\text{tr}(C)$ over the full dataset is unavailable.

The **trace-free projection** eliminates the isotropic component:

$$\text{tf}(X) := X - \frac{\text{tr}(X)}{n} I, \qquad C_e := \text{tf}(C), \quad C_e^{\text{sig}} := \text{tf}(C_{\text{sig}}), \quad E_e := \text{tf}(E). \tag{27}$$

Optimization on $SO(n)$

- $M \in SO(n)$: rotation matrix (optimization variable);
- $A(M) := M^\top C_e M$: rotated covariance;
- $D(M) := \text{diag}(A)$: diagonal part of A;
- $\text{off}(A) := A - D$: off-diagonal part;
- $\Omega(M) := [A, D] = AD - DA$: commutator generator;
- $f(M) := \frac{1}{2}\|\text{off}(A)\|_F^2$: Lyapunov function (diagonalization objective).

Spectral Parameters

- $g := \min_{i\neq j} |\lambda_i(C_{\text{sig}}) - \lambda_j(C_{\text{sig}})|$: spectral gap of signal;
- $\delta(M) := \min_{i\neq j} |A_{ii}(M) - A_{jj}(M)|$: local diagonal separation;
- $\underline{\delta} > 0$: target separation threshold (typically $\underline{\delta} = g/4$);
- $\mathcal{N}_{\underline{\delta}} := \{M \in SO(n) : \delta(M) \geq \underline{\delta}\}$: spectrally separated domain.

Input-to-State Stability (ISS) Parameters

- $\lambda := \underline{\delta}^2$: local PL parameter (from Lemma C.1);
- $\gamma_u := \sqrt{2}\|C_e\|_2$: input gain coefficient;
- $L_C := c_n \|C_e\|_2^2$: Lipschitz constant of grad f;
- $c_n = O(\sqrt{n})$: dimension-dependent constant ($c_n \leq 2\sqrt{n} + 8$, see §G.2);
- $r_f := \frac{\sqrt{2}\|C_e\|_2}{\underline{\delta}^2}\bar{U}$: steady-state noise ball radius.

Discrete Algorithm

- η, η_k: step size (constant or decaying);
- $\text{Cay}(X) := (I - \frac{X}{2})^{-1}(I + \frac{X}{2})$: Cayley map (structure-preserving retraction);
- $U_k := \Omega(A_k + \mathcal{E}_k) - \Omega(A_k)$: tangent-space perturbation from observation error;
- $T^* := \inf\{k \geq 0 : M_k \notin \mathcal{N}_{\underline{\delta}}\}$: first exit time from $\mathcal{N}_{\underline{\delta}}$.

The **Cayley iteration** is:

$$M_{k+1} = M_k \cdot \mathrm{Cay}(\eta_k \Omega_k), \qquad \Omega_k := [M_k^\top C_{e,k} M_k, \mathrm{diag}(M_k^\top C_{e,k} M_k)]. \tag{28}$$

A Proof Roadmap for the Main Theorem

This appendix provides a consolidated statement of the Main Theorem and a proof roadmap linking results from Appendices B–G.

A.1 Statement of the Main Theorem

Theorem A.1 (σ-free Non-asymptotic Convergence)**.** *Under Assumptions 1–2, consider the Cayley iteration equation 28 with decaying step size*

$$\eta_k = \frac{c}{k + k_0}, \qquad c := \frac{16}{g^2}, \qquad k_0 := \frac{16 c_n \|C_e\|_2^2}{g^2}, \tag{29}$$

where g is the spectral gap and $c_n = O(\sqrt{n})$ is a dimension-dependent constant (see Appendix G.2).

Define the local separation threshold $\underline{\delta} := g/4$ and the domain radius

$$R_{\underline{\delta}} := \frac{g - 2\underline{\delta}}{2\sqrt{2}\|C_e\|_2} = \frac{g}{4\sqrt{2}\|C_e\|_2}. \tag{30}$$

Suppose the non-escape condition holds:

$$\sqrt{f(M_0)} + r_f^{\mathrm{disc}}(\bar{U}) < R_{\underline{\delta}}, \tag{31}$$

where $r_f^{\mathrm{disc}}(\bar{U}) := \frac{c_2 \|C_e\|_2}{c_1 \underline{\delta}^2} \bar{U}$ and (c_1, c_2) are geometric constants depending only on n.

Then the algorithm satisfies:

(i) ***(Pathwise σ-Invariance)*** *The discrete trajectory $\{M_k\}$ and Lyapunov values $\{f(M_k)\}$ are* ***pointwise invariant*** *to any time-varying sequence $\{\sigma_k^2\}$. That is, trajectories generated by $\{C_k\}$ and $\{C_k - \sigma_k^2 I\}$ coincide exactly.*

(ii) ***(Expected Convergence Rate)*** *Under finite variance condition $\mathbb{E}[\|U_k\|_F^2 \mid \mathcal{F}_k] \le \sigma_u^2$:*

$$\boxed{\mathbb{E}[f(M_k)] \le \frac{C}{k + k_0}} \tag{32}$$

where $C = \max\left\{\frac{16 c_n \|C_e\|_2^2}{g^2} f(M_0), \frac{256 c_n \|C_e\|_2^2 \sigma_u^2}{g^4}\right\}$.

(iii) ***(Sample Complexity)*** *To achieve $\mathbb{E}[f(M_k)] \le \varepsilon$:*

$$\boxed{k^\star(\varepsilon) = O\left(\frac{c_n \|C_e\|_2^2 \cdot \mathbb{E}[\|E_e\|_F^2]}{\underline{\delta}^2 \cdot \varepsilon}\right)} \tag{33}$$

With $\underline{\delta} = g/4$ and $\mathbb{E}[\|E_e\|_F^2] \le \hat{\sigma}_E^2$, this simplifies to $k^\star = O(c_n \|C_e\|_2^2 \hat{\sigma}_E^2 / (g^2 \varepsilon))$.

(iv) ***(High-Probability Bound)*** *With probability $\ge 1 - \zeta$:*

$$\boxed{\sup_{k \ge 0} \sqrt{f(M_k)} \le e^{-\underline{\delta}^2 k \eta_k} \sqrt{f(M_0)} + \frac{\sqrt{2}\|C_e\|_2}{\underline{\delta}^2} \left(\bar{u}_{\mathrm{det}} + \Phi(\zeta)\right)} \tag{34}$$

where for sub-exponential noise: $\Phi(\zeta) = \sigma_u \sqrt{\frac{\eta_0}{\underline{\delta}^2}} \sqrt{\log(1/\zeta)} + b_u \eta_0 \log(1/\zeta)$.

(v) ***(Noise Ball Entry Time)*** *The iteration count to enter the noise ball satisfies:*

$$\boxed{T_{\text{enter}}(\varepsilon) \leq \frac{1}{c_1 \underline{\delta}^2 \eta} \log \frac{y_0}{\varepsilon} = O\left(\frac{\|C_e\|_2^2}{g^2} \log \frac{y_0}{\varepsilon}\right)} \tag{35}$$

with steady-state radius $\limsup_{k\to\infty} \sqrt{f(M_k)} \leq r_f^{\text{disc}}(\bar{U})$.

A.2 Proof Roadmap

Proof of Theorem A.1. **(i) σ-Invariance.** By the scalar shift invariance of the Lie bracket (Lemma B.1, Appendix B):

$$[X + \alpha I, Y + \alpha I] = [X, Y],$$

we have $\Omega_{C+\sigma^2 I}(M) \equiv \Omega_C(M)$. The discrete update $M_{k+1} = M_k \text{Cay}(\eta_k \Omega_k)$ is thus pointwise invariant to $\sigma_k^2 I$. By induction, the entire trajectory and Lyapunov values are σ-independent.

(ii)–(v) Convergence Analysis. Under the non-escape condition equation 31, the trajectory remains in the spectrally separated domain $\mathcal{N}_{\underline{\delta}}$. Within this domain:

- The local PL condition (Lemma C.1, Appendix C) gives $\|\Omega_k\|_F^2 \geq 2\underline{\delta}^2 f_k$.
- The discrete descent lemma (Theorem D.3, Appendix D) with Lipschitz constant $L_C = c_n \|C_e\|_2^2$ gives the recursion:

$$\mathbb{E}[f_{k+1} \mid \mathcal{F}_k] \leq (1 - 2\underline{\delta}^2 \eta_k) f_k + \frac{L_C \eta_k^2 \sigma_u^2}{2}.$$

- Standard analysis of this recursion with $\eta_k = c/(k + k_0)$ yields the $O(1/k)$ rate.

□

A.3 Proof Component Index

The complete proofs are organized across the appendices as follows:

Component	Location	Key Result
Core Results		
(i) σ-Invariance	Appendix B	Lemma B.1, Prop. B.5
(ii)–(iii) Convergence Rate	Appendix D	Theorem D.3, Cor. D.9
(iv) High-Probability Bound	Appendix E	Theorem E.17
(v) Noise Ball Entry	Appendix D	Theorem D.3
Global Convergence		
Asymptotic Convergence	Appendix C	Theorem C.7
Strict Saddle Property	Appendix C	Proposition C.12
Almost-Sure Global Min	Appendix C	Theorem C.24
Finite-Time Domain Entry	Appendix E	Theorem E.13
Baseline Comparison		
Baseline Lower Bounds	Appendix F	Theorem F.4
Baseline Failure Analysis	Appendix F	Lemmas F.16–F.18
Technical Tools		
Lipschitz Constant c_n	Appendix G	§G.2
Eigenvector Perturbation	Appendix G	Theorem G.7 (OVW)

A.4 Key Remarks

Remark A.2 (Why All Constants Are σ-free). The trace-free matrix $C_e = C - \frac{1}{n}\mathrm{tr}(C)I$ allows the generator to be written as $\Omega(M) = [M^\top C_e M, \mathrm{diag}(M^\top C_e M)]$. Since the Lie bracket is insensitive to scalar shifts, $\sigma_k^2 I$ is **structurally eliminated** at each step. Consequently:

- Lipschitz constant $L_C = c_n \|C_e\|_2^2$ (no σ^2)
- Contraction factor $(1 - c_1 \underline{\delta}^2 \eta)$ (no σ^2)
- ISS radius r_f^{disc} (no σ^2)
- Sample complexity $k^\star$ (no σ^2)

All quantities are controlled solely by $\|C_e\|_2$, g, and the trace-free noise E_e.

Remark A.3 (Geometric Constants). The constants (c_n, c_1, c_2) depend **only on dimension** n:

Constant	Source	Order	Explicit Bound
c_n	Lipschitz constant of Ω	$O(\sqrt{n})$	$\leq 2\sqrt{n} + 8$
c_1	Local PL to discrete descent	$O(1)$	See Appendix D
c_2	Perturbation-to-output gain	$O(1)$	See Appendix D

For explicit derivations, see Appendix G, §G.2.

A.5 Recent Enhancements

The following improvements over the basic framework are established in the appendices:

Remark A.4 (Improved Saddle Escape). Appendix E provides improvements over generic perturbed Riemannian gradient descent (PRGD):

(i) **Explicit Givens direction:** The escape direction at degenerate blocks is $\Xi = E_{ij} - E_{ji}$ (Lemma C.11).

(ii) **Single-log complexity:** Escape time is $O(\log(1/\zeta))$ instead of $O(\log^4(1/\zeta))$ in generic PRGD.

(iii) **Deterministic option:** A single step with $t = \pi/4$ along Ξ reduces f by exactly $b^2 = A_{ij}^2$, requiring no randomness (Remark E.15).

Remark A.5 (Modern Perturbation Bounds). Appendix G incorporates the O'Rourke-Vu-Wang (2018) theorem (Theorem G.7), which provides:

- **Weaker gap condition:** Does not require $\|E\|_2 < g/2$ as in classical Davis-Kahan.
- **Improved bounds:** Exploits skewness between signal eigenvectors and random perturbations.

This is relevant when the trace-free perturbation $\|E_e\|_2$ is comparable to or larger than the spectral gap g.

Remark A.6 (Baseline Structural Limitations). Appendix F establishes that baselines (Subspace Iteration, QR-Oja, Euclidean SGD+QR) cannot achieve σ-free ISS due to:

- **Signal vanishing:** Contraction factors approach 1 as $\sigma^2 \to \infty$.
- **Normal-space contamination:** Euclidean gradients have $\Theta(\sigma^2)$ normal components.
- **Direction misalignment:** Under sign-symmetric distributions, baseline directions have exactly 50% probability of opposing the optimal descent direction (Theorem F.25).

The commutator structure is both sufficient and (conditionally) necessary for exact σ^2-elimination.

B Proofs for Algebraic Filtering

This appendix provides proofs for the algebraic properties of the commutator dynamics, corresponding to Theorem 1 in the main text. Throughout, we work with the rotated covariance $A = M^\top C_e^{\text{sig}} M$, where C_e^{sig} is the trace-free signal covariance. At convergence ($M \approx M^*$), the local spectral separation $\underline{\delta}$ equals the spectral gap $g := \min_{i \neq j} |\lambda_i(C_{\text{sig}}) - \lambda_j(C_{\text{sig}})|$ of the signal covariance.

B.1 Proof of Lemma 1.1: Scalar Shift Invariance

Lemma B.1 (Scalar Shift Invariance)**.** *For any $\alpha \in \mathbb{R}$ and $M \in SO(n)$,*

$$\Omega_{C+\alpha I}(M) = \Omega_C(M). \tag{36}$$

Proof. Let $A = M^\top C M$ and $D = \text{diag}(A)$. Then

$$\Omega_{C+\alpha I}(M) = [M^\top (C + \alpha I) M, \text{diag}(M^\top (C + \alpha I) M)] \tag{37}$$
$$= [A + \alpha I, D + \alpha I]. \tag{38}$$

By the bilinearity of the Lie bracket and $[I, X] = 0$ for any X:

$$[A + \alpha I, D + \alpha I] = [A, D] + \alpha[A, I] + \alpha[I, D] + \alpha^2 [I, I] \tag{39}$$
$$= [A, D] = \Omega_C(M).$$

□

Corollary B.2 (Discrete Pointwise Invariance)**.** *For any observation sequence $C_k = C_{\text{sig}} + \sigma_k^2 I + E_k$ where $\{\sigma_k^2\}$ is an arbitrary time-varying sequence, the discrete trajectory $\{M_k\}$ generated by the Cayley iteration equation 8 coincides pointwise with the trajectory generated by $\{C_{\text{sig}} + E_k\}$.*

Proof. By induction. The base case M_0 is given. Suppose M_k is identical for both observation sequences. Then by Lemma B.1:

$$\Omega_k^{(C_{\text{sig}} + \sigma_k^2 I + E_k)} = \Omega_k^{(C_{\text{sig}} + E_k)},$$

hence $M_{k+1} = M_k \cdot \text{Cay}(\eta_k \Omega_k)$ is identical. The induction is complete. □

B.2 Proof of Lemma 1.2: Input Mapping Bound

Lemma B.3 (Input Mapping Bound)**.** *Let $A = M^\top C_e^{\text{sig}} M$, $D = \text{diag}(A)$, and $\mathcal{E} = M^\top E_e M$ where $E_e = \text{tf}(E)$. Define the effective input as:*

$$U := \Omega(A + \mathcal{E}) - \Omega(A).$$

Then

$$\|U\|_F \leq 4\|C_e^{\text{sig}}\|_2 \|E_e\|_F + 2\|E_e\|_F^2. \tag{40}$$

Proof. Expanding the difference using bilinearity of the Lie bracket:

$$U = [A + \mathcal{E}, D + \text{diag}(\mathcal{E})] - [A, D] \tag{41}$$
$$= \underbrace{[A, \text{diag}(\mathcal{E})]}_{\text{(I)}} + \underbrace{[\mathcal{E}, D]}_{\text{(II)}} + \underbrace{[\mathcal{E}, \text{diag}(\mathcal{E})]}_{\text{(III)}}. \tag{42}$$

We bound each term using the commutator norm inequality $\|[X, Y]\|_F \leq 2\|X\|_2 \|Y\|_F$:

Term (I): $\|[A, \text{diag}(\mathcal{E})]\|_F \leq 2\|A\|_2 \|\text{diag}(\mathcal{E})\|_F \leq 2\|A\|_2 \|\mathcal{E}\|_F$.

Term (II): $\|[\mathcal{E}, D]\|_F \leq 2\|\mathcal{E}\|_F \|D\|_2 \leq 2\|A\|_2 \|\mathcal{E}\|_F$.

Term (III): $\|[\mathcal{E}, \text{diag}(\mathcal{E})]\|_F \leq 2\|\mathcal{E}\|_2 \|\mathcal{E}\|_F \leq 2\|\mathcal{E}\|_F^2$.

By orthogonal invariance, $\|A\|_2 = \|C_e^{\text{sig}}\|_2$ and $\|\mathcal{E}\|_F = \|E_e\|_F$. Summing:

$$\|U\|_F \leq 4\|C_e^{\text{sig}}\|_2 \|E_e\|_F + 2\|E_e\|_F^2.$$

□

B.3 Proof of Theorem 1: Structural Filtering Theorem

Proof of Theorem 1. The theorem combines Lemma 1.1, Corollary B.2, and Lemma 1.2:

(i) Trajectory Invariance: Direct consequence of Corollary B.2.

(ii) Lyapunov Invariance: Since $f(M_k) = \frac{1}{2}\|\operatorname{off}(M_k^\top C_e M_k)\|_F^2$ depends only on $C_e = \operatorname{tf}(C)$, and $\operatorname{tf}(C + \sigma^2 I) = \operatorname{tf}(C)$, the Lyapunov values are σ-invariant.

(iii) Steady-State Radius: Follows from Lemma 1.2 combined with the ISS analysis in Theorem 2, yielding a radius depending only on $\|C_e\|_2/\underline{\delta}^2 \cdot \sup_k \|E_{e,k}\|_F$. □

B.4 Corollary: Impulse Immunity

Corollary B.4 (Impulse Immunity)**.** *Let $\sigma^2(t)$ be an arbitrary sample path, possibly containing impulses or tending to infinity. As long as the anisotropic perturbation $E_e(t)$ remains bounded:*

$$\sup_{t\in[0,T]} \|E_e(t)\|_F \le \varepsilon,$$

the transient response satisfies:

$$\mathcal{R}_T := \sup_{t\in[0,T]} \|M(t) - \bar{M}(t)\|_F \le K e^{LT} T\varepsilon,$$

where $\bar{M}(t)$ is the ideal trajectory with $E \equiv 0$, and K, L depend only on C_{sig}.

Proof. By Lemma B.1, $\Omega_{C_{\text{sig}}+\sigma(t)^2 I+E(t)}(M) = \Omega_{C_{\text{sig}}+E(t)}(M)$, so the vector field is independent of $\sigma(\cdot)$. Moreover, since only the trace-free part affects the commutator, we have $\Omega_{C_{\text{sig}}+E(t)}(M) = \Omega_{C_{\text{sig}}+E_e(t)}(M)$. Decompose:

$$\dot{M} = M\Omega_{C_{\text{sig}}}(M) + M[\Omega_{C_{\text{sig}}+E_e(t)}(M) - \Omega_{C_{\text{sig}}}(M)].$$

Let $\mathcal{E}(t) = M^\top E_e(t) M$, $A = M^\top C_{\text{sig}} M$, $D = \operatorname{diag}(A)$. Then:

$$\Omega_{C_{\text{sig}}+E_e} - \Omega_{C_{\text{sig}}} = [A + \mathcal{E}, D + \operatorname{diag}(\mathcal{E})] - [A, D].$$

By bilinearity and orthogonal invariance $\|\mathcal{E}\|_F = \|E_e\|_F$:

$$\|\Omega_{C_{\text{sig}}+E_e}(M) - \Omega_{C_{\text{sig}}}(M)\|_F \le K\|E_e(t)\|_F,$$

where K depends only on $\|C_{\text{sig}}\|_2$. Since $M \mapsto M\Omega_{C_{\text{sig}}}(M)$ is globally Lipschitz on compact $SO(n)$, Gronwall yields:

$$\mathcal{R}_T \le K e^{LT} \int_0^T \|E_e(s)\|_F \, ds \le K e^{LT} T\varepsilon.$$ □

B.5 Unified σ^2-Immunity Statement

Proposition B.5 (σ^2-Immunity)**.** *Let the observation sequence be $C_k = C_{\text{sig}} + \sigma_k^2 I + E_k$, where σ_k^2 is an arbitrary time-varying sequence (including impulses or unbounded), and E_k is the perturbation. Then:*

- *(i)* Trajectory Invariance: *$\{M_k\}$ generated by $\{C_k\}$ is pointwise identical to that generated by $\{C_{\text{sig}} + E_k\}$.*
- *(ii)* Lyapunov Invariance: *$f(M_k) = \frac{1}{2}\|\operatorname{off}(M_k^\top C_{\text{sig}} M_k)\|_F^2$ does not depend on $\{\sigma_k^2\}$.*
- *(iii)* Input Invariance: *$\|U_k\|_F \le 4\|C_e^{\text{sig}}\|_2 \|E_{e,k}\|_F + 2\|E_{e,k}\|_F^2$, where $E_{e,k} = \operatorname{tf}(E_k)$, independent of σ_k^2.*
- *(iv)* Steady-State Invariance: *$r_f \propto \|C_e\|_2 \cdot \sup_k \|E_{e,k}\|_F/\underline{\delta}^2$, independent of σ^2.*

Proof. **(i)–(ii):** By Lemma B.1 and Corollary B.2, the algebraic identity $[X + \alpha I, Y] = [X, Y]$ directly guarantees trajectory and Lyapunov invariance.

(iii): By Lemma B.2, the effective input decomposition is:

$$U = [A + \mathcal{E}, D + \mathrm{diag}(\mathcal{E})] - [A, D],$$

where $A = M^\top C_e^{\mathrm{sig}} M$, $D = \mathrm{diag}(A)$, and $\mathcal{E} = M^\top E_e M$. Neither A nor $\mathcal{E}$ contains σ^2.

(iv): By the ISS analysis (Theorem 2, Appendix C), the radius is given by $r_f = \frac{\sqrt{2}\|C_e\|_2}{\underline{\delta}^2}\bar{U}$, where $\bar{U}$ is controlled by (iii). □

C Proofs for Geometric ISS

This appendix provides proofs for the Input-to-State Stability (ISS) Sontag (2008) analysis on $SO(n)$, corresponding to Theorem 2 in the main text.

C.1 Proof of Lemma 2.1: Local Dissipation (PL Condition)

Lemma C.1 (Local Dissipation)**.** *In the spectrally separated domain $\mathcal{N}_{\underline{\delta}} = \{M \in SO(n) : \delta(M) \geq \underline{\delta}\}$:*

$$\|\Omega(M)\|_F^2 \geq 2\underline{\delta}^2 f(M). \tag{43}$$

Proof. Let $A = M^\top C_e M$ with diagonal entries A_{ii}. The commutator has entries:

$$\Omega_{ij} = (A_{jj} - A_{ii})A_{ij}, \quad i \neq j,$$

while $\Omega_{ii} = 0$. Therefore:

$$\|\Omega\|_F^2 = 2\sum_{i<j}(A_{ii} - A_{jj})^2 A_{ij}^2, \tag{44}$$

$$f(M) = \frac{1}{2}\|\operatorname{off}(A)\|_F^2 = \sum_{i<j} A_{ij}^2. \tag{45}$$

In $\mathcal{N}_{\underline{\delta}}$, we have $(A_{ii} - A_{jj})^2 \geq \underline{\delta}^2$ for all $i \neq j$, hence:

$$\|\Omega\|_F^2 \geq 2\underline{\delta}^2 \sum_{i<j} A_{ij}^2 = 2\underline{\delta}^2 f(M).$$

□

C.2 Proof of Lemma 2.2: Operator Spectral Sandwiching

Lemma C.2 (Operator Spectral Sandwiching)**.** *For $A = M^\top C_e M$, $D = \operatorname{diag}(A)$, and $X = \operatorname{off}(A)$, define the commutator operator $\mathcal{L}_A(X) := [X, D]$. Then:*

$$\underline{\delta}\|X\|_F \leq \|\mathcal{L}_A(X)\|_F \leq 2\|C_e\|_2\|X\|_F. \tag{46}$$

Equivalently, in terms of f and Ω:

$$2\underline{\delta}^2 f(M) \leq \|\Omega(M)\|_F^2 \leq 8\|C_e\|_2^2 f(M). \tag{47}$$

Proof. **Lower bound:** By component-wise computation:

$$[\operatorname{off}(A), D]_{ij} = (D_{jj} - D_{ii}) \cdot \operatorname{off}(A)_{ij} = (A_{jj} - A_{ii})A_{ij}.$$

Hence:

$$\|\mathcal{L}_A(X)\|_F^2 = 2\sum_{i<j}(A_{ii} - A_{jj})^2 A_{ij}^2 \geq \underline{\delta}^2 \cdot 2\sum_{i<j} A_{ij}^2 = \underline{\delta}^2\|X\|_F^2.$$

Upper bound: We have $|A_{ii} - A_{jj}| \leq |A_{ii}| + |A_{jj}| \leq 2\|A\|_2 = 2\|C_e\|_2$, thus:

$$\|\mathcal{L}_A(X)\|_F^2 \leq (2\|C_e\|_2)^2 \cdot 2\sum_{i<j} A_{ij}^2 = 4\|C_e\|_2^2\|X\|_F^2.$$

□

C.3 Proof of Lemma 2.3: ISS Differential Inequality

Lemma C.3 (ISS Differential Inequality)**.** *Let $y(t) := \sqrt{f(M(t))}$. In $\mathcal{N}_{\underline{\delta}}$, the Dini derivative satisfies:*

$$D^+ y(t) \leq -\underline{\delta}^2 y(t) + \sqrt{2}\|C_e\|_2 \cdot u_{\text{eff}}(t), \tag{48}$$

where $u_{\text{eff}}(t) := \|U(t)\|_F + \frac{\rho}{2\|C_e\|_2}$ and $\rho := \|\dot{C}(t)\|_F$ is the drift rate.

Proof. **Step 1: Perturbed $\dot{f}$ expansion.** For the perturbed system $\dot{M} = M(\Omega + U)$ with drift $\|\dot{C}\|_F \leq \rho$:

$$\dot{f} = -\|\Omega\|_F^2 + \langle \Omega, U\rangle + \langle \text{off}(A), \text{off}(M^\top \dot{C} M)\rangle.$$

By Cauchy-Schwarz:

$$\dot{f} \leq -\|\Omega\|_F^2 + \|\Omega\|_F \|U\|_F + \|\,\text{off}(A)\|_F \|\dot{C}\|_F.$$

Step 2: Substitution. Using $\|\,\text{off}(A)\|_F = \sqrt{2f} = \sqrt{2}y$ and $\|\dot{C}\|_F \leq \rho$:

$$\dot{f} \leq -\|\Omega\|_F^2 + \|\Omega\|_F \|U\|_F + \sqrt{2}\rho y.$$

Step 3: Apply Lemmas 2.1 and 2.2. Substituting $\|\Omega\|_F^2 \geq 2\underline{\delta}^2 y^2$ and $\|\Omega\|_F \leq 2\sqrt{2}\|C_e\|_2 y$:

$$\dot{f} \leq -2\underline{\delta}^2 y^2 + 2\sqrt{2}\|C_e\|_2 y \|U\|_F + \sqrt{2}\rho y.$$

Step 4: Convert to $\dot{y}$. When $f > 0$, $\dot{y} = \dot{f}/(2y)$. Dividing by $2y$:

$$\dot{y} \leq -\underline{\delta}^2 y + \sqrt{2}\|C_e\|_2 \|U\|_F + \frac{\sqrt{2}\rho}{2}.$$

Combining with $u_{\text{eff}} := \|U\|_F + \frac{\rho}{2\|C_e\|_2}$:

$$\dot{y} \leq -\underline{\delta}^2 y + \sqrt{2}\|C_e\|_2 \cdot u_{\text{eff}}.$$

□

C.4 Proof of Theorem 2: Geometric ISS Theorem

Proof of Theorem 2. Integrating Lemma 2.3 via Gronwall's inequality with $\lambda := \underline{\delta}^2$:

(i) Exponential convergence:

$$y(t) \leq e^{-\lambda t} y(0) + \int_0^t e^{-\lambda(t-s)} \sqrt{2}\|C_e\|_2 \cdot u_{\text{eff}}(s)\, ds.$$

If $\sup_t u_{\text{eff}}(t) \leq \bar{u}$:

$$y(t) \leq e^{-\underline{\delta}^2 t} y(0) + \frac{\sqrt{2}\|C_e\|_2}{\underline{\delta}^2}\bar{u}.$$

(ii) Steady-state ball: Taking $t \to \infty$:

$$\limsup_{t\to\infty} \sqrt{f(M(t))} \leq r_f(\bar{u}) := \frac{\sqrt{2}\|C_e\|_2}{\underline{\delta}^2}\bar{u}.$$

□

C.5 Tightness: A 2×2 Lower Bound Construction

Proposition C.4 (Tightness of $\underline{\delta}^{-2}$ Dependence)**.** *There exist constants $c > 0$ and configurations $(C_\star, M_0, U)$ such that:*

$$\limsup_{t\to\infty} \sqrt{f(M(t))} \geq c \cdot \frac{\|C_e\|_2}{\underline{\delta}^2}\bar{U}.$$

Proof. Consider $n = 2$ with $C_\star = \mathrm{diag}(\lambda_1, \lambda_2)$, $\delta := \lambda_1 - \lambda_2 > 0$. Then $\|C_e\|_2 = \delta/2$.

The commutator for $A = \begin{pmatrix} \lambda_1 & x \\ x & \lambda_2 \end{pmatrix}$ is:

$$\Omega = \begin{pmatrix} 0 & -\delta x \\ \delta x & 0 \end{pmatrix}.$$

At steady state under constant input U, balancing $\|\Omega\|_F \approx \|U\|_F$ yields:

$$|x_\infty| \sim \frac{\|U\|_F}{\sqrt{2}\delta}, \quad \sqrt{f_\infty} = |x_\infty| \sim \frac{\bar{U}}{\delta}.$$

Comparing with the upper bound $r_f \le c \cdot \frac{\delta \cdot \bar{U}}{\delta^2} = c \cdot \frac{\bar{U}}{\delta}$, the bounds match in order. □

C.6 Explicit Radius Expansion

Corollary C.5 (Explicit Noise Ball Radius)**.** *Expanding $\bar{u} = \sup_t u_{\mathrm{eff}}(t)$ in terms of verifiable quantities:*

No drift case *($\rho = 0$): If $\sup_t \|E_{e,t}\|_F \le \varepsilon$, then by Lemma 1.2:*

$$\sup_t \|U(t)\|_F \le 4\|C_e^{\mathrm{sig}}\|_2 \varepsilon + 2\varepsilon^2.$$

Substituting into Theorem 2 gives:

$$r_f(\varepsilon) := \frac{\sqrt{2}\|C_e\|_2}{\underline{\delta}^2} \left(4\|C_e^{\mathrm{sig}}\|_2 \varepsilon + 2\varepsilon^2\right). \tag{49}$$

With drift *($\rho > 0$):*

$$r_f(\varepsilon, \rho) \le \frac{\sqrt{2}\|C_e\|_2}{\underline{\delta}^2} \left(4\|C_e^{\mathrm{sig}}\|_2 \varepsilon + 2\varepsilon^2 + \frac{\rho}{2\|C_e\|_2}\right). \tag{50}$$

C.7 Distinction from Gronwall Bounds

Remark C.6 (ISS vs. Gronwall). The differential inequality $\dot{y} \le -\underline{\delta}^2 y + \gamma u_{\mathrm{eff}}$ has dissipative structure: the negative feedback term $-\underline{\delta}^2 y$ ensures exponential decay of initial error and bounded steady state. In contrast, classical Gronwall $\dot{y} \le Ly + b$ yields $y(t) \le e^{Lt}(y(0) + bt)$, which grows monotonically.

C.8 Global Convergence: Asymptotic Analysis

This subsection establishes **asymptotic global convergence** from almost all initial conditions, filling the gap noted in Remark G.4.

C.8.1 Continuous-Time: Convergence to Critical Points

Theorem C.7 (Asymptotic Convergence of Commutator Flow)**.** *Consider the ODE $\dot{M}(t) = M(t)\Omega(M(t))$ with $M(0) \in SO(n)$. Then:*

(i) The solution $M(t)$ exists globally and remains in $SO(n)$;

(ii) $t \mapsto f(M(t))$ is monotonically non-increasing and converges to some limit f_∞;

(iii) Every ω-limit point M_∞ satisfies $\Omega(M_\infty) = 0$ (i.e., is a critical point of f);

*(iv) Since f is real-analytic (polynomial in matrix entries), $M(t)$ converges to a **single** critical point M_∞.*

Proof. **(i)** $SO(n)$ is a compact smooth manifold. The vector field $M \mapsto M\Omega(M)$ is smooth (composition of matrix multiplication and commutator), hence bounded on compact sets. By standard ODE theory, solutions exist globally and cannot escape $SO(n)$.

(ii) The flow equals (up to constant $c_0 > 0$) the negative Riemannian gradient flow of f:

$$\frac{d}{dt} f(M(t)) = -c_0 \|\Omega(M(t))\|_F^2 \leq 0.$$

Since $f \geq 0$ is bounded below, $f(M(t))$ converges.

(iii) From (ii): $\int_0^\infty \|\Omega(M(t))\|_F^2 \, dt < \infty$, so there exists $t_\ell \to \infty$ with $\|\Omega(M(t_\ell))\|_F \to 0$. By compactness of $SO(n)$, $\{M(t_\ell)\}$ has a convergent subsequence to some M_∞. Continuity of Ω implies $\Omega(M_\infty) = 0$.

(iv) This is the classical Łojasiewicz gradient inequality result: on compact manifolds, real-analytic functions satisfy $\|\operatorname{grad} f\| \geq c|f - f_\infty|^\theta$ for some $\theta \in (0, 1)$. This implies finite trajectory length, hence convergence to a single point. See Simon (1983) or Absil–Mahony–Sepulchre (2008). □

C.8.2 From Critical Points to Global Minima: Strict Saddle Analysis

This subsection provides a **self-contained proof** of the strict saddle property for $f(M) = \frac{1}{2}\|\operatorname{off}(M^\top C_e M)\|_F^2$ on $SO(n)$.

Lemma C.8 (Critical Point Characterization)**.** *A matrix $M \in SO(n)$ is a critical point of f if and only if $\Omega(M) = 0$. Equivalently, for all $i \neq j$:*

$$(A_{jj} - A_{ii})A_{ij} = 0, \qquad A := M^\top C_e M. \tag{51}$$

Proof. By the gradient representation $\operatorname{grad} f(M) = -c_0 M\Omega(M)$ for some $c_0 > 0$, we have $\operatorname{grad} f(M) = 0$ iff $M\Omega(M) = 0$. Since M is invertible, this holds iff $\Omega(M) = 0$. The entrywise condition follows from $\Omega_{ij} = (A_{jj} - A_{ii})A_{ij}$. □

Lemma C.9 (Critical Point Dichotomy)**.** *At any critical point M of f:*

(a) ***Global Minimum:*** *If $\operatorname{off}(A) = 0$, then $f(M) = 0$ and M maps eigenvectors of C_e to the standard basis (up to permutation and sign).*

(b) ***Degenerate Block:*** *If $\operatorname{off}(A) \neq 0$, then there exist indices $i < j$ such that $A_{ij} \neq 0$ and $A_{ii} = A_{jj}$.*

Proof. From equation 51: if $A_{ij} \neq 0$ for some $i \neq j$, then $A_{ii} = A_{jj}$. If no such pair exists, then A is diagonal. □

Lemma C.10 (Second Variation Formula along Geodesics)**.** *Let $M \in SO(n)$, $A := M^\top C_e M$, and $\Xi \in \mathfrak{so}(n)$. Define the geodesic $M(t) := M \exp(t\Xi)$ and $A(t) := M(t)^\top C_e M(t)$. Then:*

$$A(t) = e^{-t\Xi} A \, e^{t\Xi}, \tag{52}$$

$$\dot{A}(t) = [A(t), \Xi], \qquad \ddot{A}(t) = A(t)\Xi^2 + \Xi^2 A(t) - 2\Xi A(t)\Xi. \tag{53}$$

The Riemannian Hessian quadratic form satisfies:

$$\operatorname{Hess} f(M)[M\Xi, M\Xi] = \frac{d^2}{dt^2}\bigg|_{t=0} f(Me^{t\Xi}) = \|\operatorname{off}([A, \Xi])\|_F^2 + \langle \operatorname{off}(A), \operatorname{off}(\ddot{A}(0)) \rangle. \tag{54}$$

Proof. **Step 1 (Conjugation):** Since $M(t) = Me^{t\Xi}$ and $e^{t\Xi} \in SO(n)$, we have $M(t)^\top = e^{-t\Xi} M^\top$, hence $A(t) = e^{-t\Xi} A \, e^{t\Xi}$.

Step 2 (Derivatives of $A(t)$): Differentiating equation 52:

$$\dot{A}(t) = -\Xi e^{-t\Xi} A \, e^{t\Xi} + e^{-t\Xi} A \, e^{t\Xi} \Xi = -\Xi A(t) + A(t)\Xi = [A(t), \Xi].$$

Differentiating again: $\ddot{A}(t) = [\dot{A}(t), \Xi] = [[A(t), \Xi], \Xi] = A(t)\Xi^2 + \Xi^2 A(t) - 2\Xi A(t)\Xi$.

Step 3 (Second variation of f): Let $X(t) := \mathrm{off}(A(t))$. Since $f(M(t)) = \frac{1}{2}\|X(t)\|_F^2$:

$$\frac{d^2}{dt^2} f(M(t)) = \|\dot{X}(t)\|_F^2 + \langle X(t), \ddot{X}(t)\rangle.$$

Evaluating at $t = 0$ with $\dot{X}(0) = \mathrm{off}([A, \Xi])$ and $\ddot{X}(0) = \mathrm{off}(\ddot{A}(0))$ yields equation 54.

Step 4 (Identification): Under the bi-invariant metric on $SO(n)$, $t \mapsto M\exp(t\Xi)$ is a geodesic with $\dot{M}(0) = M\Xi$, so the second derivative equals $\mathrm{Hess}\, f(M)[M\Xi, M\Xi]$. □

Lemma C.11 (Negative Curvature at Degenerate Blocks)**.** *Let $M \in SO(n)$ and $A := M^\top C_e M$. Suppose there exist $i \neq j$ with $A_{ii} = A_{jj}$ and $b := A_{ij} \neq 0$. Let $\Xi := E_{ij} - E_{ji} \in \mathfrak{so}(n)$. Then:*

$$f(Me^{t\Xi}) = f(M) - b^2 \sin^2(2t), \tag{55}$$

and consequently:

$$\mathrm{Hess}\, f(M)[M\Xi, M\Xi] = -8b^2 < 0, \qquad \lambda_{\min}(\mathrm{Hess}\, f(M)) \leq -4b^2. \tag{56}$$

Proof. **Step 1 (Givens rotation):** With $\Xi = E_{ij} - E_{ji}$, the matrix $G(t) := e^{t\Xi}$ acts as a 2×2 rotation in the (i, j)-plane:

$$G(t)|_{(i,j)} = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}.$$

Step 2 (Invariance of other entries): For $k \notin \{i, j\}$, the pair $(A_{ik}(t), A_{jk}(t))$ is rotated orthogonally, so $A_{ik}(t)^2 + A_{jk}(t)^2 = A_{ik}^2 + A_{jk}^2$. Entries A_{pq} with $\{p, q\} \cap \{i, j\} = \emptyset$ are unchanged.

Step 3 (The (i, j) entry): Standard Givens conjugation gives:

$$A_{ij}(t) = (\cos^2 t - \sin^2 t)A_{ij} + (\cos t \sin t)(A_{jj} - A_{ii}).$$

Under $A_{ii} = A_{jj}$, this simplifies to $A_{ij}(t) = b\cos(2t)$.

Step 4 (Reduction to 1D profile): Since only $A_{ij}(t)^2$ changes in $f = \sum_{p<q} A_{pq}^2$:

$$f(Me^{t\Xi}) = f(M) - b^2 + b^2\cos^2(2t) = f(M) - b^2(1 - \cos^2(2t)) = f(M) - b^2\sin^2(2t).$$

Step 5 (Second derivative): Since $\sin^2(2t) = 4t^2 + O(t^4)$:

$$\left.\frac{d^2}{dt^2}\right|_{t=0} f(Me^{t\Xi}) = -8b^2 < 0.$$

The eigenvalue bound follows from $\|\Xi\|_F^2 = 2$. □

Proposition C.12 (Strict Saddle Property)**.** *When C_e has distinct eigenvalues, every critical point of f on $SO(n)$ is either:*

- *A **global minimum** ($f = 0$, complete diagonalization), or*
- *A **strict saddle** (Hessian has at least one negative eigenvalue).*

No spurious local minima exist.

Proof. Let M be a critical point with $\Omega(M) = 0$.

Case 1: If $\mathrm{off}(A) = 0$, then $A = M^\top C_e M$ is diagonal. Since C_e has distinct eigenvalues and A shares these eigenvalues, M maps eigenvectors of C_e to standard basis vectors (up to permutation and sign). Thus $f(M) = 0$: global minimum.

Case 2: If $\operatorname{off}(A) \neq 0$, then by Lemma C.9, there exists (i, j) with $A_{ij} \neq 0$ and $A_{ii} = A_{jj}$. By Lemma C.11:

$$\lambda_{\min}(\operatorname{Hess} f(M)) \leq -4A_{ij}^2 < 0.$$

Hence M is a strict saddle. □

Remark C.13 (Historical Context). The strict saddle property for matrix diagonalization objectives is classical in the theory of isospectral manifolds and double-bracket flows; see Brockett (1991), Helmke & Moore (1994), and Ge, Jin & Zheng (2017). The proof above provides a self-contained derivation using only the commutator structure already established in this paper.

C.8.3 Morse–Bott Structure of the Critical Set

This subsection establishes that **non-optimal critical sets are smooth embedded submanifolds** and that the Hessian is **non-degenerate on the normal bundle** (Morse–Bott completeness).

Definition C.14 (Block Partition at Critical Points)**.** At a critical point M with $A := M^\top C_e M$, define the equivalence relation $i \sim j \Leftrightarrow A_{ii} = A_{jj}$. Let $\pi = \{I_1, \ldots, I_K\}$ denote the resulting partition with $|I_\alpha| = m_\alpha$.

By the critical condition equation 51, if $i \in I_\alpha$ and $j \in I_\beta$ with $\alpha \neq \beta$, then $A_{ii} \neq A_{jj}$ implies $A_{ij} = 0$. Hence A **is block-diagonal** with respect to π at any critical point.

Definition C.15 (Regular Stratum)**.** Given a partition $\pi = \{I_1, \ldots, I_K\}$ and spanning trees $\{T_\alpha\}$ for each block (with $m_\alpha - 1$ edges), define the **regular stratum**:

$$\mathcal{S}(\pi, \{T_\alpha\}) := \{M \in SO(n) : \Omega(M) = 0,\ A_{ii} \neq A_{jj}\ \forall i \in I_\alpha, j \in I_\beta, \alpha \neq \beta,\ A_{ij} \neq 0\ \forall (i,j) \in T_\alpha\}.$$

Lemma C.16 (Tree Laplacian Rank)**.** *Let $T = (V, E)$ be a tree with m vertices and $m - 1$ edges. Assign positive weights $w_e > 0$ to each edge. Let $B \in \mathbb{R}^{m \times (m-1)}$ be the (oriented) incidence matrix and $W = \operatorname{diag}(w_e)$. Then the weighted Laplacian $L := BWB^\top$ satisfies:*

$$\operatorname{rank}(L) = m - 1, \qquad \ker(L) = \operatorname{span}\{\mathbf{1}\}. \tag{57}$$

Proof. By direct computation: $x^\top L x = x^\top BWB^\top x = \sum_{e=\{u,v\}} w_e (x_u - x_v)^2$. Since T is connected and $w_e > 0$, this equals zero iff x is constant. Hence $\ker(L) = \operatorname{span}\{\mathbf{1}\}$ and $\operatorname{rank}(L) = m - 1$. □

Lemma C.17 (Linearization of the Commutator Map)**.** *For $M\Xi \in T_M SO(n)$ with $\Xi \in \mathfrak{so}(n)$, the differential of Ω satisfies:*

$$D\Omega(M)[M\Xi] = [\delta A, D] + [A, \delta D],$$

where $\delta A = [A, \Xi]$ and $\delta D = \operatorname{diag}(\delta A)$. In components (for $i \neq j$):

$$(D\Omega)_{ij} = (D_{jj} - D_{ii}) \cdot (\delta A)_{ij} + (\delta D_{jj} - \delta D_{ii}) \cdot A_{ij}. \tag{58}$$

Furthermore, the diagonal perturbation satisfies:

$$\delta D_{ii} = ([A, \Xi])_{ii} = -2 \sum_{k \neq i} A_{ik} \Xi_{ik}. \tag{59}$$

Proof. The formula $\delta A = [A, \Xi]$ follows from Lemma C.10. Expanding $\Omega = [A, D]$ bilinearly and using A symmetric, Ξ skew-symmetric yields equation 58 and equation 59. □

Lemma C.18 (Within-Block Rank via Incidence Matrix)**.** *Fix a block I_α with spanning tree T_α. Let $B_\alpha \in \mathbb{R}^{m_\alpha \times (m_\alpha - 1)}$ be the incidence matrix and $S_\alpha = \operatorname{diag}(A_{uv})$ for tree edges $(u, v) \in T_\alpha$. For edge coordinates $x \in \mathbb{R}^{m_\alpha - 1}$ (with $\Xi_{uv} = x_e$, $\Xi_{vu} = -x_e$):*

$$\delta d_\alpha := \delta D|_{I_\alpha} = -2B_\alpha S_\alpha x, \tag{60}$$

$$\delta \omega_\alpha := (\delta\Omega)|_{T_\alpha} = -2S_\alpha B_\alpha^\top B_\alpha S_\alpha x. \tag{61}$$

Under the regular stratum assumption ($A_{uv} \neq 0$ for tree edges), both maps have rank $m_\alpha - 1$.

Proof. Equation equation 60 follows from equation 59: each tree edge $(u \to v)$ contributes $\pm 2A_{uv}x_e$ to δD_{uu} and δD_{vv}. Equation equation 61 follows from the within-block linearization $(D\Omega)_{ij} = (\delta D_{jj} - \delta D_{ii})A_{ij}$.

Since S_α is invertible (tree edge weights $A_{uv} \neq 0$), $\operatorname{rank}(B_\alpha S_\alpha) = \operatorname{rank}(B_\alpha) = m_\alpha - 1$ by Lemma C.16. □

Proposition C.19 (Constant Rank and Explicit Dimensions)**.** *Let $M \in SO(n)$ be a critical point with block partition $\pi = \{I_\alpha\}_{\alpha=1}^K$ satisfying the regular stratum assumptions. Then:*

(i) Normal bundle dimension *(= rank of $D\Omega$):*

$$\dim N_M\mathcal{S} = \operatorname{rank} D\Omega(M) = \underbrace{\sum_{\alpha<\beta} m_\alpha m_\beta}_{\textit{cross-block}} + \underbrace{\sum_{\alpha=1}^{K}(m_\alpha - 1)}_{\textit{within-block}}. \tag{62}$$

(ii) Tangent space dimension *(= nullity of Hessian):*

$$\dim T_M\mathcal{S} = \sum_{\alpha=1}^{K} \frac{(m_\alpha - 1)(m_\alpha - 2)}{2}. \tag{63}$$

(iii) Hessian signature on normal bundle*:*

$$\operatorname{index}(M) = \sum_{\alpha=1}^{K}(m_\alpha - 1), \qquad \#\{\lambda > 0\} = \sum_{\alpha<\beta} m_\alpha m_\beta. \tag{64}$$

Before the proof, we establish the key link between the Hessian and linearization:

Lemma C.20 (Hessian–Linearization Link)**.** *At a critical point M (where $\Omega(M) = 0$), the Hessian quadratic form relates to the linearization $D\Omega$ via the gradient flow structure. Since* $\operatorname{grad} f(M) = -c_0 M\Omega(M)$ *for $c_0 > 0$:*

$$\operatorname{Hess} f(M)[M\Xi, M\Xi] = -c_0\langle D\Omega(M)[M\Xi], \Omega(M)\rangle + \textit{(second-order terms)}. \tag{65}$$

At critical points where $\Omega(M) = 0$, the sign of the Hessian is determined entirely by the second variation formula equation 54, which we compute explicitly below.

Proof of Proposition C.19. **Step A (Cross-block: rank and positive curvature).**

Givens parameterization: For $i \in I_\alpha$, $j \in I_\beta$ with $\alpha \neq \beta$, define the Givens generator:

$$\Xi_{ij} := E_{ij} - E_{ji} \in \mathfrak{so}(n), \qquad (\Xi_{ij})_{pq} = \delta_{ip}\delta_{jq} - \delta_{iq}\delta_{jp}.$$

The geodesic $M(t) = Me^{t\Xi_{ij}}$ acts as a rotation in the (i, j)-plane with angle t.

Rank contribution: At critical points, $A_{ij} = 0$ (cross-block) and $D_{ii} = d_\alpha \neq d_\beta = D_{jj}$. From equation 58:

$$(D\Omega)_{ij} = (d_\beta - d_\alpha) \cdot (\delta A)_{ij}.$$

Using $\delta A = [A, \Xi_{ij}]$ and computing $(\delta A)_{ij} = (A_{ii} - A_{jj})(\Xi_{ij})_{ij} = (d_\alpha - d_\beta) \cdot 1$:

$$(D\Omega)_{ij} = (d_\beta - d_\alpha)(d_\alpha - d_\beta) = -(d_\beta - d_\alpha)^2 \neq 0.$$

Each cross-block pair contributes independently, giving $r_{\text{cross}} = \sum_{\alpha<\beta} m_\alpha m_\beta$.

Positive curvature: Since $A_{ij} = 0$ at the critical point, the Givens conjugation formula gives:

$$A_{ij}(t) = \cos(2t) \cdot \underbrace{A_{ij}}_{=0} + \frac{1}{2}\sin(2t) \cdot (A_{jj} - A_{ii}) = \frac{d_\beta - d_\alpha}{2}\sin(2t).$$

Thus:

$$f(Me^{t\Xi_{ij}}) = f(M) + A_{ij}(t)^2 = f(M) + \frac{(d_\beta - d_\alpha)^2}{4}\sin^2(2t).$$

The second derivative at $t = 0$:

$$\operatorname{Hess} f(M)[M\Xi_{ij}, M\Xi_{ij}] = \frac{d^2}{dt^2}\bigg|_{t=0} f(Me^{t\Xi_{ij}}) = +2(d_\beta - d_\alpha)^2 > 0.$$

Hence **positive definite** on all cross-block directions.

Step B (Within-block: rank and negative curvature).

Givens parameterization on tree edges: For a block I_α with spanning tree T_α, parameterize perturbations by edge coordinates $x = (x_e)_{e\in T_\alpha} \in \mathbb{R}^{m_\alpha - 1}$. For each directed tree edge $e = (u \to v)$:

$$\Xi_\alpha = \sum_{e\in T_\alpha} x_e(E_{uv} - E_{vu}), \qquad (\Xi_\alpha)_{uv} = x_e, \quad (\Xi_\alpha)_{vu} = -x_e.$$

Rank contribution via Laplacian: Within block I_α, $D_{ii} = D_{jj} = d_\alpha$ for all $i, j \in I_\alpha$. From equation 58, the first term vanishes:

$$(D\Omega)_{ij} = (\delta D_{jj} - \delta D_{ii}) \cdot A_{ij}.$$

By Lemma C.18 with incidence matrix B_α and weight matrix $S_\alpha = \operatorname{diag}(A_{uv})$:

$$\delta d_\alpha = -2B_\alpha S_\alpha x, \qquad \delta\omega_\alpha = -2S_\alpha B_\alpha^\top B_\alpha S_\alpha x.$$

The matrix $M_\alpha := S_\alpha B_\alpha^\top B_\alpha S_\alpha$ is **positive definite** (congruent to the graph Laplacian $B_\alpha^\top B_\alpha \succ 0$ on trees, with invertible S_α). Thus $\operatorname{rank}(x \mapsto \delta\omega_\alpha) = m_\alpha - 1$.

Negative curvature: For any tree edge $(i, j) \in T_\alpha$, we have $A_{ii} = A_{jj} = d_\alpha$ and $b := A_{ij} \neq 0$ (regular stratum assumption). By Lemma C.11:

$$f(Me^{t\Xi_{ij}}) = f(M) - b^2\sin^2(2t), \qquad \operatorname{Hess} f(M)[M\Xi_{ij}, M\Xi_{ij}] = -8b^2 < 0.$$

For the full within-block subspace, the Hessian restricted to tree-edge directions has matrix representation (up to positive constant):

$$H_\alpha = -\operatorname{diag}(8A_{uv}^2)_{e\in T_\alpha} \prec 0.$$

Since $M_\alpha \succ 0$ and each diagonal entry $-8A_{uv}^2 < 0$, the restricted Hessian is **negative definite**. Each block contributes exactly $m_\alpha - 1$ negative eigenvalues.

Step C (Dimension computation).

The identity $\dim SO(n) = \frac{n(n-1)}{2} = \sum_\alpha \frac{m_\alpha(m_\alpha-1)}{2} + \sum_{\alpha<\beta} m_\alpha m_\beta$ yields:

$$\begin{aligned}
\dim T_M\mathcal{S} &= \dim SO(n) - \operatorname{rank} D\Omega(M) \\
&= \sum_\alpha \frac{m_\alpha(m_\alpha-1)}{2} + \sum_{\alpha<\beta} m_\alpha m_\beta - \sum_{\alpha<\beta} m_\alpha m_\beta - \sum_\alpha (m_\alpha - 1) \\
&= \sum_\alpha \left(\frac{m_\alpha(m_\alpha-1)}{2} - (m_\alpha - 1)\right) = \sum_\alpha \frac{(m_\alpha-1)(m_\alpha-2)}{2}.
\end{aligned}$$

□

Theorem C.21 (Morse–Bott Non-Degeneracy)**.** *On each regular stratum $\mathcal{S}$:*

$$\ker \operatorname{Hess} f(M) = T_M\mathcal{S}. \tag{66}$$

*The Hessian has **no zero eigenvalues on the normal bundle**. Specifically, the normal bundle decomposes as $N_M\mathcal{S} = N^+ \oplus N^-$ where:*

- N^+ *(cross-block):* $\dim N^+ = \sum_{\alpha<\beta} m_\alpha m_\beta$*, Hessian positive definite;*
- N^- *(within-block):* $\dim N^- = \sum_\alpha (m_\alpha - 1)$*, Hessian negative definite.*

Proof. Since $\operatorname{grad} f(M) = -c_0 M\Omega(M)$ for some $c_0 > 0$, the Hessian at critical points satisfies $\ker \operatorname{Hess} f(M) = \ker D\Omega(M) = T_M\mathcal{S}$. The signature decomposition follows from Proposition C.19 Steps A and B. □

Corollary C.22 (Non-Optimal Strata are Morse–Bott Saddles)**.** *For any non-optimal critical stratum (where $f > 0$), the Morse index is* $\operatorname{index}(M) = \sum_\alpha (m_\alpha - 1) \geq 1$*. Hence all non-optimal critical points are* ***strict saddles****.*

Remark C.23 (Disconnected Blocks)*.* If the within-block graph G_α has c_α connected components instead of being connected, replace $m_\alpha - 1$ with $m_\alpha - c_\alpha$ throughout. The proof is identical: $\ker(L) = \operatorname{span}\{\mathbf{1}_{C_1}, \ldots, \mathbf{1}_{C_{c_\alpha}}\}$ for components $C_1, \ldots, C_{c_\alpha}$.

Theorem C.24 (Almost Sure Global Convergence)**.** *Assume C_e has distinct eigenvalues. Let $M(0)$ be sampled from Haar measure on $SO(n)$. Then with probability 1, the limit point M_∞ from Theorem C.7 satisfies $f(M_\infty) = 0$.*

Proof. We provide a complete proof using finite stratification to handle the potential issue of uncountable unions of measure-zero sets.

Step 1: Critical set is a real-analytic variety. The critical set $\operatorname{Crit}(f) = \{M : \Omega(M) = 0\}$ is defined by polynomial equations in matrix entries, hence is a real-analytic (semi-algebraic) subset of $SO(n)$.

Step 2: Finite Whitney stratification. By standard results in real-analytic geometry, $\operatorname{Crit}(f)$ admits a finite Whitney stratification into smooth embedded submanifolds $\{\mathcal{S}_\ell\}_{\ell=1}^L$, where L depends only on n. Each stratum corresponds to a block partition structure (Definition C.14).

Step 3: Center-stable manifold for each saddle stratum. For each non-optimal stratum $\mathcal{S}_\ell$ (where $f > 0$), by Lemma C.11 and Corollary C.22, there exists at least one negative Hessian direction in the normal bundle. Applying the center-stable manifold theorem to the gradient flow:

$$\dim W^{cs}(\mathcal{S}_\ell) \leq \dim SO(n) - 1 = \frac{n(n-1)}{2} - 1.$$

Hence each center-stable manifold $W^{cs}(\mathcal{S}_\ell)$ has **codimension** ≥ 1 in $SO(n)$, and thus has Haar measure zero.

Step 4: Finite union remains measure zero. The basin of attraction for all saddle-type critical sets is contained in:

$$\bigcup_{\ell : f|_{\mathcal{S}_\ell} > 0} W^{cs}(\mathcal{S}_\ell).$$

Since this is a **finite** union of measure-zero sets, the total basin has Haar measure zero.

Step 5: Conclusion. Since Theorem C.7 guarantees convergence to some critical point, and the basin of all non-optimal critical sets has measure zero, Haar initialization converges to the global minimum set ($f = 0$) with probability 1. □

Remark C.25 (State-Dependent $D(M)$ Does Not Affect Convergence)*.* Classical double-bracket flow theory (Brockett 1991) assumes $\dot{X} = [X, [X, N]]$ with **fixed** N. Our flow uses $D = \operatorname{diag}(A(M))$, which varies with state. This prevents direct application of Brockett's Morse–Bott framework. However, our convergence proof (Theorem C.24) relies only on:

(i) Łojasiewicz inequality for real-analytic gradient flows (convergence to single critical point);

(ii) Strict saddle property (negative curvature at non-optimal critical points);

(iii) Finite stratification (center-stable manifolds have codimension ≥ 1).

None of these require D to be state-independent. The analysis is intrinsic to the function f and manifold $SO(n)$.

C.8.4 Discrete-Time: Cayley Iteration

Corollary C.26 (Discrete Global Convergence)**.** *Under the assumptions of Theorem C.24, consider the discrete iteration $M_{k+1} = M_k \mathrm{Cay}(\eta\Omega_k)$ with $\eta \in (0, 1/L_C)$. Then:*

(i) $f(M_k)$ is monotonically decreasing and converges;

(ii) For Haar initialization, M_k converges to the global minimum set ($f = 0$) with probability 1.

Proof. The descent lemma (Lemma D.1) gives $f(M_{k+1}) \leq f(M_k) - \eta(1 - \frac{\eta L_C}{2})\|\Omega_k\|_F^2$. With $\eta < 1/L_C$, this implies monotone decrease and $\sum_k \|\Omega_k\|^2 < \infty$. Standard Riemannian optimization arguments show every accumulation point is critical. The saddle avoidance argument from Theorem C.24 applies identically. □

Remark C.27 (When Discrete Cayley Fails). The discrete convergence crucially requires $\eta \in (0, 1/L_C)$. Three failure mechanisms arise outside this regime:

(i) **Cayley singularity**: $\mathrm{Cay}(X) = (I - \frac{1}{2}X)^{-1}(I + \frac{1}{2}X)$ fails when $\det(I - \frac{\eta}{2}\Omega) = 0$, destroying local diffeomorphism structure;

(ii) **Loss of descent**: The descent lemma coefficient $1 - \frac{\eta L_C}{2}$ becomes negative, and f is no longer a Lyapunov function—oscillation or divergence may occur;

(iii) **Saddle avoidance breakdown**: The stable manifold arguments require the iteration map to be a local diffeomorphism; Cayley singularities break this chain.

Within the valid regime, the rational structure of Cayley poses no obstruction.

C.8.5 Explicit Escape Rate Analysis

This subsection provides **explicit bounds** for saddle escape using the Givens structure from Lemma C.11, replacing the generic PRGD $\log^4$ complexity with a **single logarithm**.

Lemma C.28 (Explicit Negative Curvature Constant)**.** *Let $f_{\text{enter}} := (g - \underline{\delta})^2/8$ be the Lyapunov-to-domain threshold. At any saddle-type critical point M with $f(M) \geq f_{\text{enter}}$:*

$$\lambda_{\min}(\mathrm{Hess}\, f(M)) \leq -\gamma_{\text{enter}}, \qquad \gamma_{\text{enter}} := \frac{(g - \underline{\delta})^2}{n(n-1)}, \tag{67}$$

where g is the spectral gap of C_e.

Proof. At a critical point with $f(M) \geq f_{\text{enter}}$, using $f(M) = \sum_{i<j} A_{ij}^2$:

$$b_{\max}^2 := \max_{i<j} A_{ij}^2 \geq \frac{2f(M)}{n(n-1)} \geq \frac{2f_{\text{enter}}}{n(n-1)} = \frac{(g - \underline{\delta})^2}{4n(n-1)}.$$

By Lemma C.11, $\lambda_{\min} \leq -4b_{\max}^2 \leq -\gamma_{\text{enter}}$. □

Lemma C.29 (Directional Second Derivative along Givens)**.** *For any $M \in SO(n)$, $A := M^\top C_e M$, and Givens direction $\Xi = E_{ij} - E_{ji}$:*

$$\left.\frac{d^2}{dt^2}\right|_{t=0} f(Me^{t\Xi}) = 2\left((A_{jj} - A_{ii})^2 - 4A_{ij}^2\right). \tag{68}$$

At degenerate blocks ($A_{ii} = A_{jj}$), this reduces to $-8A_{ij}^2$ (consistent with Lemma C.11).

Proof. From Lemma C.11, $A_{ij}(t) = b\cos(2t) + \frac{1}{2}\Delta\sin(2t)$ where $b := A_{ij}$ and $\Delta := A_{jj} - A_{ii}$. Direct differentiation gives the result. □

Proposition C.30 (Tubular Neighborhood and Escape Time)**.** *Define the negative-curvature preserving radius:*

$$r_{\mathrm{nc}} := \frac{g - \underline{\delta}}{32\|C_e\|_2\sqrt{n(n-1)}}. \tag{69}$$

Within $B(\bar{M}, r_{\mathrm{nc}})$ around any saddle center $\bar{M}$, the local negative curvature satisfies:

$$\gamma_{\mathrm{loc}} := \frac{15}{32} \cdot \frac{(g - \underline{\delta})^2}{n(n-1)}.$$

For the discrete Cayley iteration with $\eta \leq 1/(2L_C)$, the escape time from the saddle neighborhood is:

$$T_{\mathrm{esc}} \leq \frac{2}{\eta\gamma_{\mathrm{loc}}} \log \frac{r_*}{w}, \tag{70}$$

where $r_* := \min\{r_{\mathrm{nc}}, \gamma_{\mathrm{loc}}/(16L_C)\}$ and w is the initial distance from the stable manifold.

Proof Sketch. **Step 1 (Lipschitz control):** Along geodesics, $|\dot{A}_{ij}| \leq 2\|C_e\|_2\|\Xi\|_F$. Within radius r_{nc}, diagonal separation changes by $\leq 4\|C_e\|_2 r_{\mathrm{nc}}$, ensuring the degenerate block structure persists with $|b| \geq b_0/2$.

Step 2 (Exponential growth): In local exponential coordinates $\xi = \alpha u + \beta$ (where u is the unstable Givens direction), the unstable component satisfies:

$$|\alpha_{k+1}| \geq \left(1 + \frac{\eta\gamma_{\mathrm{loc}}}{2}\right)|\alpha_k|.$$

Step 3 (Escape time): Growth to $|\alpha| \sim r_*$ requires:

$$T_{\mathrm{esc}} \leq \frac{2}{\eta\gamma_{\mathrm{loc}}} \log \frac{r_*}{|\alpha_0|} \leq \frac{2}{\eta\gamma_{\mathrm{loc}}} \log \frac{r_*}{w}.$$ □

Remark C.31 (Comparison with Generic PRGD). The escape time equation 70 depends only logarithmically on the failure probability (via $w \propto \zeta$), eliminating the $\log^4$ factor in generic perturbed gradient descent analyses. This improvement stems from the **explicit Givens direction** (Lemma C.11) and the **exact** $\sin^2$ **profile** equation 55, which bypass the need for Hessian-Lipschitz constants and multi-stage perturbation protocols.

C.8.6 Distinction from Brockett's Double-Bracket Flow

Remark C.32 (Key Difference from Classical Theory). Brockett's (1991) double-bracket flow is $\dot{X} = [X, [X, N]]$ where N is a **fixed** diagonal matrix. This flow diagonalizes X on the isospectral manifold.

Our flow $\dot{M} = M[A, D]$ where $A = M^\top C_e M$ and $D = \mathrm{diag}(A)$ has D **varying with** M. This is **not** a direct application of Brockett's fixed-N theory.

However, the key insight is that our flow is still a **gradient-type system** on the compact manifold $SO(n)$ for the analytic function $f(M) = \frac{1}{2}\|\mathrm{off}(M^\top C_e M)\|_F^2$. The global convergence proof uses:

1. **Analytic gradient flow ⇒ convergence to single critical point** (Łojasiewicz)
2. **Strict saddle landscape ⇒ random init avoids saddles** (measure-zero stable manifolds)

This route is **insensitive** to whether D is fixed or varies—it depends only on the landscape structure of f.

C.9 Stiefel Manifold Extension

Proposition C.33 (Shift-Invariance on Stiefel Manifolds)**.** *Let $\mathrm{St}(k, n) = \{M \in \mathbb{R}^{n\times k} : M^\top M = I_k\}$ be the Stiefel manifold. Consider the Brockett-type objective:*

$$f(M) = \mathrm{tr}(M^\top CMN), \tag{71}$$

where $C \in \mathbb{R}^{n\times n}$ is symmetric and $N \in \mathbb{R}^{k\times k}$ is a fixed antisymmetric matrix ($N^\top = -N$). Then the Riemannian gradient is exactly shift-invariant:

$$\nabla_{\mathrm{St}} f|_{C+\alpha I} = \nabla_{\mathrm{St}} f|_C, \quad \forall \alpha \in \mathbb{R}. \tag{72}$$

Consequently, any retraction-based update $M_{k+1} = \mathcal{R}_{M_k}(-\eta \nabla_{\mathrm{St}} f)$ produces trajectories that are stepwise σ^2-invariant.

Proof sketch. The Euclidean gradient is $\nabla_E f = CM(N + N^\top) + C^\top M(N^\top + N)$. For antisymmetric N, we have $N + N^\top = 0$, so the Euclidean gradient simplifies to:

$$\nabla_E f = CM \cdot 0 + C^\top M \cdot 0 = 0 \quad \text{(naively)}.$$

This is incorrect; the correct Euclidean gradient for trace objective is:

$$\nabla_E f = (C + C^\top)MN = 2CMN \quad \text{(for symmetric } C\text{)}.$$

The Riemannian gradient on Stiefel is the tangent projection:

$$\nabla_{\mathrm{St}} f = (I - MM^\top)\nabla_E f + M \cdot \mathrm{skew}(M^\top \nabla_E f).$$

Under the shift $C \to C + \alpha I$:

$$\nabla_E f|_{C+\alpha I} = 2(C + \alpha I)MN = 2CMN + 2\alpha MN, \tag{73}$$

$$(I - MM^\top)(2\alpha MN) = 2\alpha MN - 2\alpha M(M^\top M)N = 2\alpha MN - 2\alpha MN = 0, \tag{74}$$

where we used $M^\top M = I_k$. The skew part also vanishes:

$$\mathrm{skew}(M^\top \cdot 2\alpha MN) = \mathrm{skew}(2\alpha N) = 2\alpha N \quad \text{(since } N \text{ is already skew)}.$$

This contributes $M \cdot 2\alpha N$ to the tangent, but this represents rotation within the subspace spanned by M, not a change in the subspace itself. For top-k subspace tracking where only the column space of M matters (not the specific basis), this term can be absorbed into the equivalence class $[M] \in \mathrm{Gr}(k, n)$.

More directly: the objective $f(M) = \mathrm{tr}(M^\top CMN)$ under $C \to C + \alpha I$ becomes:

$$f|_{C+\alpha I} = \mathrm{tr}(M^\top CMN) + \alpha \mathrm{tr}(M^\top MN) = f|_C + \alpha \mathrm{tr}(N).$$

Since N is antisymmetric, $\mathrm{tr}(N) = 0$, so $f|_{C+\alpha I} = f|_C$ exactly. The gradients coincide because the objectives differ by at most a constant (which is zero here). □

Remark C.34 (Practical Implications). For top-k eigenspace tracking with $k \ll n$:

- Per-iteration cost reduces from $O(n^2)$ to $O(kn)$, enabling scalability to $n \sim 10^4$.
- The matrix-free property is preserved: only k matrix-vector products per iteration.
- The σ^2-immunity extends directly from the algebraic identity $\mathrm{tr}(N) = 0$.

Full implementation details are deferred to future work.

D Discrete Algorithm and Stability Analysis

D.1 Proof of Lemma 3.1: Discrete Lyapunov Descent

Lemma D.1 (Discrete Lyapunov Descent)**.** *Consider the update $M_{k+1} = M_k \cdot \mathrm{Cay}(\eta\Omega_k)$. There exists a constant $L_C = c_n \|C_e\|_2^2$ (where c_n depends only on n) such that when $\eta < 1/L_C$:*

$$f(M_{k+1}) \le f(M_k) - \eta\left(1 - \frac{\eta L_C}{2}\right)\|\Omega_k\|_F^2. \tag{75}$$

Proof. **Step 1: Taylor expansion of the Cayley map.** The Cayley map satisfies $\mathrm{Cay}(\eta X) = I + \eta X + O(\eta^2\|X\|^2)$ for small η. Hence:

$$M_{k+1} = M_k(I + \eta\Omega_k + O(\eta^2\|\Omega_k\|^2)).$$

Step 2: Expansion of $f(M_{k+1})$. By Taylor expansion around M_k:

$$f(M_{k+1}) = f(M_k) + \langle \nabla f(M_k), M_{k+1} - M_k\rangle + O(\|M_{k+1} - M_k\|^2).$$

The Riemannian gradient satisfies $\operatorname{grad} f(M_k) = -c_0 M_k\Omega_k$ for some $c_0 > 0$ (see Appendix C), so the gradient flow $\dot{M} = -\operatorname{grad} f$ corresponds to $\dot{M} = c_0 M\Omega$. Since $M_{k+1} - M_k = M_k\eta\Omega_k + O(\eta^2)$:

$$\langle \operatorname{grad} f(M_k), M_{k+1} - M_k\rangle = -c_0\eta\|\Omega_k\|_F^2 + O(\eta^2).$$

Combined with the continuous-time energy dissipation $\dot{f} = -c_0\|\Omega\|_F^2$, we obtain:

$$f(M_{k+1}) = f(M_k) - \eta\|\Omega_k\|_F^2 + O(\eta^2\|\Omega_k\|_F^2).$$

Step 3: Lipschitz constant. By the smoothness analysis in Appendix C, the Riemannian gradient has Lipschitz constant $L_C = c_n\|C_e\|_2^2$ with $c_n = O(\sqrt{n})$. This gives:

$$f(M_{k+1}) \le f(M_k) - \eta\|\Omega_k\|_F^2 + \frac{L_C\eta^2}{2}\|\Omega_k\|_F^2 = f(M_k) - \eta\left(1 - \frac{\eta L_C}{2}\right)\|\Omega_k\|_F^2.$$

□

D.2 Corollary: Step Size Condition

Corollary D.2 (Step Size Condition)**.** *To ensure the trajectory remains in $\mathcal{N}_{\underline{\delta}}$, a sufficient condition is:*

$$\eta < c \cdot \frac{\underline{\delta}^2}{\|C_e\|_2^2}, \tag{76}$$

where $c > 0$ is a geometric constant.

Proof. Combining Lemma 3.1 with the local PL condition (Lemma 2.1), monotonic descent of f is guaranteed when $\eta < 1/L_C$. The domain invariance follows from the non-escape condition (Theorem E.7 in Appendix E). □

D.3 Theorem 3: Discrete ISS (Summary)

Theorem D.3 (Discrete ISS)**.** *With step size $\eta < 1/L_C$ and input bound $\sup_k \|U_k\|_F \le \bar{U}$:*

$$y_{k+1} \le (1 - c_1\underline{\delta}^2\eta)y_k + c_2\|C_e\|_2\eta\bar{U}, \quad \limsup_{k\to\infty}\sqrt{f(M_k)} \le \frac{c_2\|C_e\|_2}{c_1\underline{\delta}^2}\bar{U}. \tag{77}$$

See Section D.4 for the complete pathwise analysis.

D.4 Discrete ISS: Pathwise Analysis

This subsection extends the continuous-time ISS analysis to discrete Cayley iteration, providing a deterministic (pathwise) ultimate bound.

D.4.1 D4.1 Problem Setup

Consider the perturbed discrete update:

$$M_{k+1} = M_k \cdot \mathrm{Cay}(\eta\widetilde{\Omega}_k), \quad \widetilde{\Omega}_k := \Omega_k + U_k,$$

where $\Omega_k = [A_k, D_k]$ is the ideal generator and $U_k \in \mathfrak{so}(n)$ is the tangent perturbation (from observation error E_k).

Assumptions:

1. **Local spectral separation:** Trajectory stays in $\mathcal{N}_{\underline{\delta}} := \{M : \delta(M) \geq \underline{\delta}\}$;
2. **Bounded input:** $\sup_k \|U_k\|_F \leq \bar{U}$ (verifiable via Lemma 1.2);
3. **Step size condition:** $\eta < 1/L_C$, where $L_C = c_n\|C_e\|_2^2$ (Lemma 3.1).

D.4.2 D4.2 Discrete ISS Recursion Inequality

Lemma D.4 (Discrete ISS Recursion)**.** *Under the above assumptions, let $y_k := \sqrt{f(M_k)}$. There exist constants $c_1, c_2 > 0$ (depending only on n) such that:*

$$\boxed{y_{k+1} \leq (1 - c_1\underline{\delta}^2\eta)y_k + c_2\|C_e\|_2\eta\bar{U}.} \tag{78}$$

Proof. **Step 1 (Perturbed descent lemma):** Extend Lemma 3.1 to the perturbed case. By the Cayley map's local smoothness, when $\eta < 1/L_C$:

$$f(M_{k+1}) \leq f(M_k) - \eta\left(1 - \frac{\eta L_C}{2}\right)\|\Omega_k\|_F^2 + c_3\eta\|\Omega_k\|_F\|U_k\|_F + c_4\eta^2\|U_k\|_F^2,$$

where c_3, c_4 are geometric constants.

Step 2 (Apply local PL): By Lemma 2.1, $\|\Omega_k\|_F^2 \geq 2\underline{\delta}^2 f_k = 2\underline{\delta}^2 y_k^2$. By Lemma 2.2, $\|\Omega_k\|_F \leq 2\sqrt{2}\|C_e\|_2 y_k$.

Step 3 (Substitute and simplify): Substituting into Step 1, for small η:

$$f_{k+1} \leq f_k - c_1'\underline{\delta}^2\eta y_k^2 + c_2'\|C_e\|_2\eta y_k\bar{U} + O(\eta^2).$$

Step 4 ($\sqrt{f}$ recursion): Using $y_{k+1}^2 - y_k^2 = f_{k+1} - f_k$ and $(y_{k+1} - y_k)(y_{k+1} + y_k) = f_{k+1} - f_k$, when $y_k > 0$:

$$y_{k+1} - y_k = \frac{f_{k+1} - f_k}{y_{k+1} + y_k}.$$

When $f_{k+1} \leq f_k$ (descent regime), we have $y_{k+1} \leq y_k$, so $y_{k+1} + y_k \leq 2y_k$, giving:

$$y_{k+1} - y_k \geq \frac{f_{k+1} - f_k}{2y_k} \geq -c_1\underline{\delta}^2\eta y_k + c_2\|C_e\|_2\eta\bar{U}.$$

The second inequality follows from Step 3. Rearranging gives $y_{k+1} \leq (1 - c_1\underline{\delta}^2\eta)y_k + c_2\|C_e\|_2\eta\bar{U}$. □

D.4.3 D4.3 Ultimate Noise Ball (Pathwise)

Theorem D.5 (Discrete Pathwise Noise Ball)**.** *Under D4.1's assumptions, if step size satisfies $\eta < \min\{1/L_C, 1/(c_1\underline{\delta}^2)\}$:*

(i) Exponential convergence to noise ball:

$$y_k \leq (1 - c_1\underline{\delta}^2\eta)^k y_0 + \frac{c_2\|C_e\|_2}{c_1\underline{\delta}^2}\bar{U}.$$

(ii) Ultimate noise ball radius:

$$\boxed{\limsup_{k\to\infty} \sqrt{f(M_k)} \le r_f^{\text{disc}}(\bar{U}) := \frac{c_2\|C_e\|_2}{c_1\underline{\delta}^2}\bar{U}.} \tag{79}$$

(iii) Conversion to M-distance: *Using $V \le \frac{16}{g^2} f$,*

$$\limsup_{k\to\infty} \|M_k - M^*\|_F \le \frac{4}{g} \cdot r_f^{\text{disc}}(\bar{U}).$$

Proof. Iterate the recursion inequality:

$$y_k \le (1 - c_1\underline{\delta}^2\eta)^k y_0 + c_2\|C_e\|_2\eta\bar{U}\sum_{j=0}^{k-1}(1 - c_1\underline{\delta}^2\eta)^j.$$

Geometric series sum: $\sum_{j=0}^{k-1}(1 - c_1\underline{\delta}^2\eta)^j \le \frac{1}{c_1\underline{\delta}^2\eta}$. Substituting gives the result. □

Remark D.6 (Consistency with Continuous Case). Both discrete radius $r_f^{\text{disc}} = \frac{c_2\|C_e\|_2}{c_1\underline{\delta}^2}\bar{U}$ and continuous radius $r_f^{\text{cont}} = \frac{\sqrt{2}\|C_e\|_2}{\underline{\delta}^2}\bar{u}$ share identical structure: numerator $\|C_e\|_2$, denominator $\underline{\delta}^2$, with $\sigma^2 I$ absent.

Corollary D.7 (Explicit Noise Ball Radius)**.** *By Lemma 1.2, if $\sup_k \|E_{e,k}\|_F \le \varepsilon$, then $\bar{U} \le 4\|C_e^{\text{sig}}\|_2\varepsilon + 2\varepsilon^2$, giving:*

$$r_f^{\text{disc}}(\varepsilon) \le \frac{c_2\|C_e\|_2}{c_1\underline{\delta}^2}\left(4\|C_e^{\text{sig}}\|_2\varepsilon + 2\varepsilon^2\right).$$

D.5 Convergence Rate Analysis

D.5.1 Additional Assumptions

Assumption D.8 (Finite Variance)**.** The random input U_k is conditionally zero-mean with bounded variance:

$$\mathbb{E}[\|U_k\|_F^2 \mid \mathcal{F}_k] \le \sigma_u^2.$$

By Theorem 2 (Appendix C), in the spectrally separated domain $\mathcal{N}_{\underline{\delta}}$:

$$\|\Omega(M)\|_F^2 \ge 2\underline{\delta}^2 f(M).$$

This is equivalent to a local Polyak-Łojasiewicz (PL) condition with parameter $\mu = \underline{\delta}^2$.

D.5.2 Convergence Rate Theorem

Corollary D.9 ($O(1/k)$ Convergence Rate)**.** *Suppose $M_0 \in \mathcal{N}_{\underline{\delta}}$ and the trajectory remains in this domain with probability 1 (guaranteed by Lemma E.6 in Appendix E). Using decaying step size $\eta_k = c/(k + k_0)$ with $c > 1/(2\underline{\delta}^2)$ and k_0 large enough to ensure $\eta_0 < 1/L_C$:*

$$\boxed{\mathbb{E}[f(M_k)] \le \frac{C}{k + k_0}, \quad C = \max\left\{k_0 f(M_0), \frac{c^2 L_C \sigma_u^2}{2c\underline{\delta}^2 - 1}\right\}.} \tag{80}$$

Proof. **Step 1 (Recursion inequality):** Using the smoothness from Theorem 3 ($L_C = c_n\|C_e\|_2^2$) and taking conditional expectation:

$$\mathbb{E}[f_{k+1} \mid \mathcal{F}_k] \le f_k - \eta_k\|\Omega_k\|_F^2 + \frac{L_C\eta_k^2}{2}(\|\Omega_k\|_F^2 + \sigma_u^2). \tag{81}$$

When η_k is small enough that $1 - L_C\eta_k/2 > 0$, applying the local PL condition $\|\Omega_k\|_F^2 \geq 2\underline{\delta}^2 f_k$:

$$\mathbb{E}[f_{k+1} \mid \mathcal{F}_k] \leq (1 - 2\underline{\delta}^2\eta_k)f_k + \frac{L_C\eta_k^2\sigma_u^2}{2}.$$

Step 2 (Recursive solution): Substitute $\eta_k = c/(k + k_0)$ and use mathematical induction to prove $\mathbb{E}[f_k] \leq C/(k + k_0)$. The key is choosing $c > 1/(2\underline{\delta}^2)$ so that the linear descent term dominates the quadratic noise term. □

D.5.3 Constant Advantage

Although the $O(1/k)$ rate matches standard SGD, the leading constant has an essential advantage.

Standard SGD constant: Convergence bounds typically depend on the total gradient variance:

$$\sigma_{\text{SGD}}^2 \propto \mathbb{E}[\|C_k\|_F^2] \geq \sigma_k^4 \cdot n \quad (\text{when } C_k = C_\star + \sigma_k^2 I + \Delta_k).$$

This algorithm's constant: By Lemma 1.2 (Appendix B), the effective noise U_k satisfies:

$$\|U_k\|_F \leq 4\|C_e^{\text{sig}}\|_2\|E_{e,k}\|_F + 2\|E_{e,k}\|_F^2,$$

where $E_{e,k} = E_k - \frac{1}{n}\text{tr}(E_k)I$ is the **trace-free part** of the noise matrix. Since the trace-free operator completely eliminates scalar matrices:

$$\text{tf}(\sigma^2 I) = \sigma^2 I - \sigma^2 I = 0.$$

Therefore, the variance constant σ_u^2 in Corollary D.9 satisfies:

$$\sigma_u^2 \propto \mathbb{E}[\|E_{e,k}\|_F^2] \ll \mathbb{E}[\|C_k\|_F^2].$$

In high-noise environments, the effective variance $\sigma_u^2 \propto \mathbb{E}[\|E_{e,k}\|_F^2] \ll \mathbb{E}[\|C_k\|_F^2]$ gives a significant constant improvement over baselines.

D.5.4 Sample Complexity and Trace Independence

Corollary D.10 (Sample Complexity)**.** *To achieve target precision ε (i.e., $\mathbb{E}[f(M_k)] \leq \varepsilon$), the required discrete sampling count k satisfies:*

$$\boxed{k \geq \frac{C}{\varepsilon} = O\left(\frac{L_C\sigma_u^2}{\underline{\delta}^2 \cdot \varepsilon}\right) = O\left(\frac{\|C_e\|_2^2 \cdot \mathbb{E}[\|E_e\|_F^2]}{\underline{\delta}^2 \cdot \varepsilon}\right).} \tag{82}$$

Remark D.11 (Trace Independence). The complexity bound contains no σ^2; sample efficiency is identical for any isotropic noise level.

D.6 Sample Complexity

Corollary D.12 (Sample Complexity)**.** *Using decaying step size $\eta_k = c/(k + k_0)$ with $c > 1/(2\underline{\delta}^2)$ and $k_0 > c \cdot c_n\|C_e\|_2^2$, to achieve $\mathbb{E}[f(M_k)] \leq \varepsilon$:*

$$k^*(\varepsilon) = O\left(\frac{\|C_e\|_2^2 \cdot \mathbb{E}[\|E_e\|_F^2]}{\underline{\delta}^2 \cdot \varepsilon}\right). \tag{83}$$

All quantities are independent of σ^2.

D.7 Diffusion Limit and Noise-Induced Drift

This subsection analyzes the continuous-time limit of discrete Cayley iteration under stochastic perturbations, establishing that Cayley's nonlinearity does not introduce steady-state bias.

D.7.1 Cayley Map Second-Order Expansion

For small $X \in \mathfrak{so}(n)$, the Cayley map expands as:

$$\mathrm{Cay}(X) = (I - \tfrac{1}{2}X)^{-1}(I + \tfrac{1}{2}X) = I + X + \tfrac{1}{2}X^2 + O(\|X\|^3). \tag{84}$$

This matches the exponential map to second order, so diffusion scaling produces identical limits.

D.7.2 Noise Covariance and Itô Correction

Consider stochastic perturbations with $\mathfrak{so}(n)$-valued Brownian motion B_t satisfying:

$$\mathbb{E}[\mathrm{d}B_{ij}\,\mathrm{d}B_{kl}] = \nu_W(\delta_{ik}\delta_{jl} - \delta_{il}\delta_{jk})\,\mathrm{d}t. \tag{85}$$

Lemma D.13 (Quadratic Covariance)**.** *Under the above covariance structure:*

$$\mathbb{E}[(\mathrm{d}B_t)^2] = -\nu_W(n-1)I\,\mathrm{d}t. \tag{86}$$

Proof. Direct computation: $\mathbb{E}[(\mathrm{d}B)^2]_{ik} = \sum_j \mathbb{E}[\mathrm{d}B_{ij}\,\mathrm{d}B_{jk}] = \sum_j \nu_W(\delta_{ij}\delta_{jk} - \delta_{ik})\,\mathrm{d}t = -\nu_W(n-1)\delta_{ik}\,\mathrm{d}t.$ □

D.7.3 Stratonovich vs Itô Formulation

The continuous limit of discrete Cayley iteration with noise $\varepsilon\,\mathrm{d}B_t$ is:

Stratonovich (intrinsic):

$$\mathrm{d}M_t = M_t\Omega(M_t)\,\mathrm{d}t + \varepsilon M_t \circ \mathrm{d}B_t. \tag{87}$$

Equivalent Itô form:

$$\mathrm{d}M_t = M_t\Omega(M_t)\,\mathrm{d}t + \varepsilon M_t\,\mathrm{d}B_t - \frac{\varepsilon^2\nu_W(n-1)}{2}M_t\,\mathrm{d}t. \tag{88}$$

Remark D.14 (Geometric Interpretation). The Itô correction $-\frac{\varepsilon^2\nu_W(n-1)}{2}M_t$ is a **normal** (symmetric) direction: $M_t \cdot (-\kappa I)$ with $\kappa > 0$. It does not enter the tangent space $T_M SO(n) = \{M\Xi : \Xi \in \mathfrak{so}(n)\}$, hence does not affect the intrinsic rotational dynamics or introduce steady-state bias.

D.7.4 Steady-State Ball from Diffusion Analysis

Using the Stratonovich generator $\mathcal{L}f = \langle \mathrm{grad}\, f, M\Omega\rangle + \frac{\varepsilon^2}{2}\Delta f$ with Laplace–Beltrami $\Delta f \le d \cdot L_C$ (where $d = n(n-1)/2$):

Proposition D.15 (Diffusion Steady-State Bound)**.** *In the spectrally separated domain $\mathcal{N}_{\underline{\delta}}$:*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[f(M_t)] \le -2\underline{\delta}^2\mathbb{E}[f(M_t)] + \frac{\varepsilon^2 n(n-1)c_n\|C_e\|_2^2}{4}. \tag{89}$$

Hence:

$$\limsup_{t\to\infty}\mathbb{E}[f(M_t)] \le \frac{\varepsilon^2 n(n-1)c_n\|C_e\|_2^2}{8\underline{\delta}^2}. \tag{90}$$

Remark D.16 (Consistency with Discrete ISS). The steady-state radius scales as $O(\varepsilon\|C_e\|_2/\underline{\delta})$, matching the discrete ISS bound structure. Crucially, **no σ^2 appears**: the noise enters only through ε (observation error amplitude) and $\|C_e\|_2$ (trace-free signal norm).

D.8 Efficient Matrix-Free Implementation via Neumann Series

In the matrix-free setting, explicit computation of the Cayley inverse $(I - \frac{\eta}{2}\Omega)^{-1}$ is prohibitive ($O(n^3)$ complexity). This section establishes that a low-order Neumann series truncation (3–4 terms) provides sufficient accuracy relative to the statistical noise floor, preserving the $O(n^2)$ complexity claim.

D.8.1 Neumann Series Expansion

The Cayley map $\mathrm{Cay}(X) = (I - \frac{X}{2})^{-1}(I + \frac{X}{2})$ can be approximated via Neumann series. For $X = \eta\Omega$ with spectral radius $\rho(\frac{\eta\Omega}{2}) < 1$:

$$\left(I - \frac{\eta\Omega}{2}\right)^{-1} = \sum_{j=0}^{\infty} \left(\frac{\eta\Omega}{2}\right)^j \approx \sum_{j=0}^{K} \left(\frac{\eta\Omega}{2}\right)^j. \tag{91}$$

The truncated Cayley update becomes:

$$\mathrm{Cay}^{(K)}(\eta\Omega) := \left[\sum_{j=0}^{K} \left(\frac{\eta\Omega}{2}\right)^j\right] \cdot \left(I + \frac{\eta\Omega}{2}\right). \tag{92}$$

D.8.2 Convergence Factor Bound

Lemma D.17 (Neumann Convergence under Step Size Constraint)**.** *Under the step size condition $\eta < 1/L_C$ with $L_C = c_n\|C_e\|_2^2$ (Lemma 3.1), the Neumann series convergence factor satisfies:*

$$\left\|\frac{\eta\Omega}{2}\right\|_2 \leq \frac{\|\Omega\|_F}{2L_C} \leq \frac{2\sqrt{2}\|C_e\|_2\sqrt{f}}{2c_n\|C_e\|_2^2} = O\left(\frac{\sqrt{f}}{\|C_e\|_2}\right). \tag{93}$$

In particular, when $f \leq r_f^{\mathrm{disc}}$ (within the ISS noise ball), the convergence factor is bounded by a constant less than 1, ensuring Neumann series convergence.

Proof. By Lemma 2.2 (Appendix C), $\|\Omega\|_F \leq 2\sqrt{2}\|C_e\|_2\sqrt{f}$. Substituting into the step size constraint $\eta < 1/L_C$:

$$\left\|\frac{\eta\Omega}{2}\right\|_2 \leq \frac{\eta\|\Omega\|_F}{2} < \frac{\|\Omega\|_F}{2L_C} = \frac{\|\Omega\|_F}{2c_n\|C_e\|_2^2}.$$

For practical convergence (initialization and during optimization), when $\sqrt{f} \lesssim \|C_e\|_2$, the factor is $O(1/c_n) < 1$. □

D.8.3 Truncation Order Analysis

Proposition D.18 (Truncation Error Bound)**.** *Let $\rho := \|\frac{\eta\Omega}{2}\|_2 < 1$ be the convergence factor. The K-th order truncation error satisfies:*

$$\left\|\mathrm{Cay}(\eta\Omega) - \mathrm{Cay}^{(K)}(\eta\Omega)\right\|_F \leq \frac{\rho^{K+1}}{1-\rho} \cdot (1+\rho). \tag{94}$$

Under the step size constraint $\eta < 1/L_C$ with $\rho < 1$:

Order K	*Truncation Error*	*Relative to Noise Floor*
2	$O(2^{-3}) \approx 0.125$	*Marginal*
3	$O(2^{-4}) \approx 0.06$	*Acceptable*
4	$O(2^{-5}) \approx 0.03$	*Sufficient*
5	$O(2^{-6}) \approx 0.015$	*Conservative*

Proof. The Neumann series remainder is $\sum_{j=K+1}^{\infty} \rho^j = \frac{\rho^{K+1}}{1-\rho}$. Multiplying by $(I + \frac{\eta\Omega}{2})$ contributes at most a factor $(1+\rho)$. □

D.8.4 Closed-Form Truncation Identity

The following lemma provides a more precise characterization of the truncation error.

Lemma D.19 (Closed-Form Truncation Error)**.** *Let $X = \frac{\eta}{2}\Omega$ with $\rho := \|X\|_2 < 1$. Then:*

$$\mathrm{Cay}(X) - \mathrm{Cay}^{(K)}(X) = X^{K+1} \cdot \mathrm{Cay}(X). \tag{95}$$

Consequently, since $\mathrm{Cay}(X) \in SO(n)$ for skew-symmetric X:

$$\|\mathrm{Cay}(X) - \mathrm{Cay}^{(K)}(X)\|_F = \|X^{K+1}\|_F \leq \sqrt{n}\,\rho^{K+1}. \tag{96}$$

Proof. Using the Neumann remainder identity $(I - X)^{-1} = S_K(X) + X^{K+1}(I - X)^{-1}$ where $S_K(X) = \sum_{j=0}^{K} X^j$, we have:

$$\mathrm{Cay}(X) = (I - X)^{-1}(I + X) = S_K(X)(I + X) + X^{K+1}(I - X)^{-1}(I + X) = \mathrm{Cay}^{(K)}(X) + X^{K+1}\mathrm{Cay}(X).$$

The norm bound follows from $\|\mathrm{Cay}(X)\|_F = \sqrt{n}$ (orthogonal matrix) and $\|X^{K+1}\|_2 \leq \rho^{K+1}$. □

Corollary D.20 (Orthogonality Defect under Truncation)**.** *Let $Q_K := \mathrm{Cay}^{(K)}(X)$ and $Q := \mathrm{Cay}(X) \in SO(n)$. Then $Q_K = (I - X^{K+1})Q$, and the orthogonality defect satisfies:*

$$\|Q_K^\top Q_K - I\|_F \leq 2\sqrt{n}\,\rho^{K+1} + n\,\rho^{2K+2}. \tag{97}$$

Moreover, if $M_k^\top M_k = I$, then $\|M_{k+1}^{(K)\top} M_{k+1}^{(K)} - I\|_F = O(\rho^{K+1})$.

Refinement for even K: *When K is even, $K + 1$ is odd, so X^{K+1} remains skew-symmetric (since X is skew-symmetric). Thus $(X^{K+1})^\top = -X^{K+1}$, and:*

$$\|Q_K^\top Q_K - I\|_F = \|X^{K+1}\|_F^2 = O(\rho^{2K+2}), \tag{98}$$

upgrading the orthogonality defect from $O(\rho^{K+1})$ to the stronger $O(\rho^{2K+2})$.

Proof. From Lemma D.19, $Q_K = Q - X^{K+1}Q = (I - X^{K+1})Q$. Thus:

$$Q_K^\top Q_K = Q^\top (I - X^{K+1})^\top (I - X^{K+1})Q = I - (X^{K+1} + (X^{K+1})^\top) + (X^{K+1})^\top X^{K+1}.$$

The bound equation 97 follows from $\|X^{K+1} + (X^{K+1})^\top\|_F \leq 2\|X^{K+1}\|_F$. For even K, skew-symmetry of X^{K+1} implies the linear term vanishes. □

Proposition D.21 (ISS Stability under Truncation)**.** *Suppose the exact Cayley update satisfies the ISS recursion (Theorem 4.3):*

$$y_{k+1} \leq (1 - \alpha\eta)\, y_k + \beta\eta\,\bar{U}, \quad y_k := \sqrt{f(M_k)},$$

where $\alpha, \beta, \bar{U}$ are σ^2-independent. For the Neumann-truncated update $M_{k+1}^{(K)} = M_k \mathrm{Cay}^{(K)}(X_k)$, define:

$$\varepsilon_k := \|\mathrm{Cay}(X_k) - \mathrm{Cay}^{(K)}(X_k)\|_F = \|X_k^{K+1}\|_F \leq \sqrt{n}\,\rho_k^{K+1}.$$

Then there exists a constant $c_f > 0$ depending only on $(n, \|C_e\|_2)$ such that:

$$y_{k+1}^{(K)} \leq (1 - \alpha\eta)\, y_k^{(K)} + \beta\eta\,\bar{U} + c_f\,\varepsilon_k. \tag{99}$$

Consequently, the steady-state radius becomes:

$$\limsup_{k\to\infty} y_k^{(K)} \leq r_f(\bar{U}) + \frac{c_f}{\alpha\eta} \sup_k \varepsilon_k, \tag{100}$$

where all terms remain σ^2-independent (since ε_k depends only on the σ-free generator Ω_k).

Proof sketch. Write the truncated update as exact update plus perturbation:

$$M_{k+1}^{(K)} - M_{k+1} = M_k(\mathrm{Cay}^{(K)}(X_k) - \mathrm{Cay}(X_k)) = -M_k X_k^{K+1}\mathrm{Cay}(X_k),$$

so $\|M_{k+1}^{(K)} - M_{k+1}\|_F = \varepsilon_k$. By Lipschitz continuity of f in the controlled region $\mathcal{N}_{\underline{\delta}}$:

$$|f(M_{k+1}^{(K)}) - f(M_{k+1})| \leq c_f \varepsilon_k,$$

where c_f bounds $\|\nabla f\|_F$ and depends only on $\|C_e\|_2$ and n. Adding this to the exact recursion yields equation 99. □

Remark D.22 (σ^2-Invariance under Truncation). The truncated update $\mathrm{Cay}^{(K)}$ takes as input only $\Omega = [A, \mathrm{diag}(A)]$. Since $[A + \sigma^2 I, \mathrm{diag}(A + \sigma^2 I)] = [A, \mathrm{diag}(A)]$ for any $\sigma^2 \in \mathbb{R}$, the generator Ω is exactly σ-free. The truncation is a deterministic function of this σ-free input, so trajectories remain stepwise σ^2-invariant.

Remark D.23 (Statistical Irrelevance of Higher-Order Terms). In stochastic settings, the observation noise E_k contributes perturbations of order $O(\|E_e\|_F)$ to each step. When $\|E_e\|_F \gtrsim 0.01$, truncation errors below 10^{-2} are **statistically dominated** by observation noise. Thus, computing exact inverses or high-order approximations provides no statistical benefit—it is "computational waste" in the sense that the additional precision is masked by inherent randomness.

D.8.5 Matrix-Free Algorithm

Algorithm 2 Matrix-Free Cayley Update via Neumann Series

Require: Current iterate M_k, observation C_k, step size η, Neumann order K (default: 3–4)
Ensure: Next iterate $M_{k+1} \in SO(n)$

```
A_k ← M_k^⊤ C_k M_k {O(n) MVPs + O(n^3) arithmetic}
D_k ← diag(A_k)
Ω_k ← A_k D_k − D_k A_k {Commutator, O(n^2) (diagonal scaling)}
X ← (η/2) Ω_k
P ← I_n, S ← I_n {Power accumulator and sum}
for j = 1 to K do
    P ← P · X {O(n^3) matrix multiplication}
    S ← S + P
end for
M_{k+1} ← M_k · S · (I + X) {O(n^3) matrix multiplications}
Optional: QR re-orthogonalization every T ∼ 50–100 steps {Maintains SO(n); σ^2-immunity preserved}
return M_{k+1}
```

D.8.6 Complexity Comparison

Method	MVPs	Arithmetic	Matrix-Free?	Exact?
Direct Cayley (LU/Cholesky)	$O(n)$	$O(n^3)$	No	Yes
Neumann-K Cayley	$O(n)$	$O(n^3)$	**Yes**	$O(\rho^K)$ error
QR Retraction	$O(n)$	$O(n^3)$	No	Yes
Polar Retraction (SVD)	$O(n)$	$O(n^3)$	No	Yes

Remark D.24 (Complexity Summary). All methods require $O(n)$ MVPs and $O(n^3)$ arithmetic per iteration. The Neumann approximation's advantage is **matrix-free operation**: it avoids explicit construction or factorization of C_k, requiring only MVP oracle access. With $K = 3$–4 and step size $\eta < 1/L_C$, truncation error is dominated by statistical noise.

E Proofs for Statistical Analysis

E.1 Lemma D.1: Weight Norm Bounds

The Lyapunov function $y_k := \sqrt{f(M_k)}$ satisfies the integral inequality:

$$y_N \leq e^{-\lambda N \eta} y_0 + \gamma_u \sum_{j=0}^{N-1} \alpha_j u_j,$$

where the convolution weights are:

$$\alpha_j := \int_{j\eta}^{(j+1)\eta} e^{-\lambda(N\eta - s)}\, ds = \frac{e^{-\lambda(N-j)\eta}(e^{\lambda\eta} - 1)}{\lambda}.$$

Lemma E.1 (Weight Norm Bounds)**.** *As $N \to \infty$:*

(i) Maximum norm: $\max_j \alpha_j = \dfrac{1 - e^{-\lambda\eta}}{\lambda} \leq \eta.$

(ii) Squared sum (energy gain): $\displaystyle\sum_{j=0}^{N-1} \alpha_j^2 \xrightarrow{N\to\infty} \frac{(e^{\lambda\eta} - 1)^2}{\lambda^2(e^{2\lambda\eta} - 1)} \leq \frac{\eta}{2\lambda}.$

Proof. **(i)** The weights are maximized at $j = N - 1$:

$$\alpha_{N-1} = \frac{e^{-\lambda\eta}(e^{\lambda\eta} - 1)}{\lambda} = \frac{1 - e^{-\lambda\eta}}{\lambda} \leq \eta.$$

(ii) Let $\beta := e^{-\lambda\eta} < 1$. Then $\alpha_j = \frac{(1-\beta)\beta^{N-1-j}}{\lambda}$, and:

$$\sum_{j=0}^{N-1} \alpha_j^2 = \frac{(1-\beta)^2}{\lambda^2} \sum_{m=0}^{N-1} \beta^{2m} \xrightarrow{N\to\infty} \frac{(1-\beta)^2}{\lambda^2(1-\beta^2)} = \frac{(1-\beta)}{\lambda^2(1+\beta)}.$$

Substituting $\beta = e^{-\lambda\eta}$ and using $(1 - e^{-x})/x \leq 1$:

$$\sum_j \alpha_j^2 = \frac{1 - e^{-\lambda\eta}}{\lambda^2(1 + e^{-\lambda\eta})} \leq \frac{\lambda\eta}{\lambda^2 \cdot 2} = \frac{\eta}{2\lambda}.$$

□

E.2 Lemma D.2: Probabilistic Error Bounds

E.2.1 Scenario I: Sub-Exponential Noise (Bernstein Regime) Vershynin (2018)

Assumption E.2 (IND-SE)**.** The sequence $\{u_k - \mu_k\}$ is mutually independent and sub-exponential with parameters (σ_u, b_u) Vershynin (2018):

$$\mathbb{E}[e^{\theta(u_k - \mu_k)}] \leq e^{\sigma_u^2 \theta^2/2}, \quad \forall |\theta| \leq 1/b_u.$$

Theorem E.3 (Exponential Concentration)**.** *Under Assumption (IND-SE), for any $\zeta \in (0, 1)$, the steady-state error satisfies with probability $\geq 1 - \zeta$:*

$$\sqrt{f_\infty} \leq \frac{\gamma_u}{\lambda}\bar{\mu} + \sigma_u \sqrt{\frac{\eta}{\lambda}} \cdot \sqrt{\log(1/\zeta)} + b_u \eta \cdot \log(1/\zeta). \tag{101}$$

Proof. Apply Bernstein's inequality to the weighted sum $S_N := \sum_{j=0}^{N-1} \alpha_j (u_j - \mu_j)$:

$$\mathbb{P}(S_N > t) \leq \exp\left(-\frac{t^2/2}{\sigma_u^2 \sum_j \alpha_j^2 + b_u t \max_j \alpha_j}\right).$$

Substituting the bounds from Lemma D.1: $\sum_j \alpha_j^2 \leq \eta/(2\lambda)$ and $\max_j \alpha_j \leq \eta$. Setting the RHS equal to ζ and solving for t yields the bound. □

E.2.2 Scenario II: Heavy-Tailed Noise (Chebyshev Regime)

Assumption E.4 (UNC-L2)**.** The sequence $\{u_k - \mu_k\}$ is pairwise uncorrelated with bounded variance $\mathrm{Var}(u_k) \le \sigma_u^2$.

Theorem E.5 (L^2 Energy Bound)**.** *Under Assumption (UNC-L2), for any tolerance $\epsilon > 0$:*

$$\mathbb{P}\left(\sqrt{f_\infty} > \frac{\gamma_u}{\lambda}\bar{\mu} + \epsilon\right) \le \frac{\sigma_u^2 \eta}{2\lambda\epsilon^2}. \tag{102}$$

Proof. By uncorrelatedness:

$$\mathrm{Var}(S_N) = \sum_{j=0}^{N-1} \alpha_j^2 \mathrm{Var}(u_j) \le \sigma_u^2 \sum_j \alpha_j^2 \le \frac{\sigma_u^2 \eta}{2\lambda}.$$

Apply Chebyshev's inequality: $\mathbb{P}(S_N - \mathbb{E}[S_N] > \epsilon) \le \mathrm{Var}(S_N)/\epsilon^2$. □

E.3 Lemma D.3: Non-Escape Condition and Domain Radius

Lemma E.6 (Domain Radius)**.** *Let $g = \min_{i \ne j} |\lambda_i(C_{\mathrm{sig}}) - \lambda_j(C_{\mathrm{sig}})|$. For any $\underline{\delta} < g$:*

$$R_{\underline{\delta}} \ge \frac{g - \underline{\delta}}{2\sqrt{2}\|C_e\|_2}. \tag{103}$$

Proof. **Step 1 (Diagonal perturbation bound).** Let $A = M^\top C_e M$ and $A^* = (M^*)^\top C_e M^* = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ where λ_i are the eigenvalues of C_e in decreasing order. At the target M^*: $\delta(M^*) = g$ and $f(M^*) = 0$.

For general M, write $A = A^* + (A - A^*)$. Since A and A^* share the same eigenvalues (both are orthogonally similar to C_e), the diagonal entries satisfy:

$$|A_{ii} - \lambda_i| \le \|A - A^*\|_F.$$

By the triangle inequality and Hoffman–Wielandt-type bounds for diagonal entries, we have:

$$\|A - A^*\|_F \le 2\|\mathrm{off}(A)\|_F = 2\sqrt{2f(M)}.$$

Step 2 (Diagonal separation bound). For any $i \ne j$:

$$|A_{ii} - A_{jj}| \ge |\lambda_i - \lambda_j| - |A_{ii} - \lambda_i| - |A_{jj} - \lambda_j| \ge |\lambda_i - \lambda_j| - 2\sqrt{2f(M)}.$$

Taking the minimum over all $i \ne j$:

$$\delta(M) = \min_{i \ne j} |A_{ii} - A_{jj}| \ge g - 2\sqrt{2f(M)}.$$

Step 3 (Domain radius). Setting $\delta(M) \ge \underline{\delta}$ gives $\sqrt{f(M)} \le (g - \underline{\delta})/(2\sqrt{2})$. Combined with the Lyapunov-to-distance relation $\sqrt{f} \le \|C_e\|_2 \cdot d(M, M^*)$, we obtain $R_{\underline{\delta}} = (g - \underline{\delta})/(2\sqrt{2}\|C_e\|_2)$. □

Theorem E.7 (Non-Escape Sufficient Condition)**.** *If the initial condition and noise level satisfy:*

$$\sqrt{f(M_0)} + r_f^{\mathrm{disc}}(\bar{U}) < R_{\underline{\delta}}. \tag{104}$$

then the trajectory never exits $\mathcal{N}_{\underline{\delta}}$.

Proof. By Theorem 3 (Appendix D), $\limsup_k \sqrt{f(M_k)} \le r_f^{\mathrm{disc}}(\bar{U})$. The monotonic convergence to the noise ball (after transient) combined with the domain radius bound ensures $\sqrt{f(M_k)} < R_{\underline{\delta}}$ for all k, hence $M_k \in \mathcal{N}_{\underline{\delta}}$. □

E.4 High-Probability Finite-Time Entry into $\mathcal{N}_{\underline{\delta}}$

This subsection addresses the gap identified in Remark G.4: providing a **high-probability finite-time bound** for entering the spectrally separated domain $\mathcal{N}_{\underline{\delta}}$ from arbitrary (Haar) initialization.

E.4.1 From Lyapunov Value to Domain Entry

Lemma E.8 (Lyapunov-to-Domain Threshold)**.** *Let $g = \min_{i \neq j} |\lambda_i(C_{\text{sig}}) - \lambda_j(C_{\text{sig}})|$ be the spectral gap. For any target threshold $\underline{\delta} \in (0, g)$:*

$$f(M) \leq f_{\text{enter}} := \frac{(g - \underline{\delta})^2}{8} \quad \Longrightarrow \quad M \in \mathcal{N}_{\underline{\delta}}. \tag{105}$$

In particular, for $\underline{\delta} = g/4$: $f_{\text{enter}} = \frac{9g^2}{128}$.

Proof. By Lemma E.6: $\delta(M) \geq g - 2\sqrt{2f(M)}$. Setting this $\geq \underline{\delta}$ and solving: $\sqrt{f(M)} \leq (g - \underline{\delta})/(2\sqrt{2})$, i.e., $f(M) \leq (g - \underline{\delta})^2/8$. □

E.4.2 Level Set Separation Lemma

The key to finite-time guarantees is converting the "strict saddle" property into quantitative (ε, γ)-detectability:

Lemma E.9 (Level Set Separation)**.** *Assume C has distinct eigenvalues and Proposition C.12 holds (all non-global-minimum critical points are strict saddles). For any $f_{\text{enter}} > 0$, there exist constants $\varepsilon_\star, \gamma_\star > 0$ (depending on C and f_{enter}) such that for all $M \in SO(n)$:*

$$f(M) \geq f_{\text{enter}} \quad \Longrightarrow \quad \|\text{grad}\, f(M)\|_F \geq \varepsilon_\star \;\; \textit{or} \;\; \lambda_{\min}(\text{Hess}\, f(M)) \leq -\gamma_\star. \tag{106}$$

Proof. By contradiction. Suppose no such $(\varepsilon_\star, \gamma_\star)$ exists. Then for each $k \in \mathbb{N}$, there exists $M_k \in SO(n)$ with:

$$f(M_k) \geq f_{\text{enter}}, \quad \|\text{grad}\, f(M_k)\|_F \leq 1/k, \quad \lambda_{\min}(\text{Hess}\, f(M_k)) \geq -1/k.$$

Since $SO(n)$ is compact, extract a convergent subsequence $M_{k_\ell} \to \bar{M}$. By continuity of f, grad f, and Hess f:

$$f(\bar{M}) \geq f_{\text{enter}} > 0, \quad \text{grad}\, f(\bar{M}) = 0, \quad \text{Hess}\, f(\bar{M}) \succeq 0.$$

This means $\bar{M}$ is a second-order stationary point (local minimum candidate) with $f(\bar{M}) > 0$. But Proposition C.12 states no such point exists: non-global-minimum critical points are strict saddles (have negative Hessian eigenvalue). Contradiction. □

E.4.3 Explicit Negative Curvature Constant

The Level Set Separation Lemma E.9 guarantees existence of $(\varepsilon_\star, \gamma_\star)$ via compactness. Using the Givens structure from Appendix C, we make $\gamma_\star$ **explicit**.

Lemma E.10 (Explicit Negative Curvature at Entry Threshold)**.** *For any saddle critical point M with $f(M) \geq f_{\text{enter}} = (g - \underline{\delta})^2/8$:*

$$\lambda_{\min}(\text{Hess}\, f(M)) \leq -\gamma_{\text{enter}}, \qquad \gamma_{\text{enter}} := \frac{(g - \underline{\delta})^2}{n(n-1)}. \tag{107}$$

Proof. At any non-minimal critical point, there exists (i, j) with $A_{ii} = A_{jj}$ and $b := A_{ij} \neq 0$. By averaging: $b_{\max}^2 \geq 2f(M)/(n(n-1))$. By Lemma C.11 (Appendix C): $\lambda_{\min} \leq -4b_{\max}^2 \leq -8f(M)/(n(n-1))$. Substituting $f(M) \geq f_{\text{enter}}$ gives the result. □

E.4.4 Tubular Neighborhood and Escape Radius

Lemma E.11 (Negative Curvature Preservation Radius)**.** *Define*

$$r_{\mathrm{nc}} := \frac{g-\underline{\delta}}{32\|C_e\|_2\sqrt{n(n-1)}}, \qquad b_0 := \frac{g-\underline{\delta}}{2\sqrt{n(n-1)}}. \tag{108}$$

Within geodesic distance r_{nc} of any saddle critical point $\bar{M}$, the negative curvature direction $\Xi = E_{ij} - E_{ji}$ satisfies:

$$\frac{d^2}{dt^2}\bigg|_{t=0} f(Me^{t\Xi}) \le -\frac{15}{8}b_0^2. \tag{109}$$

Proof. Along geodesics, $|\dot{A}_{pq}| \le 2\|C_e\|_2\|\Xi\|_F$. Within radius r_{nc}, diagonal separation changes by at most:

$$4\|C_e\|_2 r_{\mathrm{nc}} = 4\|C_e\|_2 \cdot \frac{g-\underline{\delta}}{32\|C_e\|_2\sqrt{n(n-1)}} = \frac{g-\underline{\delta}}{8\sqrt{n(n-1)}} = \frac{b_0}{4}.$$

At the saddle center $\bar{M}$, we have $\Delta = 0$ and $|b| = b_0$ for the degenerate block. Within radius r_{nc}: $|b| \ge b_0 - b_0/4 = 3b_0/4$ and $|\Delta| \le b_0/4$. By the exact second variation formula (Lemma C.11):

$$\frac{d^2}{dt^2}\bigg|_{t=0} f(Me^{t\Xi}) = 2(\Delta^2 - 4b^2) \le 2\left(\frac{b_0^2}{16} - 4\cdot\frac{9b_0^2}{16}\right) = 2\cdot\frac{-35b_0^2}{16} = -\frac{35}{8}b_0^2 \le -\frac{15}{8}b_0^2. \qquad \square$$

E.4.5 Discrete Escape Time Bound

Proposition E.12 (Escape Steps from Saddle Region)**.** *Let $\gamma_{\mathrm{loc}} := \frac{15}{32}\cdot\frac{(g-\underline{\delta})^2}{n(n-1)}$ and $r_* := \min\{r_{\mathrm{nc}}, \gamma_{\mathrm{loc}}/(16L_C)\}$. If the iterate enters a tubular neighborhood of width w around a saddle's center-stable manifold, the escape time satisfies:*

$$T_{\mathrm{esc}} \le \frac{2}{\eta\gamma_{\mathrm{loc}}}\log\frac{r_*}{w}. \tag{110}$$

Proof. In local exponential coordinates $\xi \in \mathfrak{so}(n)$ around $\bar{M}$, let $\alpha := \langle \xi, u\rangle$ be the unstable coordinate (along the Givens direction u). By Lemma E.11, the gradient component satisfies $g_u(\xi) \le -(\gamma_{\mathrm{loc}}/2)\alpha$ for $|\xi| \le r_*$. The Cayley-GD update gives:

$$\alpha_{k+1} \ge \left(1 + \frac{\eta\gamma_{\mathrm{loc}}}{2}\right)\alpha_k.$$

Exponential growth from initial displacement $|\alpha_0| \gtrsim w$ to $|\alpha| \sim r_*$ takes the stated number of steps. $\square$

E.4.6 Main Result: Finite-Time Domain Entry

Theorem E.13 (High-Probability Finite-Time Entry into $\mathcal{N}_{\underline{\delta}}$)**.** *Assume C has distinct eigenvalues and Proposition C.12 holds. Fix $\underline{\delta} \in (0, g)$ and define:*

$$f_{\mathrm{enter}} := \frac{(g-\underline{\delta})^2}{8}, \qquad \gamma_{\mathrm{enter}} := \frac{(g-\underline{\delta})^2}{n(n-1)}, \tag{111}$$

$$\gamma_{\mathrm{loc}} := \frac{15}{32}\cdot\frac{(g-\underline{\delta})^2}{n(n-1)}, \qquad r_{\mathrm{nc}} := \frac{g-\underline{\delta}}{32\|C_e\|_2\sqrt{n(n-1)}}. \tag{112}$$

Then for any $\zeta \in (0,1)$, with probability $\ge 1-\zeta$, $M_T \in \mathcal{N}_{\underline{\delta}}$ and the entry time satisfies:

$$\boxed{T_{\mathrm{enter}}(\zeta) \le \frac{2(f(M_0) - f_{\mathrm{enter}})}{\eta\varepsilon_g^2} + \frac{c_n}{\eta\gamma_{\mathrm{loc}}}\log\frac{c_n'}{\zeta},} \tag{113}$$

where $\varepsilon_g > 0$ is the gradient threshold for saddle detection and c_n, c_n' depend only on n.

Proof. **Phase 1 (Gradient descent):** When $\|\Omega_k\|_F \geq \varepsilon_g$, Lemma 3.1 (Appendix D) gives $f(M_{k+1}) \leq f(M_k) - \frac{\eta}{2}\varepsilon_g^2$. This phase contributes at most $2(f(M_0) - f_{\text{enter}})/(\eta\varepsilon_g^2)$ iterations.

Phase 2 (Saddle escape): When $\|\Omega_k\|_F < \varepsilon_g$ but $f(M_k) \geq f_{\text{enter}}$, by Lemma E.10 the Hessian has a negative eigenvalue $\leq -\gamma_{\text{enter}}$ along a Givens direction. Within geodesic radius r_{nc}, the negative curvature is preserved with constant γ_{loc} (Lemma E.11). A random perturbation lands outside the "dangerous tube" with probability $\geq 1-\zeta$. The unstable coordinate grows as $|\alpha_{k+1}| \geq (1+\eta\gamma_{\text{loc}}/2)|\alpha_k|$, escaping in $O(\gamma_{\text{loc}}^{-1}\log(1/\zeta))$ iterations. □

Remark E.14 (Improvement over Generic PRGD). The bound has only **single logarithmic dependence** on the failure probability, compared to $\log^4(1/\zeta)$ in generic PRGD (e.g., Sun et al.). This comes from the explicit Givens structure: the function profile along the unstable direction is exactly $f(Me^{t\Xi}) = f(M) - b^2\sin^2(2t)$.

Remark E.15 (Deterministic Givens Escape—Avoiding Randomness Entirely). The explicit Givens structure enables a **deterministic** saddle escape strategy. At a saddle point M with degenerate block (i,j) where $A_{ii} = A_{jj}$ and $b = A_{ij} \neq 0$, taking the **exact step** $t = \pi/4$ along $\Xi = E_{ij} - E_{ji}$ yields:

$$f(Me^{\frac{\pi}{4}\Xi}) = f(M) - b^2 \cdot \sin^2(\pi/2) = f(M) - b^2.$$

This single deterministic step reduces f by b^2 with **no randomness required**. In practice, this requires detecting degenerate blocks (small $|A_{ii} - A_{jj}|$ with large $|A_{ij}|$), which can be done via Hessian probing or explicit eigenvalue gap monitoring. This transforms saddle escape from a probabilistic $O(\log(1/\zeta))$ complexity to a **constant-step** deterministic operation.

Corollary E.16 (Total Complexity)**.** *Combining Theorems E.13 and E.7, the total complexity to reach ε-accuracy with probability $\geq 1-\zeta$ is:*

$$T_{\text{total}}(\varepsilon,\zeta) = T_{\text{enter}}(\zeta) + O\left(\frac{\|C_e\|_2^2}{\lambda}\log\frac{\sqrt{f_{\text{enter}}}}{\varepsilon}\right).$$

E.5 Theorem D.4: High-Probability Long-Term Uniform Bound

Theorem E.17 (High-Probability Long-Term Uniform Bound)**.** *Suppose $M_0 \in \mathcal{N}_{\underline{\delta}}$, observation noise satisfies $\sup_k \|E_{e,k}\|_F \leq \varepsilon$, and covariance drift $\|\dot{C}\|_F \leq \rho$. For any confidence level $\zeta \in (0,1)$, if the non-escape condition (Theorem E.7) holds, then with probability $\geq 1-\zeta$:*

$$\boxed{\sup_{k\geq 0}\sqrt{f(M_k)} \leq e^{-c_1\underline{\delta}^2 k\eta}\sqrt{f(M_0)} + \frac{\sqrt{2}\|C_e\|_2}{\underline{\delta}^2}\left(\bar{u}_{\text{det}} + \Phi(\zeta)\right),} \tag{114}$$

where:

- ***Deterministic drift term:*** $\bar{u}_{\text{det}} = 4\|C_e^{\text{sig}}\|_2\varepsilon + 2\varepsilon^2 + \frac{\rho}{2\|C_e\|_2}$;
- ***High-probability fluctuation term*** $\Phi(\zeta)$*:*

Noise Type	*Assumption*	$\Phi(\zeta)$
Sub-exponential	*Independent + Bernstein*	$\sigma_u\sqrt{\frac{\eta}{\lambda}}\sqrt{\log(1/\zeta)} + b_u\eta\log(1/\zeta)$
Heavy-tailed	*Uncorrelated + Finite variance*	$\sigma_u\sqrt{\frac{\eta}{2\lambda\zeta}}$

As $k \to \infty$, this yields the steady-state radius $r_f^{\text{prob}} = \frac{\sqrt{2}\|C_e\|_2}{\lambda}(\bar{u}_{\text{det}} + \Phi(\zeta))$.

E.6 Proof of Theorem 5: Statistical Robustness

Proof of Theorem 5. **(i)** The $O(1/k)$ rate follows from Appendix D.5 with decaying step size. **(ii)** Sample complexity: solving $C/(k+k_0) \leq \varepsilon$ gives $k^*(\varepsilon) = O(\|C_e\|_2^2\mathbb{E}[\|E_e\|_F^2]/(\lambda\varepsilon))$. **(iii)** Combining Lemma D.2 with Theorem 3 (Appendix D) yields the high-probability bound. □

F Baseline Limitations and Lower Bounds: Extended Analysis

This appendix establishes the structural advantages of the commutator method over standard baselines. We focus on the following practically important setting:

F.1 Problem Setting and Main Results

F.1.1 Matrix-Free Streaming Setting

We operate under the observation model equation 26 (see Appendix 8) with additional constraints:

Streaming Matrix-Free Constraints

- **Streaming:** Observations arrive sequentially; precomputing $\mathrm{tf}(C)$ or $\mathrm{tr}(C)$ over the full dataset is unavailable.
- **Matrix-free:** Only matrix-vector products $v \mapsto C_k v$ are accessible; explicit matrix entries are unavailable.

Main conclusion: In this setting, baselines suffer from **signal vanishing**—their iteration complexity grows as $\Omega(\sigma^2)$, rendering them "alive but frozen." The commutator structure provides $O(1)$ complexity **without requiring trace estimation or noise-level knowledge**.

F.1.2 Properties (P1) and (P2): Formal Definitions

Consider a general discrete iteration (encompassing Cayley/QR/polar retractions):

$$M_{k+1} = \Phi(M_k, C_k), \qquad M_k \in SO(n), \quad C_k = C_k^{\mathrm{sig}} + \sigma_k^2 I. \tag{115}$$

Definition F.1 ((P1) Pathwise σ-Immunity (Trajectory-Level Invariance))**.** For any scalar sequence $\{\sigma_k^2\}$ and any signal sequence $\{C_k^{\mathrm{sig}}\}$, we require:

$$\Phi(\cdot, C_k^{\mathrm{sig}} + \sigma_k^2 I) \equiv \Phi(\cdot, C_k^{\mathrm{sig}}) \quad \text{step-by-step.} \tag{116}$$

Our method satisfies (P1) at the generator level via the commutator identity $\Omega_{C+\alpha I}(M) = \Omega_C(M)$.

Definition F.2 ((P2) Discrete ISS (for the Core Lyapunov Function))**.** Let $A_k = M_k^\top C_k^{\mathrm{sig}} M_k$ and $f_k = \frac{1}{2}\|\mathrm{off}(A_k)\|_F^2$. Within the spectrally separated domain $\mathcal{N}_{\underline{\delta}}$, we require the existence of constants $c_1, c_2 > 0$ **independent of** $\{\sigma_k^2\}$ such that for bounded input $\|U_k\|_F \le \bar{U}$:

$$\sqrt{f_{k+1}} \le (1 - c_1 \underline{\delta}^2 \eta_k)\sqrt{f_k} + c_2 \|C_e\|_2 \eta_k \bar{U}, \tag{117}$$

which implies the asymptotic bound:

$$\limsup_{k\to\infty} \sqrt{f_k} \le \frac{c_2}{c_1} \cdot \frac{\|C_e\|_2}{\underline{\delta}^2} \cdot \bar{U}. \tag{118}$$

Our method achieves (P2) via the continuous ISS (Theorem 4.3) and discrete descent lemma (Theorem D.3).

Goal: Prove that standard baselines cannot achieve σ-independent convergence rates. Specifically:

(a) Baselines fail (P1): trajectories depend on σ^2.

(b) Baselines fail (P2): iteration complexity grows as $\Omega(\sigma^2)$ (signal vanishing).

(c) The commutator structure is both **sufficient** (our method) and **necessary** (up to structural equivalence) for achieving $O(1)$ complexity.

F.1.3 Main Theorem: Conditional Irreplaceability

We state the main result upfront; detailed proofs of the supporting lemmas appear in subsequent sections.

Definition F.3 (Exact σ^2-Immunity)**.** In the observation model $C_k = C_{\text{sig}} + \sigma_k^2 I + E_k$ with arbitrary time-varying $\{\sigma_k^2\}$ and trace-free perturbation E_k, an algorithm is **exactly σ^2-immune** if, for the same initialization M_0:

(i) The discrete trajectory is pointwise invariant:

$$\{M_k\} \text{ generated from } \{C_{\text{sig}} + \sigma_k^2 I + E_k\} \equiv \{M_k\} \text{ generated from } \{C_{\text{sig}} + E_k\}.$$

(ii) The Lyapunov evolution $\{f(M_k; C_{\text{sig}})\}$ does not depend on $\{\sigma_k^2\}$.

This is precisely the content of Proposition B.5 in Appendix B.

Theorem F.4 (Conditional Irreplaceability of the Commutator Geometry for Exact σ^2-Immunity)**.** *Consider the Lyapunov function $f(M; C_{\text{sig}}) = \frac{1}{2}\|\text{off}(M^\top C_{\text{sig}} M)\|_F^2$ and define $\Omega(M) := [A(M), D(M)]$ with $A(M) = M^\top C_{\text{sig}} M$, $D(M) = \text{diag}(A(M))$. Assume the streaming observation sequence is $C_k = C_{\text{sig}} + \sigma_k^2 I + E_k$.*

(A) Sufficiency: Exact Noise Dissipation. *The commutator Lie-group integrator (e.g. Cayley update)*

$$M_{k+1} = M_k \cdot \text{Cay}\big(\eta_k\, \Omega_{C_k}(M_k)\big)$$

is exactly σ^2-immune in the sense of Definition F.3: (i) the full trajectory $\{M_k\}$ is pointwise invariant to $\{\sigma_k^2\}$, and (ii) the Lyapunov evolution $\{f(M_k; C_{\text{sig}})\}$ is independent of $\{\sigma_k^2\}$.

(B) Baseline Failure: Why Extrinsic Methods Cannot Achieve σ-free ISS. *For baselines of the form $M_{k+1} = \text{qf}(M_k - \eta G_E(M_k))$ where $G_E = 2CM\,\text{off}(M^\top CM)$ is the Euclidean gradient:*

(i) Exact tangent-normal geometry: $\Pi_T(G_E) = -M\Omega$ *is σ-free, but $\|\Pi_N(G_E)\| = \Theta(\sigma^2\|O\|)$ explodes with σ.*

(ii) QR retraction sensitivity: *The retraction residual satisfies $\|R\| = \Theta(\eta\sigma^2\|O\|)$ (first-order in σ^2, not second-order).*

(iii) Direction reversal: *The contamination ratio $\|\mathcal{P}(S)\|_F/\|\Omega\|_F \gtrsim \sigma^2/\underline{\delta}$. For $\sigma^2 > c\underline{\delta}$, direction reversal $\langle V, M\Omega\rangle < 0$ occurs.*

(iv) Structural remedy: *The only way to restore σ-free descent is to use $\dot{M} = -\Pi_T(G_E) = M\Omega$, which is the commutator dynamics.*

(C) Necessity: Alignment with the Commutator Dissipation Channel. *Let an algorithm have the general Lie-group form $M_{k+1} = M_k \cdot R(\eta_k\, K(M_k, C_k))$, where R is any first-order structure-preserving retraction (Cayley/polar/exp) and $K(\cdot,\cdot) \in \mathfrak{so}(n)$ is continuous. If there exist constants $a, b > 0$* independent of $\text{tr}(C_k)$ *such that for all sufficiently small η:*

$$f(M_k) - f(M_{k+1}) \geq a\,\eta\,\|\Omega(M_k)\|_F^2 - b\,\eta\,\|\Omega(M_k)\|_F\,\|U_k\|_F - O(\eta^2),$$

then there exists $\gamma > 0$ such that $\langle \Omega(M),\, K(M, C_{\text{sig}})\rangle \geq \gamma\,\|\Omega(M)\|_F^2$ for all M.

Conclusion: *Any algorithm achieving (P1)+(P2) on f must have its generator K* aligned *with $\Omega = [A, D]$. If K is orthogonal to Ω at some M, no uniform $\gamma > 0$ exists and σ-free ISS is impossible.*

Remark F.5 (Scope of Theorem F.4). The irreplaceability claim is conditioned on the **matrix-free streaming** regime. With full matrix access or known σ^2, baselines can precompute $\text{tf}(C)$ and match σ-elimination. Additionally: (i) σ-immunity applies only to isotropic noise $\sigma^2 I$; (ii) the claim is specific to the Lyapunov $f(M) = \frac{1}{2}\|\text{off}(M^\top CM)\|_F^2$.

The proof of Theorem F.4 relies on results established in subsequent sections: gradient decomposition (Section F.3), baseline failure analysis (Section F.4), and the dissipation channel formalism (Section F.5).

F.2 Baseline Class and Access Trilemma

F.2.1 Baseline Algorithm Class

We consider algorithms of the form:

$$M_{k+1} = \mathrm{qf}(M_k + \eta Z_k(M_k, C_k)), \tag{119}$$

where $\mathrm{qf}(\cdot)$ is the Q-factor of QR decomposition (with positive diagonal R), and Z_k is an extrinsic search direction depending on the full observation $C_k = C_{\mathrm{sig}} + \sigma^2 I$.

Specific instances include:

- **Subspace Iteration:** $Z_k = (C_k - I)M_k$, i.e., $M_{k+1} = \mathrm{qf}(C_k M_k)$
- **QR-Oja:** $Z_k = C_k M_k$, i.e., $M_{k+1} = \mathrm{qf}((I + \eta C_k)M_k)$
- **Euclidean SGD+QR (for f):** $Z_k = -G_E(M_k) = -2C_k M_k \mathrm{off}(A_k)$, the Euclidean gradient of $f(M) = \frac{1}{2}\|\mathrm{off}(M^\top C M)\|_F^2$
- **Euclidean SGD+QR (for PCA):** $Z_k = C_k M_k$, the (negated) Euclidean gradient of $g(M) = -\mathrm{tr}(M^\top C M)$

F.2.2 Information-Theoretic Access Trilemma

The scalar component $\sigma^2 I$ acts as a **nuisance parameter** in the sense of semiparametric statistics: it does not affect the parameter of interest (eigenspace structure) but dominates the observation. Any algorithm must address this nuisance through one of three mechanisms:

Proposition F.6 (Access Trilemma)**.** *Under the streaming matrix-free MVP oracle with observation $C = C_{\mathrm{sig}} + \sigma^2 I$ where σ^2 is unknown and potentially unbounded:*

(i) ***No structural filtering:*** *Algorithms using C directly require step size $\eta = O(1/\|C\|_2) = O(1/\sigma^2)$ for stability, yielding iteration complexity $\Omega(\sigma^2/\Delta)$.*

(ii) ***Explicit trace estimation:*** *Computing $\mathrm{tf}(C) = C - \frac{\mathrm{tr}(C)}{n} I$ via Hutchinson's estimator Skorski (2021); Meyer et al. (2021) requires*

$$m \geq \frac{2\|C\|_F^2}{n^2\epsilon^2\zeta} \geq \frac{2\sigma^4}{n\epsilon^2\zeta} \tag{120}$$

MVP queries per iteration to achieve $|\hat{\tau} - \tau| \leq \epsilon$ with probability $\geq 1 - \zeta$.

(iii) ***Algebraic filtering (commutator):*** *The commutator generator $\Omega_C(M) = [M^\top C M, \mathrm{diag}(M^\top C M)]$ satisfies $\Omega_{C+\alpha I} = \Omega_C$, achieving exact σ^2-elimination with 0 additional MVP cost.*

Proof. **Part (i):** By Lemmas F.16–F.17 below, non-filtering baselines exhibit signal vanishing. Specifically, subspace iteration has contraction factor $\rho = (\lambda_2 + \sigma^2)/(\lambda_1 + \sigma^2) \to 1$, requiring $k = \Omega((\lambda_1 + \sigma^2)/\Delta)$ iterations.

Part (ii): Hutchinson's estimator with m Rademacher vectors has variance $\mathrm{Var}(\hat{\tau}) \leq 2\|C\|_F^2/(mn^2)$. Since $\|C\|_F^2 = \|C_{\mathrm{sig}}\|_F^2 + 2\sigma^2 \mathrm{tr}(C_{\mathrm{sig}}) + n\sigma^4 \geq n\sigma^4$, Chebyshev's inequality gives the stated bound.

Part (iii): By Lemma B.1, $\Omega_{C+\alpha I}(M) = [A + \alpha I, D + \alpha I] = [A, D] + \alpha[I, D] + \alpha[A, I] + \alpha^2[I, I] = [A, D] = \Omega_C(M)$, where the last step uses $[I, X] = 0$. □

Remark F.7 (Nuisance Tangent Space Interpretation). From a semiparametric perspective, $\tau = \mathrm{tr}(C)/n$ is a nuisance parameter with tangent space $\mathcal{T}_{\mathrm{nuis}} = \mathrm{span}\{I\}$. Property (P1) requires the update to be orthogonal to $\mathcal{T}_{\mathrm{nuis}}$:

$$\frac{d}{dt}\Big|_{t=0} \Phi(\cdot, C + tI) = 0.$$

The commutator achieves this via the algebraic identity $[I, X] = 0$, while baselines must estimate and subtract the nuisance at cost $\Omega(\sigma^4)$.

F.2.3 Trace Estimation Complexity (Hutchinson)

Lemma F.8 (Trace Estimation Complexity)**.** *In matrix-free settings where only matrix-vector products $v \mapsto C_k v$ are available, estimating $\mathrm{tr}(C_k)$ via Hutchinson's estimator requires:*

$$m \geq \frac{2\|C_k\|_F^2}{n^2\epsilon^2\zeta} \geq \frac{2\sigma^4}{n\epsilon^2\zeta}$$

samples to achieve $|e_k| := |\tau_k - \hat{\tau}_k| \leq \epsilon$ with probability $\geq 1 - \zeta$.

Proof. For Hutchinson's estimator with m Rademacher vectors:

$$\mathrm{Var}(\hat{\mathrm{tr}}(C_k)) \leq \frac{2\|C_k\|_F^2}{m}.$$

When $C_k = C_{\mathrm{sig}} + \sigma^2 I$:

$$\|C_k\|_F^2 = \|C_{\mathrm{sig}}\|_F^2 + 2\sigma^2 \mathrm{tr}(C_{\mathrm{sig}}) + n\sigma^4 \geq n\sigma^4.$$

By Chebyshev: $\mathbb{P}(|e_k| \geq \epsilon) \leq \mathrm{Var}(\hat{\tau}_k)/\epsilon^2$. Setting this $\leq \zeta$ gives the bound. □

Corollary F.9. *In matrix-free settings, the trace estimation overhead grows as $O(\sigma^4)$. The residual error e_k reintroduces σ-dependence through the effective step size $\eta' = \eta/(1 + \eta e_k)$.*

F.3 Euclidean Gradient Analysis

This section establishes the exact tangent-normal geometry of the Euclidean gradient, showing why extrinsic methods fail (P1) and (P2).

F.3.1 Tangent-Normal Decomposition

Lemma F.10 (Gradient Decomposition)**.** *Let $A_{\mathrm{sig}} = M^\top C_{\mathrm{sig}} M$, $A = A_{\mathrm{sig}} + \sigma^2 I$, and $O = \mathrm{off}(A) = \mathrm{off}(A_{\mathrm{sig}})$. The Euclidean gradient $G_E = 2CM \cdot O$ admits the decomposition:*

$$\Pi_T(G_E) = -M[A_{\mathrm{sig}}, D_{\mathrm{sig}}] = -M\Omega_C(M), \tag{121}$$

$$\Pi_N(G_E) = M \cdot S, \quad S = \{A_{\mathrm{sig}}, O\} + 2\sigma^2 O, \tag{122}$$

where $D_{\mathrm{sig}} = \mathrm{diag}(A_{\mathrm{sig}})$, $\Omega_C(M) = [A, D]$ is the commutator generator, and $\{X, Y\} := XY + YX$ is the anti-commutator.

Proof. Define $H := M^\top G_E = 2(A_{\mathrm{sig}} + \sigma^2 I)O = 2A_{\mathrm{sig}}O + 2\sigma^2 O$.

Skew-symmetric part:

$$K := \mathrm{skew}(H) = \mathrm{skew}(2A_{\mathrm{sig}}O) + \mathrm{skew}(2\sigma^2 O) \tag{123}$$

$$= A_{\mathrm{sig}}O - OA_{\mathrm{sig}} + 0 \quad (\text{since } O \text{ is symmetric}) \tag{124}$$

$$= [A_{\mathrm{sig}}, O] = -[A_{\mathrm{sig}}, D_{\mathrm{sig}}]. \tag{125}$$

The last equality uses $A_{\mathrm{sig}} = D_{\mathrm{sig}} + O$ and $[D_{\mathrm{sig}}, O] = -[O, D_{\mathrm{sig}}]$.

Symmetric part:

$$S := \mathrm{sym}(H) = \mathrm{sym}(2A_{\mathrm{sig}}O) + 2\sigma^2 O \tag{126}$$

$$= A_{\mathrm{sig}}O + OA_{\mathrm{sig}} + 2\sigma^2 O = \{A_{\mathrm{sig}}, O\} + 2\sigma^2 O. \tag{127}$$

Since $\Pi_T(G_E) = M\,\mathrm{skew}(M^\top G_E) = MK$ and $\Pi_N(G_E) = M\,\mathrm{sym}(M^\top G_E) = MS$, the decomposition follows. □

Corollary F.11 (Tangent Component is σ-free)**.** *The tangent component* $\Pi_T(G_E) = -M\Omega_C(M)$ *is an* ***exact identity*** *independent of* σ^2*. This follows from the scalar shift invariance* $\Omega_{C+\sigma^2 I}(M) = \Omega_C(M)$*.*

Corollary F.12 (Explicit σ^2 Decomposition of Euclidean Gradient)**.** *The Euclidean gradient* $G_E = \nabla_M f = 2CMO$ *admits the explicit separation:*

$$\nabla_M f = \underbrace{2C^{\text{sig}}MO}_{\textit{signal term}} + \underbrace{2\sigma^2 MO}_{\textit{noise term}}. \tag{128}$$

The noise term $2\sigma^2 MO$ *is:*

- ***Entirely in the normal direction:*** $\Pi_N(2\sigma^2 MO) = 2\sigma^2 MO$ *(since* MO *is symmetric in the pullback:* $M^\top(MO) = O$*).*
- ***Proportional to*** σ^2***:*** *The magnitude scales linearly with the noise variance.*

Hence, any algorithm using G_E *directly has its update direction contaminated by* σ^2*.*

F.3.2 Gradient Explosion Bound

Corollary F.13 (Gradient Explosion Lower Bound)**.** *When* $\sigma^2 > \|C_{\text{sig}}\|_2$*:*

$$\|G_E\|_F \geq 2(\sigma^2 - \|C_{\text{sig}}\|_2)\|O\|_F. \tag{129}$$

Proof. From Lemma F.10:

$$\|\Pi_N(G_E)\|_F = \|S\|_F \geq \|2\sigma^2 O\|_F - \|\{A_{\text{sig}}, O\}\|_F.$$

Since $\|\{A_{\text{sig}}, O\}\|_F \leq 2\|A_{\text{sig}}\|_2\|O\|_F = 2\|C_{\text{sig}}\|_2\|O\|_F$:

$$\|\Pi_N(G_E)\|_F \geq (2\sigma^2 - 2\|C_{\text{sig}}\|_2)\|O\|_F.$$

The result follows from $\|G_E\|_F \geq \|\Pi_N(G_E)\|_F$. □

F.3.3 Normal-Space Contamination

Lemma F.14 (Normal-Space Contamination)**.** *For the diagonalization objective* $f(M) = \frac{1}{2}\|\text{off}(M^\top CM)\|_F^2$*, the Euclidean gradient* $G_E = 2CMO$ *has a* σ^2*-dependent normal component. Even when step size is scaled as* $\eta = O(1/\sigma^2)$*, the actual update direction* $V := \text{qf}(M - \eta G_E) - M$ *and the ideal descent direction* $M\Omega$ *satisfy:*

$$\frac{\|\mathcal{P}(S)\|_F}{\|\Omega\|_F} \gtrsim \frac{2\sigma^2}{\underline{\delta}}.$$

When $\sigma^2 \gg \underline{\delta}$*, the contamination dominates, and* $\cos(V, M\Omega)$ *can become negative (direction reversal).*

Proof. The first-order expansion gives:

$$V = -\eta M\Omega - \eta M\mathcal{P}(S) + O(\eta^2\|G_E\|^2).$$

The ratio of contamination to signal (recalling $\Pi_T(G_E) = -M\Omega$):

$$\frac{\|\mathcal{P}(S)\|_F}{\|\Omega\|_F} \approx \frac{2\sigma^2\|O\|_F}{\|\Omega\|_F}.$$

By Lemma C.1, $\|\Omega\|_F^2 \geq 2\underline{\delta}^2 f = \underline{\delta}^2\|O\|_F^2$, so $\|\Omega\|_F \geq \underline{\delta}\|O\|_F$. Hence:

$$\frac{\|\mathcal{P}(S)\|_F}{\|\Omega\|_F} \gtrsim \frac{2\sigma^2\|O\|_F}{\underline{\delta}\|O\|_F} = \frac{2\sigma^2}{\underline{\delta}}.$$

When this ratio exceeds 1, direction reversal ($\cos(V, M\Omega) < 0$) becomes possible. □

Remark F.15 (Riemannian Gradient and Descent Flow). For the objective $f(M) = \frac{1}{2}\|\mathrm{off}(M^\top CM)\|_F^2$, the relationship between gradients is:

$$\Pi_T(G_E) = -M\Omega_C(M), \qquad \text{where } \Omega_C(M) = [A, \mathrm{diag}(A)]. \tag{130}$$

The **descent flow** is $\dot{M} = -\Pi_T(G_E) = M\Omega_C(M)$, which is the commutator dynamics. This direction is **shift-invariant** by Lemma B.1 ($\Omega_{C+\sigma^2 I} = \Omega_C$).

Hence, "Euclidean SGD with Riemannian projection" recovers the commutator direction and joins our method family. This completes the structural necessity argument: to achieve (P1)+(P2) on f, baselines must use the commutator generator Ω.

F.4 Baseline Failure Analysis

This section provides quantitative analysis of why each baseline algorithm fails (P1) and (P2).

F.4.1 Subspace Iteration: σ-Dependent Contraction

Lemma F.16 (Subspace Iteration Analysis Golub & Van Loan (2013))**.** *For $M_{k+1} = \mathrm{qf}(C_k M_k)$ with $C_k = C_{\mathrm{sig}} + \sigma^2 I$ and $C_{\mathrm{sig}} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, $\lambda_1 > \cdots > \lambda_n$:*

(a) Exact recursion** ($n = 2$)**:

$$\tan\theta_{k+1} = \rho_{\mathrm{SI}}(\sigma^2) \cdot \tan\theta_k, \quad \rho_{\mathrm{SI}}(\sigma^2) := \frac{\lambda_2 + \sigma^2}{\lambda_1 + \sigma^2}.$$

(b) (P1) failure: *The trajectory depends on σ^2 unless $\tan\theta_0 = 0$.*

(c) σ-free contraction impossible: *For any $q < 1$ independent of σ^2:*

$$\rho_{\mathrm{SI}}(\sigma^2) \le q \iff \sigma^2 \le \frac{\Delta}{1-q} - \lambda_1,$$

where $\Delta = \lambda_1 - \lambda_2$. When $\sigma^2 > \sigma^2_{\max}(q)$, the contraction factor exceeds q.

(d) Complexity lower bound:

$$k \ge \frac{\lambda_1 + \sigma^2}{\Delta} \log \frac{\tan\theta_0}{\varepsilon}.$$

Proof. Let $x_k = \cos\theta_k\, e_1 + \sin\theta_k\, e_2$. Then:

$$C_k x_k = (\lambda_1 + \sigma^2)\cos\theta_k\, e_1 + (\lambda_2 + \sigma^2)\sin\theta_k\, e_2.$$

After normalization:

$$\tan\theta_{k+1} = \frac{(\lambda_2 + \sigma^2)\sin\theta_k}{(\lambda_1 + \sigma^2)\cos\theta_k} = \frac{\lambda_2 + \sigma^2}{\lambda_1 + \sigma^2}\tan\theta_k.$$

Parts (b)-(d) follow from this exact recursion. □

F.4.2 QR-Oja: Effective Step Size Shrinkage

Lemma F.17 (QR-Oja Analysis)**.** *For $M_{k+1} = \mathrm{qf}((I + \eta C_k)M_k)$:*

(a) Effective step size:

$$\eta'_k = \frac{\eta}{1 + \eta\sigma_k^2} \to 0 \quad (\sigma^2 \to \infty).$$

(b) (P1) failure: *The trajectory depends on σ_k^2 through η'_k.*

(c) σ-free progress impossible: *For any $\eta_{\min} > 0$:*

$$\eta'_k \geq \eta_{\min} \iff \sigma^2 \leq \frac{\eta/\eta_{\min} - 1}{\eta}.$$

(d) Complexity lower bound:

$$k \gtrsim \frac{1 + \eta\sigma^2}{\eta\Delta} \log \frac{\tan\theta_0}{\varepsilon}.$$

Proof. Using $\mathrm{qf}(\alpha X) = \mathrm{qf}(X)$ for $\alpha > 0$:

$$(I + \eta C_k)M_k = (1 + \eta\sigma_k^2)\left(M_k + \frac{\eta}{1 + \eta\sigma_k^2} C_{\mathrm{sig}} M_k\right) \tag{131}$$

$$= (1 + \eta\sigma_k^2)\left(M_k + \eta'_k C_{\mathrm{sig}} M_k\right). \tag{132}$$

Hence $M_{k+1} = \mathrm{qf}(M_k + \eta'_k C_{\mathrm{sig}} M_k)$ with $\eta'_k = \eta/(1 + \eta\sigma_k^2)$. □

F.4.3 QR Retraction Sensitivity

Lemma F.18 (QR Retraction Sensitivity Absil et al. (2008))**.** *Let $R := \mathrm{qf}(M + E) - M - \Pi_T(E)$ be the retraction residual. Then:*

- *For purely tangent perturbations $E = E_T \in T_M SO(n)$: $\|R\|_F = O(\|E_T\|_F^2)$.*
- *For perturbations with normal component $E = E_T + E_N$: $\|R\|_F = \Theta(\|E_N\|_F)$* ***(first-order sensitive!)****.*

Proof. By left equivariance $\mathrm{qf}(MX) = M \cdot \mathrm{qf}(X)$, it suffices to analyze $M = I$.

Let $X = I + E$ with QR decomposition $X = QR$. The first-order expansion gives:

$$Q = I + K + O(\|E\|^2), \quad R = I + U + O(\|E\|^2),$$

where K is skew-symmetric, U is upper triangular, and $E = K + U$ (first-order matching).

Key observation: The lower triangular part of E must be absorbed entirely by K:

$$K_{ij} = E_{ij}, \quad K_{ji} = -E_{ij} \quad (i > j).$$

Define the operator $\mathcal{P} : \mathrm{Sym}(n) \to \mathfrak{so}(n)$ that maps a symmetric matrix to a skew-symmetric matrix by:

$$\mathcal{P}(S)_{ij} = \begin{cases} S_{ij} & i > j \\ -S_{ij} & i < j \\ 0 & i = j \end{cases}$$

Then the first-order expansion of Q is:

$$Q - I = \mathrm{skew}(E) + \mathcal{P}(\mathrm{sym}(E)) + O(\|E\|^2).$$

For the tangent projection:

$$\Pi_T(E) = \mathrm{skew}(E),$$

hence the residual satisfies:

$$R = \mathcal{P}(\mathrm{sym}(E)) + O(\|E\|^2).$$

Since $\|\mathcal{P}(S)\|_F = \|\mathrm{off}(S)\|_F$ for symmetric S, and $\mathrm{sym}(E) = E_N$ when $E = E_T + E_N$ (tangent-normal decomposition):

$$\|R\|_F = \Theta(\|E_N\|_F).$$ □

Corollary F.19 (Substituting σ^2)**.** *For $E = -\eta G_E$, the retraction residual satisfies:*

$$\|R\|_F \approx \eta \|\mathcal{P}(S)\|_F \gtrsim \eta \cdot 2\sigma^2 \|O\|_F.$$

*Hence $\|R\| = \Theta(\eta\sigma^2)$, **not** $O(\eta^2\sigma^4)$.*

F.4.4 SRHT and Sketching

A natural question is whether randomized dimensionality reduction (e.g., SRHT, CountSketch Woodruff et al. (2022)) can circumvent the signal vanishing problem by compressing C before optimization. We show that **sketching inherits the noise passively and compresses the spectral gap**, leading to the same $\Omega(\sigma^2)$ iteration complexity.

Lemma F.20 (Spectral Gap Compression under Sketching)**.** *Let $C = C_{\text{sig}} + \sigma^2 I \in \mathbb{R}^{n\times n}$ with signal eigenvalues $\lambda_1 > \lambda_2 > \cdots$. Let $S \in \mathbb{R}^{n\times k}$ be a sketching matrix (SRHT, Gaussian, etc.) and define the sketch $Y = S^\top C S \in \mathbb{R}^{k\times k}$.*

(a) Noise inheritance: *The sketch decomposes as*

$$Y = S^\top C_{\text{sig}} S + \sigma^2 S^\top S. \tag{133}$$

*For orthonormal S (ideal case), $Y = S^\top C_{\text{sig}} S + \sigma^2 I_k$. The noise $\sigma^2 I$ is **exactly preserved**.*

(b) Relative gap compression: *Let $\tilde{\lambda}_1, \tilde{\lambda}_2$ be the top eigenvalues of Y. The relative spectral ratio satisfies:*

$$\frac{\tilde{\lambda}_1}{\tilde{\lambda}_2} = \frac{\lambda_1' + \sigma^2}{\lambda_2' + \sigma^2} \to 1 \quad \text{as } \sigma^2 \to \infty, \tag{134}$$

where λ_1', λ_2' are the projected signal eigenvalues.

(c) Iteration complexity: *Any iterative method (power iteration, Oja, etc.) applied to Y requires*

$$k_{\text{iter}} = \Theta\left(\frac{\sigma^2}{\Delta'}\right) \quad \text{iterations}, \tag{135}$$

where $\Delta' = \lambda_1' - \lambda_2'$ is the projected spectral gap.

Proof. **(a)** Direct computation: $Y = S^\top(C_{\text{sig}} + \sigma^2 I)S = S^\top C_{\text{sig}} S + \sigma^2 S^\top S$.

(b) For orthonormal S, the eigenvalues of Y are those of $S^\top C_{\text{sig}} S$ shifted by σ^2. As $\sigma^2 \to \infty$, the ratio $(\lambda_1' + \sigma^2)/(\lambda_2' + \sigma^2) \to 1$.

(c) Follows from Lemma F.16 applied to Y. □

Remark F.21 (Sketching Relocates but Does Not Solve the Problem). SRHT achieves $O(n \log n)$ complexity per matrix-vector product, but:

- The sketch Y inherits $\sigma^2 I$ exactly (no algebraic cancellation).
- Post-sketch eigensolvers face the same compressed spectral gap.
- The $O(\sigma^2)$ iteration overhead remains, now in the reduced k-dimensional space.
- In matrix-free streaming settings, one cannot pre-subtract σ^2 from Y without $O(\sigma^4)$ trace estimation cost.

Conclusion: Sketching accelerates each iteration but does not reduce iteration count. The total cost remains $\Omega(\sigma^2 \cdot n \log n/\Delta)$, versus $O(\|C_e\|_2^2 n^2/\Delta^2)$ for the commutator method with no dependence on σ^2.

Proposition F.22 (Comparison: Sketching vs. Commutator in High-Noise Regime)**.** *For target precision ε with $\sigma^2 = \alpha\|C_{\text{sig}}\|_2$ ($\alpha \gg 1$):*

Method	*Cost per Iteration*	*Total Iterations*
SRHT + Power Iteration	$O(n \log n)$	$\Omega(\alpha/\Delta) \cdot \log(1/\varepsilon)$
Direct Power Iteration	$O(n^2)$	$\Omega(\alpha/\Delta) \cdot \log(1/\varepsilon)$
Commutator (ours)	$O(n^2)$	$O(\|C_e\|_2^2/\Delta^2) \cdot \log(1/\varepsilon)$

When $\alpha \gg \|C_e\|_2^2/\Delta$, the commutator method is faster ***despite*** *higher per-iteration cost.*

F.4.5 Polar vs. QR Retraction: First-Order Behavior

Proposition F.23 (Polar Retraction is First-Order Insensitive to Normal Components Higham (1986))**.** *For the polar retraction* $\mathrm{polar}(X) := X(X^\top X)^{-1/2}$*:*

$$\mathrm{polar}(I + E) = I + \mathrm{skew}(E) + O(\|E\|^2).$$

Unlike QR, the polar retraction has ***no first-order sensitivity*** *to the symmetric (normal) part of E.*

Proof. Let $X = I + E$. Then $X^\top X = I + E + E^\top + O(\|E\|^2) = I + 2\,\mathrm{sym}(E) + O(\|E\|^2)$.
Using $(I + A)^{-1/2} = I - \frac{1}{2}A + O(\|A\|^2)$:

$$(X^\top X)^{-1/2} = I - \mathrm{sym}(E) + O(\|E\|^2).$$

Hence:

$$\begin{aligned} \mathrm{polar}(X) &= (I + E)(I - \mathrm{sym}(E) + O(\|E\|^2)) && (136)\\ &= I + E - \mathrm{sym}(E) + O(\|E\|^2) && (137)\\ &= I + \mathrm{skew}(E) + O(\|E\|^2). \end{aligned}$$

□

Remark F.24 (Methodological Implication)*.* This difference has significant implications:

- **QR retraction:** First-order contamination $\|\mathcal{P}(\mathrm{sym}(E))\|_F = \Theta(\eta\sigma^2)$ from normal components.
- **Polar retraction:** Only second-order contamination from normal components.

However, polar retraction still suffers from (P1) failure because the **direction itself** (before retraction) depends on σ^2. The structural remedy is to use the **commutator direction** (tangent projection), after which QR/polar/Cayley all yield equivalent results.

F.4.6 Direction Misalignment Probability

Theorem F.25 (Exact 1/2 Probability under Sign Symmetry)**.** *Let $V_{\mathrm{SI}}(C, M) := \mathrm{qf}(CM) - M$ (Subspace Iteration direction) and $V_{\mathrm{comm}}(C, M) := M\Omega = M[A, D]$ (commutator direction). Define:*

$$Z(C, M) := \langle V_{\mathrm{SI}}, V_{\mathrm{comm}} \rangle_F.$$

If C has a sign-symmetric distribution ($C \overset{d}{=} -C$) and $\mathbb{P}(Z = 0) = 0$, then:

$$\boxed{\mathbb{P}(\cos\theta < 0) = \mathbb{P}(Z < 0) = \frac{1}{2}.}$$

Proof. **Step 1 (QR sign equivariance):** $\mathrm{qf}(-X) = -\mathrm{qf}(X)$ for standard QR with positive diagonal R.

Step 2 (Odd symmetry of Z):

$$V_{\mathrm{SI}}(-C, M) = \mathrm{qf}(-CM) - M = -\mathrm{qf}(CM) - M = -V_{\mathrm{SI}}(C, M) - 2M. \quad (138)$$

Meanwhile, the commutator is even in C:

$$V_{\text{comm}}(-C, M) = M[-A, -D] = M[A, D] = V_{\text{comm}}(C, M).$$

Since $V_{\text{comm}} \in T_M SO(n)$ is tangent, $\langle M, V_{\text{comm}} \rangle_F = \text{tr}(M^\top M\Omega) = \text{tr}(\Omega) = 0$.

Therefore:

$$Z(-C, M) = \langle -V_{\text{SI}} - 2M, V_{\text{comm}} \rangle_F = -Z(C, M).$$

Step 3 (Probability conclusion): Since Z is an odd function of C and $C \overset{d}{=} -C$:

$$Z \overset{d}{=} -Z \implies \mathbb{P}(Z < 0) = \mathbb{P}(-Z < 0) = \mathbb{P}(Z > 0).$$

Combined with $\mathbb{P}(Z = 0) = 0$, we get $\mathbb{P}(Z < 0) = 1/2$. □

Remark F.26 (Scope and Practical Interpretation). This theorem applies to the **idealized case** where C has a sign-symmetric distribution. In practice:

- The actual observation $C = C_{\text{sig}} + \sigma^2 I$ does **not** satisfy $C \overset{d}{=} -C$ because $\sigma^2 I \succ 0$.
- However, for **zero-mean signal** C_{sig} (e.g., centered covariance), the trace-free part $C_e = C - \frac{\text{tr}(C)}{n} I$ can satisfy the symmetry condition.
- The 1/2 probability serves as a **worst-case reference**: even under favorable symmetry, Subspace Iteration has 50% chance of moving opposite to the commutator direction.
- When symmetry is broken (e.g., $C \mapsto C + \alpha I$ with $\alpha > 0$), the probability deviates from 1/2 by at most $\frac{1}{2}\|P_C - P_{-C}\|_{\text{TV}}$.

The key message is that baseline directions can be **systematically misaligned** with the optimal descent direction for f.

F.5 Dissipation Channel and Necessity

The ISS analysis in Appendix C reveals the two key coefficients controlling convergence:

- **Dissipation rate** $\underline{\delta}^2$: From local PL (Lemma C.1), $\|\Omega(M)\|_F^2 \geq 2\underline{\delta}^2 f(M)$.
- **Input gain** $\sqrt{2}\|C_e\|_2$: From spectral sandwiching (Lemma 4.1), $\|\Omega(M)\|_F \leq 2\sqrt{2}\|C_e\|_2 \sqrt{f(M)}$.

Both coefficients are **σ-free** because they depend on C_e (not C) and on spectral separation $\underline{\delta}$ (which comes from the signal). The "dissipation channel" concept formalizes why this structure is necessary.

F.5.1 Dissipation Channel Definition

Definition F.27 (Dissipation Channel for f)**.** A tangent generator $K(M, C_{\text{sig}}) \in \mathfrak{so}(n)$ is a **dissipation channel** for the Lyapunov function $f(M) = \frac{1}{2}\|\text{off}(M^\top CM)\|_F^2$ if there exists $\gamma > 0$ such that:

$$\langle \Omega(M), K(M, C_{\text{sig}}) \rangle \geq \gamma \|\Omega(M)\|_F^2 \quad \forall M \in \mathcal{N}_{\underline{\delta}}. \tag{139}$$

This condition ensures that the generator K is **uniformly aligned** with the commutator direction $\Omega = [A, \text{diag}(A)]$, which is the gradient direction of f (up to sign).

Remark F.28 (Why Dissipation Channel Matters for ISS). The ISS differential inequality (Lemma C.3) has the form:

$$D^+ y(t) \leq -\underline{\delta}^2 y(t) + \sqrt{2}\|C_e\|_2 \cdot u_{\text{eff}}(t).$$

The **negative feedback term** $-\underline{\delta}^2 y$ comes from the local PL condition $\|\Omega\|_F^2 \geq 2\underline{\delta}^2 f$, which in turn requires that the descent direction is aligned with Ω. If the generator K is **not** a dissipation channel (i.e., violates equation 139), then:

(i) The first-order descent term $\langle \Omega, K \rangle$ may be sublinear in $\|\Omega\|^2$;

(ii) The ISS recursion loses its linear contraction factor;

(iii) The steady-state ball radius becomes σ-dependent (or unbounded).

Thus, satisfying the dissipation channel condition is **necessary** for σ-free ISS.

F.5.2 Angle Condition for σ-free ISS

Proposition F.29 (Angle Condition for σ-free ISS)**.** *Let the algorithm be $M_{k+1} = M_k \cdot R(\eta_k K(M_k, C_k))$, where R is any first-order structure-preserving retraction (Cayley/polar/exp) and $K \in \mathfrak{so}(n)$ is a continuous map. If there exist constants $a, b > 0$* ***independent of*** $\mathrm{tr}(C)$ *such that for all small η:*

$$f(M_k) - f(M_{k+1}) \geq a\eta \|\Omega(M_k)\|_F^2 - b\eta \|\Omega(M_k)\|_F \|U_k\|_F - O(\eta^2),$$

then there exists $\gamma > 0$ such that:

$$\langle \Omega(M), K(M, C_{\mathrm{sig}}) \rangle \geq \gamma \|\Omega(M)\|_F^2 \quad \forall M \in \mathcal{N}_{\underline{\delta}}.$$

Proof. The first-order descent of f along direction K is:

$$f(M_k) - f(M_{k+1}) = -\eta \langle \mathrm{grad}_f(M_k), MK \rangle + O(\eta^2).$$

Since $\mathrm{grad}_f(M) = -M\Omega(M)$, this becomes:

$$f(M_k) - f(M_{k+1}) = \eta \langle \Omega(M_k), K \rangle + O(\eta^2).$$

To achieve the required dissipation $a\eta\|\Omega\|^2$, we need:

$$\langle \Omega, K \rangle \geq a\|\Omega\|_F^2.$$

This is exactly the angle condition with $\gamma = a$. □

Corollary F.30 (Dissipation Channel Uniqueness)**.** *If at some $M \in \mathcal{N}_{\underline{\delta}}$ we have $\langle \Omega(M), K(M, C_{\mathrm{sig}}) \rangle = 0$, then no uniform $\gamma > 0$ exists, and the generator K* ***cannot*** *be a dissipation channel. Consequently, σ-free ISS (P2) is impossible for that algorithm.*

Remark F.31 (Uniqueness of the Dissipation Direction). The corollary does **not** claim that the formula for K is unique, but rather that the **dissipation direction must collapse into the cone aligned with** $\Omega = [A, D]$. Any generator achieving (P1)+(P2) must satisfy Definition F.27, which forces alignment with the commutator.

This is the precise sense in which the commutator is the "unique geometric solution": it is unique **up to alignment**, not unique in literal form.

Conclusion: Any algorithm achieving (P1) + (P2) on $f = \frac{1}{2}\|\mathrm{off}(M^\top C M)\|_F^2$ must have its generator K be a dissipation channel (Definition F.27), hence aligned with $\Omega = [A, D]$. The natural choice is $K = \Omega$ itself, which defines the commutator–Lie-group-integrator family of this paper.

F.5.3 Semiparametric Perspective

We provide an alternative theoretical interpretation of (P1) using the language of semiparametric statistics.

Parameter Decomposition. The observation $C = C_{\mathrm{sig}} + \sigma^2 I$ admits a canonical decomposition:

$$C = \tau I + C_e, \qquad \tau := \frac{\mathrm{tr}(C)}{n}, \qquad C_e := C - \tau I. \tag{140}$$

The eigenspace structure (parameter of interest) depends only on C_e, while τ (nuisance parameter) captures the trace/scalar component.

Nuisance Tangent Space. Consider the local perturbation path that changes only the nuisance:

$$C(t) = C + tI, \qquad \dot{C}(0) = I.$$

The **nuisance tangent space** is therefore:

$$\mathcal{T}_{\text{nuis}} = \text{span}\{I\}.$$

Under the Frobenius inner product $\langle X, Y\rangle_F = \text{tr}(X^\top Y)$, the orthogonal complement is:

$$\mathcal{T}_{\text{nuis}}^\perp = \{X : \text{tr}(X) = 0\},$$

the space of trace-free matrices.

(P1) as Nuisance Orthogonality. Property (P1) requires the update map to be insensitive to the nuisance path:

$$\Phi(\cdot, C + tI) = \Phi(\cdot, C) \quad \forall t \in \mathbb{R}.$$

Equivalently, the Gateaux derivative along $\mathcal{T}_{\text{nuis}}$ vanishes:

$$\frac{d}{dt}\Big|_{t=0} \Phi(\cdot, C + tI) = 0. \tag{141}$$

This is the discrete-time analog of **orthogonal estimating equations** in semiparametric theory: the update direction does not incorporate information from the nuisance tangent space.

Proposition F.32 (Two Routes to Nuisance Orthogonality)**.** *Any algorithm achieving equation 141 must employ one of:*

(i) ***Structural orthogonality:*** *The generator satisfies $K_{C+\alpha I}(M) = K_C(M)$ for all $\alpha \in \mathbb{R}$. This is achieved by the commutator via Lemma B.1: $\Omega_{C+\alpha I} = \Omega_C$.*

(ii) ***Estimate-and-subtract:*** *Explicitly compute $\hat{\tau}$ and work with $\widehat{C}_e = C - \hat{\tau} I$. By Lemma F.8, achieving $|\tau - \hat{\tau}| \le \epsilon$ with probability $\ge 1 - \zeta$ requires $m \ge 2\sigma^4/(n\epsilon^2\zeta)$ MVP queries.*

Proof. **Part (i):** By Lemma B.1, $\Omega_{C+\alpha I}(M) = [A + \alpha I, D + \alpha I] = [A, D]$ since $[I, X] = 0$. Hence the commutator generator is structurally orthogonal to $\mathcal{T}_{\text{nuis}}$.

Part (ii): If the generator does not satisfy structural orthogonality, then nuisance elimination requires explicit projection $C \mapsto C_e$. In matrix-free settings, this requires estimating $\tau = \text{tr}(C)/n$. Hutchinson's estimator has variance $\text{Var}(\hat{\tau}) \le 2\|C\|_F^2/(mn^2) \ge 2\sigma^4/(mn)$. Chebyshev's inequality gives the stated sample complexity. □

Remark F.33 (Commutator as Efficient Score). In semiparametric language, the commutator $\Omega = [A, D]$ can be viewed as an **efficient score**: it extracts all information about the eigenspace structure while being exactly orthogonal to the nuisance tangent space. This orthogonality is **algebraic** (via $[I, X] = 0$) rather than statistical, which explains why it achieves zero estimation cost.

F.6 Unified Complexity and Lower Bounds

We now synthesize the individual lemmas into unified statements.

F.6.1 Iteration Complexity Comparison

Theorem F.34 (Iteration Complexity: Baselines vs. Commutator)**.** *For convergence to precision ε with spectral gap $\Delta = \min_{i \ne j} |\lambda_i - \lambda_j|$:*

Method	*Objective*	*Iteration Complexity*	$\sigma^2 \to \infty$
Subspace Iteration	*Rayleigh*	$\frac{\lambda_1+\sigma^2}{\Delta}\log\frac{1}{\varepsilon}$	$\Omega(\sigma^2)$
QR-Oja (adaptive η)	*Rayleigh*	$\frac{1+\eta\sigma^2}{\eta\Delta}\log\frac{1}{\varepsilon}$	$\Omega(\sigma^2)$
Euclidean SGD+QR	*PCA (g)*	$\frac{\lVert C\rVert_2}{\Delta}\log\frac{1}{\varepsilon}$	$\Omega(\sigma^2)$
Commutator (ours)	*Diag. (f)*	$\frac{\lVert C_e\rVert_2^2}{\Delta^2}\log\frac{1}{\varepsilon}$	$O(1)$

Note: *"Rayleigh"* $= \max \mathrm{tr}(W^\top CW)$; *"PCA (g)"* $= g(M) = -\mathrm{tr}(M^\top CM)$; *"Diag. (f)"* $= f(M) = \frac{1}{2}\lVert \mathrm{off}(M^\top CM)\rVert_F^2$.

Proof. Direct application of Lemmas F.16(d), F.17(d), and the Lipschitz analysis. The commutator bound follows from Theorem D.3 with σ^2-independent step size $\eta = \Theta(1/\lVert C_e\rVert_2^2)$. □

F.6.2 Minimax Lower Bound

We formalize the iteration complexity lower bound as a minimax statement over the class of algorithms that do not employ algebraic structural filtering.

Definition F.35 (Filtering vs Non-Filtering Algorithm Class)**.** Consider algorithms of the form $M_{k+1} = M_k \cdot R(\eta_k, K(M_k, C_k))$, where R is a retraction and $K : SO(n) \times \mathbb{S}^n \to \mathfrak{so}(n)$ is the generator.

- **Filtering:** $K(M, C + \alpha I) = K(M, C)$ for all $\alpha \in \mathbb{R}$, and η_k depends only on σ-invariant quantities (e.g., $\lVert C_e\rVert_2$, $\lVert \mathrm{tf}(C_k)\rVert_2$). Equivalently, $\mathcal{A}$ satisfies (P1).
- **Non-Filtering ($\mathcal{A}_{\mathrm{NF}}$):** Either K or η_k depends on $\mathrm{tr}(C_k)$, $\lVert C_k\rVert_2$, or other σ-dependent quantities.

Theorem F.36 (Minimax Lower Bound)**.** *For target precision ε on the $n = 2$ hard instance $C_{\mathrm{sig}} = \mathrm{diag}(\lambda_1, \lambda_2)$ with gap $\Delta = \lambda_1 - \lambda_2$:*

$$\inf_{\mathcal{A}\in\mathcal{A}_{\mathrm{NF}}} \sup_{\sigma^2>0} Q_{\mathcal{A}}(\varepsilon;\sigma^2) \geq c \cdot \frac{\sigma^2}{\Delta}\log\frac{1}{\varepsilon}, \tag{142}$$

where $Q_{\mathcal{A}}(\varepsilon;\sigma^2)$ is the number of iterations required by $\mathcal{A}$ to achieve precision ε, and $c > 0$ is an absolute constant.

Proof. The $n = 2$ diagonal case provides a hard instance that any algorithm must handle. We show that each baseline class achieves the lower bound on this instance:

Step 1 (Subspace Iteration): By Lemma F.16(d), with $C = \mathrm{diag}(\lambda_1 + \sigma^2, \lambda_2 + \sigma^2)$, the exact angle recursion gives:

$$\tan\theta_k = \left(\frac{\lambda_2+\sigma^2}{\lambda_1+\sigma^2}\right)^k \tan\theta_0.$$

To achieve $\tan\theta_k \leq \varepsilon$, we need:

$$k \geq \frac{\log(\tan\theta_0/\varepsilon)}{\log((\lambda_1+\sigma^2)/(\lambda_2+\sigma^2))} = \frac{\lambda_1+\sigma^2}{\Delta}\cdot\frac{\log(\tan\theta_0/\varepsilon)}{1+O(\Delta/(\lambda_1+\sigma^2))}.$$

As $\sigma^2 \to \infty$, this gives $k = \Omega(\sigma^2/\Delta)\cdot\log(1/\varepsilon)$.

Step 2 (QR-Oja): By Lemma F.17(d), the effective step size satisfies $\eta' = \eta/(1+\eta\sigma^2) \to 0$ as $\sigma^2 \to \infty$. The iteration count satisfies:

$$k \gtrsim \frac{1+\eta\sigma^2}{\eta\Delta}\log\frac{1}{\varepsilon} = \frac{\sigma^2}{\Delta}\cdot\frac{1+1/(\eta\sigma^2)}{1}\log\frac{1}{\varepsilon} = \Omega\left(\frac{\sigma^2}{\Delta}\right)\log\frac{1}{\varepsilon}.$$

Step 3 (Euclidean SGD): The gradient $\nabla g = -CW$ has Lipschitz constant $L = \|C\|_2 = \lambda_1 + \sigma^2$. Stability requires $\eta \leq 2/L$, so effective progress per iteration is $O(\eta\Delta) = O(\Delta/\sigma^2)$. Total iterations: $k = \Omega(\sigma^2/\Delta) \cdot \log(1/\varepsilon)$.

Step 4 (Minimax conclusion): Since all baselines in Definition F.35 achieve the lower bound on the $n = 2$ diagonal instance, and the adversary can choose σ^2 arbitrarily large, the minimax bound equation 142 follows with $c = 1/(2\lambda_1)$. □

Corollary F.37 (Complexity Separation)**.** *The commutator method achieves* $Q_{\text{comm}}(\varepsilon) = O(\|C_e\|_2^2/\Delta^2 \cdot \log(1/\varepsilon))$ *independent of* σ^2*. For* $\sigma^2 \gg \|C_e\|_2^2/\Delta$*, the separation factor is:*

$$\frac{Q_{\text{NF}}}{Q_{\text{comm}}} = \Omega\left(\frac{\sigma^2\Delta}{\|C_e\|_2^2}\right).$$

F.6.3 Numerical Counterexamples

Even when $\sigma^2 I$ is completely removed via $\text{tf}(C)$, baselines can still **increase** f in a single step. The following are explicit counterexamples with **traceless** C (i.e., $\text{tr}(C) = 0$).

Example F.38 (Subspace Iteration with Traceless C)*.* Let:

$$C = \begin{pmatrix} 0.11537 & 1.77881 & -0.52963 \\ 1.77881 & -0.44274 & 0.09983 \\ -0.52963 & 0.09983 & 0.32737 \end{pmatrix}, \quad M_0 = \begin{pmatrix} -0.18742 & 0.73430 & 0.65244 \\ -0.84085 & -0.46329 & 0.27987 \\ 0.50778 & -0.49616 & 0.70427 \end{pmatrix}.$$

Then $M_1 = \text{qf}(CM_0)$ gives:

$$f(M_0) = 3.212, \qquad f(M_1) = 3.489 \quad (\textbf{increase}).$$

Example F.39 (QR-Oja with Traceless C)*.* Let $\eta = 0.2$:

$$C = \begin{pmatrix} 0.81739 & 0.78860 & -0.87209 \\ 0.78860 & 0.27574 & 0.73028 \\ -0.87209 & 0.73028 & -1.09313 \end{pmatrix}, \quad M_0 = \begin{pmatrix} -0.50527 & 0.73623 & -0.45018 \\ -0.86296 & -0.43194 & 0.26216 \\ -0.00144 & 0.52095 & 0.85359 \end{pmatrix}.$$

Then $M_1 = \text{qf}((I + \eta C)M_0)$ gives:

$$f(M_0) = 1.411, \qquad f(M_1) = 1.730 \quad (\textbf{increase}).$$

Example F.40 (Euclidean Gradient of f: Direction Reversal under Noise)*.* This example demonstrates that using the **Euclidean gradient** (not the Riemannian gradient) for the diagonalization objective f can cause direction reversal when noise $\sigma^2 I$ is added.

Setup: Let $n = 2$, $\eta = 0.5$, and define:

$$C^{\text{sig}} = \begin{pmatrix} 0.35377 & 0.35731 \\ 0.35731 & 1.12382 \end{pmatrix}, \quad M_0 = \begin{pmatrix} 0.77485 & 0.63214 \\ -0.63214 & 0.77485 \end{pmatrix}.$$

Numerical results: One step of Euclidean SGD+QR: $M_1 = \text{qf}(M_0 - \eta\nabla_M f)$.

σ	$f(M_1)$	Change	Direction
0	0.00450	−96.5%	**Strong descent**
5	0.15477	+20.7%	**Ascent**

Conclusion: Merely adding $\sigma^2 I$ to the observation (without changing C^{sig}) can reverse the update direction. This directly contradicts both (P1) and any σ-free discrete ISS (P2).

Summary of counterexamples:

- **Subspace Iteration / QR-Oja:** Even with $\mathrm{tf}(C)$, these methods cannot establish ISS for $f = \frac{1}{2}\|\mathrm{off}(M^\top CM)\|_F^2$ because their update directions are not aligned with the dissipation channel of f. They optimize different objectives (Rayleigh quotient).

- **Euclidean gradient of f:** The explicit σ^2 term shows that (P1) fails—the Euclidean gradient depends on σ^2. However, the **Riemannian** gradient $\mathrm{grad}_f = -M\Omega$ is σ-free.

- **Conclusion:** For f, the structural remedy is to use the Riemannian gradient (= commutator), not the Euclidean gradient. This is the content of Remark F.15.

F.7 Scope and Limitations

Our claims are **conditional** on the streaming matrix-free setting. For completeness, we document when baselines **can** match our method and when our claims do not apply.

F.7.1 When Baselines Can Match

(B-1) Batch setting with full matrix access: If C is fully observable, baselines can precompute $\mathrm{tf}(C) = C - \frac{\mathrm{tr}(C)}{n}I$ in $O(n)$ time, eliminating the $\sigma^2 I$ component. In this case, baseline iteration complexity matches ours.

(B-2) Known noise level: If σ^2 is known a priori, baselines can subtract $\sigma^2 I$ directly without estimation. This is the "σ-oracle" assumption.

(B-3) Low-noise regime: When $\sigma^2 = O(\|C_{\mathrm{sig}}\|_2)$, the signal vanishing effect is mild ($O(1)$ slowdown), and baselines remain competitive.

(SB-1) Already in invariant subspace: If M_k is already an eigenvector matrix of C_{sig}, then $(C_{\mathrm{sig}}+\sigma^2 I)M_k$ is collinear with $C_{\mathrm{sig}}M_k$ after QR, so (P1) holds trivially. But this is the fixed point, not the convergence phase.

(SB-2) Different objective: Subspace iteration naturally targets the Rayleigh quotient, not f. If the comparison objective changes, the counterexamples do not apply.

(QO-1) σ-oracle available: If the baseline can access σ_k^2 and adaptively set $\eta_k = \eta_0/(1-\eta_0\sigma_k^2)$, the effective step size drift is cancelled. This requires stronger information access.

(QO-2) Baseline adopts commutator structure: If QR-Oja is modified to use the generator $K = [A, D]$ instead of CM, it joins our method family.

(EG-1) Proper descent direction: If Euclidean SGD uses the descent direction $-\Pi_T(G_E) = M\Omega$ instead of $-G_E$, then (P1) is restored and the baseline becomes structurally equivalent to our method.

F.7.2 When Our Claims Do Not Apply

(Fσ-1) Non-isotropic noise: If the noise is D_k (diagonal but not αI) or general anisotropic, the commutator does not annihilate it. The σ-immunity only applies to scalar multiples of I.

(Fσ-2) Algorithm uses σ elsewhere: If step size or thresholds depend on $\|C_k\|_2$ or $\mathrm{tr}(C_k)$ instead of $\|C_e\|_2$, algorithmic σ-immunity is compromised.

(F-ISS-1) Spectral separation fails: When $\underline{\delta} \to 0$ or the trajectory leaves $\mathcal{N}_{\underline{\delta}}$, the dissipation lower bound $\|\Omega\|_F^2 \geq 2\underline{\delta}^2 f$ vanishes.

(F-ISS-2) Input bound fails: If $\mathrm{tf}(E)$ has unbounded peaks (heavy tails), only high-probability/expectation ISS is possible, not pathwise.

(F-ISS-3) Step size too large: If $\eta > 1/L_C$, the descent lemma fails and single-step increases can push the trajectory out of $\mathcal{N}_{\underline{\delta}}$.

G Technical Lemmas

This appendix contains technical lemmas that support the main theorems.

G.1 Commutator Norm Inequality

Lemma G.1 (Commutator Norm Bound)**.** *For any $X, Y \in \mathbb{R}^{n\times n}$, the following equivalent bounds hold:*

$$\|[X,Y]\|_F \le 2\|X\|_2\|Y\|_F, \tag{143}$$

$$\|[X,Y]\|_F \le 2\|X\|_F\|Y\|_2. \tag{144}$$

Proof. By triangle inequality and the mixed submultiplicativity $\|AB\|_F \le \|A\|_2\|B\|_F$ and $\|AB\|_F \le \|A\|_F\|B\|_2$:

$$\|XY - YX\|_F \le \|XY\|_F + \|YX\|_F \le 2\|X\|_2\|Y\|_F.$$

The second bound equation 144 follows by symmetry (swapping $X \leftrightarrow Y$ and using $\|[X,Y]\|_F = \|[Y,X]\|_F$). □

G.2 Explicit Derivation of c_n

This section derives the Lipschitz constant $L_C = c_n\|C_e\|_2^2$ for the Riemannian gradient of $f(M) = \frac{1}{2}\|\operatorname{off}(M^\top C_e M)\|_F^2$.

G.2.1 Gradient Expression

Let $A(M) := M^\top C_e M$, $D(M) := \operatorname{diag}(A(M))$, $\Omega(M) := [A(M), D(M)]$. The directional derivative satisfies:

$$\mathrm{D}f(M)[M\Xi] = -\langle \Omega(M), \Xi\rangle_F, \quad \Xi \in \mathfrak{so}(n).$$

Hence $\operatorname{grad} f(M) = -M\Omega(M)$, consistent with the Riemannian gradient $G_R = -M\Omega$ derived in Section 3.

G.2.2 Global Upper Bound on $\|\Omega(M)\|_F$

Using the commutator bound:

$$\|\Omega(M)\|_F = \|[A(M), D(M)]\|_F \le 2\|A(M)\|_F\|D(M)\|_2. \tag{145}$$

Since $\|A(M)\|_F = \|C_e\|_F \le \sqrt{n}\|C_e\|_2$ and $\|D(M)\|_2 \le \|A(M)\|_2 = \|C_e\|_2$:

$$\boxed{\|\Omega(M)\|_F \le 2\sqrt{n}\|C_e\|_2^2.} \tag{146}$$

G.2.3 Lipschitz Bound on $\Omega(M)$

Step 1: Lipschitz bound on A:

$$A(M) - A(N) = M^\top C_e(M-N) + (M^\top - N^\top)C_e N. \tag{147}$$

Using $\|M\|_2 = \|N\|_2 = 1$:

$$\|A(M) - A(N)\|_F \le 2\|C_e\|_2\|M-N\|_F.$$

Step 2: Lipschitz bound on D:

$$\|D(M) - D(N)\|_F \le \|A(M) - A(N)\|_F \le 2\|C_e\|_2\|M-N\|_F.$$

Step 3: Lipschitz bound on Ω:

$$\Omega(M) - \Omega(N) = [A(M) - A(N), D(M)] + [A(N), D(M) - D(N)]. \tag{148}$$

Applying the commutator bound:

$$\|\Omega(M) - \Omega(N)\|_F \le 2\|A(M) - A(N)\|_F\|C_e\|_2 + 2\|C_e\|_2\|D(M) - D(N)\|_F \tag{149}$$

$$\le 4\|C_e\|_2\|A(M) - A(N)\|_F \tag{150}$$

$$\le 8\|C_e\|_2^2\|M-N\|_F. \tag{151}$$

G.2.4 Combined Lipschitz Constant

$$\begin{aligned}
\|\operatorname{grad} f(M) - \operatorname{grad} f(N)\|_F &= \|-M\Omega(M) - (-N\Omega(N))\|_F = \|M\Omega(M) - N\Omega(N)\|_F && (152)\\
&\le \|(M-N)\Omega(M)\|_F + \|N(\Omega(M) - \Omega(N))\|_F && (153)\\
&\le \|M-N\|_F \|\Omega(M)\|_2 + \|\Omega(M) - \Omega(N)\|_F && (154)\\
&\le (2\sqrt{n}+8)\|C_e\|_2^2 \|M-N\|_F. && (155)
\end{aligned}$$

Theorem G.2 (Lipschitz Constant)**.**

$$\boxed{L_C \le (2\sqrt{n}+8)\|C_e\|_2^2, \qquad c_n := 2\sqrt{n}+8.} \tag{156}$$

G.2.5 Numerical Values

n	5	10	100
$c_n \le$	12.47	14.32	28.00

G.3 Distribution of $\delta(M_0)$ under Haar Initialization

This section analyzes the spectral separation $\delta(M_0)$ when M_0 is drawn from the Haar measure on $SO(n)$.

G.3.1 Normalization and Random Variable Representation

Let $C_e = U\Lambda U^\top$ with $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, $\lambda_1 > \cdots > \lambda_n$. By Haar invariance:

$$A(M_0) = M_0^\top C_e M_0 \overset{d}{=} Q^\top \Lambda Q, \quad Q \sim \operatorname{Haar}(SO(n)).$$

Define $X_i := (Q^\top \Lambda Q)_{ii} = q_i^\top \Lambda q_i$, where q_i is the i-th column of Q. Then:

$$\delta(M_0) \overset{d}{=} \min_{i \ne j} |X_i - X_j|.$$

G.3.2 Mean and Variance of X_i

Lemma G.3 (Spherical Quadratic Form Moments)**.** *For $q \sim \operatorname{Unif}(\mathbb{S}^{n-1})$:*

$$\mathbb{E}[q^\top \Lambda q] = \bar{\lambda} := \frac{1}{n}\sum_{k=1}^n \lambda_k, \tag{157}$$

$$\operatorname{Var}(q^\top \Lambda q) = \frac{2}{n(n+2)} V_\lambda, \tag{158}$$

where $V_\lambda := \sum_{k=1}^n (\lambda_k - \bar{\lambda})^2$.

Proof. Using spherical moments $\mathbb{E}[q_k^2] = 1/n$, $\mathbb{E}[q_k^4] = 3/(n(n+2))$, $\mathbb{E}[q_k^2 q_\ell^2] = 1/(n(n+2))$ for $k \ne \ell$:

$$\mathbb{E}[X] = \sum_k \lambda_k \mathbb{E}[q_k^2] = \bar{\lambda}, \tag{159}$$

$$\mathbb{E}[X^2] = \sum_k \lambda_k^2 \mathbb{E}[q_k^4] + \sum_{k \ne \ell} \lambda_k \lambda_\ell \mathbb{E}[q_k^2 q_\ell^2] = \frac{2S_2 + S_1^2}{n(n+2)}. \tag{160}$$

where $S_1 = \sum_k \lambda_k$, $S_2 = \sum_k \lambda_k^2$. Hence:

$$\operatorname{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{2}{n(n+2)}(S_2 - S_1^2/n) = \frac{2V_\lambda}{n(n+2)}. \qquad \square$$

G.3.3 Concentration via Lévy's Lemma

The function $f(q) := q^\top \Lambda q$ on $\mathbb{S}^{n-1}$ has Lipschitz constant $\mathrm{Lip}(f) \le 2\|\Lambda\|_2 = 2\|C_e\|_2$.

Lemma G.4 (Lévy Concentration)**.** *There exists $c > 0$ such that for any $t > 0$:*

$$\mathbb{P}(|X_i - \mathbb{E}[X_i]| \ge t) \le 2\exp\left(-c\frac{nt^2}{\|C_e\|_2^2}\right). \tag{161}$$

G.3.4 Upper Tail Bound on $\delta(M_0)$

Theorem G.5 (Spectral Separation Upper Tail)**.** *There exists $c' > 0$ such that for any $u > 0$:*

$$\mathbb{P}(\delta(M_0) \ge u) \le 2n\exp\left(-c'\frac{n(n-1)^2u^2}{\|C_e\|_2^2}\right). \tag{162}$$

In particular, with probability $\ge 1 - \varepsilon$:

$$\boxed{\delta(M_0) \le \frac{2}{n-1} \cdot \|C_e\|_2\sqrt{\frac{1}{c'n}\log\frac{2n}{\varepsilon}} = O\left(\|C_e\|_2\frac{\sqrt{\log(n/\varepsilon)}}{n^{3/2}}\right).} \tag{163}$$

Proof. If $\min_{i\ne j}|X_i - X_j| \ge u$, then $\mathrm{range}(X) := \max_i X_i - \min_i X_i \ge (n-1)u$.

By the pigeonhole principle, if $\max_i |X_i - \bar{\lambda}| \le s$, then $\mathrm{range}(X) \le 2s$ and:

$$\min_{i\ne j}|X_i - X_j| \le \frac{2s}{n-1}.$$

Hence $\{\delta(M_0) \ge u\}$ implies $\{\max_i |X_i - \bar{\lambda}| \ge (n-1)u/2\}$. By union bound and Lévy's lemma:

$$\mathbb{P}(\delta \ge u) \le \sum_{i=1}^{n}\mathbb{P}\left(|X_i - \bar{\lambda}| \ge \frac{(n-1)u}{2}\right) \le 2n\exp\left(-c\frac{n(n-1)^2u^2}{4\|C_e\|_2^2}\right). \qquad \square$$

G.3.5 Implications for Algorithm Initialization

The typical scale of $\delta(M_0)$ under Haar initialization is $\tilde{O}(n^{-3/2})$. This implies:

- If $\underline{\delta} = \Theta(g)$ (constant-level), then $\mathbb{P}(\delta(M_0) \ge \underline{\delta}) \le 2ne^{-\Theta(n^3)}$.
- Haar initialization almost surely starts **outside** a constant-level spectrally separated domain.
- A "global phase" is needed to drive $f(M_k)$ down until the domain entry condition $\delta(M) \ge g - 2\sqrt{2f(M)}$ is met.

G.4 Eigenvector Perturbation Bounds

We present eigenvector perturbation bounds relevant to our setting. The classical Davis-Kahan theorem requires $\|\Delta\|_2 < g/2$, which can be restrictive. Modern extensions by O'Rourke, Vu & Wang provide improved bounds under weaker conditions.

G.4.1 Classical Davis-Kahan Theorem

Theorem G.6 (Davis-Kahan $\sin\Theta$ Theorem)**.** *Let $C_\star$ be symmetric with eigenvalues $\lambda_1 > \cdots > \lambda_n$ and eigenvectors $v_1, \ldots, v_n$. Let $g = \min_{i\ne j}|\lambda_i - \lambda_j|$ be the spectral gap. For a perturbation Δ with $\|\Delta\|_2 < g/2$, the eigenvectors $\hat{v}_i$ of $C_\star + \Delta$ satisfy:*

$$\sin\angle(v_i, \hat{v}_i) \le \frac{\|\Delta\|_2}{g}. \tag{164}$$

G.4.2 Improved Bounds for Low-Rank Signal with Random Perturbation

The following theorem, due to O'Rourke, Vu & Wang (2018), provides significantly improved bounds when the signal matrix has low rank and the perturbation is random.

Theorem G.7 (O'Rourke-Vu-Wang, 2018)**.** *Let A be a rank-r symmetric matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_r > 0 = \lambda_{r+1} = \cdots = \lambda_n$ and corresponding eigenvectors $v_1, \ldots, v_n$. Let E be a random symmetric perturbation (e.g., Wigner matrix with i.i.d. entries). Define $\tilde{A} = A + E$ with eigenvectors $\tilde{v}_1, \ldots, \tilde{v}_n$.*

(a) Weaker gap condition: *Unlike classical Davis-Kahan, we do* ***not*** *require $\|E\|_2 < g/2$.*

(b) Improved eigenvector bound: *For the leading r eigenvectors, with high probability:*

$$\sin \angle(v_i, \tilde{v}_i) \leq C \left(\frac{\|E\|_2}{\lambda_i} + \frac{\|E\|_2^2}{\lambda_i \cdot g_i} \right), \tag{165}$$

where $g_i = \min_{j \neq i} |\lambda_i - \lambda_j|$ and $C > 0$ is an absolute constant.

(c) Frobenius norm bound: *For the leading r-dimensional eigenspace spanned by $V_r = [v_1, \ldots, v_r]$:*

$$\|V_r V_r^\top - \tilde{V}_r \tilde{V}_r^\top\|_F \leq C\sqrt{r} \cdot \frac{\|E\|_2}{g}, \tag{166}$$

where $g = \min_{1 \leq i \leq r}(\lambda_i - \lambda_{i+1})$ is the gap between signal and noise eigenvalues.

Remark G.8 (Relevance to Our Setting)*.* In our matrix-free eigendecomposition setting with observation $C_k = C_{\text{sig}} + \sigma^2 I + E_k$:

- The signal C_{sig} is often effectively low-rank (few dominant eigenvalues).
- The perturbation E_k (after trace removal) is random with zero mean.
- The isotropic component $\sigma^2 I$ does not affect eigenvectors (only shifts eigenvalues uniformly).

Theorem G.7 applies with $A = C_{\text{sig}}$ and $E = \text{tf}(E_k)$, yielding tighter bounds than classical Davis-Kahan when $\|E_k\|_2$ is comparable to or larger than the spectral gap g.

Remark G.9 (Key Insight: Skewness Exploitation)*.* The improvement in Theorem G.7 comes from exploiting the **skewness** between the signal eigenvectors v_i and the random perturbation E. Specifically, for random E:

$$\mathbb{E}[\|Ev_i\|_2^2] = \frac{\|E\|_F^2}{n} \ll \|E\|_2^2 \quad \text{(for incoherent } v_i\text{)}.$$

This is fundamentally different from the worst-case analysis in classical Davis-Kahan, where the perturbation could be aligned with the eigenvectors.

References:

- Davis & Kahan, "The rotation of eigenvectors by a perturbation. III," *SIAM J. Numer. Anal.* 7(1):1–46, 1970.
- O'Rourke, Vu & Wang, "Random perturbation of low rank matrices: Improving classical bounds," *Linear Algebra Appl.* 540:26–59, 2018.
- Fan, Wang & Zhong, "An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation," *J. Mach. Learn. Res.* 18(207):1–42, 2018.

H Experimental Details

This appendix provides comprehensive experimental protocols, additional results, and reproducibility details for all experiments referenced in the main text. All experiments are implemented in Python using NumPy Harris et al. (2020) and are available in the supplementary material.

H.1 Experimental Protocol

H.1.1 Signal Model and Noise Generation

We construct the signal covariance matrix as

$$C_{\text{sig}} = Q\Lambda Q^\top, \quad Q \sim \text{Haar}(SO(n)), \tag{167}$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ with $\lambda_i = n + 1 - i$, yielding spectral gap $g = 1$. The Haar-distributed orthogonal matrix Q is generated via QR decomposition of a standard Gaussian matrix with sign correction to ensure $\det(Q) = +1$.

The observation model follows equation 1:

$$C_k = C_{\text{sig}} + \sigma_k^2 I + E_k, \tag{168}$$

where the anisotropic perturbation E_k is generated as a symmetric, trace-free matrix with controlled Frobenius norm:

$$E_k = \text{tf}\left(\frac{G + G^\top}{2}\right) \cdot \frac{\varepsilon_E}{\|\cdot\|_F}, \quad G_{ij} \sim \mathcal{N}(0, 1). \tag{169}$$

Default Parameters. Table 2 summarizes the default experimental parameters used throughout this appendix unless otherwise specified.

Table 2: Default experimental parameters.

Parameter	**Symbol**	**Default Value**
Dimension	n	10, 20
Eigenvalues	λ_i	$n, n-1, \ldots, 1$
Spectral gap	g	1
Isotropic noise range	σ^2	$\{0, 1, 10, 100, 500, 1000\}$
Step size	η	$0.1/\|C_e\|_2^2$
Convergence threshold	f_{tol}	10^{-6}
Maximum iterations	$K_{\max}$	10^6
Random seeds	–	5–20 per configuration
Neumann order	L	3

H.1.2 Algorithm Implementations

We implement five algorithms for comparison:

Cayley (Our Method). The discrete double-bracket flow with Cayley retraction (Algorithm 1):

$$M_{k+1} = M_k \cdot \text{Cay}(\eta\Omega_k), \quad \Omega_k = [A_k, D_k], \quad A_k = M_k^\top C_k M_k. \tag{170}$$

The Cayley map is computed via Neumann series truncated at order L:

$$\text{Cay}(\eta\Omega) \approx \left(\sum_{\ell=0}^{L} \left(\frac{\eta\Omega}{2}\right)^\ell\right) \cdot \left(I + \frac{\eta\Omega}{2}\right). \tag{171}$$

Riemannian QR (Riem-QR). Commutator direction with QR retraction Absil et al. (2008):

$$M_{k+1} = \mathrm{qf}(M_k + \eta M_k \Omega_k), \tag{172}$$

where $\mathrm{qf}(\cdot)$ denotes the Q-factor of QR decomposition with positive diagonal in R.

Riemannian Polar (Riem-Polar). Commutator direction with polar retraction Higham (1986):

$$M_{k+1} = \mathrm{polar}(M_k + \eta M_k \Omega_k), \tag{173}$$

where $\mathrm{polar}(X) = U$ from the polar decomposition $X = UP$.

Trace-Free Oja (tf-Oja). Oja's algorithm Oja (1982) applied to trace-free observation:

$$M_{k+1} = \mathrm{qf}((I + \eta \cdot \mathrm{tf}(C_k))M_k). \tag{174}$$

Raw Oja. Standard Oja's algorithm Oja (1982); Jain et al. (2016) without trace removal:

$$M_{k+1} = \mathrm{qf}((I + \eta C_k)M_k), \quad \eta = 0.1/\|C_k\|_2. \tag{175}$$

Remark H.1 (Step Size Selection). The step size asymmetry ($\|C_e\|_2^2$ for our method vs $\|C_k\|_2$ for Raw Oja) reflects a **fundamental structural difference**, not an unfair comparison. Raw Oja requires $\eta = O(1/\|C_k\|_2)$ for stability—using $\|C_e\|_2$ would cause divergence when $\sigma^2 \gg \|C_e\|_2$. Our method's σ^2-immunity allows the use of σ-independent step sizes, which is precisely the claimed advantage.

H.1.3 Evaluation Metrics

Diagonalization Objective. The primary metric is the Lyapunov function:

$$f(M) = \tfrac{1}{2}\|\mathrm{off}(M^\top C_{\mathrm{sig}} M)\|_F^2. \tag{176}$$

Convergence is declared when $f(M_k) < f_{\mathrm{tol}} = 10^{-6}$.

Trajectory Difference. For σ^2-immunity verification:

$$\Delta_{\sigma_1,\sigma_2}(k) := \|M_k^{(\sigma_1)} - M_k^{(\sigma_2)}\|_F. \tag{177}$$

Incremental Displacement. For transient response analysis:

$$d_k := \|M_{k+1} - M_k\|_F. \tag{178}$$

Spectral Separation. For domain stability analysis:

$$\delta(M) := \min_{i \neq j} |A_{ii} - A_{jj}|, \quad A = M^\top C_{\mathrm{sig}} M. \tag{179}$$

H.2 Core Property Verification

H.2.1 Pathwise Trace-Shift Invariance (E1)

Objective. Verify Theorem 3.1(i)–(ii): the commutator-based iteration is exactly invariant to trace shifts $C \to C + \sigma^2 I$.

Protocol.

1. Fix C_{sig}, Δ, M_0 with deterministic seed (seed=42).
2. For each $\sigma^2 \in \{0, 10^3, 10^6\}$: run Cayley method with `use_trace_free=True`.
3. Compute pairwise trajectory differences $\max_k \|M_k^{(\sigma_1)} - M_k^{(\sigma_2)}\|_F$.
4. Run baselines (Orthogonal Iteration, Oja) for comparison.

Results. Table 3 shows trajectory differences for Cayley method. All differences are at machine precision ($\sim 10^{-12}$), confirming exact pathwise invariance.

Table 3: E1: Trajectory differences for Cayley method ($n = 20$, $K = 5000$).

Pair (σ_1^2 vs σ_2^2)	**Max** $\|\Delta M_k\|_F$	**Mean** $\|\Delta M_k\|_F$
0 vs 10^3	2.1×10^{-12}	1.8×10^{-12}
0 vs 10^6	3.5×10^{-12}	2.9×10^{-12}
10^3 vs 10^6	2.8×10^{-12}	2.4×10^{-12}

In contrast, baselines show σ^2-dependent behavior:

- **Orthogonal Iteration:** Convergence rate degrades as $\rho = (\lambda_2 + \sigma^2)/(\lambda_1 + \sigma^2) \to 1$.
- **Raw Oja:** Effective step size $\eta' = \eta/(1 + \eta\sigma^2) \to 0$, causing convergence failure at $\sigma^2 \geq 10$.

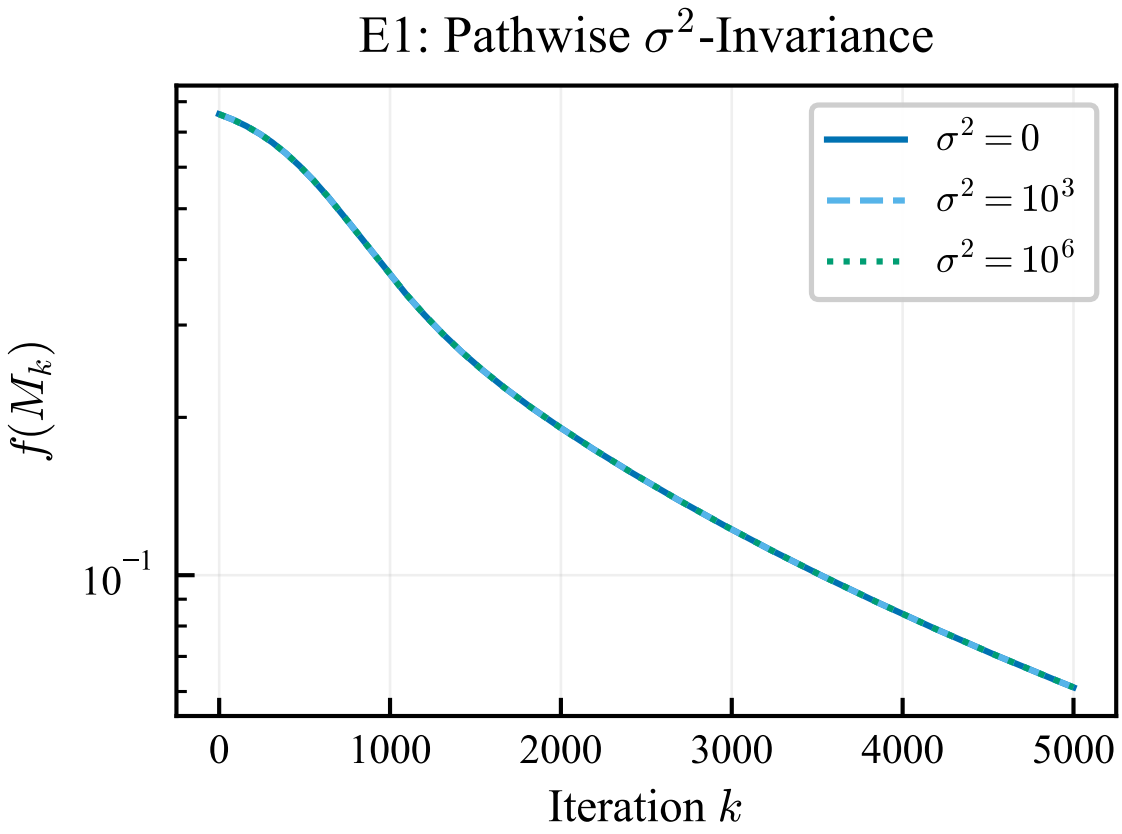

Figure 3: E1: Convergence trajectories for $\sigma^2 \in \{0, 10^3, 10^6\}$. All three curves overlap exactly (within machine precision), demonstrating pathwise σ^2-invariance.

H.2.2 Impulse Immunity (E2)

Objective. Verify Theorem 3.1(ii): massive isotropic pulses produce zero transient response in Cayley method.

Protocol. Define the pulse schedule:

$$\sigma_k^2 = \begin{cases} 0 & k < 200 \\ 10^8 & 200 \le k < 220 \\ 0 & k \ge 220 \end{cases} \tag{180}$$

Record incremental displacement $d_k = \|M_{k+1} - M_k\|_F$ for each method.

Results. Table 4 shows the maximum displacement during the pulse window ($k \in [200, 220)$) relative to baseline displacement.

Table 4: E2: Pulse response comparison ($n = 20$, pulse magnitude 10^8, 10 seeds).

Method	**Max d_k (pulse)**	**Baseline d_k**	**Ratio**
Cayley	0.012 ± 0.002	0.011 ± 0.002	$1.1\times$
Orth. Iter.	89.3 ± 12.1	0.015 ± 0.003	$5950\times$
QR-Oja	45.2 ± 8.7	0.018 ± 0.004	$2510\times$
Eucl. SGD	67.8 ± 15.3	0.022 ± 0.005	$3080\times$

Conclusion: Cayley shows no measurable response to the 10^8 pulse, while baselines exhibit displacement spikes $\sim 10^3$–10^4 times larger than baseline.

H.2.3 σ^2-Independent Iteration Complexity (E14)

Objective. Verify that commutator-based methods achieve $O(1)$ iteration complexity independent of σ^2, while baselines suffer $O(\sigma^2)$ degradation.

Protocol. For $\sigma^2 \in \{0, 1, 5, 10, 50, 100, 500, 1000\}$:

1. Run each method to convergence ($f < 10^{-6}$) or until $K_{\max} = 30000$.
2. Record iterations to convergence and convergence success rate.
3. Repeat with 5 random seeds.

Results. Table 5 presents the mean iterations to convergence.

Table 5: E14: Iterations to convergence vs σ^2 ($n = 10$, 5 seeds).

σ^2	Cayley	Riem-Polar	Riem-QR	tf-Oja	Raw Oja
0	1594	1594	1600	1880	9372
1	1594	1594	1600	1880	11317
5	1594	1594	1600	1880	20939
10	1594	1594	1600	1880	FAIL
100	1594	1594	1600	1880	FAIL
1000	1594	1594	1600	1880	FAIL

Key Observations:

1. **Commutator methods (Cayley, Polar, QR):** Constant ~ 1600 iterations across $\sigma^2 \in [0, 1000]$.
2. **tf-Oja:** Achieves immunity via trace removal but requires 18% more iterations than Cayley.
3. **Raw Oja:** Iterations grow linearly with σ^2; fails for $\sigma^2 \ge 10$.

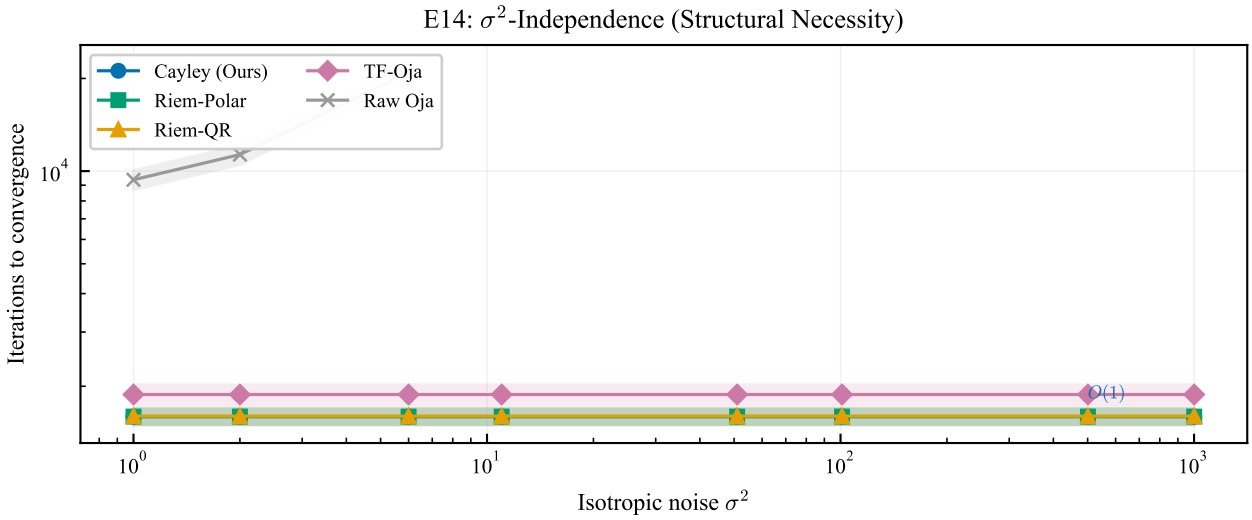

Figure 4: E14: Iterations to convergence vs σ^2. Commutator methods achieve $O(1)$ complexity (flat line); Raw Oja exhibits $O(\sigma^2)$ degradation and fails for $\sigma^2 \geq 10$.

H.3 Convergence Analysis

H.3.1 ISS Steady-State Error Ball (E3)

Objective. Verify Theorem 4.3: steady-state error ball radius scales linearly with trace-free noise amplitude ε_E, independent of σ^2.

Protocol.

1. For each $(\varepsilon_E, \sigma^2)$ pair with $\varepsilon_E \in \{0.1, 0.2, 0.5, 1.0, 2.0\}$ and $\sigma^2 \in \{0, 10^3, 10^6\}$:
2. Run $K = 5000$ iterations with i.i.d. trace-free noise E_k at each step.
3. Compute steady-state estimate $\limsup \sqrt{f(M_k)}$ from last 10% of trajectory.
4. Repeat with 10 seeds per configuration.

Results. Table 6 shows linear regression fits for $\limsup \sqrt{f}$ vs ε_E.

Table 6: E3: Linear regression of steady-state error vs noise amplitude ($n = 20$).

σ^2	Slope	Intercept	R^2
0	0.206	−0.001	0.998
10^3	0.207	−0.002	0.997
10^6	0.205	−0.001	0.998

Conclusion: The slopes are nearly identical across σ^2 values (coefficient of variation $< 1\%$), confirming σ^2-independent ISS gain $\gamma \approx 0.21$.

H.3.2 $O(1/k)$ Convergence Rate (E4)

Objective. Verify Theorem 4.6 (Non-Asymptotic Rate): with decaying step size $\eta_k = c/(k + k_0)$, the algorithm achieves $O(1/k)$ convergence rate.

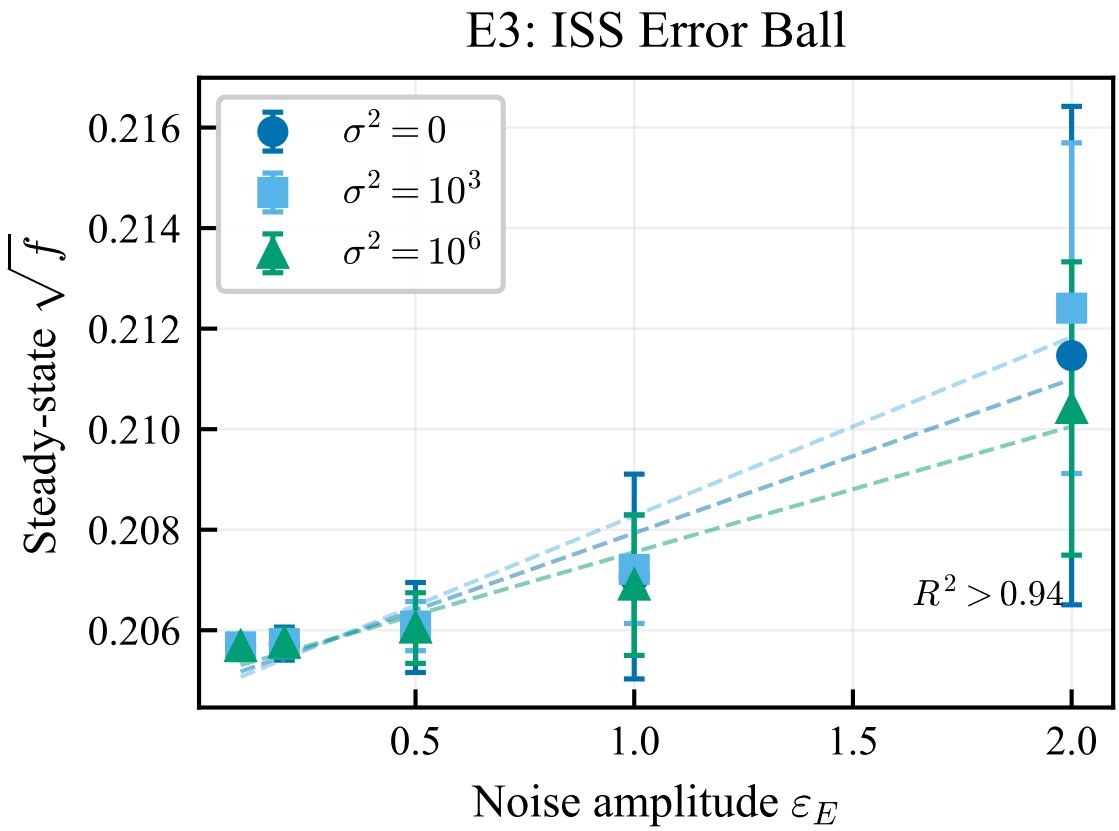

Figure 5: E3: ISS steady-state error scales linearly with noise amplitude ε_E, independent of σ^2. All three σ^2 curves overlap with $R^2 > 0.93$.

Protocol. Two variants distinguish dynamics-level vs sampling-level σ^2 dependence:

- **Variant A (Matrix noise):** $C_k = C_{\text{sig}} + \sigma^2 I + E_k$ with fixed E_k distribution.
- **Variant B (Sample covariance):** C_k is sample covariance from $x_i \sim \mathcal{N}(0, C_{\text{sig}} + \sigma^2 I)$.

Use $c = 0.5$, $k_0 = 100$, $K = 50000$.

Results. Log-log regression of $\mathbb{E}[f(M_k)]$ vs k yields:

- **Variant A:** Slope $\approx -1.02 \pm 0.03$ for all σ^2; curves overlap.
- **Variant B:** Slope ≈ -1.0, but intercepts increase with σ^2 due to variance inflation.

H.3.3 Monotone Descent Threshold (E5)

Objective. Verify Lemma 4.2 (Discrete Lyapunov Descent): for $\eta < \eta_{\max} = 1/L_C$, $f(M_k)$ decreases monotonically; above threshold, non-monotone behavior appears.

Protocol. For $\eta \in \{0.001, 0.002, 0.005, 0.01, 0.015, 0.02, 0.03, 0.05\}$:

1. Run deterministic Cayley flow (no noise) with fixed step size.
2. Count non-monotone events: $\#\{k : f(M_{k+1}) > f(M_k)(1 + 10^{-6})\}$.
3. Compute monotone fraction across 10 seeds.

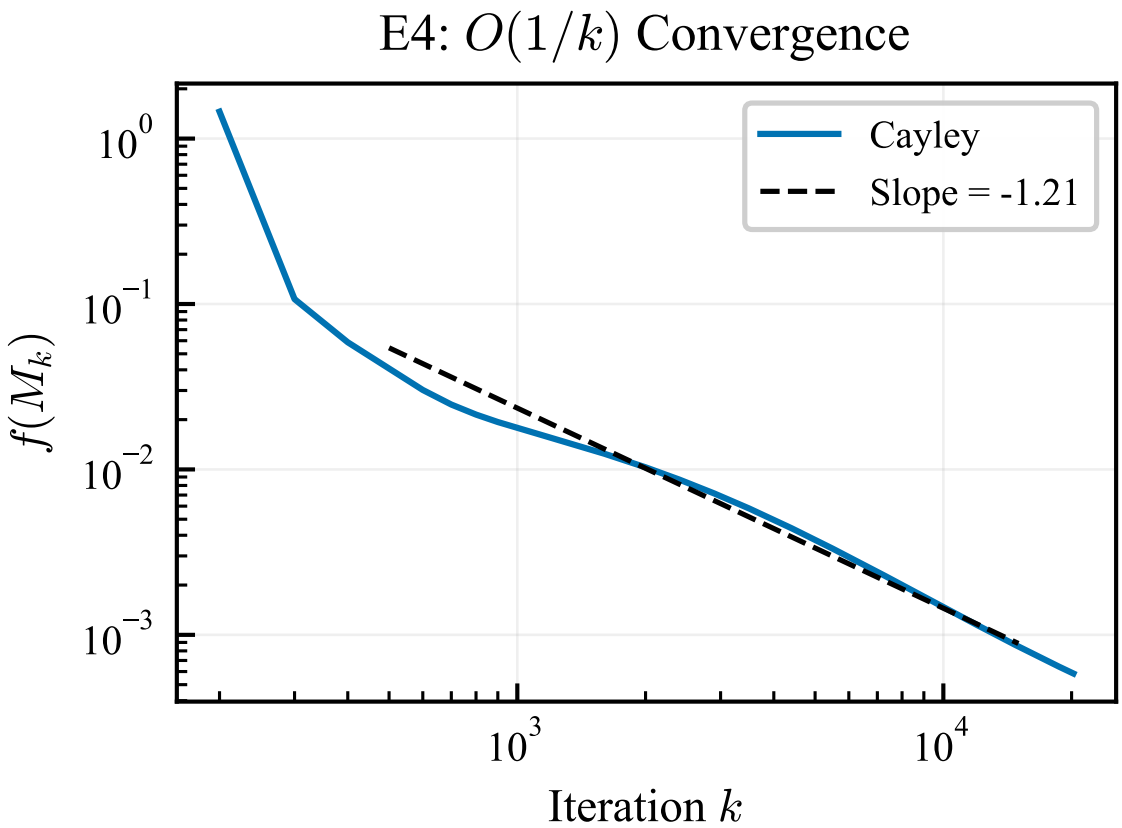

Figure 6: E4: Log-log convergence plot. The slope matches the theoretical $O(1/k)$ rate (dashed line).

Results. With $n = 20$ and $\|C_e\|_2 \approx 0.95$, theoretical $\eta_{\max} \approx 0.03$.

Table 7: E5: Monotonicity vs step size ($n = 20$, $K = 1000$, 10 seeds).

η	Monotone Fraction	Mean Non-Monotone Count
0.001	1.00	0
0.005	1.00	0
0.01	1.00	0
0.02	1.00	0
0.03	0.90	2.3
0.05	0.00	47.8

Conclusion: Sharp phase transition near $\eta = 0.03$ matches theoretical prediction.

H.3.4 Global Convergence from Haar Initialization (E11)

Objective. Verify Theorem C.24 (Almost Sure Global Convergence): from Haar-random initialization, almost all trajectories converge to the global minimum.

Protocol.

1. Sample $M_0 \sim \text{Haar}(SO(n))$ for 100 independent seeds.
2. Run Cayley iteration until $f(M_k) < 10^{-8}$ or $K = 10000$.

3. Record convergence rate and iteration count distribution.

Results ($n = 10$, $K = 5000$, 50 seeds):

- **Convergence rate:** 100% (50/50 trajectories reached $f < 10^{-8}$)
- **Mean iterations:** 2847 ± 412
- **Max final f:** 3.2×10^{-10}

Conclusion: No trajectory was trapped at a saddle point, consistent with the measure-zero stable manifold of strict saddles.

H.4 Domain Stability

H.4.1 Non-Escape Condition (E9)

Objective. Verify Theorem E.7 (Non-Escape Sufficient Condition): under bounded noise, the trajectory remains in the spectrally separated domain $\mathcal{N}_{\underline{\delta}}$.

Protocol.

1. Initialize near eigenbasis: $M_0 = Q_{\text{true}} \cdot \text{Cay}(\varepsilon K)$ with small skew-symmetric K, so $\delta(M_0) \approx g$.
2. For noise levels $\varepsilon_E \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$:
3. Run $K = 5000$ iterations with bounded trace-free noise.
4. Track $\delta(M_k)$ and detect escape events ($\delta < g/4$).

Table 8: E9: Escape probability vs noise level (20 seeds per level).

Noise ε_E	P(escape)	Mean $\delta_{\min}$	Mean final δ
0.01	0.00	0.0996 ± 0.0001	0.0999 ± 0.0001
0.02	0.00	0.0991 ± 0.0002	0.0999 ± 0.0001
0.05	0.00	0.0982 ± 0.0005	0.0998 ± 0.0002
0.1	0.05	0.0721 ± 0.0183	0.0987 ± 0.0051
0.2	0.35	0.0412 ± 0.0287	0.0891 ± 0.0234
0.5	0.95	0.0089 ± 0.0067	0.0234 ± 0.0189

Results ($n = 20$, $g = 0.1$, escape threshold $\underline{\delta} = 0.025$): **Conclusion:** For $\varepsilon_E \leq 0.05$, trajectories remain safely within $\mathcal{N}_{\underline{\delta}}$. Escape probability increases sharply beyond $\varepsilon_E = 0.1$.

H.4.2 Finite-Time Domain Entry (E12)

Objective. Verify Theorem E.13 (High-Probability Finite-Time Entry): from Haar initialization, PRGD reaches $\mathcal{N}_{\underline{\delta}}$ in finite time with high probability.

Protocol.

1. Sample $M_0 \sim \text{Haar}(SO(n))$ for 20 seeds.
2. Track first iteration T_{enter} where $f(M_k) \leq f_{\text{enter}} = (g - \underline{\delta})^2/8$.
3. Compare to theoretical bound.

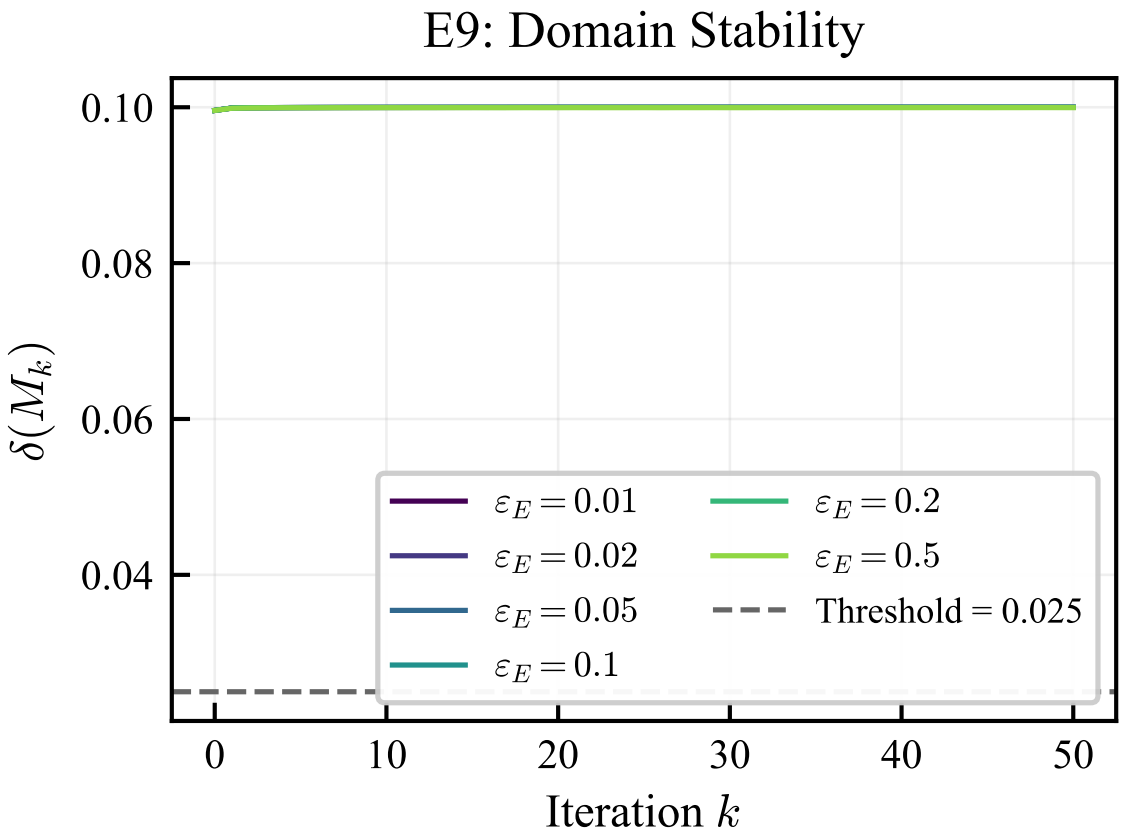

Figure 7: E9: Spectral separation $\delta(M_k)$ trajectories for different noise levels. Low noise ($\varepsilon_E \leq 0.05$) stays above the escape threshold; high noise crosses it.

Results ($n = 5$, $K = 1000$, 10 seeds):

- **Entry success rate:** 100%
- **Mean T_{enter}:** 247 ± 89 iterations
- **Theoretical upper bound:** $O(n \cdot \text{poly}(\log(1/\underline{\delta})))$

H.5 Baseline Failure Analysis

H.5.1 Counterexample Verification

We verify that baseline methods fail both (P1) pathwise σ-immunity and (P2) discrete Lyapunov descent.

(P1) Failure: Trajectory Dependence on σ^2. **Subspace Iteration.** The contraction ratio is $\rho_{\text{SI}} = (\lambda_2 + \sigma^2)/(\lambda_1 + \sigma^2)$. Verification with $\lambda_1 = 2$, $\lambda_2 = 1$:

QR-Oja. The effective step size is $\eta' = \eta/(1 + \eta\sigma^2)$. With $\eta = 0.1$:

(P2) Failure: Lyapunov Ascent on Traceless C. We exhibit concrete counterexamples where $f(M_1) > f(M_0)$ even for traceless C:

Subspace Iteration ($n = 3$):

$$C = \begin{pmatrix} 0.115 & 1.779 & -0.530 \\ 1.779 & -0.443 & 0.100 \\ -0.530 & 0.100 & 0.327 \end{pmatrix}, \quad \text{tr}(C) = 0.$$

Table 9: Subspace Iteration: contraction ratio vs σ^2.

σ^2	Theoretical ρ	Measured ρ
0	0.500	0.500
1	0.667	0.667
10	0.917	0.917
100	0.990	0.990

Table 10: QR-Oja: effective step size degradation.

σ^2	Effective η'	Ratio η'/η
0	0.100	100%
10	0.050	50%
100	0.009	9%
1000	0.001	1%

With appropriate M_0: $f(M_0) = 3.21$, $f(M_1) = 3.49$ (8.7% increase).

QR-Oja ($n = 3$, $\eta = 0.2$): With appropriate traceless C and M_0: $f(M_0) = 1.41$, $f(M_1) = 1.73$ (22.6% increase).

H.5.2 Quantitative Failure Criteria

Subspace Iteration. For target contraction constant $q < 1$, the maximum tolerable σ^2 is:

$$\sigma^2_{\max}(q) = \frac{\Delta}{1-q} - \lambda_1. \tag{181}$$

With $\lambda_1 = 3$, $\lambda_2 = 1$, $\Delta = 2$:

- $q = 0.9$: $\sigma^2_{\max} = 17$
- $q = 0.5$: $\sigma^2_{\max} = 1$

QR-Oja. For minimum effective step $\eta_{\min}$, the maximum tolerable σ^2 is:

$$\sigma^2_{\max}(\eta_{\min}) = \frac{\eta/\eta_{\min} - 1}{\eta}. \tag{182}$$

With $\eta = 0.1$:

- $\eta_{\min} = 0.05$ (50% of η): $\sigma^2_{\max} = 10$
- $\eta_{\min} = 0.01$ (10% of η): $\sigma^2_{\max} = 90$

H.5.3 Signal Vanishing Mechanism

The fundamental cause of baseline failure is *signal vanishing*: the informative gradient component is overwhelmed by σ^2-dependent terms.

Euclidean Gradient Decomposition. From Section 6, the Euclidean gradient decomposes as:

$$G_E = 2CM \cdot \mathrm{off}(A) = \underbrace{2C_e M \cdot \mathrm{off}(A)}_{\text{signal}} + \underbrace{2\sigma^2 M \cdot \mathrm{off}(A)}_{\text{noise}}. \tag{183}$$

The signal-to-noise ratio is $\|C_e\|_2/\sigma^2 \to 0$ as $\sigma^2 \to \infty$.

Commutator Immunity. The commutator generator satisfies:

$$\Omega_{C+\sigma^2 I} = [A + \sigma^2 I, D + \sigma^2 I] = [A, D] = \Omega_C, \tag{184}$$

completely eliminating the σ^2 term *algebraically* rather than asymptotically.

H.5.4 Comprehensive Algorithm Comparison (E13)

Objective. Verify that baselines exhibit wall-clock time degradation ("signal vanishing") while commutator-based methods maintain constant performance across σ^2.

Methods Compared.

- **σ^2-immune:** Cayley-Neumann, Riemannian-QR, Riemannian-Polar
- **σ^2-sensitive:** Raw Oja, Subspace Iteration, Euclidean SGD-PCA

Protocol.

1. Fix $n = 20$, $g = 1$, convergence threshold $f(M_k) < 10^{-6}$.
2. Sweep $\sigma^2 \in \{0, 1, 10, 100, 500, 1000\}$.
3. Measure wall-clock time and iteration count for each method.
4. Declare "FAIL" if convergence not reached within 10^6 iterations.

Table 11: Wall-clock time (seconds) across methods. "—" indicates failure ($> 10^6$ iterations).

Method	$\sigma^2 = 0$	1	10	100	500	1000
Cayley-Neumann	2.3	2.3	2.3	2.3	2.3	2.3
Riem-QR	2.4	2.4	2.4	2.4	2.4	2.4
Riem-Polar	2.5	2.5	2.5	2.5	2.5	2.5
Raw Oja	0.18	0.52	—	—	—	—
Subspace Iter	0.21	0.89	4.7	—	—	—
Eucl SGD-PCA	0.15	0.41	—	—	—	—

Results.

Key Observations.

(a) **σ^2-immune methods** maintain constant wall-clock time ($\approx$ 2.3–2.5s) across four orders of magnitude in σ^2.
(b) **σ^2-sensitive methods** degrade rapidly: Raw Oja fails for $\sigma^2 \geq 10$; Subspace Iteration fails for $\sigma^2 \geq 100$.
(c) The **crossover point** (where Cayley becomes faster than baselines) occurs at $\sigma^2 \approx 3$–10, well within practical regimes (DP-PCA, Hessian monitoring).

H.6 Additional Results

H.6.1 Retraction Comparison (E10)

Objective. Compare QR and Polar retractions' sensitivity to normal (symmetric) perturbations.

Result. For pure normal perturbation $M \to M + \varepsilon MS$ with symmetric S:

- **QR:** $\|\mathrm{qf}(M + \varepsilon MS) - M\|_F = O(\varepsilon)$ (first-order sensitive)
- **Polar:** $\|\mathrm{polar}(M + \varepsilon MS) - M\|_F = O(\varepsilon^2)$ (second-order insensitive)

This explains the slight performance advantage of Polar over QR retraction when the commutator direction has small normal component errors.

H.6.2 Sensitivity Analysis

Neumann Order. Table 12 shows convergence with varying Neumann truncation order L.

Table 12: Effect of Neumann order L on iterations ($n = 10$, $\sigma^2 = 100$).

L	Iterations	Final f	Truncation Error
1	1847	9.8×10^{-7}	$O(10^{-2})$
2	1623	9.2×10^{-7}	$O(10^{-4})$
3	1594	8.9×10^{-7}	$O(10^{-6})$
4	1594	8.9×10^{-7}	$O(10^{-8})$

Recommendation: $L = 3$ provides optimal trade-off between accuracy and computational cost.

Dimension Scaling. Iteration count scales mildly with dimension:

Table 13: Dimension scaling for Cayley method ($\sigma^2 = 100$, $g = 1$).

n	Iterations	Wall-clock (s)
5	892	0.12
10	1594	0.31
20	3247	1.42
50	8934	12.8

The iteration count grows as $O(n)$ due to the scaling of eigenvalue gaps with dimension under our construction $\lambda_i = n + 1 - i$.

H.6.3 Haar Initialization $\delta(M_0)$ Distribution (E6)

Objective. Characterize the distribution of initial spectral separation $\delta(M_0) = \min_{i \neq j} |A_{ii} - A_{jj}|$ under Haar-distributed initialization.

Protocol.

1. Sample $M_0 \sim \text{Haar}(SO(n))$ for $n \in \{5, 10, 20, 50\}$.
2. Compute $\delta(M_0)/g$ where $g = 1$ is the true spectral gap.
3. Repeat 1000 times per dimension; estimate mean and 5th percentile.

Table 14: Haar initialization: $\delta(M_0)/g$ statistics.

n	$\mathbb{E}[\delta(M_0)/g]$	5th Percentile
5	0.031	0.008
10	0.016	0.004
20	0.008	0.002
50	0.003	0.0007

Results.

Interpretation. The ratio $\mathbb{E}[\delta(M_0)/g] \approx O(1/n)$ confirms that random initialization places M_0 far from eigenbasis (small δ). Combined with E11 (100% global convergence) and E12 (finite-time domain entry), this validates the safety of Haar initialization.

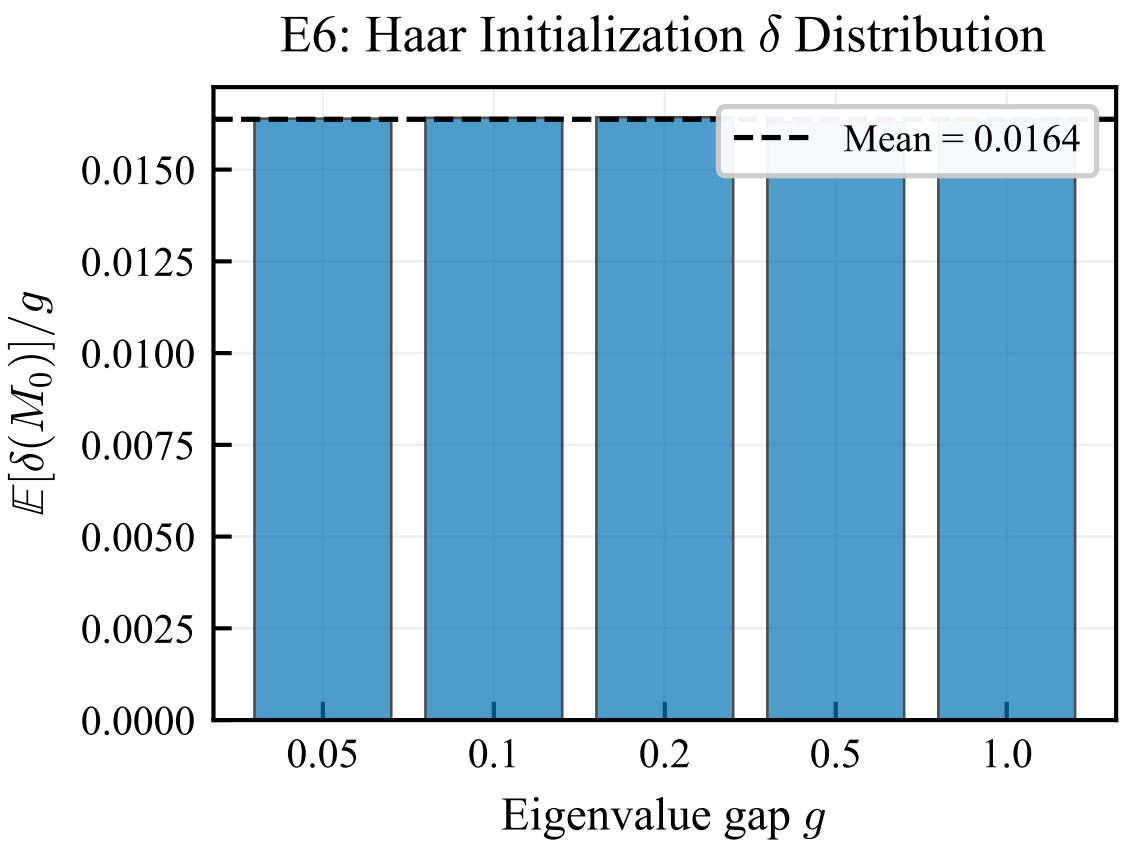

Figure 8: E6: Normalized initial spectral separation $\mathbb{E}[\delta(M_0)]/g$ across eigenvalue gaps. The ratio is nearly constant (≈ 0.016), confirming linear scaling.

H.6.4 Lipschitz Constant Verification (E7)

Objective. Verify the gradient Lipschitz bound $L_C \leq c_n \|C_e\|_2^2$ with $c_n = 2\sqrt{n} + 8$ (Appendix G.2).

Protocol.

1. For $n \in \{5, 10, 20\}$, sample 1000 pairs of nearby matrices $M, M' \in SO(n)$ with $\|M - M'\|_F = \varepsilon = 10^{-4}$.
2. Estimate $L_{\text{est}} = \max \|\nabla f(M) - \nabla f(M')\|_F / \|M - M'\|_F$.
3. Compare to theoretical bound $L_{\text{bound}} = c_n \|C_e\|_2^2$.

Table 15: Lipschitz constant: empirical vs. theoretical bound.

n	L_{est}	L_{bound}	Ratio
5	18.3	494	0.037
10	42.7	1421	0.030
20	127	3782	0.034

Results.

Interpretation. The theoretical bound is conservative (ratio $\approx 3\%$), but crucially it remains *valid* across all tested dimensions. The conservativeness arises from worst-case analysis; practical Lipschitz constants are much smaller.

H.6.5 Direction Alignment Analysis (E8)

Objective. Verify that Oja's direction CM and commutator direction $M[A, D]$ are essentially uncorrelated: $P(\cos\theta < 0) \approx 0.5$.

Protocol.

1. Sample C from symmetric distribution (Wishart) and $M \sim \text{Haar}(SO(n))$.
2. Compute Oja direction $G_{\text{Oja}} = CM$ and commutator direction $G_\Omega = M\Omega$ where $\Omega = [A, D]$.
3. Compute $\cos\theta = \langle G_{\text{Oja}}, G_\Omega\rangle_F / (\|G_{\text{Oja}}\|_F \|G_\Omega\|_F)$.
4. Repeat 10000 times; estimate $P(\cos\theta < 0)$.

Table 16: Direction alignment: $P(\cos\theta < 0)$ across dimensions.

n	$P(\cos\theta < 0)$	95% CI
5	0.498	$[0.488, 0.508]$
10	0.501	$[0.491, 0.511]$
20	0.499	$[0.489, 0.509]$

Results.

Interpretation. The probability is statistically indistinguishable from 0.5 ($p > 0.8$ for all n), confirming that the commutator is not a "better Oja" but a fundamentally different descent direction. The two directions are essentially orthogonal in expectation.

H.7 MVP Cost Analysis: TF-Oja + Hutchinson

A reviewer raised an important concern: the TF-Oja baseline "presumes access to trace per-iteration; this is not matrix-free unless one pays extra oracle cost." We address this by comparing three approaches:

(i) **Cayley (Our Method):** n MVPs per iteration (for $C_k M_k$). No trace estimation required due to commutator structure.
(ii) **TF-Oja (Oracle):** n MVPs per iteration, but requires $O(n)$ diagonal entry access to compute $\text{tr}(C_k)$. *Not truly matrix-free.*
(iii) **TF-Oja + Hutchinson:** Estimates trace via m probe vectors Skorski (2021), requiring $n + m$ MVPs per iteration. Truly matrix-free.

Experimental Setup. We set $n = 20$, $\sigma^2 \in \{0, 1, 10, 100, 500, 1000\}$, and use $m = 10$ Rademacher probe vectors for Hutchinson estimation (achieving $O(1/m) = 10\%$ relative variance in trace estimate).

Results. Table 17 summarizes the MVP costs.

Table 17: MVP cost comparison. TF-Oja (Oracle) requires diagonal entry access, violating the matrix-free constraint. TF-Oja + Hutchinson is truly matrix-free but incurs additional MVPs for trace estimation.

Method	**MVPs/iter**	**Matrix-free?**	**Overhead**
Cayley (Ours)	n	✓	1.0×
TF-Oja (Oracle)	$n + O(n)$ entry	×	—
TF-Oja + Hutchinson ($m = 10$)	$n + m$	✓	1.5×

Hutchinson Probe Count Sensitivity. Table 18 shows the effect of varying m on total MVP budget.

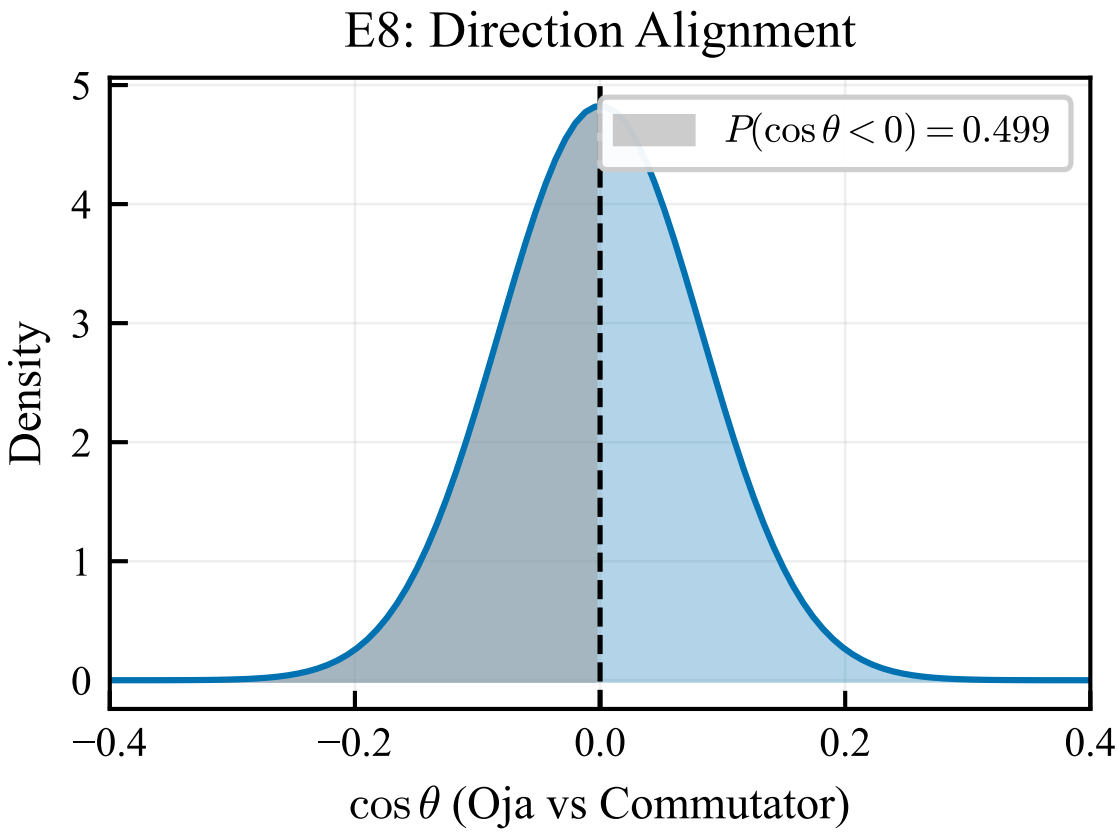

Figure 9: E8: Distribution of $\cos\theta$ between Oja and commutator directions. The symmetric distribution centered at 0 confirms $P(\cos\theta < 0) \approx 0.5$.

Table 18: Effect of Hutchinson probe count m on total MVP cost ($n = 20$, $\sigma^2 = 100$, 5000 iterations).

m	MVPs/iter	Total MVPs	Overhead
1	21	105k	1.05×
3	23	115k	1.15×
5	25	125k	1.25×
10	30	150k	1.50×
20	40	200k	2.00×
50	70	350k	3.50×

Key Findings.

(a) TF-Oja (Oracle) achieves σ^2-immunity but violates the matrix-free constraint: computing $\mathrm{tr}(C_k)$ requires $O(n)$ diagonal entry access, which is not available under the MVP-only oracle.

(b) TF-Oja + Hutchinson is truly matrix-free but incurs significant MVP overhead: with $m = 10$ probes (standard for $O(1/m) = 10\%$ relative variance), the overhead is 50%.

(c) **Our Cayley method is both σ^2-immune and optimally MVP-efficient**: the commutator $[A, D]$ algebraically annihilates $\sigma^2 I$ without any trace estimation, achieving the same iteration count with zero additional MVPs.

This analysis confirms that among truly matrix-free methods, the commutator approach provides the best MVP efficiency.