SPECTRAL ANALYSIS OF THE DYNAMICS OF GRADIENTS AND UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Let $G_T = \{g_1, \ldots, g_T\}$ and $U_T = \{u_1, \ldots, u_T\}$ be the sequences of gradient and update vectors, where (g_i, u_i) are from the i-th iteration of an optimization algorithm or the entire state of a deep neural network (DNN) at the i-th step characterising it on input (x_i, y_i), including backpropagation for minimizing the error between y_i and its prediction. We study the spectral properties of G_T and U_T, for a collection of convex optimization problems and DNNs, such as convolutional neural networks (CNNs) and residual networks (ResNet). Our experiments show that they are sharply energy-concentrated in a small dimension in the spectral domain.

Now, focusing only on optimization algorithms, for a given choice of descent such as Nesterov accelerated gradient descent (NAG), we define a corresponding Nesterov-Ramanujan descent where the vectors in the gradient and update histories are modified with a transformation $v \mapsto Mv$. Here, M is a sparse adjacency matrix of a randomly directed version of a Ramanujan graph that is ergodic, rapidly mixing and has a large girth and the edges are 2-colored with random directions. Thus for the original NAG $M = Id$. Our update rule uses U_t, G_t and computes a step size Δ_t and (g_{t+1}, u_{t+1}). Then, spectral characteristics of one of U_t or G_t improve and we show how to use it to stably increase the step sizes, typically enabling faster convergence. Surprisingly, U_T is more informative, while the momentum methods use mainly G_T.

1 INTRODUCTION

Nesterov showed that a time-varying two-step recurrence achieves an $O(t^{-2})$ objective-gap rate (Nesterov, 1983), whereas fixed-coefficient stationary linear first-order methods are limited to worst-case decay of order t^{-1} (Arjevani & Shamir, 2016). This spawned momentum methods that use U_T and Adam variants that implicitly control the evolution in U_T and G_T (Kingma & Ba, 2015). The empirical observation motivating this work is that these history matrices frequently exhibit spectral collapse: most of their energy is concentrated in a small number of singular directions.

Most descent procedures over $\mathbb{R}^d$ are necessarily **local** in that they depend on Taylor-series expansions on a locally convex neighborhood. The line connecting x and $x + h$ must lie in a sufficiently small convex neighborhood in which the expansion remains valid. Within such a neighborhood one writes

$$f(x + h) = f(x) + \nabla f(x)^\top h + \frac{1}{2} h^\top H(x) h + o(\|h\|^2).$$

Usual convexity or strong-convexity assumptions control the behavior of the Hessian and therefore the size of admissible steps. This creates two competing requirements: (1) h must be small enough that the Taylor expansion remains accurate and does not create instability; and (2) h must be large enough to achieve meaningful progress. If a region is essentially flat, then f is nearly constant along the segment joining x and $x + h$, and one may take a comparatively large step while producing only a small change in objective value. On the other hand, if local curvature is large, meaning that the Hessian possesses large eigenvalues in relevant directions, then admissible step sizes must necessarily be much smaller to maintain stability and the validity of the local approximation.

In general, given G_t, g_{t+1} may be uniquely defined, but not u_{t+1} which may be probabilistically chosen. In fact, non-local methods may be gradient-free in the sense that $G_T = \emptyset$ and they bound

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

progress directly through $f(x_i) - \alpha^\star$, where $\alpha^\star = \min_x f(x)$; random-subspace global optimization on Grassmannians provides one such outer objective-gap analysis (Gupta et al., 2026).

Our methods can be viewed as a universal, optimizer-independent template for weighting U_T and G_T while preserving the same stationary optimum under certain conditions.

The key observation concerns how the graph structure allows gradients and updates at one location or level to affect those at other locations without the effect immediately cycling back. In this sense, ResNets create shorter computational paths and fare better than otherwise comparable plain networks (He et al., 2016). Our graph has diameter and girth logarithmic in the circuit size (number of vertices). Our constructions derive their graphs from Ramanujan graphs due to Lubotzky, Phillips, and Sarnak (Lubotzky et al., 1988), or can use nearly Ramanujan graphs based on GRH (Jao et al., 2009).

Contributions

1. **Optimizer-history framework.** Nesterov's results motivated the use of gradient and update history in optimizers such as NAG and later adaptive momentum methods. We formulate the dynamics through the paired histories G_T and U_T and provide an approach for applying weighted transformations to both histories. The framework is universal, applicable to all optimizers and neural network architectures.
2. **Spectral-collapse motivation.** The framework is motivated by the spectral collapse of G_T and U_T onto a small number of dimensions: we experimentally discovered this phenomenon across several convex problems and DNN training runs.
3. **RDLPS graph family.** We identify the properties needed and construct a class of sparse directed graphs, called RDLPS (Randomized Directed LPS graphs), derived from undirected LPS graphs which have an optimal spectral gap, and a logarithmic diameter and girth. In general, spectral gap in a directed case does not imply rapid mixing, but we show that it is true for RDLPS. These graphs have $O(1)$ clique size which avoids certain instability issues in experiments. The construction is summarized in section 3.
4. **Nesterov–Ramanujan Descent and adaptive rule.** We design Nesterov–Ramanujan Descent by modifying each vector v in $G_T \cup U_T$ with a transformation $v \mapsto Mv$, where M is derived from the adjacency of the RDLPS graph. We use an algorithm that takes (G_t, U_t) as input and selects the increment Δ_t and the next pair (g_{t+1}, u_{t+1}). Under suitable hypotheses, this leads to convergence in fewer iterations; see appendix section E. Since our experimental scale is admittedly small, the experiments can only serve as a heuristic.

2 Related Work

Acceleration and history geometry. Heavy-ball and Nesterov acceleration introduced influential momentum recurrences (Polyak, 1964; Nesterov, 1983); continuous-time, variational, Lyapunov, and lower-bound analyses clarify their rates and energy decay (Su et al., 2016; Wibisono et al., 2016; Wilson et al., 2021; Arjevani & Shamir, 2016). On quadratics, Krylov and polynomial views connect iterates to spectral filtering, while spectral-density and extrapolation methods use curvature or iterate histories algorithmically (Hestenes & Stiefel, 1952; d'Aspremont et al., 2021; Pedregosa & Scieur, 2020; Scieur et al., 2016). These traditions characterize convergence and spectral filtering, but generally do not treat the spectra of finite-window gradient and realized-update histories as empirical state variables. We extend this modal perspective to the geometry of the realized trajectory subspace: the relevant quantities are the Hessian curvature exposed along the history, its restricted condition number, and whether the history covers directions needed to reach the target. Spectral concentration is therefore not inherently harmful; it becomes a potential obstruction only when persistent concentration aligns with unfavorable curvature or omits target-relevant directions. This connection motivates the target-relative restricted-curvature and reachability analysis in section E, but does not by itself establish that curvature or spectral broadening causes Nesterov acceleration.

Low-dimensional optimization histories. We use the raw finite-window histories G_T and U_T as distinct diagnostics and intervention targets. Entropy-based rank (Roy & Vetterli, 2007) complements evidence that neural optimization and gradients can occupy low-dimensional subspaces (Li

et al., 2018; 2023; Gur-Ari et al., 2018; Azam et al., 2022; Jaiswal et al., 2025). The dynamical interpretation and trajectory-subspace rescue test are detailed in appendix section D.

Directed and decentralized mixing. Nonreversible chains and directed Laplacians require tools beyond a symmetric eigenvalue-gap argument (Fill, 1991; Chung, 2005). Topology and spectral gap also govern communication in decentralized optimization, including time-varying directed and push-sum methods (Duchi et al., 2012; Scaman et al., 2017; Nedić & Olshevsky, 2015; Assran et al., 2019). Decentralized SGD and momentum methods provide additional algorithmic precedents (Lian et al., 2017; Yuan et al., 2021), but they mix information across workers, whereas our operator routes optimizer history within one trajectory.

Ramanujan and large-girth routing. Sparse expanders provide global connectivity with communication cost linear in the number of vertices, and Ramanujan graphs attain the optimal nontrivial-eigenvalue scale for regular graphs (Lubotzky et al., 1988; Chow et al., 2016). Large girth delays short returns and reduces overlap among local routing neighborhoods. These properties motivate the undirected support used by RDLPS before orientation and balancing; by themselves, they do not guarantee rapid mixing for the resulting nonnormal directed operators, which we show separately.

3 Construction of the RDLPS Graph

1. Choose a finite, connected, sparse undirected LPS graph $G = (V, E)$.
2. Choose an unoriented edge $\{i, j\}$ at random and randomly orient it. If the proposed direction creates a directed cycle, reverse it.
3. Continue until every edge is oriented and color all oriented edges green.
4. On the same undirected support, use the same randomized acyclic orientation procedure to propose the red edges. Accept the proposal if every source in green has an incoming red edge and every sink in green has an outgoing red edge, and if the combined green–red directed graph is strongly connected. After a fixed declared number of unsuccessful proposals, orient every red edge opposite to its green copy.
5. Verify that the combined green–red directed graph is strongly connected and has no source or sink.

Thus green and red are two complete orientations of the same edge set, not a partition of it. The red orientation removes sources and sinks left by the green orientation, so alternating colors does not leave a vertex permanently without an incoming or outgoing route.

Termination. The edge-by-edge acyclic-orientation procedure always completes. Suppose the already oriented edges form a directed acyclic graph and consider an unoriented edge $\{i, j\}$. If adding $i \to j$ creates a cycle, then the current graph contains a directed path from j to i. The reverse orientation $j \to i$ cannot also create a cycle, because that would require a directed path from i to j, and the two existing paths would already form a directed cycle. Thus at least one orientation is admissible for every edge. By induction, green is completed after exactly $|E|$ edge assignments; the same argument applies to each red proposal.

The number of random red proposals is bounded by the declared cutoff. If none is accepted, the fallback assigns to each edge the direction opposite to its green direction, requiring exactly another $|E|$ assignments. Reversing all edges preserves acyclicity. For every undirected edge $\{i, j\}$, the green–red pair then contains both arcs $i \to j$ and $j \to i$. Since G is connected, the combined directed graph is strongly connected. Since every vertex has positive undirected degree, every vertex has at least one incoming and one outgoing arc in the combined graph. Thus it has no source or sink, the final check passes, and the procedure stops.

This procedure is randomized, but it need not sample uniformly from all acyclic orientations of G.

Randomness of the two colors. At each green step, choose the next unoriented edge uniformly and choose its proposed direction by an independent fair coin; reverse that direction only when necessary to preserve acyclicity. Thus the completed green orientation is random, although not

necessarily uniform over all acyclic orientations. Use fresh independent edge choices and direction coins for every red proposal. Each proposal is therefore random, but the accepted red orientation is conditioned on repairing the sources and sinks of green and is not independent of green. If the deterministic fallback is invoked, red is exactly the reverse of green and is no longer random conditional on green.

Remark We discovered that even though random regular graphs are almost Ramanujan, they are unsuitable for our purposes since they have small girth. As we note later, we can construct similar graphs using randomized directed constructions starting with GRH based expanders modulo p: each $x < p$, x is connected to $x \cdot p_i$, where p_i is the i-th prime, and $i \leq O(\log(p))^2$. In appendix section B, we can see a slightly different version and also how to concretely verify the properties of an RDLPS graph of a given fixed size. □

Next, we balance each directed adjacency matrix to obtain a doubly stochastic operator W_c, where $c \in \{\mathrm{g}, \mathrm{r}\}$. Add an algebraic self-loop:

$$P_c = (1-\gamma)I + \gamma W_c.$$

The self-loop is a diagonal weight, not an additional communication edge. At iteration t, alternate

$$c_t = \mathrm{g}, \mathrm{r}, \mathrm{g}, \mathrm{r}, \ldots$$

and define

$$M_{c_t} = (1-\alpha)I + \alpha P_{c_t}.$$

Nesterov baseline and RDLPS modification. Write the previous realized displacement as $u_t = x_t - x_{t-1}$. In displacement form, ordinary NAG is

$$m_t^{\mathrm{NAG}} = \beta u_t, \qquad g_t^{\mathrm{NAG}} = \nabla f(x_t + m_t^{\mathrm{NAG}}), \qquad u_{t+1}^{\mathrm{NAG}} = m_t^{\mathrm{NAG}} - \eta_t g_t^{\mathrm{NAG}}.$$

Nesterov–Ramanujan Descent (NRD) uses the RDLPS transport and changes only the entering momentum by replacing βu_t with $\beta M_{c_t} u_t$:

$$m_t = \beta M_{c_t} u_t, \qquad g_t = \nabla f(x_t + m_t),$$

followed by

$$u_{t+1} = m_t - \eta_t g_t, \qquad x_{t+1} = x_t + u_{t+1}.$$

Thus the identity-transport arm $W_{c_t} = I$, and hence $M_{c_t} = I$, recovers ordinary NAG. The evaluated NAG and NRD runs use $\beta = 0.6$.

Adaptive step size. For the evaluated convex controller, f is L-smooth and both L and the minimum $f^\star$ are known. Write $\eta_t = \rho_t / L$; both colored steps of a transaction use the same ρ_t. The controller starts at the largest factor in the authenticated NAG bank $\{0.125, 0.25, 0.5\}$, trials continuation every cycle, and at scheduled checks after every eight completed cycles may also trial the prescribed successor factor when the current factor is below the top of the bank, as after a restart. Among admissible trials it selects the smaller endpoint objective gap, with continuation winning an exact tie. This choice realizes the step-size increment $\Delta_t = \eta_t - \eta_{t-1}$.

The objective-plus-momentum Lyapunov value used by the gate is

$$E(x, u) = f(x) - f^\star + \frac{L}{8}\|u\|_2^2.$$

Admissibility requires finite states and no increase, beyond the declared binary64 tolerance, in the endpoint objective gap or E, together with the registered magnitude and bank-membership checks. If no ordinary trial is admissible, the controller attempts a zero-momentum restart at the smallest factor; if that fails, it holds x_t, clears the momentum, and retains the preceding factor. Only committed accepted or restart pairs (g_t, u_{t+1}) extend G_{t+1} and U_{t+1}; a hold instead writes two zero update columns and marks the corresponding gradient columns unavailable. Gradients from rejected trials are charged but never substituted into the reported histories.

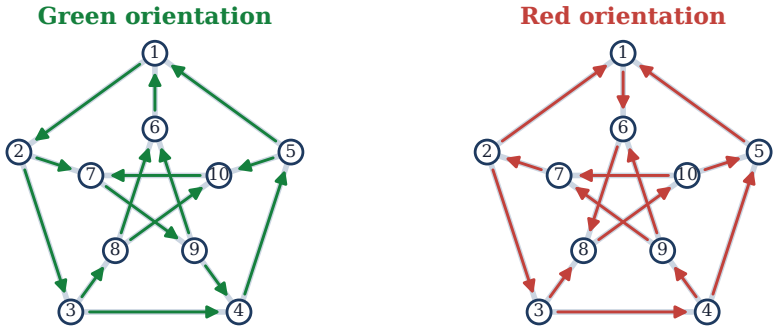


Figure 1: Schematic of the RDLPS routing construction. Green and red are two strongly connected, source/sink-free orientations of the same Petersen support, each with directed girth five and support for doubly stochastic balancing. The support is 3-regular and Ramanujan, with girth five, diameter two, and clique number two. Alternating colors changes the directed routing pattern without changing the support.

4 Study Design and Overview

4.1 Effective spectral rank and energy-coverage rank

We quantify the dimensionality of the raw optimization histories using two complementary spectral statistics: the effective spectral rank D_1 and the 90% energy rank r_{90}. For either the raw gradient history $G_T = [g_1, \ldots, g_T]$ or the raw emitted-update history $U_T = [u_1, \ldots, u_T]$, let

$$H = [h_1, \ldots, h_T] \tag{1}$$

denote the corresponding uncentered history matrix. Let the singular values of H be $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$. We normalize their squared values to obtain the spectral-energy distribution

$$p_i = \frac{\sigma_i^2}{\sum_j \sigma_j^2}, \qquad \sum_i p_i = 1. \tag{2}$$

Effective spectral rank. The order-one effective spectral rank is

$$D_1(H) = \exp\left(-\sum_i p_i \log p_i \right), \tag{3}$$

with the convention $0 \log 0 = 0$. Thus, D_1 is the exponential of the Shannon entropy of the spectral-energy distribution. It can be interpreted as the effective number of directions over which the history energy is distributed. For example, if all energy lies in one direction, then $D_1 = 1$. If the energy is distributed equally over m directions, then $D_1 = m$. In general, D_1 is continuous and need not be an integer.

The 90% energy rank. The raw 90% energy rank is

$$r_{90}(H) = \min\left\{ k : \sum_{i=1}^{k} p_i \geq 0.90 \right\}. \tag{4}$$

It is the minimum number of leading singular directions required to explain at least 90% of the total history energy. Unlike D_1, r_{90} is an integer threshold statistic. It changes only when the cumulative energy crosses the fixed 90% boundary.

Neither D_1 nor r_{90} is the numerical matrix rank. Numerical rank counts singular values larger than a numerical tolerance, whereas D_1 and r_{90} describe how the observed history energy is distributed among the singular directions. The greater sensitivity of D_1 to gradual spectral redistribution is detailed in section A.

4.2 CONVEX LADDER

The adaptive D_1 comparison in table 1 combines eight quadratic families with one smooth nonlinear convex family in dimension $d = 2184$. The quadratic campaign contains 14 family–condition settings with 10 seeds per setting, giving 140 matched NAG cases; its companion HB campaign gives 280 trajectories across both optimizers. The separately preregistered regularized-logistic extension contributes 10 additional matched NAG cases, for 150 NAG comparisons across 15 family–condition settings. Every trajectory runs for 512 updates, and the displayed D_1 table uses raw, uncentered G_T and U_T from the first 128-step window.

Clustered spectrum. For $\kappa \in \{10^2, 10^4, 10^6\}$, let

$$f(x) = \frac{1}{2}x^\top Hx, \qquad H = \mathrm{diag}(\lambda_1, \ldots, \lambda_d),$$

where the first $d/2$ eigenvalues are geometrically spaced from $1/\kappa$ to $2/\kappa$, and the remaining $d/2$ are geometrically spaced from $1/2$ to 1. This produces two separated curvature clusters.

Dense least squares.

$$f(x) = \|Ax - b\|_2^2,$$

where $A \in \mathbb{R}^{4d\times d}$ is generated with condition parameter $\kappa = 50$. This is the single moderate data-induced quadratic in the intervention ladder.

Geometric spectrum. For $\kappa \in \{10^2, 10^4, 10^6\}$, use

$$f(x) = \frac{1}{2}x^\top Hx, \qquad H = \mathrm{diag}(\mathrm{geomspace}(1/\kappa, 1, d)).$$

The spectrum fills the complete curvature interval geometrically.

Ill-conditioned least squares.

$$f(x) = \|Ax - b\|_2^2,$$

with $A \in \mathbb{R}^{4d\times d}$ generated using condition parameter $\kappa = 500$. This preserves the least-squares objective while increasing the separation between its slow and fast directions.

Randomly rotated SPD quadratic.

$$f(x) = \frac{1}{2}x^\top Hx + c^\top x, \qquad H = Q\,\mathrm{diag}\big(\mathrm{linspace}(1, 100, d)\big)Q^\top,$$

where Q is obtained from a seeded Gaussian QR factorization. This is the single random-Q, condition-100 setting.

Ridge regression.

$$f(x) = \|Ax - b\|_2^2 + 0.1\|x\|_2^2,$$

where $A \in \mathbb{R}^{4d\times d}$ is generated with condition parameter $\kappa = 100$.

Sharp spectrum. For $\kappa \in \{10^2, 10^4, 10^6\}$, use

$$f(x) = \frac{1}{2}x^\top Hx, \qquad H = \mathrm{diag}(\lambda_1, \ldots, \lambda_d),$$

where $\lambda_1, \ldots, \lambda_{d-1}$ are geometrically spaced from $1/\kappa$ to $2/\kappa$ and $\lambda_d = 1$. This isolates one sharp high-curvature direction from a low-curvature bulk.

Prescribed-spectrum "Wishart"

$$f(x) = \frac{1}{2}x^\top Hx + c^\top x, \qquad H = Q\,\mathrm{diag}\big(\mathrm{linspace}(1, 200, d)\big)Q^\top.$$

This is a seeded, randomly rotated prescribed-spectrum quadratic with condition parameter 200.

Table 1: Raw-history effective spectral rank before and after adaptive RDLPS. The post-RDLPS column reports the trajectory selected by adaptive RDLPS together with its independently evolved fixed-RDLPS shadow safeguard. Statistics use raw, uncentered 128-step histories under NAG. Higher D_1 indicates that spectral energy is distributed more broadly.

			Original NAG		Adaptive + shadow RDLPS		
Problem	History	Active	Mean D_1	Median D_1	Mean D_1	Median D_1	Improved
Clustered spectrum	G_T	30	1.099	1.082	**1.146**	**1.129**	30/30
	U_T	30	1.106	1.082	**1.163**	**1.122**	30/30
Dense least squares	G_T	10	1.771	1.771	**1.988**	**1.992**	10/10
	U_T	10	1.826	1.825	**2.192**	**2.197**	10/10
Geometric spectrum	G_T	30	1.571	1.569	**1.682**	**1.684**	30/30
	U_T	30	1.622	1.621	**1.835**	**1.836**	30/30
Ill-conditioned least squares	G_T	10	1.397	1.403	**1.475**	**1.478**	10/10
	U_T	10	1.442	1.447	**1.594**	**1.601**	10/10
Random Q	G_T	10	1.975	1.977	**2.219**	**2.219**	10/10
	U_T	10	2.006	2.008	**2.343**	**2.346**	10/10
Ridge	G_T	10	1.619	1.617	**1.761**	**1.759**	10/10
	U_T	10	1.684	1.684	**1.974**	**1.975**	10/10
Regularized logistic regression	G_T	10	1.233	1.232	**1.289**	**1.290**	10/10
	U_T	10	1.229	1.228	**1.274**	**1.273**	10/10
Sharp spectrum	G_T	30	1.075	1.008	**1.101**	**1.016**	30/30
	U_T	30	1.061	1.010	**1.134**	**1.102**	26/30
Wishart	G_T	10	1.756	1.765	**1.923**	**1.927**	10/10
	U_T	10	1.850	1.859	**2.229**	**2.229**	10/10
All nine problems	G_T	150	1.399	1.404	**1.496**	**1.480**	150/150
	U_T	150	1.427	1.447	**1.600**	**1.602**	146/150

Note. $D_1 = \exp(-\sum_i p_i \log p_i)$, where $p_i = \sigma_i^2 / \sum_j \sigma_j^2$. "Improved" means that the adaptive-plus-shadow trajectory has larger D_1 than the matched original-NAG trajectory. Both G_T and U_T improve in 146 of 150 cases. In the remaining four cases, G_T improves while U_T does not; therefore at least one raw history improves in all 150 cases.

Regularized logistic regression.

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \log\bigl(1 + \exp(-b_i a_i^\top x)\bigr) + \frac{\lambda}{2}\|x\|^2.$$

This ninth family replaces fixed quadratic curvature by a Hessian that changes along the trajectory while retaining smoothness, convexity, and regularization. Figure 2 visualizes the matched changes underlying the regularized-logistic rows of table 1.

5 A Simple DNN Spectral-Collapse Study Using Raw Rank and D_1

For a raw, uncentered history $H \in \mathbb{R}^{D\times T}$, let $q := \min(D, T)$ denote its algebraic rank ceiling. We report the numerical raw rank

$$r_{\mathrm{raw}}(H) = \#\Bigl\{ i : \sigma_i > \sqrt{\max(D,T)\epsilon_{64}}\,\sigma_1 \Bigr\}, \qquad r_{\mathrm{raw}} \le q,$$

using a fixed FP64 Gram-spectrum tolerance. The singular values σ_i, energy weights p_i, and entropy dimension D_1 are defined earlier. Raw rank asks whether directions are numerically present. D_1 asks how many directions carry substantial energy. We call a history algebraically collapsed when r_{raw}/q is small, and spectrally collapsed when D_1/q is small even if $r_{\mathrm{raw}}/q \approx 1$. For this study the declared descriptive rule is $r_{\mathrm{raw}}/q \ge .95$ and $D_1/q \le .10$.

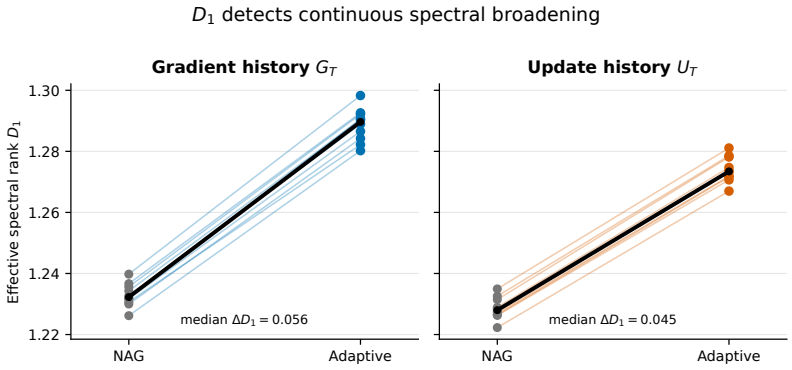


Figure 2: Matched effective-rank changes for the 10-seed regularized-logistic extension reported in table 1. Each thin line joins original NAG to its adaptive-plus-shadow counterpart; the black line joins the medians. The left and right panels report the raw, uncentered 128-step gradient and realized-update histories separately. We picked logistic as an example for visualization.

Table 2: Raw numerical rank and energy-entropy dimension. The span-to-energy gap r_{raw}/D_1 reports how many numerically present directions there are per effectively energetic direction. For example, $50\times$ means that the numerical span is 50 times larger than the effective energetic dimension.

Architecture	Optimizer	History	Raw rank/q	D_1	D_1/q	Span/energy gap
Shallow	SGD	G	256/256	16.86	0.066	15.18
Shallow	SGD	U	256/256	16.86	0.066	15.18
Shallow	ADAMW	G	256/256	9.27	0.036	27.62
Shallow	ADAMW	U	256/256	16.60	0.065	15.43
Deep	SGD	G	256/256	6.76	0.026	37.86
Deep	SGD	U	256/256	6.76	0.026	37.86
Deep	ADAMW	G	256/256	5.06	0.020	50.60
Deep	ADAMW	U	256/256	7.46	0.029	34.30

We train a one-hidden-layer MLP $(64, 128, 10)$ and a four-hidden-layer MLP $(64, 128, 128, 64, 32, 10)$ on the sklearn digits data. Each architecture is trained with plain SGD and AdamW for 1024 minibatch updates, using batch size 64 and three recorded seeds for reproducibility. The train/validation split is stratified independently by seed. After 768 warmup updates, we record gradients and realized parameter updates for 256 steps, producing raw full-network histories $G, U \in \mathbb{R}^{D\times 256}$. We verified that training resulted in high validation accuracy.

ResNet-18 replication. We also tested a CIFAR-10 ResNet-18 using exact full-network temporal Gram matrices over the final 512 training updates. Literal Nesterov acceleration (NAG), heavy-ball momentum, and equal-rate SGD were each run with three recorded seeds. Because the parameter dimension exceeds 512, the algebraic rank ceiling is $q = 512$. The archived analysis retains D_1 but not the thresholded raw rank, so Table 3 reports no inferred raw rank. The capacity-to-energy gap q/D_1 compares the available temporal dimension with the effective energetic dimension; unlike r_{raw}/D_1, it does not assert that all 512 directions are numerically present.

The ResNet-18 result separates gradients from realized updates. NAG and heavy-ball gradients remain spectrally broad, while their update histories satisfy the declared $D_1/q \leq 0.10$ collapse threshold. Equal-rate SGD is less concentrated: its update history has $D_1/q = 0.293$.

6 Discussion and Future Work

We suggest using update history U_T in addition to the gradient history G_T (which is customary in momentum methods) for computing for the next step updates and step sizes.

Table 3: ResNet-18 CIFAR-10 results at $T = q = 512$. Values are medians across three recorded seeds. The archived raw-rank field is unavailable.

Optimizer	History	q	D_1	D_1/q	Capacity/energy gap q/D_1
NAG	G	512	392.70	0.767	1.30
NAG	U	512	46.57	0.091	11.00
Heavy ball	G	512	398.44	0.778	1.29
Heavy ball	U	512	46.18	0.090	11.09
SGD	G	512	241.48	0.472	2.12
SGD	U	512	150.04	0.293	3.41

We discovered and documented the spectral collapse of optimization algorithms such as Nesterov or Adam in the case of both convex optimization as well as certain DNNs.

Finally, we provided a modification using our proposed RDLPS graphs that was able to increase the spectral rank, i.e., avoid spectral collapse in the case of convex optimization problems, which typically leads to convergence in fewer iterations. We also gave a heuristically motivated proof that this would work for convex optimization in general.

In terms of future work, we would like to document the collapse for a wider variety of neural networks at a larger scale. Apart from convex optimization problems, we believe that RDLPS graphs may be useful in intervening in the spectral collapse of neural networks. Using the proposed RDLPS graphs, we could construct a new class of networks as an alternative to standard deep neural network architectures, where each node in the graph contains a small depth standard DNN, and whose input is defined by incoming green and red edges (the output nodes are defined similarly). The network would behave like a typical DNN on training inputs $(x_i, y_i), i = 1, \ldots, T$.

AI Assistance Disclosure

The authors began with mathematical intuitions, manually generated manuscripts and whiteboard sessions containing algorithmic details, conjectures, and proofs. AI-based tools assisted with manuscript formatting, LaTeX preparation, and experimental verification. The computational resources for this project were deliberately very limited, 2 Lenovo workstations with RTX 3080 GPUs. The authors discovered that M365 Copilot does not permit large calculations and cannot directly communicate or share details with other engines such as VSCode: we use this cryptographic isolation to audit one engine's work with the other, in particular experimental results. The authors reviewed and verified the resulting text, code, analyses, and claims and take full responsibility for the paper.

REFERENCES

Yossi Arjevani and Ohad Shamir. On the iteration complexity of oblivious first-order optimization algorithms. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 908–916. PMLR, 2016. URL https://proceedings.mlr.press/v48/arjevani16.html.

Mahmoud Assran, Nicolas Loizou, Nicolas Ballas, and Mike Rabbat. Stochastic gradient push for distributed deep learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 344–353. PMLR, 2019. URL https://proceedings.mlr.press/v97/assran19a.html.

Sheikh Shams Azam, Seyyedali Hosseinalipour, Qiang Qiu, and Christopher Brinton. Recycling model updates in federated learning: Are gradient subspaces low-rank? In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2202.00280.

Steven L. Brunton, Marko Budišić, Eurika Kaiser, and J. Nathan Kutz. Modern Koopman theory for dynamical systems. *SIAM Review*, 64(2):229–340, 2022. doi: 10.1137/21M1401243. URL https://doi.org/10.1137/21M1401243.

Anherutowa Calvo. Predictable gradient manifolds in deep learning: Temporal path-length and intrinsic rank as a complexity regime. *arXiv preprint arXiv:2601.04270*, 2026. URL https://arxiv.org/abs/2601.04270.

Yat-Tin Chow, Wei Shi, Tianyu Wu, and Wotao Yin. Expander graph and communication-efficient decentralized optimization. In *2016 50th Asilomar Conference on Signals, Systems and Computers*, pp. 1715–1720, 2016. doi: 10.1109/ACSSC.2016.7869675. URL https://doi.org/10.1109/ACSSC.2016.7869675.

Fan Chung. Laplacians and the Cheeger inequality for directed graphs. *Annals of Combinatorics*, 9(1):1–19, 2005. doi: 10.1007/s00026-005-0237-z. URL https://doi.org/10.1007/s00026-005-0237-z.

Alexandre d'Aspremont, Damien Scieur, and Adrien Taylor. Acceleration methods. *Foundations and Trends in Optimization*, 5(1–2):1–245, 2021. doi: 10.1561/2400000036.

John C. Duchi, Alekh Agarwal, and Martin J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57 (3):592–606, 2012. doi: 10.1109/TAC.2011.2161027.

James Allen Fill. Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process. *The Annals of Applied Probability*, 1(1): 62–87, 1991. doi: 10.1214/aoap/1177005981. URL https://doi.org/10.1214/aoap/1177005981.

Kartik Gupta, Stephen D. Miller, Pradeep Ravikumar, and Ramarathnam Venkatesan. Stochastic global optimization of continuous functions via random walks on grassmannians. *arXiv preprint arXiv:2605.14151*, 2026. URL https://arxiv.org/abs/2605.14151.

Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018. URL https://arxiv.org/abs/1812.04754.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Magnus R. Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49(6):409–436, 1952. doi: 10.6028/jres.049.044. URL https://doi.org/10.6028/jres.049.044.

Ajay Kumar Jaiswal, Yifan Wang, Lu Yin, Shiwei Liu, Runjin Chen, Jiawei Zhao, Ananth Grama, Yuandong Tian, and Zhangyang Wang. From low rank gradient subspace stabilization to low-rank weights: Observations, theories, and applications. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning*

Research, pp. 26740–26756. PMLR, 2025. URL https://proceedings.mlr.press/v267/jaiswal25a.html.

David Jao, Stephen D. Miller, and Ramarathnam Venkatesan. Expander graphs based on GRH with an application to elliptic curve cryptography. *Journal of Number Theory*, 129(6):1491–1504, 2009. doi: 10.1016/j.jnt.2008.11.006.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Bernard O. Koopman. Hamiltonian systems and transformation in Hilbert space. *Proceedings of the National Academy of Sciences of the United States of America*, 17(5):315–318, 1931. doi: 10.1073/pnas.17.5.315. URL https://doi.org/10.1073/pnas.17.5.315.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1804.08838.

Tao Li, Lei Tan, Zhehao Huang, Qinghua Tao, Yipeng Liu, and Xiaolin Huang. Low dimensional trajectory hypothesis is true: DNNs can be trained in tiny subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3411–3420, 2023. doi: 10.1109/TPAMI.2022.3178101. URL https://doi.org/10.1109/TPAMI.2022.3178101.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5330–5340. Curran Associates, Inc., 2017. URL https://arxiv.org/abs/1705.09056.

Alexander Lubotzky, Ralph Phillips, and Peter Sarnak. Ramanujan graphs. *Combinatorica*, 8(3):261–277, 1988. doi: 10.1007/BF02126799.

Angelia Nedić and Alex Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615, 2015. doi: 10.1109/TAC.2014.2364096. URL https://doi.org/10.1109/TAC.2014.2364096.

Yurii E. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Doklady Akademii Nauk SSSR*, 269(3):543–547, 1983. URL https://www.mathnet.ru/eng/dan46009.

Fabian Pedregosa and Damien Scieur. Acceleration through spectral density estimation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7553–7562. PMLR, 2020. URL https://proceedings.mlr.press/v119/pedregosa20a.html.

Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964. doi: 10.1016/0041-5553(64)90137-5. URL https://doi.org/10.1016/0041-5553(64)90137-5.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *Proceedings of the 15th European Signal Processing Conference*, pp. 606–610, 2007. URL https://infoscience.epfl.ch/handle/20.500.14299/10320.

Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3027–3036. PMLR, 2017. URL https://proceedings.mlr.press/v70/scaman17a.html.

Damien Scieur, Alexandre d'Aspremont, and Francis Bach. Regularized nonlinear acceleration. In *Advances in Neural Information Processing Systems*, volume 29, pp. 712–720. Curran Associates, Inc., 2016. URL https://arxiv.org/abs/1606.04133.

Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17 (153):1–43, 2016. URL https://jmlr.org/papers/v17/15-084.html.

Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016. doi: 10.1073/pnas.1614734113. URL https://doi.org/10.1073/pnas.1614734113.

Ashia C. Wilson, Benjamin Recht, and Michael I. Jordan. A Lyapunov analysis of accelerated methods in optimization. *Journal of Machine Learning Research*, 22(113):1–34, 2021. URL https://jmlr.org/papers/v22/20-195.html.

Kun Yuan, Yiming Chen, Xinmeng Huang, Yingya Zhang, Pan Pan, Yinghui Xu, and Wotao Yin. DecentLaM: Decentralized momentum SGD for large-batch deep training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3009–3019, 2021. doi: 10.1109/ICCV48922.2021.00302. URL https://doi.org/10.1109/ICCV48922.2021.00302.

A Spectral-Diagnostic Sensitivity

Why D_1 is more sensitive to spectral broadening. The principal difference is that D_1 uses the complete spectral-energy distribution, while r_{90} retains only the location at which cumulative energy first reaches 90%. Consequently, D_1 responds continuously when energy moves away from a dominant direction, even if the number of directions needed to capture 90% remains unchanged.

For example, consider the two energy distributions

$$p^{\text{before}} = (0.89, 0.11), \qquad p^{\text{after}} = (0.55, 0.45). \tag{5}$$

Both distributions satisfy

$$r_{90}^{\text{before}} = r_{90}^{\text{after}} = 2, \tag{6}$$

because neither leading direction alone contains 90% of the energy. Nevertheless,

$$D_1^{\text{before}} \approx 1.41, \qquad D_1^{\text{after}} \approx 1.99. \tag{7}$$

The increase in D_1 correctly reflects that the second distribution is substantially less concentrated, whereas r_{90} cannot distinguish the two.

Conversely, r_{90} can change discontinuously after a small redistribution. For

$$p^{\text{before}} = (0.91, 0.09), \qquad p^{\text{after}} = (0.89, 0.11), \tag{8}$$

a two-percentage-point change gives

$$r_{90}^{\text{before}} = 1, \qquad r_{90}^{\text{after}} = 2. \tag{9}$$

This discrete jump is caused by crossing the 90% threshold, not by a correspondingly abrupt change in the underlying spectrum. The effective rank D_1, by contrast, changes smoothly.

For this reason, we use D_1 as the primary diagnostic of spectral broadening. It doesn't require a threshold, uses all of the spectral energy, and detects gradual redistribution that is invisible to an integer energy-coverage count. We retain r_{90} as a complementary statistic because it has a simple operational interpretation: the number of leading directions required to explain 90% of the observed energy. Thus, D_1 is not universally superior to r_{90}; rather, it is better suited to the specific question of whether RDLPS continuously redistributes energy away from a small number of dominant directions.

The essential distinction is that r_{90} measures whether spectral broadening is large enough to cross a fixed 90% energy threshold, whereas D_1 continuously measures the redistribution of energy across the complete spectrum. Consequently, D_1 can reveal meaningful spectral broadening even when the integer-valued r_{90} remains unchanged.

B A Concrete RDLPS Construction

We now present a slightly different construction than mentioned in section 3, and show how to check its properties on a graph of given size. NRD uses a deterministic, directed green/red construction derived from an authenticated LPS$(5, 13)$ graph. The base has 2,184 vertices, degree six, 6,552 edges, exact undirected girth eight, and a verified bipartition. We add exactly 1,024 undirected edges sequentially. Each proposal draws two uniform vertices, rejects loops and existing edges, and is accepted only when the endpoints have current graph distance at least six. The final instance has 7,576 edges, degree range six through nine, is connected and non-bipartite, and has exact undirected girth seven.

Green and red are independent seeded orientations of this same complete support, not a partition of its edges. Base edges are oriented by independent shuffled Euler circuits and added edges by independent direction bits. Each color must contain exactly one arc per support edge, be strongly connected, have no source or sink, have in- and outdegree between three and six, and have directed girth at least seven. Figure 1 illustrates the shared-support, independent- orientation principle on a small explanatory Ramanujan support.

For each color c, let $K_c[i, j] = 1$ exactly when the arc $j \to i$ exists. The graph support and K_c have no self-loops. Sinkhorn row/column scaling produces a nonnegative, zero-diagonal, doubly

stochastic receiver-by-sender operator W_c without changing its 7,576-arc support. The construction exposes two choices:

$$\text{No self-loop: } W_c, \qquad \text{With self-loop: } \widetilde{P}_c = 0.55I + 0.45W_c. \tag{10}$$

The diagonal of $\widetilde{P}_c$ is local retention and does not add a communication edge. NRD subsequently forms $M_c = 0.75I + 0.25\widetilde{P}_c$. Thus every state has a Markov self-loop: $\widetilde{P}_c[i,i] = 0.55$ and $M_c[i,i] = 0.8875$. Together with strong connectivity, these positive diagonal entries guarantee aperiodicity and hence ergodicity of the finite routing chain. Every structural, stochasticity, support, stationary-vector, disagreement-contraction, provenance-hash, and replay gate must pass before an optimizer run is authorized.

The construction was selected because it combines the principal structural requirements for sparse optimization transport of gradient and update information. First, the authenticated LPS base is a constant-degree Ramanujan graph, so before augmentation its nontrivial adjacency spectrum has the extremal expander scale available to regular graphs. This gives rapid global information propagation per communication edge while retaining only $O(n)$ edges (Lubotzky et al., 1988; Chow et al., 2016). Second, large girth suppresses short-cycle duplication: early routing neighborhoods explore distinct vertices rather than repeatedly recycling the same local information. Together, rapid expansion and large girth allow gradient and update information at any vertex to influence every other vertex quickly, while delaying its return through short cycles; this limits premature information recirculation that can produce experimental instability. GRH-conditional ray-class Cayley expanders provide a less rigid arithmetic range of graph orders than the standard LPS parameterization; for example, the specialization on $(\mathbb{Z}/q\mathbb{Z})^\times$ has $\varphi(q)$ vertices for arbitrary moduli q (Jao et al., 2009). Their nontrivial eigenvalues obey a near-square-root bound, but their abelian Cayley structure forces short commutator cycles. For the $(\mathbb{Z}/q\mathbb{Z})^\times$ specialization, the construction provides logarithmic nonabelian and odd girth rather than the large ordinary-girth guarantee used here. This tradeoff is why the authenticated LPS base is selected when locally nonrepeating routing is required, despite its more restricted family of graph orders.

Third, the original LPS instance is bipartite. The $1{,}024 = \Theta(n)$ accepted additions preserve linear sparsity while the 839 same-partition edges make the graph non-bipartite; the exact girth remains seven. This removes the base graph's parity obstruction without replacing sparse transport by dense averaging. Fourth, the independently seeded green and red orientations retain the same complete support but provide two strongly connected, high-girth directional flows. Alternating them supplies temporal routing diversity at the smallest nontrivial number of operators, while using one fixed orientation would repeatedly impose the same directional bias.

Fifth, Sinkhorn scaling changes weights but not support, making each directed operator doubly stochastic. Hence constants are fixed and the uniform mean is preserved despite unequal directed in- and outdegrees. The positive identity component places a self-loop at every Markov state. Strong connectivity gives irreducibility, and the positive diagonal gives period one, so both lazy routing operators are primitive and ergodic. More precisely, every eigenvalue λ of W_c satisfies $|\lambda| \le 1$, while the corresponding eigenvalues of the lazy graph-package and optimizer operators are

$$0.55 + 0.45\lambda \quad \text{and} \quad 0.8875 + 0.1125\lambda, \tag{11}$$

whose real parts are at least 0.10 and 0.775, respectively. Thus the operators used for routing have no negative real eigenvalues, even though the directed W_c spectrum can be negative or complex. The sparse off-diagonal component controls communication cost.

C Directed Mixing: Finite Certificate and Conditional Family Theorem

For a reversible walk, an eigenvalue gap controls mixing directly. A directed operator need not be normal, however, so its nontrivial eigenvalue moduli alone do not suffice. We therefore state the guarantee in the induced 2-norm on the disagreement subspace. Let

$$J_n = \frac{1}{n}\mathbf{1}\mathbf{1}^\top.$$

Every doubly stochastic Q satisfies $QJ_n = J_nQ = J_n$, and $\sigma_\perp(Q) := \|Q - J_n\|_{2\to 2}$ is its worst-case one-step disagreement contraction.

Theorem C.1 (Directed disagreement contraction)**.** *Let $Q_1, \dots, Q_k$ be nonnegative doubly stochastic operators of the same dimension. Then*

$$\|Q_k \cdots Q_1 - J_n\|_{2\to 2} \le \prod_{t=1}^{k} \sigma_\perp(Q_t).$$

In particular, if $\sigma_\perp(Q_t) \le r < 1$ for every t, every deterministic or adaptively selected sequence obeys

$$\|Q_k \cdots Q_1 - J_n\|_{2\to 2} \le r^k.$$

No reversibility, normality, diagonalizability, or commutativity assumption is required.

Proof. Doubly stochasticity gives

$$AB - J_n = (A - J_n)(B - J_n)$$

for any two such operators A, B. Repeated application of this identity and submultiplicativity of the induced norm give the result. The same argument is pathwise when the selected color depends on the preceding iterates. □

Theorem C.2 (Family theorem under a uniform symmetrized gap)**.** *For each n and color c, let $P_{n,c}$ be nonnegative and doubly stochastic, and define*

$$S_{n,c} = \frac{P_{n,c} + P_{n,c}^\top}{2}, \qquad L_{n,c} = I - S_{n,c}.$$

Order the eigenvalues of $L_{n,c}$ increasingly. Suppose that, uniformly in n and c,

$$\lambda_2(L_{n,c}) \ge \gamma_* > 0.$$

For a fixed $0 < \alpha < 1$, set

$$Q_{n,c} = (1-\alpha)I + \alpha P_{n,c}, \qquad r_* = \sqrt{1 - 2\alpha(1-\alpha)\gamma_*} < 1.$$

Then every color sequence satisfies

$$\|Q_{n,c_k} \cdots Q_{n,c_1} - J_n\|_{2\to 2} \le r_*^k.$$

Moreover, under the receiver-by-sender Markov convention, $\|z\|_{\mathrm{TV}} := \frac{1}{2}\|z\|_1$, and

$$\max_i \left\| Q_{n,c_k} \cdots Q_{n,c_1} e_i - \frac{\mathbf{1}}{n} \right\|_{\mathrm{TV}} \le \frac{\sqrt{n-1}}{2}\, r_*^k.$$

Consequently a sufficient worst-case total-variation mixing time is

$$\tau_{\mathrm{mix}}(\varepsilon) \le \left\lceil \frac{\log\big(\sqrt{n-1}/(2\varepsilon)\big)}{-\log r_*} \right\rceil,$$

which is $O(\log(n/\varepsilon))$.

Proof. For $x \perp \mathbf{1}$, doubly stochasticity gives $\|P_{n,c}x\|_2 \le \|x\|_2$, while the gap assumption gives

$$x^\top P_{n,c} x = x^\top S_{n,c} x \le (1 - \gamma_*)\|x\|_2^2.$$

Expanding the squared norm yields

$$\|Q_{n,c}x\|_2^2 \le \big[1 - 2\alpha(1-\alpha)\gamma_*\big]\|x\|_2^2.$$

Theorem C.1 gives the switched product bound. Finally, the product fixes J_n, while $\|z\|_1 \le \sqrt{n}\|z\|_2$ and $\|e_i - \mathbf{1}/n\|_2 = \sqrt{1 - 1/n}$ give the displayed total-variation inequality. □

This mixing time concerns products of the homogeneous routing operators. In NRD those operators act on entering displacements inside a gradient-driven recursion, so it is not an optimization-convergence rate.

Unaugmented LPS specialization. The uniform-gap premise has a direct LPS specialization, but only before the augmentation and Sinkhorn weighting used by NRD. Let G be a connected d-regular Ramanujan graph with even d, adjacency matrix A, and a balanced Eulerian orientation. Under the receiver-by-sender convention, put $P_{vu} = 2/d$ when $u \to v$. Then P is doubly stochastic and

$$\frac{P + P^\top}{2} = \frac{A}{d}.$$

Writing $\lambda_2(A)$ for the second-largest *algebraic* adjacency eigenvalue,

$$\gamma_{\mathrm{add}} := 1 - \lambda_{\max}\left(\frac{P + P^\top}{2}\bigg|_{\mathbf{1}^\perp} \right) = 1 - \frac{\lambda_2(A)}{d} \geq 1 - \frac{2\sqrt{d-1}}{d}$$

by the Ramanujan bound (Lubotzky et al., 1988). For a bipartite base, the trivial eigenvalue $-d$ prevents a non-lazy absolute-gap mixing claim but does not change this algebraic upper-edge bound; the lazy theorem above removes the periodicity. In the authenticated LPS$(5, 13)$ base, $d = p+1 = 6$, so this lower bound is $1 - 2\sqrt{5}/6$.

C.1 Trajectory Evolution, Forced Dynamics, and History Windows

The conceptual role of routing is deeper than a one-step mixing interpretation. We study how repeated, ordered transitions filter the trajectory over a window: previously generated components can decay, current gradients inject new components, and their balance determines the singular geometry of the realized history. Thus the relevant progression is from transition operators to ordered products, from ordered products to finite-window trajectory geometry, and from that geometry to a target-relevant optimization test. Only the first two links are supplied by the contraction theory; whether the resulting subspace is more useful must be measured separately.

At complete green–red transaction boundaries, let $H_k = (x_{2k}, u_{2k}, \zeta_k)$ denote a state that includes the optimizer variables and the controller state ζ_k: color phase, authenticated rate and bank position, cycle counter, and restart/hold status. The objective, $f^\star$, L, authenticated rate pool, initialization gap, and graph artifacts are fixed instance data. Relative to these constants, H_k determines the next committed transaction; the shorter triple (x_t, g_t, u_t) does not. Thus the deterministic controller defines one piecewise state-complete map F, with $H_k = F^k(H_0)$. For the pathwise analysis below, let F_j denote the branch realized at transaction j and define

$$\mathcal{F}_{k:j} = F_{k-1} \circ \cdots \circ F_j, \qquad \mathcal{F}_{j:j} = \mathrm{Id}\,.$$

These maps satisfy the evolution relation $\mathcal{F}_{k:\ell} \circ \mathcal{F}_{\ell:j} = \mathcal{F}_{k:j}$ for $j \leq \ell \leq k$. This two-parameter notation records the realized branch composition; it does not assert exogenous nonautonomy. At the complete-state level, $\mathcal{F}_{k:j}$ is the restriction of F^{k-j} to the realized orbit. At the reduced displacement level, however, powers of one routing matrix do not represent the adaptive NRD controller.

For an observable ψ of the state, the pullback

$$\mathcal{K}^\tau \psi = \psi \circ F^\tau$$

is a Koopman-style evolution of that observable; equivalently, its restriction along a realized interval is $\mathcal{K}_{k:j}\psi = \psi \circ \mathcal{F}_{k:j}$. The composition-operator viewpoint originates with Koopman (1931); modern extensions and control-affected formulations are surveyed by Brunton et al. (2022). We use this only as an organizing viewpoint. We do not estimate a Koopman operator or claim that finite-window singular vectors are Koopman eigenfunctions, invariant modes, or asymptotic Lyapunov directions.

In the displacement coordinate space $\mathbb{R}^n$, which is the 2,184-dimensional parameter axis rather than the full state or observable space, the only common invariant splitting established here belongs to the routing component:

$$\mathbb{R}^n = \mathrm{span}\{\mathbf{1}\} \oplus \mathbf{1}^\perp.$$

Every exactly doubly stochastic M_c fixes the first subspace and leaves the second invariant. For ideal exactly balanced versions of the sealed operators, the audited one-step estimate $\sigma_\perp(M_c) \leq 0.969411$ makes the second subspace contractive, and theorem C.1 compounds the corresponding exact bound along any color schedule. The stored binary64 matrices retain the numerical qualification stated above. After multiplication by the momentum coefficient a_t, even the consensus

component of the homogeneous momentum channel is multiplied by a_t; holds and restarts can erase all earlier homogeneous memory. No other common persistent eigenspace follows from the audit because the directed green and red operators need not commute, be normal, or be simultaneously diagonalizable.

The preceding mixing result acts on the graph-coordinate axis, not directly on the temporal index of a stored history matrix. In the evaluated NRD instance, this graph-node axis is the 2,184-dimensional parameter axis. Its role in NRD is therefore better expressed through the forced update recurrence. Let $\Pi_\perp = I - J_n$, and let $a_t \in [0, \bar{\beta}]$, where $\bar{\beta} := \sup_t a_t \leq 0.9 < 1$, denote the momentum multiplier at step t. An ordinary accepted NRD step has $a_t = \beta$ and $e_t = -\eta_t g_t$, where $\eta_t = \rho_t / L$. A zero-momentum restart has $a_t = 0$ on its entering green step and $e_t = -\eta_t g_t$, after which the ordinary multiplier resumes. Here t indexes the two reported colored steps in every controller cycle. A hold writes two zero realized-update columns, represented by $a_t = 0, e_t = 0$, and marks the corresponding raw-gradient columns unavailable; gradients from rejected trials are charged but never substituted into the reported history. Thus

$$u_{t+1} = a_t M_{c_t} u_t + e_t. \tag{12}$$

The forcing may depend on the entire preceding trajectory, as it does when $g_t = \nabla f(q_t)$. In exact arithmetic the proposed and realized displacements coincide; the measured U_T differs only through the finite-precision rounding convention described below.

For $t > s$, define

$$\Phi(t, s) = M_{c_{t-1}} \cdots M_{c_s}, \qquad a_{t:s} = \prod_{\ell=s}^{t-1} a_\ell,$$

and set $\Phi(t, t) = I$ and $a_{t:t} = 1$.

Theorem C.3 (Pathwise forced-history decomposition)**.** *For every realized sequence of operators and forcings,*

$$u_t = a_{t:0} \Phi(t, 0) u_0 + \sum_{s=0}^{t-1} a_{t:s+1} \Phi(t, s+1) e_s. \tag{13}$$

If every selected M_{c_t} is exactly doubly stochastic, the consensus component satisfies

$$J_n u_{t+1} = a_t J_n u_t + J_n e_t.$$

If, in addition, every selected operator obeys $\|M_{c_t} - J_n\|_{2\to 2} \leq r < 1$, then

$$\|\Pi_\perp u_t\|_2 \leq (\bar{\beta} r)^t \|\Pi_\perp u_0\|_2 + \sum_{s=0}^{t-1} (\bar{\beta} r)^{t-1-s} \|\Pi_\perp e_s\|_2. \tag{14}$$

For ordinary accepted NRD steps, $\Pi_\perp e_s = -\eta_s \Pi_\perp g_s$. Hence old disagreement is forgotten geometrically, while each newly generated gradient injects fresh disagreement that is subsequently filtered by the ordered green–red products.

Proof. Equation equation 13 follows by induction. Doubly stochasticity gives

$$\Pi_\perp M_c = M_c - J_n = (M_c - J_n) \Pi_\perp.$$

Applying $\Pi_\perp$ to the unrolled recurrence, using, for $t > s$,

$$\Phi(t, s) - J_n = (M_{c_{t-1}} - J_n) \cdots (M_{c_s} - J_n),$$

and then applying Theorem C.1 yields equation (14). The derivation is pathwise and does not assume that the gradients are exogenous or independent. □

Returning to equation (14), if $\|\Pi_\perp e_t\|_2 \leq \epsilon$ uniformly, then

$$\|\Pi_\perp u_t\|_2 \leq (\bar{\beta} r)^t \|\Pi_\perp u_0\|_2 + \epsilon \frac{1 - (\bar{\beta} r)^t}{1 - \bar{\beta} r}.$$

Thus persistent heterogeneous forcing generally leaves a nonzero disagreement floor; contraction and fresh excitation are two terms in the same bound rather than independent benefits.

Proposition C.4 (Shared stationary point and conditional limit)**.** *Let $f \in C^1$ and let the accepted step size be positive. For either NAG ($M_c = I$) or NRD, an accepted colored-step map has a fixed point $(\bar{x}, \bar{u})$ if and only if*

$$\bar{u} = 0, \qquad \nabla f(\bar{x}) = 0.$$

Thus routing does not change the stationary fixed-point set. If f is μ-strongly convex, this set is the single point $(x^\star, 0)$. Moreover, if a NAG or NRD trajectory converges, has infinitely many accepted or restart steps, and their step sizes are bounded below by $\eta_{\min} > 0$, then its limit is $x^\star$. The evaluated finite rate banks verify this lower bound, with $\eta_{\min} = 0.125/L$ for NAG and $0.5/L$ for HB. Hence NAG and NRD cannot converge to different points on the same strongly convex objective under these conditions. The hold map is excluded: it fixes every state $(x, 0)$, including nonstationary ones, which is why infinitely many non-hold steps are required.

Proof. At a fixed point, $x_{t+1} = x_t$ and $u_{t+1} = u_t$ imply $\bar{u} = 0$. The routed momentum then vanishes, the NAG query point is $\bar{x}$, and the positive-step update gives $\nabla f(\bar{x}) = 0$; the converse is immediate. For the limit statement, convergence of x_t implies $u_t = x_t - x_{t-1} \to 0$. Along the infinite accepted/restart subsequence, $\|M_c\|_{2\to 2} = 1$ makes routed momentum tend to zero, so $\eta_t g_t \to 0$. The positive lower bound on η_t and continuity of ∇f give $\nabla f(\bar{x}) = 0$, and strong convexity gives $\bar{x} = x^\star$. □

Towards unconditional convergence. For a quadratic

$$f(x) = \frac{1}{2}(x - x^\star)^\top H (x - x^\star)$$

and a fixed colored step with $B = aM_c$, the state $z_t = (x_t - x^\star, u_t)$ evolves under NAG or HB by

$$z_{t+1} = T^{\mathrm{NAG}}_{B,\eta} z_t, \qquad T^{\mathrm{NAG}}_{B,\eta} = \begin{pmatrix} I - \eta H & (I - \eta H)B \\ -\eta H & (I - \eta H)B \end{pmatrix}, \qquad T^{\mathrm{HB}}_{B,\eta} = \begin{pmatrix} I - \eta H & B \\ -\eta H & B \end{pmatrix}.$$

A sufficient arbitrary-switching certificate is a common $P \succ 0$ and $\gamma < 1$ satisfying, for $\mathrm{opt} \in \{\mathrm{NAG}, \mathrm{HB}\}$,

$$(T^{\mathrm{opt}}_{B,\eta})^\top P T^{\mathrm{opt}}_{B,\eta} \preceq \gamma^2 P$$

for each applicable optimizer and every

$$(B, \eta) \in \{\beta I, \beta M_{\mathrm{g}}, \beta M_{\mathrm{r}}, 0\} \times \{\rho/L : \rho \in \mathcal{B}_{\mathrm{opt}}\}.$$

Together with a nonexpansive certificate for the hold map and infinitely many non-hold transitions, this would imply convergence of NAG and NRD to the shared $x^\star$, linearly in the number of non-hold transitions and linearly in elapsed steps when their asymptotic fraction is bounded below. For $B = \beta I$, the NAG matrix reduces modewise to the standard untreated NAG companion matrix. The common certificate has not been established for the current controller. The present endpoint gates guarantee accepted endpoint nonincrease, not sufficient decrease, and the finite rate bank may hold at a nonstationary point; they therefore do not by themselves prove convergence.

D Trajectory Dynamics

D.1 Persistent and Decaying Modes

For an associated linear transport A, suppose $\mathbb{R}^d = V_{\mathrm{pers}} \oplus V_{\mathrm{dec}}$ and $\|A^s y\| \le \varepsilon_s \|y\|$ for $y \in V_{\mathrm{dec}}$, with $\varepsilon_s \to 0$. Directions in V_{dec} are transient, whereas directions in V_{pers} remain visible over the horizon; this is a dynamical hypothesis, not an automatic property of a finite-window singular subspace.

D.2 Spectral Collapse as a Dynamical Phenomenon

If $U_T = \sum_{i=1}^{r} \sigma_i a_i b_i^\top$ and $\sigma_1, \ldots, \sigma_k$ dominate the tail, then $V_T = \mathrm{span}\{a_1, \ldots, a_k\}$ summarizes the directions carrying most realized update energy. We treat V_T as a finite-window observable rather than necessarily an invariant or asymptotic mode.

D.3 Trajectory Subspaces and Curvature

For the orthogonal projector P_T onto V_T, the compressed operator $P_T H(x) P_T$ records the curvature exposed within that subspace. Equal-dimensional trajectories may therefore encounter different curvature when their subspaces have different orientations, although curvature alone does not determine optimization quality.

D.4 RDLPS as Trajectory-Subspace Transport

The principal effect is not necessarily a change in ambient dimension. Rather, graph transport changes which directions become persistent and which directions decay, leading to a different low-dimensional trajectory subspace.

D.5 Connection with the Subspace-Rescue Experiments

Let V_I and V_R be equal-dimensional subspaces extracted from untreated and routed trajectories. A matched rescue test asks whether

$$\inf_{x \in x_0 + V_R} f(x) < \inf_{x \in x_0 + V_I} f(x),$$

together with projected optimizer reruns in the two affine spaces. Such an outcome would indicate that routing changes not merely spectral breadth but the target-relevant directions represented by the trajectory. The held-out experiment reported below tests exactly this criterion. The theory motivates the test but does not determine its outcome; spectral broadening alone is insufficient.

There is recent precedent for low-dimensional gradient geometry: gradients concentrate in a small, persistent top-Hessian subspace, and gradient/update subspaces across training can be strongly compressible (Gur-Ari et al., 2018; Azam et al., 2022; Jaiswal et al., 2025). A recent preprint also studies predictable temporal gradient manifolds (Calvo, 2026). These observations establish that low-dimensional histories can occur, but they do not determine whether the concentration is benign or obstructive.

E Conditional Convex Acceleration from RDLPS Trajectory Geometry

The following terms are used throughout the note.

Identity baseline. This is the original optimizer without RDLPS routing or mixing; its transport operator is $M = I$. Its trajectory subspace is denoted by S_I.

RDLPS trajectory. This is the optimizer trajectory after RDLPS routing. Its trajectory subspace is denoted by S_R.

Trajectory subspace. A trajectory subspace S is the span of selected dominant directions in an optimizer's gradient or update history. Starting from x_0, optimization restricted to these directions can visit only the affine space

$$x_0 + S = \{x_0 + s : s \in S\}.$$

Representation floor. The representation floor is the smallest objective gap attainable inside the discovered affine space:

$$\Delta_S = \min_{x \in x_0 + S} f(x) - f(x^\star).$$

Thus $\Delta_{S_R} < \Delta_{S_I}$ means that the RDLPS subspace can represent a point closer in objective value to the true optimum. A lower floor means a better attainable final value, even if optimization inside both spaces were perfect.

Restricted curvature and condition number. Restricted curvature measures the smallest and largest Hessian curvature only along directions $v \in S$:

$$\mu_{S,\mathcal{N}} \le v^\top \nabla^2 f(x) v \le L_{S,\mathcal{N}}, \qquad \|v\| = 1,$$

throughout a declared neighborhood $\mathcal{N}$. The restricted condition number is

$$\kappa_{S,\mathcal{N}} = \frac{L_{S,\mathcal{N}}}{\mu_{S,\mathcal{N}}}.$$

Thus $\kappa_{S_R,\mathcal{N}_R} < \kappa_{S_I,\mathcal{N}_I}$ means that the objective is less ill-conditioned along the RDLPS directions. This gives a better certified contraction factor for projected gradient descent: optimization within the RDLPS subspace is easier and potentially faster.

The distinction. A lower representation floor concerns *where the optimizer can ultimately reach*. Better restricted conditioning concerns *how rapidly it can optimize within the represented space*. Neither property follows from spectral rank or spectral broadening alone.

E.1 Setup and the C1 hypothesis

Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable on a convex neighborhood $\mathcal{N}$, and suppose f has a global minimizer $x^\star$. Given a history matrix, for example

$$U_T = [u_0, \dots, u_{T-1}],$$

let S be a declared k-dimensional trajectory subspace, such as the span of the first k left singular vectors of U_T. Its affine trajectory space is

$$\mathcal{A}_S = x_0 + S,$$

and P_S denotes orthogonal projection onto S. The same extraction rule, history source, horizon, and dimension must be used in every matched identity versus routed comparison.

Let S_I and S_R denote spaces extracted from matched identity and RDLPS trajectories. The experimentally testable condition remembered as "C1" is the following geometry condition.

Assumption E.1 (C1: matched trajectory-geometry improvement)**.** *The fixed, equal-dimensional spaces S_I and S_R obey all admissibility and restricted-curvature hypotheses below. There exist $\delta_\Delta, \delta_\kappa \ge 0$, not both zero, such that*

$$\Delta_{S_R} \le \Delta_{S_I} - \delta_\Delta, \qquad \kappa_{S_R,\mathcal{N}_R} \le \kappa_{S_I,\mathcal{N}_I} - \delta_\kappa. \tag{15}$$

Here Δ_S measures what the affine space can represent, while $\kappa_{S,\mathcal{N}}$ measures how difficult optimization is within that space. They are defined next. Spectral broadening or an increase in effective rank does not imply C1.

E.2 Restricted geometry and representation floor

For a subspace S and convex neighborhood $\mathcal{N}$, define the segment

$$\mathcal{T}_{S,\mathcal{N}} = \{(x,h) : x \in \mathcal{N},\ h \in S,\ \{x + \tau h : 0 \le \tau \le 1\} \subseteq \mathcal{N}\}. \tag{16}$$

Definition E.2 (Segmentwise restricted curvature)**.** *Define*

$$L_{S,\mathcal{N}} = \sup_{\substack{z \in \mathcal{N},\ v \in S \\ \|v\|=1}} v^\top \nabla^2 f(z) v, \tag{17}$$

$$\mu_{S,\mathcal{N}} = \inf_{\substack{z \in \mathcal{N},\ v \in S \\ \|v\|=1}} v^\top \nabla^2 f(z) v. \tag{18}$$

When $\mu_{S,\mathcal{N}} > 0$, set

$$\kappa_{S,\mathcal{N}} = \frac{L_{S,\mathcal{N}}}{\mu_{S,\mathcal{N}}}.$$

For every $(x, h) \in \mathcal{T}_{S,\mathcal{N}}$, Taylor's theorem with integral remainder gives

$$f(x+h) = f(x) + \nabla f(x)^\top h + \int_0^1 (1-\tau) h^\top \nabla^2 f(x+\tau h) h \, d\tau.$$

Consequently,

$$\frac{\mu_{S,\mathcal{N}}}{2}\|h\|^2 \le f(x+h) - f(x) - \nabla f(x)^\top h \le \frac{L_{S,\mathcal{N}}}{2}\|h\|^2. \tag{19}$$

Assume the restricted minimum is attained at

$$x_S^\star \in \arg\min_{x \in x_0 + S} f(x).$$

Definition E.3 (Representation floor)**.** *The representation floor of $x_0 + S$ is*

$$\Delta_S = f(x_S^\star) - f(x^\star) \ge 0. \tag{20}$$

It is the objective error remaining after exact optimization in the affine space discovered by the trajectory.

A small $\kappa_{S,\mathcal{N}}$ does not imply a small Δ_S, and $\Delta_S = 0$ does not imply favorable restricted conditioning.

E.3 Projected trajectory-subspace convergence

Theorem E.4 (Trajectory-subspace principle)**.** *Suppose $x_S^\star \in (x_0 + S) \cap \mathcal{N}$ and*

$$0 < \mu_{S,\mathcal{N}} \le L_{S,\mathcal{N}} < \infty.$$

Run projected gradient descent in $x_0 + S$:

$$x_{t+1} = x_t - \eta P_S \nabla f(x_t), \qquad \eta = \frac{2}{L_{S,\mathcal{N}} + \mu_{S,\mathcal{N}}}. \tag{21}$$

Assume $x_t \in \mathcal{N}$ and $(x_t, x_{t+1} - x_t) \in \mathcal{T}_{S,\mathcal{N}}$ for every iterate considered. With

$$q_S = \frac{\kappa_{S,\mathcal{N}} - 1}{\kappa_{S,\mathcal{N}} + 1},$$

one has

$$f(x_t) - f(x^\star) \le \Delta_S + \frac{L_{S,\mathcal{N}}}{2} q_S^{2t} \|x_0 - x_S^\star\|^2. \tag{22}$$

Equivalently,

$$f(x_t) - f(x^\star) \le \Delta_S + \kappa_{S,\mathcal{N}} q_S^{2t} \big(f(x_0) - f(x_S^\star)\big). \tag{23}$$

Proof. The first-order condition for the affine restricted minimizer is $P_S \nabla f(x_S^\star) = 0$. Define

$$\overline{H}_t = \int_0^1 \nabla^2 f\big(x_S^\star + s(x_t - x_S^\star)\big) \, ds.$$

Convexity of $\mathcal{N}$ places the segment in $\mathcal{N}$. Therefore $(P_S \overline{H}_t P_S)|_S$ has spectrum in $[\mu_{S,\mathcal{N}}, L_{S,\mathcal{N}}]$. The mean-value identity gives

$$x_{t+1} - x_S^\star = \big(I_S - \eta (P_S \overline{H}_t P_S)|_S\big)(x_t - x_S^\star).$$

For the step size in equation 21, the operator norm is at most q_S. Thus

$$\|x_t - x_S^\star\| \le q_S^t \|x_0 - x_S^\star\|.$$

Restricted smoothness yields

$$f(x_t) - f(x_S^\star) \le \frac{L_{S,\mathcal{N}}}{2} q_S^{2t} \|x_0 - x_S^\star\|^2.$$

Adding $f(x_S^\star) - f(x^\star) = \Delta_S$ proves equation 22. Restricted strong convexity gives

$$\|x_0 - x_S^\star\|^2 \le \frac{2}{\mu_{S,\mathcal{N}}} \big(f(x_0) - f(x_S^\star)\big),$$

which proves equation 23. □

E.4 Consequences of C1

For $J \in \{I, R\}$, write

$$C_J = \frac{L_{S_J,\mathcal{N}_J}}{2}\|x_0 - x^\star_{S_J}\|^2, \qquad q_J = \frac{\kappa_{S_J,\mathcal{N}_J} - 1}{\kappa_{S_J,\mathcal{N}_J} + 1}.$$

Theorem E.5 (Conditional matched-subspace advantage)**.** *Under Assumption E.1, run the projected method equation 21 separately in $x_0 + S_I$ and $x_0 + S_R$. The routed affine space has an attainable floor at least δ_Δ lower whenever $\delta_\Delta > 0$. The routed projected method has a strictly smaller certified contraction factor whenever $\delta_\kappa > 0$. Moreover, if $\delta_\Delta > 0$, then for every*

$$t \geq T^\star := \min\{s \in \mathbb{N}_0 : C_R q_R^{2s} < \delta_\Delta\}, \tag{24}$$

the routed projected iterate obeys

$$f(x_{R,t}) < f(x_{I,t}). \tag{25}$$

Proof. The floor comparison is the first inequality in equation 15. Since $\kappa \mapsto (\kappa - 1)/(\kappa + 1)$ is strictly increasing on $[1, \infty)$, the second inequality gives $q_R < q_I$ whenever $\delta_\kappa > 0$. For $t \geq T^\star$, Theorem E.4 gives

$$f(x_{R,t}) - f(x^\star) \leq \Delta_{S_R} + C_R q_R^{2t} < \Delta_{S_I} \leq f(x_{I,t}) - f(x^\star),$$

which proves equation 25. □

E.5 Conditional transfer to actual RDLPS iterates

Let x_t^I, x_t^R denote matched actual identity-baseline and RDLPS iterates, respectively, and let y_t^I, y_t^R denote their projected-gradient reference trajectories in $x_0 + S_I$ and $x_0 + S_R$.

Assumption E.6 (C2: actual-to-projected acceleration margin)**.** *Suppose $\delta_\Delta := \Delta_{S_I} - \Delta_{S_R} > 0$, both projected references satisfy Theorem E.4, and nonnegative sequences $\varepsilon_I(t), \varepsilon_R(t)$ satisfy*

$$\left|f(x_t^J) - f(y_t^J)\right| \leq \varepsilon_J(t), \qquad J \in \{I, R\}. \tag{26}$$

Assume there is a finite $T^\star_{\text{actual}}$ such that, for every $t \geq T^\star_{\text{actual}}$,

$$\Delta_{S_R} + C_R q_R^{2t} + \varepsilon_R(t) < \Delta_{S_I} - \varepsilon_I(t). \tag{27}$$

Theorem E.7 (Conditional RDLPS acceleration)**.** *Under Assumption E.6, for every $t \geq T^\star_{\text{actual}}$,*

$$f(x_t^R) - f(x^\star) < f(x_t^I) - f(x^\star). \tag{28}$$

Moreover, let τ satisfy $\Delta_{S_R} < \tau < \Delta_{S_I}$. If at some $t_R \geq T^\star_{\text{actual}}$,

$$\Delta_{S_R} + C_R q_R^{2t_R} + \varepsilon_R(t_R) \leq \tau, \tag{29}$$

then the actual RDLPS trajectory reaches objective gap τ by t_R, whereas the identity-baseline trajectory cannot reach τ at any time t for which $\varepsilon_I(t) < \Delta_{S_I} - \tau$.

Proof. The projected theorem and equation 26 imply

$$f(x_t^R) - f(x^\star) \leq \Delta_{S_R} + C_R q_R^{2t} + \varepsilon_R(t).$$

Since $y_t^I \in x_0 + S_I$, restricted optimality gives $f(y_t^I) - f(x^\star) \geq \Delta_{S_I}$. Hence

$$f(x_t^I) - f(x^\star) \geq \Delta_{S_I} - \varepsilon_I(t).$$

The strict margin equation 27 proves equation 28. Equation equation 29 proves the RDLPS hitting statement, while the identity-baseline lower bound is strictly above τ whenever $\varepsilon_I(t) < \Delta_{S_I} - \tau$. □