PAPF: PHASE-AWARE POLYNOMIAL FILTERING FOR FIXED-POINT ACCELERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose phase-aware polynomial filtering (PAPF), a simple method for accelerating fixed-point iterations by exploiting the phases of local Jacobian eigenvalues. PAPF applies the degree-two filter $S_\alpha = \alpha I + (1-\alpha)T^2$, with $\alpha \in (0,1)$, using two evaluations of the original operator per iteration. For firmly nonexpansive operators with a nonempty fixed-point set, the filter preserves averagedness and fixed points, guaranteeing global convergence for every fixed mixing parameter. Beyond this setting, we characterize precisely when PAPF improves the local convergence factor over the equal-cost baseline T^2. For a differentiable map with Jacobian $G = J_T(z^\star)$ and spectral radius $\rho \in (0,1)$, such an improvement is possible if and only if every dominant eigenvalue $\rho e^{i\theta_j}$ satisfies $\cos(2\theta_j) < \rho^2$. We derive closed-form optimal parameters for individual spectral modes and formulate parameter selection for the full spectrum as a one-dimensional convex problem. When the spectrum is unavailable, we propose an empirical parameter-selection procedure based on short residual probes, requiring no explicit Jacobian information. We apply the framework to primal–dual hybrid gradient iterations for quadratically regularized optimal transport and to the power method for nonsymmetric eigenvalue problems. Experiments on these applications and synthetic operators demonstrate reductions in operator evaluations in favorable spectral regimes and support the predicted dependence of acceleration on eigenvalue phase.

1 INTRODUCTION

Fixed-point iterations of firmly nonexpansive (FNE) operators are a foundational building block of modern optimization and machine learning: proximal point (Rockafellar, 1976), Douglas–Rachford (DR) splitting (Lions & Mercier, 1979), ADMM (Boyd et al., 2011), and the primal–dual hybrid gradient (PDHG) method (Chambolle & Pock, 2011) power optimal transport (Peyré & Cuturi, 2019) and image recovery (Combettes & Pesquet, 2011), and they converge under remarkably mild assumptions (Bauschke & Combettes, 2011; Ryu & Boyd, 2016). Their local behavior is another story: residual trajectories often decay slowly and non-monotonically.

This behavior has a structural origin: when T is FNE and differentiable at a fixed point $z^\star$, the Jacobian $G = J_T(z^\star)$ has all eigenvalues in the *FNE disk* $\{\lambda : |\lambda - \frac{1}{2}| \le \frac{1}{2}\}$ (Bauschke & Combettes, 2011), which contains complex conjugate pairs; when the dominant eigenvalue is such a pair, the local iteration rotates while it contracts, and measured residuals can oscillate when the geometry is nonnormal, the norm is not adapted to the rotation, or several modes superpose. The same disk limits how far the nonnegative degree-2 filters studied here can help (Proposition 2).

Existing accelerations leave open *when* a static filter strictly accelerates beyond the equal-cost baseline T^2 (related work below). This paper answers the question for the minimal-degree candidate $\mathsf{S}_\alpha = \alpha I + (1-\alpha)T^2$: for any differentiable map with $\rho(J_T(z^\star)) < 1$, S_α beats T^2 for some $\alpha \in (0,1)$ iff every dominant Jacobian eigenvalue $\rho\, e^{i\theta_j}$ satisfies $\cos(2\theta_j) < \rho^2$ (Theorem 2). When T is FNE the filter also preserves averagedness and global convergence, but the FNE disk caps its asymptotic T-call speedup at $B(\rho) < 3/2$ ($\to 9/8$ as $\rho \to 1$; Proposition 2), so the large accelerations arise for convergent maps whose dominant modes lie outside the disk. On a non-symmetric power-method instance ($\theta = \pi/3$, $\rho = 0.99$) adaptive PAPF gives a nearly $10\times$ speedup over T^2 and the best fixed α $30\times$; no-extrapolation PDHG instances for optimal transport yield 1.2–$1.9\times$, while on the FNE Chambolle–Pock map the best filter degenerates to T^2, as Theorem 2 and the cap predict.

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.

Related work (extended discussion in App. A). Two-stage KM constructions, from the Ishikawa scheme (Ishikawa, 1974) to two-evaluation KM with adaptive momentum (He et al., 2025) and inertial or perturbed KM (Cortild & Peypouquet, 2025), converge under method-specific conditions but do not characterize when a static blend of T and T^2 beats T^2. Nesterov/FISTA (Nesterov, 1983; Beck & Teboulle, 2009), inertial methods for monotone inclusions (Alvarez & Attouch, 2001), Halpern acceleration (Park & Ryu, 2022), and Anderson acceleration (Anderson, 1965; Walker & Ni, 2011; Toth & Kelley, 2015) target objective gaps, global residual complexity, or local R-linear rates under their own assumptions. Classical Chebyshev constructions (Golub & Varga, 1961; Varga, 1962; d'Aspremont et al., 2021) exploit real spectral intervals, whereas polynomial and Krylov methods also treat complex spectra and nonsymmetric systems (Saad, 2011; Saad & Schultz, 1986), Chebyshev over-relaxation factors have been applied to nonlinear fixed-point maps with real Jacobian spectra (Wadayama & Takabe, 2021), and regularized nonlinear acceleration to nonsymmetric primal–dual iterations (Bollapragada et al., 2019); none of these asks when a stationary blend of T and T^2 beats T^2 (App. A). Linear convergence of splitting methods under error bounds or partial smoothness (Liang et al., 2017; Davis & Yin, 2017a; Boley, 2013) does not by itself give the differentiability we assume; adaptive stepsize rules (Malitsky & Pock, 2018; Fercoq, 2024) change the operator, whereas PAPF tunes a filter coefficient that preserves fixed points and averagedness.

1.1 Contributions

The filter is a KM relaxation of T^2; our contribution is not a new acceleration principle but an exact account of what this minimal nonnegative family can and cannot do, and where it stands against relaxed KM.

1. **Sharp local criterion, and what FNE allows.** S_α improves the linearized factor over the equal-cost baseline T^2 iff every dominant Jacobian eigenvalue $\rho\, e^{i\theta_j}$ satisfies $\cos(2\theta_j) < \rho^2$ (Theorem 2), with a closed-form per-mode optimum; for FNE T it is averaged for every $\alpha \in (0,1)$ (Theorem 1) but capped at the per-step factor $2\rho/\sqrt{1+3\rho^2}$, i.e., at $B(\rho) < 3/2$ in asymptotic T-calls (Proposition 2).
2. **The equal-cost degree-2 family.** Compared mode by mode with convex relaxed KM ($\beta \in [0,1)$), the degree-1 member of the convex family $\{aI + bT + cT^2 : a,b,c \geq 0,\ a+b+c=1\}$, PAPF is never worse for a dominant pair in the right half-plane and wins up to a crossover angle tending to $\arccos(-1/3) \approx 109.5°$, beyond which relaxed KM wins (Proposition 3; the full-spectrum winner can differ, Remark 12).
3. **Residual-only adaptive selection.** A golden-section search over residual probes, guarded by an $\alpha = 0$ acceptance test, selects $\hat{\alpha}$ from data alone (Algorithm 1). We prove consistency of the probe rate for locally affine maps via the Riesz projection (Proposition 4) and a finite-window bound on the selected factor (Proposition 8), exhibit a smooth nonlinear map that misleads it, and document both failure modes of the search: missing a good parameter and accepting a worse one (Remark 14).
4. **Applications and experiments.** We instantiate the filter for PDHG on quadratically regularized optimal transport with explicit stepsizes (Theorem 3) and for the power method on nonsymmetric matrices (Theorem 4). On nine instances PAPF is compared per T-call with T^2, tuned relaxed KM, and (on subsets) TKMA, Anderson, Halpern, inertial, and GMRES, with matched-accuracy, equal-budget, and repeated-initialization controls; a boundary sweep confirms the criterion's sharpness in both directions.

2 Preliminaries and the PAPF Iteration

An operator $T : \mathbb{R}^d \to \mathbb{R}^d$ is *nonexpansive* (NE) if $\|Tz - Tw\| \leq \|z - w\|$ for all z, w; *firmly nonexpansive* (FNE) if $\|Tz - Tw\|^2 \leq \langle z - w,\, Tz - Tw\rangle$ for all z, w; and *κ-averaged* ($\kappa \in (0,1)$) if $T = (1-\kappa)I + \kappa R$ with an NE R. FNE is equivalent to $\frac{1}{2}$-averagedness, which in turn is equivalent to the reflector $2T - I$ being NE (Bauschke & Combettes, 2011). We write $\operatorname{Fix} T = \{z : Tz = z\}$, $\rho(\cdot)$ for spectral radius, and $J_T(z)$ for the Jacobian of T at z. All of these notions may be taken with respect to any inner product on $\mathbb{R}^d$. For primal–dual splitting maps such as PDHG, FNE holds—when it holds at all—in a preconditioned inner product: the extrapolated Chambolle–Pock map is the resolvent of a maximally monotone operator in such a metric (Chambolle & Pock, 2011; He & Yuan,

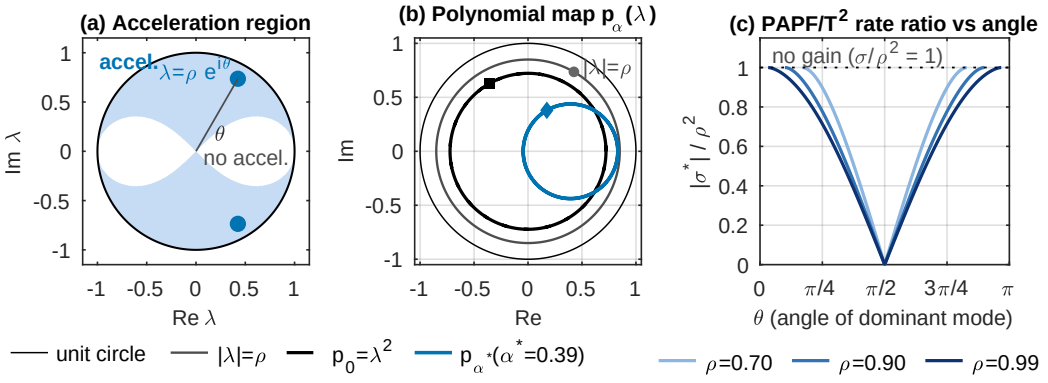

Figure 1: **Theory of phase-aware polynomial filtering.** (a) The PAPF acceleration region $\cos(2\theta) < \rho^2$ (shaded) for a dominant eigenvalue $\lambda = \rho\, e^{i\theta}$, together with the FNE disk: FNE maps are confined to the disk, where the region is thin (Proposition 2); the large gains lie outside it. (b) p_α maps the circle $|\lambda| = \rho$ to the circle of radius $(1-\alpha)\rho^2$ centered at α (solid blue, $\alpha = \alpha^\star$; dashed black: the T^2 image, radius ρ^2): squaring doubles the phase and the convex combination with 1 shifts the image toward the real axis, so it lies outside the dashed circle near the positive real axis and inside it near $\pm\pi/2$, where the phase-doubled mode is nearly cancelled. (c) Per-step factor ratio $|\sigma^\star|/\rho^2$ (lower is better) vs. angle θ for $\rho \in \{0.7, 0.9, 0.99\}$: the ratio decreases toward $\theta = \pi/2$ and tends to $|\cos\theta|$ as $\rho \to 1$ (Corollary 3); the corresponding asymptotic T-call speedup is $\log|\sigma^\star|/\log\rho^2$.

2012) (Section 5.1). The eigenvalues of $J_T(z^\star)$ do not depend on the inner product, so the FNE-disk constraint below is a metric-independent *necessary* condition; the no-extrapolation PDHG map used in Section 5 violates it and is therefore convergent but not FNE in any metric.

Calculus of averaged operators. If T_1 is κ_1-averaged and T_2 is κ_2-averaged, then $T_1 \circ T_2$ is κ-averaged with $\kappa = (\kappa_1+\kappa_2-2\kappa_1\kappa_2)/(1-\kappa_1\kappa_2)$ (Combettes & Yamada, 2015), and convex combinations $\sum_i \omega_i T_i$ are $\sum_i \omega_i\kappa_i$-averaged. Hence T^2 is $\frac{2}{3}$-averaged for FNE T and, treating I as 0-averaged, $\mathsf{S}_\alpha = \alpha I + (1-\alpha)T^2$ is $\frac{2(1-\alpha)}{3}$-averaged for $\alpha \in (0,1)$.

Global convergence (KM theory). If T is averaged with $\operatorname{Fix} T \neq \emptyset$, the Picard iteration is Fejér-monotone w.r.t. $\operatorname{Fix} T$ and asymptotically regular, and Opial's lemma yields convergence to a point of $\operatorname{Fix} T$ (Opial, 1967; Bauschke & Combettes, 2011).

Assumption 1 (Local regularity and linearization)**.** There exists $z^\star \in \operatorname{Fix} T$ and a neighborhood U such that T is Fréchet differentiable on U with Jacobian $G = J_T(z^\star)$ satisfying $\rho(G) < 1$.

Under Assumption 1, the Picard iteration is locally R-linearly convergent at rate $\rho(G) + \varepsilon$ for any $\varepsilon > 0$ (Kelley, 1995; Ortega & Rheinboldt, 1970).

FNE disk. If T is FNE with respect to some inner product $\langle\cdot,\cdot\rangle_M$ and differentiable at $z^\star$, every eigenvalue of $G = J_T(z^\star)$ lies in the closed disk $\{\lambda \in \mathbb{C} : |\lambda - \frac{1}{2}| \le \frac{1}{2}\}$ (write $T = \frac{1}{2}(I+R)$ with R nonexpansive in $\|\cdot\|_M$; then $G = \frac{1}{2}(I + J_R(z^\star))$ with $\|J_R(z^\star)\|_M \le 1$). In polar form, $\lambda = re^{i\theta}$ lies in the disk iff $r \le \cos\theta$; in particular $\operatorname{Re}(\lambda) \in [0,1]$, and (since G is real) non-real eigenvalues come in conjugate pairs: complex modes are compatible with firm nonexpansiveness, and Proposition 2 shows why the nonnegative PAPF family cannot remove them cheaply.

The PAPF iteration. Given a fixed-point map T, choose a mixing parameter $\alpha \in (0,1)$ and iterate

$$z^{k+1} \;=\; \mathsf{S}_\alpha(z^k) \;=\; \alpha\, z^k + (1-\alpha)\, T\big(T(z^k)\big), \qquad \text{stopping when } \|T(z^k) - z^k\| \le \varepsilon. \quad (1)$$

Each step costs two T-evaluations, and $T(z^k)$ yields the residual for free. The optimal α follows from the Jacobian spectrum (Section 3.2) or, when the spectrum is unknown, from residuals (Section 4).

Proposition 1 (Averagedness and fixed points)**.** *Let T be FNE. Then T^2 is averaged, and for any $\alpha \in (0,1)$ the map $\mathsf{S}_\alpha = \alpha I + (1-\alpha)T^2$ is averaged with $\operatorname{Fix}\mathsf{S}_\alpha = \operatorname{Fix} T$.*

Theorem 1 (Global convergence)**.** *Let T be FNE with $\operatorname{Fix} T \neq \emptyset$, and let $\alpha \in (0,1)$. Then the sequence generated by* (1) *converges to some $z^\star \in \operatorname{Fix} T$.*

By Proposition 1 this is a KM iteration of an averaged map with $\operatorname{Fix} \mathsf{S}_\alpha = \operatorname{Fix} T$ (Bauschke & Combettes, 2011; Combettes & Pesquet, 2021). Its limit need not be the $z^\star$ of Assumption 1; the local results apply once iterates enter U. We restrict the analysis to $\alpha \in [0,1)$ (Remark 10).

3 Local Rate Analysis

Let $z^\star \in \operatorname{Fix} T$ and suppose Assumption 1 holds with Jacobian $G = J_T(z^\star)$. By the chain rule, $J_{T^2}(z^\star) = G^2$ and

$$J_{\mathsf{S}_\alpha}(z^\star) \;=\; \alpha I + (1-\alpha)G^2 \;=\; p_\alpha(G), \qquad p_\alpha(\lambda) = \alpha + (1-\alpha)\lambda^2. \tag{2}$$

We call $q(\alpha) := \rho(p_\alpha(G)) = \rho(J_{\mathsf{S}_\alpha}(z^\star))$ the *linearized asymptotic spectral factor*: when $q(\alpha) < 1$, iterates started close enough to $z^\star$ converge R-linearly with any rate above $q(\alpha)$, while the root rate of a particular trajectory can differ from $q(\alpha)$ (Remark 13). The cost-honest comparison is against T^2: both schemes use two T-evaluations per step, T^2 already attains rate $\rho^2 < \rho$, and the meaningful question is when $q(\alpha) < \rho^2$. We answer this first. Let $\{\mu_j = r_j e^{i\theta_j}\}$ denote the eigenvalues of G and let $\mathcal{D} = \{j : r_j = \rho\}$ be the set of *dominant modes*, where $\rho = \rho(G)$.

3.1 When does PAPF beat T^2?

Theorem 2 (Necessary and sufficient condition for PAPF to beat T^2)**.** *Let G have spectral radius $\rho \in (0,1)$. The following are equivalent:* (i) $\exists\, \alpha \in (0,1)$ *such that* $\rho\big(\alpha I + (1-\alpha)G^2\big) < \rho^2$*;* (ii) $\cos(2\theta_j) < \rho^2$ *for every* $j \in \mathcal{D}$.

Proof sketch (full proof in Appendix F). For an eigenvalue $\mu = \rho\, e^{i\theta}$ of G, the corresponding eigenvalue of S_α is $\sigma(\alpha) = \alpha + (1-\alpha)\mu^2$, and direct expansion gives

$$|\sigma(\alpha)|^2 - \rho^4 \;=\; \alpha\big[\,\alpha\, A + B\,\big], \qquad A := |1-\rho^2 e^{2i\theta}|^2 > 0, \qquad B := 2\rho^2\big(\cos 2\theta - \rho^2\big). \tag{3}$$

If $\cos 2\theta \geq \rho^2$ then $B \geq 0$ and $|\sigma(\alpha)| \geq \rho^2$ for every α, so one dominant mode failing (ii) rules out (i); if $\cos 2\theta < \rho^2$ then $B < 0$, the bracket is negative on $(0, -B/A)$ with minimizer $\alpha^\star = -B/(2A)$, and subdominant modes are uniformly contracted for small α, giving (i).

Lemma 1 (Single-mode improvement over T^2)**.** *Let $\mu = \rho\, e^{i\theta}$ with $0 < \rho < 1$ be an eigenvalue of G and $\sigma(\alpha) = \alpha + (1-\alpha)\mu^2$ the corresponding eigenvalue of S_α. Then* (a) $|\sigma(\alpha)| < \rho^2$ *for some $\alpha \in (0,1)$ iff* $\cos(2\theta) < \rho^2$*;* (b) *when* $\cos(2\theta) < \rho^2$*, $|\sigma(\alpha)|$ is minimized at*

$$\alpha^\star(\rho,\theta) = \frac{\rho^2(\rho^2 - \cos 2\theta)}{|1-\rho^2 e^{2i\theta}|^2} \;\in\; (0,1), \tag{4}$$

and $|\sigma(\alpha)| < \rho^2$ for all $\alpha \in (0,\, 2\alpha^\star)$; (c) *at $\alpha = \alpha^\star$ the minimized eigenvalue and its ratio to the T^2 eigenvalue ρ^2 are*

$$|\sigma^\star| = \frac{\rho^2\,|\sin 2\theta|}{|1-\rho^2 e^{2i\theta}|}, \qquad \frac{|\sigma^\star|}{\rho^2} = \frac{|\sin 2\theta|}{|1-\rho^2 e^{2i\theta}|} \;<\; 1. \tag{5}$$

3.2 Optimal parameters

For the full operator, the optimal α minimizes a one-dimensional convex minimax over the Jacobian spectrum (Corollary 1 in Appendix F); when G has a unique dominant conjugate pair and the filtered subdominant modes satisfy the gap condition (31) of Corollary 2, the minimax reduces to the closed form $\alpha^\star(\rho,\theta)$ in (4). PAPF beats T^2 on a whole interval $(0, \bar\alpha)$, whose width has no uniform lower bound (it shrinks to zero as the dominant phase approaches the boundary θ_b). As $\rho \to 1^-$ the ratio (5) tends to $|\cos\theta|$ (Corollary 3): 0 at $\theta = \pi/2$, $1/\sqrt{2}$ at $\pi/4$, 1 as $\theta \to 0$. The full-spectrum minimizer lies in $[0, \rho^2/(1+\rho^2)] \subset [0, \tfrac{1}{2}]$ and is unique (Lemma 3, App. F), so the search interval $[0, \tfrac{1}{2}]$ is sufficient.

3.3 WHAT THE FNE DISK ALLOWS

For FNE maps the disk of Section 2 restricts every dominant angle to $\theta \le \arccos\rho$, which with Lemma 1 caps what the family $\alpha \in [0,1)$ can achieve.

Proposition 2 (FNE gain cap). *Let T be FNE with respect to some inner product and let Assumption 1 hold with $\rho = \rho(G) \in (0,1)$. Then for every $\alpha \in [0,1)$,*

$$\frac{q(\alpha)}{\rho^2} \;\ge\; \kappa_{\mathrm{FNE}}(\rho) := \frac{2\rho}{\sqrt{1+3\rho^2}}, \qquad q(\alpha) = \rho\big(\alpha I + (1-\alpha)G^2\big). \tag{6}$$

The bound is sharp: the planar map $T(z) = \rho R(\theta)z$ with $\rho = \cos\theta$ ($R(\theta)$ the rotation by θ) is FNE and attains equality at $\alpha = \alpha^\star$. Since two schemes with per-step factors q and ρ^2 (two T-calls each) need asymptotically $\log q/\log\rho^2$ times fewer T-calls when both attain their spectral factors, the asymptotic T-call speedup is at most $B(\rho) := \log(\rho^2\kappa_{\mathrm{FNE}}(\rho))/\log\rho^2$, with $9/8 < B(\rho) < 3/2$, $B(\rho) \to 9/8$ as $\rho \to 1$, $B(0.99) = 1.126$, and $B(\frac{1}{2}) = 1.202$ (attained by $T = \frac{1}{2}R(\pi/3)$).

The per-step factor ratio (bounded below by $\kappa_{\mathrm{FNE}} \to 1$), the asymptotic T-call speedup (bounded above by $B(\rho)$, assuming both trajectories attain their spectral factors and fixed start-up costs are negligible as the tolerance tends to zero), and the measured speedup of a finite run are distinct quantities. For FNE maps and $\alpha \in [0,1)$ the asymptotic equal-cost advantage is thus bounded by $B(\rho) - 1 < 50\%$, and by about 12.5% in the slow-contraction limit $\rho \to 1$, whereas outside the disk it is unbounded: Theorem 1 is a *safety* guarantee, not a source of large speedups; the cap concerns this family and FNE geometry only (Remark 10).

3.4 PAPF VERSUS RELAXED KM WITHIN THE DEGREE-TWO FAMILY

The canonical one-parameter, averagedness-preserving modification of Picard iteration is convex relaxed KM, $\mathsf{R}_\beta = \beta I + (1-\beta)T$ with $\beta \in [0,1)$ (over-relaxation $\beta < 0$ lies outside the convex family; Remark 10), at one T-call per step; PAPF is the $b = 0$ member of the convex degree-2 family $\{aI + bT + cT^2 : a,b,c \ge 0,\ a+b+c = 1\}$ at two T-calls per step (within FNE geometry the cap of Proposition 2 holds for the whole family, Remark 11). The comparison is made mode by mode at the level of Jacobians (two relaxed-KM steps have Jacobian $(\beta I + (1-\beta)G)^2$, although $\mathsf{R}_\beta \circ \mathsf{R}_\beta$ is not a polynomial in T). For a dominant pair $\rho e^{\pm i\theta}$, the same sign-of-B argument as in Lemma 1 gives the per-mode optimal moduli $\kappa_{\mathrm{KM}}(\rho,\theta)$ of relaxed KM and $\kappa_{\mathrm{P}}(\rho,\theta)$ of PAPF, stated in (33) (App. F); the equal-cost comparison is κ_{KM}^2 (two T-calls of relaxed KM) against κ_{P} (one PAPF step).

Proposition 3 (Single-pair equal-cost comparison with convex relaxed KM). *Let $\rho \in (0,1)$, $\theta \in (0,\pi)$, $c = \cos\theta$, and $g(\rho,c) := (1+\rho^2)(1-5c^2) + 2\rho c + 6\rho c^3$.*

(i) (Nesting.) *For $\theta \in (0,\pi/2]$, $\cos\theta < \rho$ implies $\cos 2\theta < \rho^2$: wherever relaxed KM accelerates, so does PAPF, but not conversely.*

(ii) (Right half-plane.) *For $\theta \in (0,\pi/2]$, $\kappa_{\mathrm{P}} \le \kappa_{\mathrm{KM}}^2$, strictly iff $\cos 2\theta < \rho^2$ (where relaxed KM accelerates, $g(\rho,c) \ge (1-\rho^2)^2 > 0$); at $\theta = \pi/2$, $\kappa_{\mathrm{P}} = 0$ while $\kappa_{\mathrm{KM}}^2 = \rho^2/(1+\rho^2)$.*

(iii) (Left half-plane.) *For $\theta \in (\pi/2,\pi)$ relaxed KM always accelerates, and $\kappa_{\mathrm{P}} < \kappa_{\mathrm{KM}}^2$ iff $c > c_\times(\rho)$, the unique zero of $g(\rho,\cdot)$ on $(-1,0)$; $c_\times(\rho) \to -1/3$ as $\rho \to 1$ (crossover angle $\to \arccos(-1/3) \approx 109.47^\circ$), and as $\theta \to \pi$ relaxed KM annihilates the real eigenvalue $-\rho$ at $\beta = \rho/(1+\rho)$ whereas PAPF cannot improve on T^2.*

The two filters are thus complementary mode by mode (Figure 5, App. F): sign cancellation (relaxed KM) owns modes near the negative real axis, phase-doubled cancellation (PAPF) the right half-plane and the modes near $\pm\pi/2$. The per-mode comparison does not decide the full-spectrum winner by itself (Remark 12); the instances of Section 5, each filter tuned as stated in Table 1, are consistent with it.

Proposition 7 (App. F) completes the picture: within $a+b+c = 1$ the degree-2 annihilator of a dominant pair is convex iff $\cos\theta \le 0$, independently of ρ, and at $\theta = \pi/2$ it *is* PAPF's optimal filter; for $\theta < \pi/2$ exact cancellation is incompatible with averagedness-preserving convexity.

4 ADAPTIVE PARAMETER SELECTION

The optimal $\alpha^\star$ depends on the spectrum of $G = J_T(z^\star)$, typically unknown. When the operator structure makes the spectrum explicit (e.g., PDHG-OT, App. C), $\alpha^\star$ can be set analytically; for black-box operators we propose an empirical procedure that identifies an effective $\hat{\alpha}$ from residual observations alone. We do *not* claim recovery of $\alpha^\star$: a short probe can miss a good parameter on flat or transient landscapes, or accept an asymptotically worse one when a slow mode is initially under-excited (Remark 14); the acceptance test below reduces but does not eliminate the latter.

4.1 OBSERVABLE PROXY AND CONSISTENCY

Observable proxy. The *fixed-point residual* $r_k := \|T(z^k) - z^k\|$ is computable without knowing $z^\star$ (the intermediate $T(z^k)$ is produced in every PAPF step). Under the hypotheses of Proposition 4 the *empirical root rate* $\hat{q}_k(\alpha) := (r_k/r_0)^{1/k}$ converges to $q(\alpha)$; the single-step ratio r_{k+1}/r_k can oscillate for a complex dominant pair (not necessarily: for an isotropic rotation in the Euclidean norm it is constant), so we track the k-step geometric mean, which is less sensitive to any single ratio.

Proposition 4 (Consistency of the empirical contraction rate, locally affine case)**.** *Let Assumption 1 hold and suppose T is affine on a neighborhood of $z^\star$, i.e., $T(z) = z^\star + G(z - z^\star)$ near $z^\star$. Fix $\alpha \in [0, 1)$ with $q(\alpha) := \rho(A_\alpha) > 0$, $A_\alpha := \alpha I + (1-\alpha)G^2$, and let Π_α be the spectral (Riesz) projection onto the invariant subspace of A_α associated with its eigenvalues of modulus $q(\alpha)$. If z^0 is sufficiently close to $z^\star$ and $\Pi_\alpha(z^0 - z^\star) \neq 0$, then $\left(r_k/r_0\right)^{1/k} \to q(\alpha)$ as $k \to \infty$.*

For diagonalizable A_α the condition means a nonzero coefficient of $z^0 - z^\star$ (against *left* eigenvectors if nonnormal) on an eigenvector attaining $q(\alpha)$; under Assumption 2 the PDHG map of App. C is affine near $z^\star$; for smooth nonlinear T the condition is *not* sufficient (Remark 13), and we make no nonlinear claim.

4.2 GOLDEN-SECTION SEARCH

Proposition 5 (Unimodality of $q(\alpha)$)**.** *The contraction factor $q(\alpha) = \rho(\alpha I + (1-\alpha)G^2) = \max_j |\alpha + (1-\alpha)\mu_j^2|$ is a convex function of $\alpha \in [0, 1]$, hence unimodal on $[0, 1]$.*

Unimodality of q makes golden-section search (Kiefer, 1953) applicable; the finite-probe proxy need not inherit it, one reason for the acceptance test. These results motivate Algorithm 1 without certifying it: after a *burn-in* phase intended to reduce transient effects (no fixed length guarantees the local regime), golden-section search narrows a bracket using short probes from the common point $\bar{z}$, and the final midpoint $\hat{\alpha}$ is accepted only if it beats an explicit $\alpha = 0$ probe by the relative margin δ_{acc}—an *empirical* safeguard, not a guarantee (Section 5.4). Proposition 4 only says that sufficiently long probes recover the order of two fixed parameters with different q; it gives no uniformity over the interval and does not say that the short probes of the search locate $\alpha^\star_G$. For an affine map and a finite candidate set, the factor of the parameter selected by a K-step probe is within $(\kappa/\gamma)^{1/K}$ of the best candidate's, with κ the eigenbasis conditioning and γ the relative excitation of the least-excited factor-deciding mode (Proposition 8, App. G.6): this quantifies why non-normality, under-excitation, and short windows defeat short probes, but certifies nothing, as κ and γ are unobservable.

Cost. With the defaults $(K_{\text{burn}}, K_{\text{probe}}, \delta, \delta_{\text{acc}}) = (30, 5, 0.005, 0.02)$ the search needs $\lceil \log(0.5/\delta)/\log\varphi \rceil = 10$ contractions after two initial probes, plus two acceptance probes, at $2K_{\text{probe}} = 10$ T-calls each: 140 T-calls of overhead beyond the burn-in (which is ordinary T^2 progress). The locked phase thus starts after at least 200 T-calls, more than the entire best-fixed-α run on the cheap instances of Table 1 (32–186 calls), so the search pays off only when the remaining decay is long enough to amortize it (a break-even rule and an accepted $\hat{\alpha}$ that raises the total cost: App. G.3). The residual proxy, burn-in, and memory are discussed in App. G.6.

Global convergence. For FNE T the retained trajectory (burn-in, then one fixed averaged map from $\bar{z}$; probes discarded) converges by Theorem 1 for any $\hat{\alpha}$, which does not imply a speedup (App. G.6); for non-FNE T Algorithm 1 is a local heuristic supported by Theorem 2.

Algorithm 1 PAPF with adaptive α: golden-section search with an $\alpha = 0$ acceptance test

Require: T, z^0, ε; K_{burn}, K_{probe}, bracket tolerance δ, acceptance margin δ_{acc}

1: **Phase 1 (burn-in):** run K_{burn} steps of T^2 from z^0 (stop if $r_k \le \varepsilon$); let $\bar{z}$ be the result
2: **Phase 2 (golden-section search):** $[a, b] \leftarrow [0, \frac{1}{2}]$; probe the two interior golden-section points; **while** $b - a > \delta$: keep the sub-bracket containing the smaller probe value and probe its one new interior point; then $\hat{\alpha} \leftarrow (a + b)/2$
3: **Acceptance test:** $f_{\hat{\alpha}} \leftarrow \textsc{Probe}(\hat{\alpha})$, $f_0 \leftarrow \textsc{Probe}(0)$; **if** $f_{\hat{\alpha}} \ge (1 - \delta_{\mathrm{acc}})\, f_0$ **then** $\hat{\alpha} \leftarrow 0$ (fall back to T^2)
4: **Phase 3 (locked):** run $\mathsf{S}_{\hat{\alpha}}$ from $\bar{z}$ until $r_k \le \varepsilon$
5: **Subroutine** $\textsc{Probe}(\alpha)$: from $\bar{z}$ run K_{probe} filtered updates $z \leftarrow \alpha z + (1-\alpha)T(T(z))$ ($2K_{\mathrm{probe}}$ T-calls); return the residual $r_{K_{\mathrm{probe}}-1}$ computed inside the last update

Table 1: T-calls to tolerance on the nine non-FNE instances. Relaxed KM and PAPF use their best fixed parameter (in parentheses); "adaptive" is Algorithm 1. Bold: better one-parameter filter. Best fixed parameters: spectral minimax for the small OT, power-method, and synthetic instances; finite-tolerance full-run grid for the two 50×50 OT instances (App. E). †Plus 40 (36) eigenvalues at 1 from non-unique duals (App. C.4). ‡Warm-start local test ($z^0 = z^\star + 10^{-6}\xi$, stop on $\|z - z^\star\|$).

Instance	Spectrum	T^2	rel. KM (β)	PAPF (α)	speedup	adaptive
Quad-OT (6×5)‡	$\sqrt{a}$=0.997, $\theta\in[0.23, 1]$	6516	3625 (0.11)	**3416** (0.06)	1.9×	5408 (0.39)
Quad-OT (50×50)	$\sqrt{a}$=0.999, $\theta\le0.24^{\dagger}$	33356	31008 (0.11)	**27884** (0.15)	1.2×	33496 (0)
Iris OT (50×50)	$\sqrt{a}$=0.999, $\theta\le0.22^{\dagger}$	21264	16864 (0.13)	**14970** (0.27)	1.4×	15392 (0.38)
Power method (8×8)	$0.99e^{\pm i\pi/3}$	3564	245 (0.49)	**118** (0.38)	30×	368 (0.25)
Markov chain (3 states)	$0.98e^{\pm i2\pi/3}$	1558	**46** (0.49)	90 (0.49)	17×	288 (0.49)
PageRank (15 nodes)	$0.98e^{\pm i2\pi/3}$	1616	**49** (0.49)	96 (0.49)	17×	290 (0.49)
Synthetic $\pi/4$ (30×30)	$0.99e^{\pm i\pi/4}$	3200	386 (0.48)	**186** (0.49)	17×	382 (0.50)
Toy linear ($\mathbb{R}^3$)	$0.98e^{\pm i\pi/2}$, real 0.8	1404	157 (0.17)	**156** (0.14)	9.0×	326 (0.49)
Complex-dom. ($\mathbb{R}^{50}$)	$r_j e^{\pm i\pi/2}$, $r_j\in[0.7, 0.99]$	2798	82 (0.50)	**32** (0.42)	87×	220 (0.49)

5 Numerical Experiments

Setup and conventions. All counts are T-calls to the stated stopping criterion (the residual, except the state error $\|z - z^\star\|$ on the small OT instance, a *warm-start local test* from $z^\star + 10^{-6}\xi$). "Best fixed α" is the minimax optimum of Corollary 1 on the computed spectrum for the small OT, power-method, and synthetic instances, and the best value on a finite-tolerance full-run grid (0.01 step, including $\alpha = 0$) for the two 50×50 OT instances; relaxed KM is tuned the same way. Adaptive PAPF is Algorithm 1 with $(K_{\mathrm{burn}}, K_{\mathrm{probe}}, \delta, \delta_{\mathrm{acc}}) = (30, 5, 0.005, 0.02)$. For the optimal-transport (OT) instances, n, m are the numbers of source and target points (App. E gives each construction), ε the quadratic regularization weight, and $a = 1/(1 + \tau\varepsilon)$ the primal contraction factor ($\sqrt{a}$ is the modulus of the coupled complex eigenvalues); we use the stability stepsizes $\tau = \sigma = 0.99/\|\mathcal{K}_r\|$ and verify Theorem 2 on the computed spectrum of the connected small instance rather than through the (sufficient) centering rule of Theorem 3; the two 50×50 instances are empirical tests (App. C.4). The nine instances (three PDHG-OT problems, four power-method matrices, two synthetic operators) are listed with their spectra in Table 1 and specified in full in App. E; tolerances: 10^{-14} on $\|z - z^\star\|$ (small OT), 10^{-12} on the residual (other OT and synthetic), 10^{-14} (power-method family, where T^2, relaxed KM, PAPF in all phases, and TKMA renormalize after every update: the sphere algorithm $\widetilde{\mathsf{S}}_\alpha$ of App. D).

5.1 PDHG for quadratically regularized OT: a non-FNE map and an FNE control

Spectrum. For no-extrapolation PDHG (9) on the small instance the Jacobian at $z^\star$ is computable: the 20 coupled complex eigenvalues lie on the circle of radius $\sqrt{a} = 0.997$ with angles $\theta \in [0.23, 1.00]$, the 18 null-space modes are real at exactly a, and two are zero. The largest angle violates the FNE disk ($\max_\lambda |\lambda - \frac{1}{2}| - \frac{1}{2} = +0.342$), so this map is convergent but *not FNE in any metric*; the binding

mode $\theta = 0.232$ gives $\cos 2\theta = 0.894 < a = 0.994$, so Theorem 2 holds for every dominant mode; the minimax of Corollary 1 over the computed spectrum gives $\alpha^\star = 0.058$ (Table 1), balancing the binding coupled mode against the 18 real modes at a. At $n{=}m{=}50$ the exact Jacobians have the same structure plus 40 and 36 eigenvalues exactly 1 from non-unique duals (disconnected active-set graph, so Assumption 3 fails and $p_\alpha(1) = 1$); a diagnostic minimax over the non-unit spectrum gives 0.165 and 0.276, while on the full state space $q(\alpha) = 1$ for every α (App. C.4). The extrapolated Chambolle–Pock map T_{cp} on the same instance (App. G.2) is the resolvent of the KKT operator in a preconditioned metric $M \succ 0$ (Chambolle & Pock, 2011; He & Yuan, 2012); its computed spectrum lies inside the disk with real dominant modes, so Theorem 2 rules out any gain (and Proposition 2 would cap it at $B(0.997) - 1 \approx 12.5\%$ for complex modes).

Results (Table 1, Fig. 2c). On the non-FNE map PAPF reaches tolerance in 3416 T-calls vs. 6516 for T^2 on the small instance (a warm-start local test; 1.9× at the spectral minimax $\alpha^\star = 0.058$), 27,884 vs. 33,356 at $n{=}m{=}50$ (1.2×), and 14,970 vs. 21,264 on Iris (1.4×), at the grid optima 0.15 and 0.27, ahead of tuned relaxed KM in all three cases, consistent with the per-mode comparison of Proposition 3(ii) for right-half-plane spectra (both filters tuned as stated in Table 1: spectral minimax on the small instance, full-run grid at $n{=}m{=}50$). On the FNE control T_{cp} the best filter is $\alpha = 0$ and the adaptive rule falls back to it, paying only the search overhead (3354 vs. 3214 T-calls; 22,286 vs. 22,146 at $n{=}m{=}50$). T_{cp}^2 is itself faster than PAPF on T (3214 vs. 3416; 22,146 vs. 27,884): extrapolation is the better algorithm here, and PAPF recovers most of that gap on the small instance and about half at $n{=}m{=}50$ without extrapolation or memory; the ranking is unchanged at matched KKT accuracy (Table 5, App. G.2). TKMA (He et al., 2025) stalls on the non-FNE map in both the Euclidean and the M-metric on all three instances and converges on T_{cp} in both while trailing T_{cp}^2 (App. G.2); TKMA's convergence theorem (He et al., 2025, arXiv v3, Thm. 4.4) assumes a κ-averaged operator, with unrestricted mixing only for $\kappa \le \frac{1}{2}$ (i.e., FNE) and a κ-dependent restriction otherwise; the no-extrapolation map is not FNE and we have not verified the remaining case, so the comparison is empirical and the stall is not a theoretical counterexample. At an $n{=}m{=}200$ probe (40,399 variables, tolerance 10^{-10}) the best fixed-α gain shrinks to 1.08× (26,636 vs. 28,904; relaxed KM 27,126) and the adaptive rule falls back to $\alpha = 0$ (29,044). Its exact Jacobian spectrum (App. C.4) has 75 unit eigenvalues, 70 at a, and 324 conjugate pairs all of modulus $\sqrt{a}$ with angles in $[0.003, 0.136]$; 19 pairs lie below $\theta_b = \frac{1}{2}\arccos a = 0.022$, so the criterion of Theorem 2 fails, the minimax over the non-unit spectrum is $\alpha = 0$, the 1.08× is a finite-horizon effect, and the fallback is spectrally correct. KKT residuals track the fixed-point residual on every PAPF run (App. G.4).

5.2 Non-FNE instances: power method and synthetic operators

Far outside the disk the gains are large (Table 1): 30× over T^2 on the power-method matrix ($\theta = \pi/3$; Theorem 4), 17× on the synthetic $\pi/4$ matrix, and 9× and 87× on the two $\theta = \pi/2$ synthetics, where PAPF approaches annihilation while relaxed KM can only halve the dominant pair (Proposition 3(ii)). On the toy instance the single-pair closed form gives $\alpha = 0.49$ but the filtered real mode 0.8 becomes dominant, and the full-spectrum minimax $\alpha = 0.14$ (Corollary 1) matches the full-run grid optimum (Table 1). The two left-half-plane instances go the other way, consistent with Proposition 3(iii) for their dominant pair ($2\pi/3 = 120^\circ > 109.5^\circ$): relaxed KM needs 46 and 49 T-calls against PAPF's 90 and 96.

5.3 Other accelerators, Krylov methods, and wall-clock

Figure 2 and Table 3 (App. G.1) compare PAPF with safeguarded Anderson acceleration (AA, $m{\in}\{5, 10\}$), Halpern, heavy-ball inertial, and TKMA iterations on six instances: on the PDHG-OT problems all four *underperform* T^2 (0.1–0.7×; TKMA stalls) while PAPF wins; AA($m{=}10$) wins the power method (85 calls) and TKMA wins PageRank (56 calls against 96 for the best fixed α), whereas on the complex-dominant instance PAPF beats TKMA by 5× and AA by 70×. In wall-clock PAPF's overhead is below timing noise, whereas AA's least-squares solve costs 50–100 μs per T-call on the OT instance and 7 μs on the 8×8 power instance (T-call: 0.35 μs), where AA($m{=}10$) takes 0.61 ms against 0.054 ms for fixed PAPF and 0.24 ms for the adaptive rule (App. G.5). On an affine variant $x \mapsto Ax + b$ of the complex-dominant instance (one matrix–vector product per T-call; App. G.1), full-memory GMRES (Saad & Schultz, 1986) needs 24 products, PAPF 30, restarted GMRES(5)/(2) 34/43, relaxed KM 76, T^2 2582.

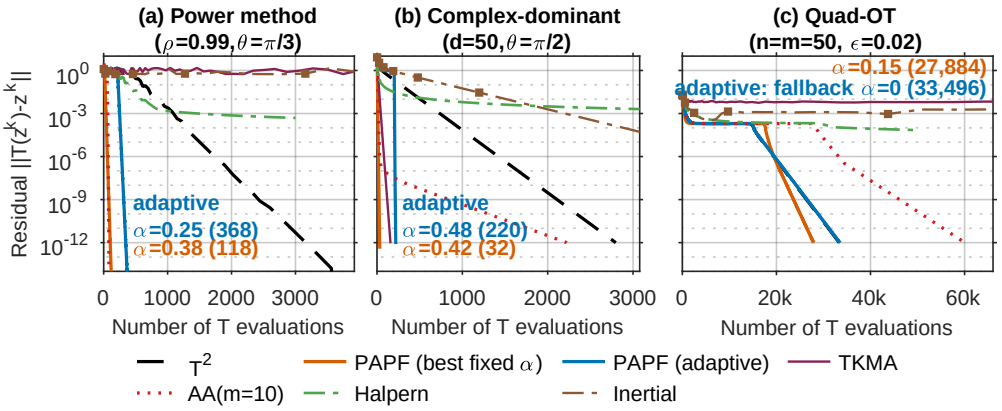

Figure 2: **Per-T-call convergence.** (a) Power method (ρ=0.99, θ=$\pi/3$): best fixed α reaches 10^{-14} in 118 T-calls, adaptive in 368, T^2 in 3,564; AA(m=10) wins (85); TKMA stalls. (b) Complex-dominant synthetic (d=50, θ=$\pi/2$): PAPF with the best fixed α dominates; AA(m=10) reaches only 1.3×. (c) Quad-OT (n=m=50): PAPF wins (1.2×); AA, Halpern, and inertial all underperform T^2. Speedup summary and wall-clock: Figure 6 (App. G.1).

5.4 Boundary sweep and robustness of the adaptive rule

Sweeping a dominant pair $\rho e^{\pm i\theta}$ ($\rho = 0.99$, subdominant real modes 0.6ρ, 0.3ρ) across the boundary $\theta_b = \frac{1}{2}\arccos\rho^2$ (Table 6, App. G.3): outside and on the boundary the tuned optimum is $\alpha = 0$, as Theorem 2 states, and the rule falls back at the cost of ≈140 search T-calls; just inside it falls back conservatively; at $\theta = \pi/2$ it selects $\alpha = 0.495$ (210 T-calls vs. 2340). On the large OT map the defaults fall back (+0.4% vs. T^2, forgoing the 16% fixed-α gain; pre-asymptotic probes, App. C.4), while a longer search $(200, 30)$ accepted a harmful $\alpha = 0.44$ (−21%). At the Table 1 starts the accepted $\hat{\alpha}$ never increased total T-calls or the tail rate relative to T^2, but the test is no guarantee (Remark 14). Over 20 random starts per instance (App. G.8) this holds for the power-method and synthetic families (no fallback or harmful selection in 160 runs, median speedups 4.4–12.8×) but not for the warm-started small OT instance, where the rule is neutral in the median (1.03×; 8 of 20 selections cost up to 2.3% more than T^2) and the Table 1 start is among the more favorable ones; the Iris acceptance is marginal in every start (probe ratio 0.9793–0.9798 against the threshold 0.98). Equally budgeted alternatives (App. G.7): a uniform-grid PAPF selector makes the same acceptance decisions as golden section and agrees with it within 1.3% except on the power method, where its endpoint choice is cheaper (306 vs. 368 T-calls) because the five-step probe objective is bimodal and neither selector finds the best fixed α; an adaptive relaxed-KM rule wins only where relaxed KM is the better filter.

6 Conclusion

PAPF strictly improves the local rate over T^2 iff $\cos(2\theta_j) < \rho^2$ on every dominant mode; for FNE T the gain is globally safe (Theorem 1) but capped at an asymptotic T-call speedup $B(\rho) < 3/2$, $B \to 9/8$ as $\rho \to 1$ (Proposition 2); PAPF and relaxed KM own complementary phase regimes mode by mode (Proposition 3). The experiments agree: 1.2–1.9× over T^2 on the non-FNE PDHG-OT maps, 9–87× on the power-method and synthetic instances, no gain on the FNE map, a loss to relaxed KM at $\theta = 2\pi/3$, and no advantage of the automatic rule over the extrapolated Chambolle–Pock map on OT.

Limitations. The analysis is local; PAPF cannot improve real spectra, and its FNE-map advantage is bounded by $B(\rho)$; proxy consistency is proved for locally affine maps only; the adaptive rule can miss a good parameter or accept a worse one (Remark 14) and its gains are instance-dependent (App. G.8); the experiments are small-scale ($n \leq 200$). *Outlook:* convex $b > 0$ filters (Proposition 7; they cannot beat the FNE cap, Remark 11), higher-degree averaged filters, and matrix-free iterations such as phase retrieval (Fienup, 1982; Bauschke et al., 2002; Marchesini, 2007).

AI USE STATEMENT

In this work, we used generative AI tools for assisting in checking some closed-form claims (the comparison results of Section 3.4, the cap of Proposition 2 and its extension in Remark 11, and the finite-window bound of Proposition 8), for feedback on the design of the experiments (the boundary sweep and the matched-accuracy, equal-budget, and repeated-initialization controls), for implementing experiment and diagnostic scripts (including the equal-budget selectors of App. G.7), and for drafting the interpretation of diagnostic results; every claim, proof, script, and interpretation produced this way was verified line by line, rerun, or checked numerically by the authors. We have not used generative AI tools to generate data sets beyond the documented random instance generators, to assist with translation, or to clean or reformat data, and the remaining tasks with required disclosure (qualitative and thematic data analysis) are not applicable to this work. Additionally, we used generative AI tools to draft and edit parts of the paper for readability, to suggest its structure, to create and modify figure scripts, and to identify and format references. We have reviewed all AI-assisted work; in particular, every bibliography entry was checked by the authors against publisher records, original papers, or authoritative archives, and persistent identifiers or official source URLs are given where available. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All theoretical results are stated with their assumptions in Sections 2–4 and proved in full in Appendices C, D, and F. Appendix E gives every instance generator, seed, stepsize, tolerance, and adaptive-search setting, and Appendix G tabulates all auxiliary measurements (baselines, TKMA in both metrics, boundary sweep, wall-clock, ablations).

REFERENCES

Felipe Alvarez and Hedy Attouch. An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping. *Set-Valued Analysis*, 9(1–2):3–11, 2001. doi: 10.1023/A:1011253113155.

Donald G. Anderson. Iterative procedures for nonlinear integral equations. *Journal of the ACM*, 12 (4):547–560, 1965. doi: 10.1145/321296.321305.

Edgar Anderson. The irises of the Gaspé Peninsula. *Bulletin of the American Iris Society*, 59:2–5, 1935.

Timo Aspelmeier, C. Charitha, and D. Russell Luke. Local linear convergence of the ADMM/Douglas–Rachford algorithms without strong convexity and application to statistical imaging. *SIAM Journal on Imaging Sciences*, 9(2):842–868, 2016. doi: 10.1137/15M103580X.

Felipe Atenas, Farhana Ahmed Simi, and Matthew K. Tam. Linear convergence of relocated fixed-point iterations. *arXiv preprint arXiv:2512.12954*, 2025. URL https://arxiv.org/abs/2512.12954.

Felipe Atenas, Heinz H. Bauschke, Minh N. Dao, and Matthew K. Tam. Relocated fixed-point iterations with applications to variable stepsize resolvent splitting. *SIAM Journal on Optimization*, 36(3):1297–1325, 2026a. doi: 10.1137/25M1776810.

Felipe Atenas, Minh N. Dao, and Matthew K. Tam. Variable stepsize distributed forward-backward splitting methods as relocated fixed-point iterations. *arXiv preprint arXiv:2601.15531*, 2026b. URL https://arxiv.org/abs/2601.15531.

Hedy Attouch and Juan Peypouquet. Convergence of inertial dynamics and proximal algorithms governed by maximally monotone operators. *Mathematical Programming*, 174(1–2):391–432, 2019. doi: 10.1007/s10107-018-1252-x.

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer New York, New York, 2011. doi: 10.1007/978-1-4419-9467-7.

Heinz H. Bauschke, Patrick L. Combettes, and D. Russell Luke. Phase retrieval, error reduction algorithm, and Fienup variants: a view from convex optimization. *Journal of the Optical Society of America A*, 19(7):1334–1345, 2002. doi: 10.1364/JOSAA.19.001334.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. doi: 10.1137/080716542.

Daniel Boley. Local linear convergence of the alternating direction method of multipliers on quadratic or linear programs. *SIAM Journal on Optimization*, 23(4):2183–2207, 2013. doi: 10.1137/120878951.

Raghu Bollapragada, Damien Scieur, and Alexandre d'Aspremont. Nonlinear acceleration of primal-dual algorithms. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 739–747. PMLR, 2019. URL https://proceedings.mlr.press/v89/bollapragada19a.html.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. doi: 10.1561/2200000016.

Radu Ioan Boţ, Ernö Robert Csetnek, and Christopher Hendrich. Inertial Douglas–Rachford splitting for monotone inclusion problems. *Applied Mathematics and Computation*, 256:472–487, 2015. doi: 10.1016/j.amc.2015.01.017.

Claude Brezinski and Michela Redivo-Zaglia. *Extrapolation Methods: Theory and Practice*, volume 2 of *Studies in Computational Mathematics*. North-Holland, Amsterdam, 1991. ISBN 0-444-88814-4.

Maciej J. Capiński and Anna Wasieczko-Zając. Geometric proof of strong stable/unstable manifolds with application to the restricted three body problem. *Topological Methods in Nonlinear Analysis*, 46(1):363–399, 2015. doi: 10.12775/TMNA.2015.051.

Antonin Chambolle and Charles Dossal. On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". *Journal of Optimization Theory and Applications*, 166(3): 968–982, 2015. doi: 10.1007/s10957-015-0746-4.

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011. doi: 10.1007/s10851-010-0251-1.

Antonin Chambolle, Claire Delplancke, Matthias J. Ehrhardt, Carola-Bibiane Schönlieb, and Junqi Tang. Stochastic primal-dual hybrid gradient algorithm with adaptive step sizes. *Journal of Mathematical Imaging and Vision*, 66(3):294–313, 2024. doi: 10.1007/s10851-024-01174-1.

Patrick L. Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In Heinz H. Bauschke, Regina S. Burachik, Patrick L. Combettes, Veit Elser, D. Russell Luke, and Henry Wolkowicz (eds.), *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, volume 49 of *Springer Optimization and Its Applications*, pp. 185–212. Springer New York, New York, 2011. doi: 10.1007/978-1-4419-9569-8_10.

Patrick L. Combettes and Jean-Christophe Pesquet. Fixed point strategies in data science. *IEEE Transactions on Signal Processing*, 69:3878–3905, 2021. doi: 10.1109/TSP.2021.3069677.

Patrick L. Combettes and Isao Yamada. Compositions and convex combinations of averaged nonexpansive operators. *Journal of Mathematical Analysis and Applications*, 425(1):55–70, 2015. doi: 10.1016/j.jmaa.2014.11.044.

Daniel Cortild and Juan Peypouquet. Krasnoselskii–Mann iterations: Inertia, perturbations and approximation. *Journal of Optimization Theory and Applications*, 204(2), 2025. doi: 10.1007/s10957-024-02600-5. Article 35.

Damek Davis and Wotao Yin. Faster convergence rates of relaxed Peaceman–Rachford and ADMM under regularity assumptions. *Mathematics of Operations Research*, 42(3):783–805, 2017a. doi: 10.1287/moor.2016.0827.

Damek Davis and Wotao Yin. A three-operator splitting scheme and its optimization applications. *Set-Valued and Variational Analysis*, 25(4):829–858, 2017b. doi: 10.1007/s11228-017-0421-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.

Haw-ren Fang and Yousef Saad. Two classes of multisecant methods for nonlinear acceleration. *Numerical Linear Algebra with Applications*, 16(3):197–221, 2009. doi: 10.1002/nla.617.

Olivier Fercoq. Monitoring the convergence speed of PDHG to find better primal and dual step sizes. *arXiv preprint arXiv:2403.19202*, 2024. URL https://arxiv.org/abs/2403.19202.

J. R. Fienup. Phase retrieval algorithms: a comparison. *Applied Optics*, 21(15):2758–2769, 1982. doi: 10.1364/AO.21.002758.

Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 (2):179–188, 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x.

Anqi Fu, Junzi Zhang, and Stephen Boyd. Anderson accelerated Douglas–Rachford splitting. *SIAM Journal on Scientific Computing*, 42(6):A3560–A3583, 2020. doi: 10.1137/19M1290097.

Casey Garner, Gilad Lerman, and Teng Zhang. Improved convergence factor of windowed Anderson acceleration for symmetric fixed-point iterations. *SIAM Journal on Numerical Analysis*, 64(5): 1891–1922, 2026. doi: 10.1137/24M1643591.

Pontus Giselsson and Stephen Boyd. Linear convergence and metric selection for Douglas–Rachford splitting and ADMM. *IEEE Transactions on Automatic Control*, 62(2):532–544, 2017. doi: 10.1109/TAC.2016.2564160.

Gene H. Golub and Richard S. Varga. Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second order Richardson iterative methods. Part I. *Numerische Mathematik*, 3(1):147–156, 1961. doi: 10.1007/BF01386013.

A. R. Gourlay. The acceleration of the Peaceman–Rachford method by Chebyshev polynomials. *The Computer Journal*, 10(4):378–382, 1968. doi: 10.1093/comjnl/10.4.378.

Bingsheng He and Xiaoming Yuan. Convergence analysis of primal-dual algorithms for a saddle-point problem: From contraction perspective. *SIAM Journal on Imaging Sciences*, 5(1):119–149, 2012. doi: 10.1137/100814494.

Yongxin He, Jingyuan Li, Yizun Lin, and Deren Han. A two-step Krasnosel'skiĭ–Mann algorithm with adaptive momentum and its applications to image denoising and matrix completion. *arXiv preprint arXiv:2510.24544*, 2025. doi: 10.48550/arXiv.2510.24544. URL https://arxiv.org/abs/2510.24544.

Morris W. Hirsch, Charles C. Pugh, and Michael Shub. *Invariant Manifolds*, volume 583 of *Lecture Notes in Mathematics*. Springer, Berlin, Heidelberg, 1977. doi: 10.1007/BFb0092042.

Shiro Ishikawa. Fixed points by a new iteration method. *Proceedings of the American Mathematical Society*, 44(1):147–150, 1974. doi: 10.1090/S0002-9939-1974-0336469-5.

C. T. Kelley. *Iterative Methods for Linear and Nonlinear Equations*, volume 16 of *Frontiers in Applied Mathematics*. SIAM, Philadelphia, PA, 1995. doi: 10.1137/1.9781611970944.

Jack Kiefer. Sequential minimax search for a maximum. *Proceedings of the American Mathematical Society*, 4(3):502–506, 1953. doi: 10.1090/S0002-9939-1953-0055639-3.

Mark Aleksandrovich Krasnosel'skii. Two remarks on the method of successive approximations. *Uspekhi Matematicheskikh Nauk*, 10(1(63)):123–127, 1955. URL https://www.mathnet.ru/eng/rm7954. In Russian.

Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Local convergence properties of Douglas–Rachford and alternating direction method of multipliers. *Journal of Optimization Theory and Applications*, 172(3):874–913, 2017. doi: 10.1007/s10957-017-1061-z.

Pierre-Louis Lions and Bertrand Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979. doi: 10.1137/0716071.

Dirk A. Lorenz and Thomas Pock. An inertial forward-backward algorithm for monotone inclusions. *Journal of Mathematical Imaging and Vision*, 51(2):311–325, 2015. doi: 10.1007/s10851-014-0523-2.

Yura Malitsky and Thomas Pock. A first-order primal-dual algorithm with linesearch. *SIAM Journal on Optimization*, 28(1):411–432, 2018. doi: 10.1137/16M1092015.

W. Robert Mann. Mean value methods in iteration. *Proceedings of the American Mathematical Society*, 4(3):506–510, 1953. doi: 10.1090/S0002-9939-1953-0054846-3.

S. Marchesini. Invited article: A unified evaluation of iterative projection algorithms for phase retrieval. *Review of Scientific Instruments*, 78(1):011301, 2007. doi: 10.1063/1.2403783.

Yurii E. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983. Translation of Dokl. Akad. Nauk SSSR 269 (1983), 543–547.

Zdzisław Opial. Weak convergence of the sequence of successive approximations for nonexpansive mappings. *Bulletin of the American Mathematical Society*, 73(4):591–597, 1967. doi: 10.1090/S0002-9904-1967-11761-0.

James M. Ortega and Werner C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Computer Science and Applied Mathematics. Academic Press, New York, 1970. ISBN 978-0-12-528550-6. Reprinted by SIAM, Philadelphia, PA, 2000, doi:10.1137/1.9780898719468.

Jisun Park and Ernest K. Ryu. Exact optimal accelerated complexity for fixed-point iterations. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17420–17457. PMLR, 2022. URL https://proceedings.mlr.press/v162/park22c.html.

Gabriel Peyré and Marco Cuturi. Computational optimal transport with applications to data sciences. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. doi: 10.1561/2200000073.

R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976. doi: 10.1137/0314056.

Ernest K. Ryu and Stephen Boyd. A primer on monotone operator methods. *Applied and Computational Mathematics*, 15(1):3–43, 2016. URL https://web.stanford.edu/~boyd/papers/monotone_primer.html.

Youcef Saad and Martin H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7(3):856–869, 1986. doi: 10.1137/0907058.

Yousef Saad. *Numerical Methods for Large Eigenvalue Problems*, volume 66 of *Classics in Applied Mathematics*. SIAM, Philadelphia, PA, revised edition, 2011. doi: 10.1137/1.9781611970739.

Damien Scieur, Alexandre d'Aspremont, and Francis Bach. Regularized nonlinear acceleration. *Mathematical Programming*, 179(1–2):47–83, 2020. doi: 10.1007/s10107-018-1319-8.

Alex Toth and C. T. Kelley. Convergence analysis for Anderson acceleration. *SIAM Journal on Numerical Analysis*, 53(2):805–819, 2015. doi: 10.1137/130919398.

Richard S. Varga. *Matrix Iterative Analysis*. Prentice-Hall Series in Automatic Computation. Prentice-Hall, Englewood Cliffs, NJ, 1962.

Tadashi Wadayama and Satoshi Takabe. Chebyshev periodical successive over-relaxation for accelerating fixed-point iterations. *IEEE Signal Processing Letters*, 28:907–911, 2021. doi: 10.1109/LSP.2021.3073620. Preprint: arXiv:2001.03280, "Chebyshev inertial iteration for accelerating fixed-point iterations".

Homer F. Walker and Peng Ni. Anderson acceleration for fixed-point iterations. *SIAM Journal on Numerical Analysis*, 49(4):1715–1735, 2011. doi: 10.1137/10078356X.

Zheng Xu, Mário A. T. Figueiredo, and Tom Goldstein. Adaptive ADMM with spectral penalty parameter selection. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 718–727. PMLR, 2017. URL https://proceedings.mlr.press/v54/xu17a.html.

Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. Discerning the linear convergence of ADMM for structured convex optimization through the lens of variational analysis. *Journal of Machine Learning Research*, 21(83):1–75, 2020. URL https://jmlr.org/papers/v21/18-562.html.

Yunkai Zhou, James R. Chelikowsky, and Yousef Saad. Chebyshev-filtered subspace iteration method free of sparse diagonalization for solving the Kohn–Sham equation. *Journal of Computational Physics*, 274:770–782, 2014. doi: 10.1016/j.jcp.2014.06.056.

A Extended Related Work

Averaged operator theory and two-step KM schemes. Convex combinations and compositions of averaged maps remain averaged and converge under KM iteration; sharp averagedness constants (Combettes & Yamada, 2015) and Opial-type arguments (Opial, 1967; Krasnosel'skii, 1955; Mann, 1953) supply the structural toolkit (Bauschke & Combettes, 2011; Ryu & Boyd, 2016; Combettes & Pesquet, 2021). Two-stage fixed-point iterations include the Ishikawa scheme $x_{n+1} = (1-a_n)x_n + a_n T\big((1-b_n)x_n + b_n T x_n\big)$ (Ishikawa, 1974), which for nonlinear T is not a polynomial in T. More recent work studies two-evaluation KM constructions with adaptive momentum (TKMA, He et al., 2025) and inertial or perturbed KM iterations (Cortild & Peypouquet, 2025); these establish convergence under specific parameter conditions, with rate statements that concern distinct quantities (last-iterate differences vs. best-iterate squared residuals, under their respective assumptions), whereas whether a static blend of T and T^2 can provably beat T^2 for a given dominant-mode structure was not previously characterized. We resolve this for the canonical degree-2 blend by giving a sharp iff condition, and we compare it in closed form with the degree-1 member of the same family, relaxed KM $\beta I + (1-\beta)T$ (Section 3.4). TKMA stalls on our no-extrapolation PDHG instances in both the Euclidean and the preconditioned metric (Section 5); TKMA's convergence theorem (He et al., 2025, arXiv v3, Thm. 4.4 of the PDF, Thm. 9 of the HTML rendering) is stated for κ-averaged operators, with an unrestricted mixing parameter only for $\kappa \le \frac{1}{2}$ (i.e., FNE) and a κ-dependent restriction otherwise, and its rate theorem requires $\kappa \le \frac{1}{2}$; the no-extrapolation map is not FNE, so it lies outside the $\kappa \le \frac{1}{2}$ class, and we have not verified the remaining case, so we treat the comparison as empirical only.

Inertial, Nesterov, Halpern, and Anderson accelerations. Nesterov-type acceleration and FISTA attain $\mathcal{O}(1/k^2)$ objective-gap bounds under suitable convex-optimization assumptions (Nesterov, 1983; Beck & Teboulle, 2009; Chambolle & Dossal, 2015); inertial methods for general monotone inclusions (Alvarez & Attouch, 2001; Attouch & Peypouquet, 2019; Lorenz & Pock, 2015; Boţ et al., 2015) use a two-term recursion that is not in general an averaged map of the original operator and provide convergence guarantees under method-specific assumptions. For nonexpansive operators, Halpern-type acceleration achieves the optimal $\mathcal{O}(1/k)$ fixed-point residual norm (equivalently $\mathcal{O}(1/k^2)$ for the squared residual); Park & Ryu (2022) also analyze contractive operators, and their focus is worst-case global complexity rather than the local linear rate. Anderson acceleration (Anderson, 1965; Walker & Ni, 2011; Fang & Saad, 2009; Scieur et al., 2020; Fu et al., 2020) mixes past residuals via least squares and is locally R-linearly convergent for contractive nonlinear maps under bounded mixing coefficients (Toth & Kelley, 2015); for symmetric contractive linear maps, strict improvement of the convergence factor by windowed AA is established under window-size

or spectral conditions by Garner et al. (2026), who also give a qualified nonlinear extension with a modified algorithm, complementary to PAPF's focus on complex (non-self-adjoint) modes. PAPF instead applies a static operator polynomial that stays in the averaged class for *any* FNE T and $\alpha \in (0, 1)$, with fixed overhead and a local theory normalized to the equal-cost baseline T^2.

Chebyshev, polynomial filtering, and extrapolation. Classical Chebyshev semi-iteration (Golub & Varga, 1961; Varga, 1962; Gourlay, 1968) and the unified view of momentum and Chebyshev methods on quadratics (d'Aspremont et al., 2021) are built on real spectral intervals. Polynomial methods also address complex spectra: Chebyshev filters on enclosing ellipses and polynomial-filtered subspace iteration for eigenvalue problems (Saad, 2011; Zhou et al., 2014), Krylov methods such as GMRES for general nonsymmetric linear systems (Saad & Schultz, 1986), and sequence-transformation (extrapolation) methods (Brezinski & Redivo-Zaglia, 1991). These are higher-degree or memory-based constructions and carry no averagedness guarantee for nonlinear operators. Two closer works address nonlinear or nonsymmetric fixed-point maps directly. Wadayama & Takabe (2021) apply the inverse roots of a Chebyshev polynomial as periodic over-relaxation factors to nonlinear fixed-point iterations, with a local linearization analysis under the assumption that the Jacobian at the fixed point has real eigenvalues; Bollapragada et al. (2019) apply regularized nonlinear acceleration (an extrapolated nonlinear average of stored iterates, with weights from a small linear system) to nonsymmetric fixed-point operators such as momentum and primal–dual methods, including Chambolle–Pock. Neither considers the stationary degree-two blend of T and T^2 or the complex-phase question we ask, and we do not claim that polynomial or nonlinear acceleration of complex-spectrum maps is new; what is new here is the exact equal-cost phase criterion for the nonnegative family S_α, the sharp FNE cap, and the crossover with convex relaxed KM. PAPF is a low-degree ($d = 2$) operator polynomial designed for *complex* dominant modes via phase cancellation, with an averagedness guarantee for FNE nonlinear operators; it targets the memory-1, safeguard-free niche, and on our affine benchmark the tested GMRES configurations show the benefit of additional memory (Section 5.3).

Local rates and adaptive parameters in operator splitting. Linear convergence of DR/ADMM has been established globally under strong convexity and smoothness (Giselsson & Boyd, 2017) and locally under error bounds, metric subregularity, partial smoothness, or polyhedrality (Liang et al., 2017; Davis & Yin, 2017a; Aspelmeier et al., 2016; Yuan et al., 2020; Boley, 2013), with extensions to three-operator splitting (Davis & Yin, 2017b). Our spectral analysis additionally assumes differentiability at the fixed point (Assumption 1); this does not follow from an error bound alone, although in the polyhedral and partly smooth settings the iteration becomes locally affine once the active manifold is identified (Boley, 2013; Liang et al., 2017). Adaptive parameter rules for splitting include Malitsky–Pock (Malitsky & Pock, 2018), spectral penalty rules for ADMM (Xu et al., 2017), spectral-radius-based PDHG step search (Fercoq, 2024), and adaptive stochastic PDHG (Chambolle et al., 2024). These tune parameters that *change the underlying operator*; PAPF instead tunes a filter coefficient inside a fixed operator polynomial, preserves fixed points and averagedness for all $\alpha \in (0, 1)$, and uses residual-only golden-section search over a scalar. Relocated fixed-point iterations (Atenas et al., 2026a; 2025; 2026b) iterate a parameter-varying family of averaged (in Atenas et al., 2026b, conically averaged) operators (e.g., resolvent splittings with variable stepsizes) and insert a "relocator" that maps the fixed points of one operator onto those of the next, so that global convergence survives the variation; this goal is orthogonal to PAPF's local spectral cancellation.

B NOTATION

For convenience we collect the major symbols used throughout the paper. We disambiguate three uses of the letter "a": the bold OT marginals $\mathbf{a}, \mathbf{b}$, the PAPF mixing weight α, and the scalar $a := 1/(1 + \tau\varepsilon)$ governing the OT spectral analysis in App. C.2.

PAPF and spectral analysis (§2–§3)	
T	FNE (or differentiable) operator on $\mathbb{R}^d$
T^2	two-step Picard map, $T \circ T$
S_α	PAPF map $\alpha I + (1-\alpha)T^2$
$\alpha \in (0,1)$	PAPF mixing weight (free parameter)
$\alpha^\star(\rho,\theta)$	per-mode optimal α, Eq. (4)
$\hat{\alpha}$	adaptive estimate from Alg. 1
$G = J_T(z^\star)$	Jacobian of T at the fixed point
$\mu_j = r_j e^{i\theta_j}$	eigenvalues of G
$\rho = \rho(G)$	spectral radius of G
θ	dominant-mode angle, $\theta \in [0,\pi]$
$p_\alpha(\lambda)$	filter polynomial $\alpha + (1-\alpha)\lambda^2$
$\sigma(\alpha)$	filtered eigenvalue $p_\alpha(\mu)$
$q(\alpha)$	contraction factor $\rho(p_\alpha(G))$
$\hat{q}_k(\alpha)$	empirical contraction rate $(r_k/r_0)^{1/k}$
r_k	residual $\|T(z^k) - z^k\|$
Quad-reg OT and PDHG (App. C)	
$\mathbf{a}, \mathbf{b}$	OT marginals (boldface, vectors)
X	transport plan $\in \mathbb{R}_+^{n\times m}$
C	cost matrix
ε	quadratic OT regularization parameter
τ, σ	PDHG primal/dual stepsizes
$\mathcal{K}_r, \mathcal{K}_\Omega$	reduced / restricted constraint operators
$a := 1/(1+\tau\varepsilon)$	PDHG primal-prox shrinkage scalar (non-bold)
$s_{\min}, s_{\max}$	smallest/largest singular values of $\mathcal{K}_\Omega$
$\kappa = s_{\max}^2/s_{\min}^2$	restricted condition number
Adaptive selection (§4)	
K_{burn}	burn-in length (default 30)
K_{probe}	probe length per α (default 5)
δ	golden-section bracket tolerance (default 0.005)
φ	golden ratio $(1+\sqrt{5})/2$

C PDHG FOR QUADRATICALLY REGULARIZED OT

We derive *explicit* conditions on the regularization (through $a = 1/(1+\tau\varepsilon)$) and on the conditioning of the active-set coupling operator (through κ) that guarantee PAPF beats the equal-cost baseline T^2 when applied to no-extrapolation PDHG for quadratically regularized optimal transport. (Throughout this appendix the boldface vectors $\mathbf{a}, \mathbf{b}$ denote the OT marginals; the unrelated scalar $a := 1/(1+\tau\varepsilon)$ introduced in Section C.2 is non-bold and reserved for the spectral analysis.)

C.1 PROBLEM FORMULATION AND PDHG ITERATION

Let $C \in \mathbb{R}^{n\times m}$ be a cost matrix, $\mathbf{a} \in \mathbb{R}_+^n$ and $\mathbf{b} \in \mathbb{R}_+^m$ marginal distributions with $\mathbf{1}_n^\top \mathbf{a} = \mathbf{1}_m^\top \mathbf{b}$ (equal total mass), and $\varepsilon > 0$ a regularization parameter. The quadratically regularized optimal transport problem is

$$\min_{X \in \mathbb{R}_+^{n\times m}} \langle C, X\rangle + \frac{\varepsilon}{2}\|X\|_F^2 \quad \text{s.t.} \quad X\mathbf{1}_m = \mathbf{a}, \qquad X^\top \mathbf{1}_n = \mathbf{b}. \tag{7}$$

Here X_{ij} represents the mass transported from source i to target j, the marginal constraints enforce that the row and column sums match $\mathbf{a}$ and $\mathbf{b}$, and the quadratic term makes the objective strictly convex (ensuring a unique optimal plan).

Constraint reduction. The $n+m$ marginal constraints are linearly dependent: summing all row-sum equations and all column-sum equations both yield $\sum_{i,j} X_{ij}$ (using $\mathbf{1}_n^\top \mathbf{a} = \mathbf{1}_m^\top \mathbf{b}$), so the system has rank at most $n+m-1$. Dropping the last column-sum constraint (which is implied by the remaining $n+m-1$), we define the *reduced constraint operator* $\mathcal{K}_r : \mathbb{R}^{n\times m} \to \mathbb{R}^{n+m-1}$ by

$$\mathcal{K}_r(X) = \begin{pmatrix} X\mathbf{1}_m \\ (X^\top \mathbf{1}_n)_{1:m-1} \end{pmatrix}, \qquad c_r = \begin{pmatrix} \mathbf{a} \\ \mathbf{b}_{1:m-1} \end{pmatrix},$$

with adjoint $\mathcal{K}_r^*(u,v) = u\,\mathbf{1}_m^\top + \mathbf{1}_n\,[v^\top, 0]$. Since $\mathcal{K}_r$ has full row rank, the normal operator $\mathcal{K}_r\mathcal{K}_r^*$ is invertible, enabling closed-form projections onto the affine constraint set. Problem (7) is equivalently

$$\min_{X\in\mathbb{R}^{n\times m}} \langle C, X\rangle + \frac{\varepsilon}{2}\|X\|_F^2 + \delta_{\mathbb{R}_+^{n\times m}}(X) \quad \text{s.t.} \quad \mathcal{K}_r(X) = c_r. \tag{8}$$

PDHG iteration. We solve (8) with no-extrapolation PDHG. For stepsizes $\tau, \sigma > 0$, each iteration updates a dual variable $y \in \mathbb{R}^{n+m-1}$ (enforcing the marginal constraints) and a primal variable $X \in \mathbb{R}_+^{n\times m}$ (the transport plan):

$$\begin{aligned} y^{k+1} &= y^k + \sigma\big(\mathcal{K}_r X^k - c_r\big) && \text{(dual: penalize constraint violation)},\\ X^{k+1} &= \Big[\frac{X^k - \tau\big(\mathcal{K}_r^* y^{k+1} + C\big)}{1+\tau\varepsilon}\Big]_+ && \text{(primal: gradient step + projection)}, \end{aligned} \tag{9}$$

where $[\cdot]_+$ denotes entrywise projection onto $\mathbb{R}_+$. We write T for the induced fixed-point map on $z := (X, y)$.

C.2 Linearization at the fixed point

To apply PAPF, we need the Jacobian $G = J_T(z^\star)$ at a fixed point $z^\star = (X^\star, y^\star)$. The dual update in (9) is affine, hence automatically differentiable. The primal update involves the entrywise projection $[\cdot]_+$, which is *not* differentiable at entries equal to zero. The following assumption ensures differentiability.

Assumption 2 (Strict complementarity)**.** Define the pre-projection matrix $W^\star := (X^\star - \tau(\mathcal{K}_r^* y^\star + C))/(1+\tau\varepsilon)$, so that $X^\star = [W^\star]_+$. We assume there exists $\delta > 0$ such that every entry satisfies either $W_{ij}^\star \geq \delta$ (active: $X_{ij}^\star \geq \delta$) or $W_{ij}^\star \leq -\delta$ (inactive: $X_{ij}^\star = 0$).

In the OT context, this means every route either carries strictly positive mass or has a strictly positive reduced cost (multiplier). The assumption is standard in linear/quadratic programming and holds generically (e.g., when C and the marginals are in general position).

Under Assumption 2, the active set $\Omega := \{(i,j) : X_{ij}^\star > 0\}$ is locally constant and $[\cdot]_+$ is differentiable near $W^\star$ with Jacobian P_Ω (the diagonal projector onto Ω). We now linearize the full PDHG iteration.

Step 1: linearize each update. Write $X^k = X^\star + \delta X^k$, $y^k = y^\star + \delta y^k$. The dual step is already affine; since $\mathcal{K}_r X^\star = c_r$, it linearizes exactly to

$$\delta y^{k+1} = \delta y^k + \sigma\,\mathcal{K}_r\,\delta X^k.$$

For the primal step, expanding the argument of $[\cdot]_+$ gives the perturbation $\delta W^k = a(\delta X^k - \tau\,\mathcal{K}_r^*\,\delta y^{k+1})$, where

$$a := \frac{1}{1+\tau\varepsilon} \in (0,1).$$

Applying $J_{[\cdot]_+}(W^\star) = P_\Omega$:

$$\delta X^{k+1} = P_\Omega\,\delta W^k = a\,P_\Omega\big(\delta X^k - \tau\,\mathcal{K}_r^*\,\delta y^{k+1}\big).$$

Step 2: Restrict to the active set. The primal projection confines perturbations to the active set: $P_\Omega\,\delta X^k = \delta X^k$ (inactive entries stay zero). Define the *restricted coupling operator* $\mathcal{K}_\Omega := \mathcal{K}_r \circ P_\Omega$, which maps $\mathbb{R}^{|\Omega|}$ (active primal entries) to $\mathbb{R}^{n+m-1}$ (dual space). Then $\mathcal{K}_r\,\delta X^k = \mathcal{K}_\Omega\,\delta X^k$ and $P_\Omega\,\mathcal{K}_r^* = \mathcal{K}_\Omega^*$, so the linearized system becomes

$$\begin{aligned} \delta y^{k+1} &= \delta y^k + \sigma\,\mathcal{K}_\Omega\,\delta X^k,\\ \delta X^{k+1} &= a\big(\delta X^k - \tau\,\mathcal{K}_\Omega^*\,\delta y^{k+1}\big). \end{aligned} \tag{10}$$

Step 3: Combined Jacobian. Substituting the dual equation into the primal equation and writing the system in matrix form:

$$\begin{pmatrix}\delta X^{k+1}\\ \delta y^{k+1}\end{pmatrix} = \underbrace{\begin{pmatrix} a(I - \tau\sigma\,\mathcal{K}_\Omega^*\mathcal{K}_\Omega) & -a\tau\,\mathcal{K}_\Omega^*\\ \sigma\,\mathcal{K}_\Omega & I\end{pmatrix}}_{=:\,G}\begin{pmatrix}\delta X^k\\ \delta y^k\end{pmatrix}. \tag{11}$$

Step 4: SVD block-diagonalization. The coupling operator $\mathcal{K}_\Omega$ is *rectangular* ($|\Omega|$ columns, $n+m-1$ rows), so it does not have eigenvalues. Its SVD $\mathcal{K}_\Omega = U\Sigma V^\top$, however, simultaneously diagonalizes the products $\mathcal{K}_\Omega^*\mathcal{K}_\Omega$ and $\mathcal{K}_\Omega\mathcal{K}_\Omega^*$ that appear in G, block-diagonalizing the Jacobian.

We assume that $\mathcal{K}_\Omega$ has full row rank $p := n+m-1$ (Assumption 3); Lemma 2 characterizes this condition combinatorially and shows that a per-vertex condition is not enough.

Assumption 3 (Full row rank of the active coupling operator)**.** $\mathcal{K}_\Omega$ has full row rank $n+m-1$.

Lemma 2 (Connectivity characterization of Assumption 3)**.** *Let B_Ω be the bipartite graph with vertex set $\{1,\dots,n\} \sqcup \{1,\dots,m\}$ (sources and targets) and edge set Ω, and let k be its number of connected components. Then* $\operatorname{rank}\mathcal{K}_\Omega = n+m-k$. *In particular, Assumption 3 holds iff B_Ω is connected. The condition "every source and every one of the first $m-1$ targets has an active route" is* not *sufficient: for $n=m=2$ and $\Omega = \{(1,1),(2,2)\}$ (an example we owe to an anonymous reviewer), $\mathcal{K}_\Omega = \left[\begin{smallmatrix}1&0\\0&1\\1&0\end{smallmatrix}\right]$ has rank $2<3$, and indeed B_Ω has $k=2$ components. Conversely, any Ω containing a spanning tree of the complete bipartite graph—for instance the support of a nondegenerate basic feasible transport plan—satisfies the assumption.*

Proof. In coordinates, for $x = (x_{ij})_{(i,j)\in\Omega}$, $(\mathcal{K}_\Omega x)_i = \sum_{j:(i,j)\in\Omega} x_{ij}$ for $i \le n$ and $(\mathcal{K}_\Omega x)_{n+j} = \sum_{i:(i,j)\in\Omega} x_{ij}$ for $j \le m-1$: $\mathcal{K}_\Omega$ is the unsigned vertex–edge incidence matrix $N \in \mathbb{R}^{(n+m)\times|\Omega|}$ of B_Ω with the row of target m deleted. A vector $(u,v) \in \mathbb{R}^n \times \mathbb{R}^m$ lies in the left null space of N iff $u_i + v_j = 0$ for every edge $(i,j) \in \Omega$; following paths, this forces u to equal a constant c_ℓ on the sources and v to equal $-c_\ell$ on the targets of each component ℓ, and every such vector is a left null vector. Hence the left null space of N has dimension k and $\operatorname{rank} N = n+m-k$. The left null vectors of $\mathcal{K}_\Omega$ are the left null vectors of N with $v_m = 0$; this pins the constant of the component containing target m to 0 and leaves the other $k-1$ constants free, so the left null space of $\mathcal{K}_\Omega$ has dimension $k-1$ and $\operatorname{rank}\mathcal{K}_\Omega = (n+m-1)-(k-1) = n+m-k$. □

When B_Ω is disconnected, $\mathcal{K}_\Omega$ has a zero singular value, the corresponding coupled block $M(0) = \operatorname{diag}(a,1)$ in (12) has eigenvalue 1, and Assumption 1 fails (the dual solution is then non-unique); Theorem 3 excludes this case. Let $\Sigma = \operatorname{diag}(s_1,\dots,s_p)$ with $s_{\min} := \min_i s_i > 0$, $s_{\max} := \max_i s_i$. Writing $\delta X = V\xi + V_\perp\xi_\perp$ and $\delta y = U\eta$ in the SVD basis:

- **Null-space modes** ($\xi_\perp$, primal directions invisible to the constraints): $\xi_\perp^{k+1} = a\,\xi_\perp^k$, contracting at rate a.
- **Coupled modes** (ξ_i, η_i for $i = 1,\dots,p$): each pair evolves via the 2×2 matrix

$$M(s_i) = \begin{pmatrix} a(1-\tau\sigma s_i^2) & -a\tau s_i \\ \sigma s_i & 1 \end{pmatrix}. \tag{12}$$

The eigenvalues of G are therefore $\{a\}$ (multiplicity $|\Omega|-p$) together with the $2p$ eigenvalues of $M(s_1),\dots,M(s_p)$. This is the spectrum of the active block; the full Jacobian on all $nm+n+m-1$ coordinates, with I the inactive indices, has a zero row block for X_I (the projection maps inactive entries to zero), so after permuting the coordinates to (X_Ω, y, X_I) it is block upper triangular: its spectrum is that of the restriction to $\{\delta X_I = 0\}$, i.e., the active-block spectrum above, together with $nm - |\Omega|$ zeros. PAPF maps these zeros to α, and they are included in the full-spectrum minimax; the affine representation of T near $z^\star$ must hold for both T-calls of a filtered step, which a smaller neighborhood guarantees by continuity.

Eigenvalues of each coupled mode. Each 2×2 block has

$$\operatorname{tr}\big(M(s)\big) =: t(s) = 1 + a - a\tau\sigma s^2, \qquad \det\big(M(s)\big) = a. \tag{13}$$

When $|t(s)| < 2\sqrt{a}$ (the discriminant is negative), the eigenvalues form a complex conjugate pair

$$\lambda_\pm(s) = \sqrt{a}\, e^{\pm i\theta(s)}, \qquad \cos\theta(s) = \frac{t(s)}{2\sqrt{a}}.$$

A key structural property: every complex mode has the *same modulus* $\sqrt{a}$ (since $\det M(s) = a$), regardless of the singular value s_i. Only the *angle* $\theta(s)$ varies across modes, determined by the trace $t(s)$. Consequently, the spectral radius of the coupled part is $\sqrt{a}$, the baseline T^2 contracts each mode by a, and PAPF can improve over T^2 when the angles satisfy the condition of Theorem 2.

Remark 1 (The null-space eigenvalue a is subdominant)**.** The real eigenvalue a from the null-space modes does not obstruct PAPF: since $a < \sqrt{a}$ for $a \in (0, 1)$, it lies strictly below the spectral radius $\rho(G) = \sqrt{a}$ and is not a dominant mode. The PAPF-filtered value $p_\alpha(a) = \alpha + (1 - \alpha)a^2$ satisfies $p_\alpha(a) < a$ whenever $\alpha < a/(1 + a)$, which is guaranteed by the bound $\alpha^\star \le a/(1 + a)$ from Remark 7 (with equality only at $\theta = \pi/2$, excluded here). The null-space modes do, however, bound the attainable gain: $q(\alpha) \ge \alpha + (1 - \alpha)a^2$ for every α, which is why on the small instance of Section 5.1 the minimax optimum is $\alpha \approx 0.059$ rather than the coupled-mode value 0.47.

C.3 Explicit stepsize and mixing-parameter design

We now choose the PDHG stepsizes (τ, σ) and the PAPF mixing parameter α to ensure that *every* coupled mode satisfies the condition of Theorem 2. The key idea is a *centering* stepsize rule that symmetrizes the range of traces $t(s)$ across modes, followed by verifying the PAPF condition uniformly.

Theorem 3 (Explicit (τ, σ) and α guaranteeing PAPF beats T^2)**.** *Assume Assumptions 2 and 3 hold at $z^\star$ (by Lemma 2, the latter means that the active-route graph B_Ω is connected), and let $s_{\min} > 0$ and $s_{\max}$ be the extreme singular values of $\mathcal{K}_\Omega$. Assume the non-degeneracy condition $\kappa := s_{\max}^2/s_{\min}^2 > 1$ (equivalently, $\mathcal{K}_\Omega$ is not a multiple of an isometry on its row space). Fix any $\tau > 0$ and set $a := 1/(1 + \tau\varepsilon) \in (0, 1)$. Choose $\sigma > 0$ by the* centering *rule*

$$\tau\sigma = \frac{2(1 + a)}{a\,(s_{\min}^2 + s_{\max}^2)}. \tag{14}$$

Then $t(s_{\min}) = -t(s_{\max})$ and $\max_i |t(s_i)| \le \gamma$, where

$$\gamma := \frac{(1 + a)\,(s_{\max}^2 - s_{\min}^2)}{s_{\max}^2 + s_{\min}^2} > 0. \tag{15}$$

If in addition

$$\gamma \;<\; \sqrt{2a(1 + a)}, \tag{16}$$

then all coupled modes are complex and there exists $\alpha \in (0, 1)$ with $\rho(J_{\mathsf{S}_\alpha}(z^\star)) < a$, i.e., PAPF is locally strictly faster than the two-call baseline T^2.

An explicit valid choice of α is

$$\alpha = \frac{a^2 - a\,c}{1 - 2ac + a^2}, \qquad c := \frac{\gamma^2}{2a} - 1. \tag{17}$$

Remark 2 (Degenerate case $\kappa = 1$)**.** The assumption $\kappa > 1$ excludes the edge case $s_{\min} = s_{\max}$ ($\mathcal{K}_\Omega$ acts isometrically on its row space), where all coupled modes are $\pm i\sqrt{a}$, $\gamma = 0$, $c = -1$, and formula (17) gives $\alpha = a/(1 + a)$. This particular α annihilates the coupled pair but maps a primal null-space mode (present iff $|\Omega| > p$) to $p_\alpha(a) = a$, so it does not improve strictly on T^2; the degeneracy is that of this formula, not of the problem. A smaller parameter such as $\alpha = a/(2(1 + a))$ gives $|p_\alpha(\pm i\sqrt{a})| = a/2$, $p_\alpha(a) = a(1+a)/2 < a$, and $p_\alpha(0) = \alpha < a$, hence a strict full-spectrum improvement, and when $|\Omega| = p$ there is no null-space mode and $\alpha = a/(1 + a)$ itself improves strictly. Theorem 3 therefore extends to $\kappa = 1$ with this choice of α. The experimental instance satisfying Assumption 3 (the small OT instance) has $\kappa = 17.3 > 1$; for the disconnected 50×50 instances κ is undefined (Section C.4).

Proof. Under Assumption 2, the Jacobian of the primal prox is aP_Ω and the coupled spectrum is governed by $\mathcal{K}_\Omega$; hence the 2×2 mode matrices are $M(s)$ with (13). With the centering choice (14), $t(s)$ is affine in s^2 and satisfies $t(s_{\min}) = -t(s_{\max})$, giving $\max_i |t(s_i)| = |t(s_{\max})| = \gamma$ as in (15).

Condition (16) implies $|t(s)| < 2\sqrt{a}$ for all modes, hence each has a complex conjugate pair $\lambda_\pm(s) = \sqrt{a}\, e^{\pm i\theta(s)}$. Moreover, $\gamma < \sqrt{2a(1 + a)}$ is equivalent to $\cos(2\theta(s)) < a$ for every coupled mode (using $\cos(2\theta) = t^2/(2a) - 1$). By Theorem 2, there exists $\alpha \in (0, 1)$ with strict spectral-radius improvement.

Since $\cos(2\theta(s)) \le c := \gamma^2/(2a) - 1 < a$ for all modes, the worst-case angle has $\cos(2\theta) = c$. Choosing α as in (17) (the per-mode minimizer at $\cos(2\theta) = c$) yields strict contraction uniformly over all coupled modes. The real null-space modes at a (and the zero modes) are also contracted below a, because this α equals $\alpha^\star(\sqrt{a}, \theta)$ at $\cos 2\theta = c > -1$ and therefore satisfies $\alpha < a/(1 + a)$ (Remark 7, strict since $\kappa > 1$), which gives $p_\alpha(a) = \alpha + (1 - \alpha)a^2 < a$ and $p_\alpha(0) = \alpha < a$. □

Remark 3 (Condition-number interpretation)**.** Let $\kappa := s_{\max}^2/s_{\min}^2 \geq 1$. Then $\gamma = (1+a)(\kappa-1)/(\kappa+1)$ and condition (16) is equivalent to

$$\kappa < \kappa_{\max}(a) := \frac{1+\omega(a)}{1-\omega(a)}, \qquad \omega(a) := \sqrt{\frac{2a}{1+a}} \in (0,1). \tag{18}$$

The threshold $\kappa_{\max}(a)$ increases with a: when $a \approx 1$ (weak regularization, slow convergence), the condition allows a fairly ill-conditioned restricted operator; when a is small, it demands good conditioning. In the regime $a = 1-\delta$ with $\delta \ll 1$, $\kappa_{\max}(a) \approx 8/(1-a) + O(1)$, so the condition can accommodate large κ provided a is sufficiently close to 1.

Remark 4 (Admissible ε under stability stepsizes)**.** With the standard (non-centering) choice $\tau = \sigma = \eta/L$, where $L \geq s_{\max}$ (the experiments use $L = \|\mathcal{K}_r\|$) and $\eta \in (0,1)$, set $h := \eta^2 s_{\min}^2/L^2$ and $b := \eta\varepsilon/L$, so that $a = 1/(1+b)$ and the bottleneck trace is $t(s_{\min}) = 1 + a - ah > 0$. The condition $t(s_{\min}) < \sqrt{2a(1+a)}$ is equivalent to $(2+b-h)^2 < 2(2+b)$, i.e., to

$$0 < b < h - 1 + \sqrt{1+2h} \qquad \Longleftrightarrow \qquad \varepsilon < \frac{L}{\eta}\big(h - 1 + \sqrt{1+2h}\big), \tag{19}$$

which can be checked directly. For small h the right-hand side equals $2\eta s_{\min}^2/L + O(h^2 L/\eta)$; with $L = s_{\max}$, so that $h = \eta^2/\kappa$ with $\kappa = s_{\max}^2/s_{\min}^2$, this is the asymptotic ceiling $\varepsilon \lesssim 2\eta s_{\max}/\kappa$. The relevant κ is that of the reduced active operator, not of the full marginal operator: for full support and $m \geq 2$, $\mathcal{K}_r\mathcal{K}_r^* = \left[\begin{smallmatrix} mI_n & \mathbf{1}_n\mathbf{1}_{m-1}^\top \\ \mathbf{1}_{m-1}\mathbf{1}_n^\top & nI_{m-1} \end{smallmatrix}\right]$ has eigenvalues m (multiplicity $n-1$), n (multiplicity $m-2$), and $\lambda_\pm = \frac{1}{2}\big(n+m \pm \sqrt{(n+m)^2 - 4n}\big)$, so $\kappa = \lambda_+/\lambda_- \approx 4N$ for $n = m = N$ (198.0 at $N = 50$); for $m = 1$, $\mathcal{K}_r = I_n$. Smaller stepsize fractions η admit smaller ε.

Remark 5 (Estimating $s_{\min}, s_{\max}$, and what the centering rule does not tolerate)**.** The centering rule and the ceiling (19) depend on the extreme singular values of $\mathcal{K}_\Omega$. (i) $s_{\max}$ can be estimated by power iteration on the row-space Gram operator $\mathcal{K}_\Omega\mathcal{K}_\Omega^*$ (or `svds`); the number of iterations needed depends on the spectral gap and on the accuracy sought, and we make no general count. (ii) $s_{\min}$ must be the smallest *positive* singular value: since $|\Omega| > p$ in general, $\mathcal{K}_\Omega^*\mathcal{K}_\Omega$ has a nontrivial null space and a smallest-magnitude eigenvalue solver applied to it returns (near) zero; after verifying full row rank (Lemma 2) one should instead take the square root of the smallest eigenvalue of the $p \times p$ operator $\mathcal{K}_\Omega\mathcal{K}_\Omega^*$. (On the small instance of Section 5, where $|\Omega| = 28 > p = 10$, `eigs` with `'smallestabs'` on $\mathcal{K}_\Omega^*\mathcal{K}_\Omega$ returns $-2.5 \cdot 10^{-18}$ with an ill-conditioning warning, whereas the smallest eigenvalue of $\mathcal{K}_\Omega\mathcal{K}_\Omega^*$ gives $s_{\min} = 0.756$.) (iii) Robustness: the centering rule is *not* forgiving of scale errors. With $a = 0.99$, $s_{\min} = 1$, $s_{\max} = 10$, the exact centering product $\tau\sigma = 2(1+a)/(a(s_{\min}^2 + s_{\max}^2)) \approx 0.0398$ satisfies condition (16); underestimating both singular values by 10% leaves κ unchanged but inflates the product to ≈ 0.0491, and the true extreme block then has trace ≈ -2.875 and eigenvalues ≈ -2.475 and -0.400: the iteration is unstable, and no $\alpha \in [0,1)$ repairs a real eigenvalue below -1. If certified bounds $0 < \ell \leq s_{\min}^2 \leq s_{\max}^2 \leq u$ are available, then $\tau\sigma = 2(1+a)/(a(\ell+u))$ together with $[(1+a)(u-\ell)/(u+\ell)]^2 < 2a(1+a)$ certifies $|t(s)| \leq (1+a)(u-\ell)/(u+\ell)$ for every true coupled mode, because $t(s) = (1+a)(1-2s^2/(\ell+u))$; without such bounds the estimates are heuristic, and the selected stepsizes should be checked on the computed spectrum, as we do in Section 5.

C.4 Active-set diagnostics on the experimental instances

For each PDHG-OT instance of Table 1 we computed $z^\star$ by running PAPF to residual 10^{-13}, reading off the active set $\Omega = \{(i,j) : X_{ij} > 10^{-10}\}$ (no entry lies in $(0, 10^{-10}]$ on any instance), and solving the active-set KKT system exactly (the minimum-norm KKT correction of the iterate when $\mathcal{K}_\Omega$ is rank deficient, which keeps $y^\star$ in the dual class the iteration converged to); the result satisfies $\|T(z^\star) - z^\star\| \leq 4 \cdot 10^{-17}$ and lies within $2 \cdot 10^{-11}$ of the iterate. The exact block-triangular Jacobian of Section C was checked against directional differences of the implemented T at $z^\star$ (agreement to $3 \cdot 10^{-7}$ with step 10^{-9}, i.e., rounding level), and its spectrum was computed by dense `eig` ($\dim z = nm+n+m-1 \leq 2599$). Table 2 lists the results; the counts obey $\#\{\lambda{=}1\} = p - \operatorname{rank}\mathcal{K}_\Omega$, $\#\{\lambda{=}a\} = |\Omega| - \operatorname{rank}\mathcal{K}_\Omega$, $\#\{\lambda{=}0\} = nm - |\Omega|$, $\#\{\text{coupled}\} = 2\operatorname{rank}\mathcal{K}_\Omega$, and $\operatorname{rank}\mathcal{K}_\Omega = n + m - \#\{\text{components of } B_\Omega\}$ (Lemma 2) exactly. On the small instance 28 of the 30 entries are active; stationarity on a fully active rectangle forces $C_{ij} + C_{kl} - C_{il} - C_{kj} = -\varepsilon(X_{ij}^\star + X_{kl}^\star - X_{il}^\star - X_{kj}^\star)$, which holds to 10^{-18} on all 114 such rectangles, with $\max|C_{ij} + C_{kl} - C_{il} - C_{kj}| = 2.5 \cdot 10^{-3} <$

$2\varepsilon \min(1/n, 1/m) = 6.7 \cdot 10^{-3}$; this near-additive cost structure is possible because the costs are scaled by 10^{-3} (App. E). For the $n{=}m{=}200$ probe of Section 5.1 (40,399 variables) we found $|\Omega| = 394$, 76 components, $\operatorname{rank} \mathcal{K}_\Omega = 324 < 399$, and $\delta = 7.4 \cdot 10^{-8}$ (smallest active entry $1.7 \cdot 10^{-6}$). Dense `eig` is unnecessary there: the rows of J off Ω vanish, so the spectrum reduces to a $793{\times}793$ block, and an SVD of $\mathcal{K}_\Omega$ splits it into the $2{\times}2$ blocks of Section C. The exact spectrum is 1 ($\times 75 = p - \operatorname{rank} \mathcal{K}_\Omega$; the unit eigenspace $\{(0, \delta y) : \mathcal{K}_\Omega^* \delta y = 0\}$ was built explicitly and J fixes it to 10^{-16}), a ($\times 70$), 0 ($\times 39{,}606$), and 324 conjugate pairs of modulus exactly $\sqrt{a} = 0.99951$ with angles in $[0.0034, 0.136]$ (dense `eig` of the reduced block agrees to $3 \cdot 10^{-14}$). An earlier Arnoldi call on the full J requesting six eigenvalues had returned six eigenvalues from these pairs (three conjugate pairs with angles 0.129, 0.130, 0.136; Ritz residuals $\leq 3 \cdot 10^{-11}$); they are genuine eigenvalues but not the largest ones, since all pairs share one modulus and the unit modes are larger. Because 19 pairs have angles below $\theta_b = \frac{1}{2} \arccos a = 0.022$, the criterion of Theorem 2 fails on them and the minimax over the non-unit spectrum is $\alpha = 0$ ($q(0.02) = 0.99903 > a = 0.99901$): the $1.08\times$ fixed-α gain of Section 5.1 at tolerance 10^{-10} is a finite-horizon effect, and the adaptive fallback to $\alpha = 0$ is spectrally correct. Assumption 1 fails at this fixed point as well.

Table 2: Active-set diagnostics at $z^\star$ ($p = n + m - 1$). δ: strict-complementarity margin of Assumption 2; $s_{\min}$: smallest *positive* singular value of $\mathcal{K}_\Omega$; coupled modes have modulus exactly $\sqrt{a}$, angles in radians; "at a/1/0": multiplicities of the real eigenvalues; $\alpha_G^\star$: minimax of Corollary 1 over the non-unit spectrum (10^{-3} grid) with its rate, a diagnostic where unit modes exist (Table 1 uses 0.058 from this minimax and the full-run grid optima 0.15, 0.27).

Instance	$\lvert\Omega\rvert$	comp.	rank/p	δ	$s_{\min}, s_{\max}$	coupled: #, $\lvert\theta\rvert$	at a / 1 / 0	$\alpha_G^\star$ (rate)
Quad-OT ($6{\times}5$)	28	1	10/10	$1.0 \cdot 10^{-4}$	0.756, 3.14	20, $[0.23, 1.00]$	18 / 0 / 2	0.058 (0.9886)
Quad-OT ($50{\times}50$)	67	41	59/99	$2.5 \cdot 10^{-7}$	0.781, 2.39	118, $[0.08, 0.24]$	8 / 40 / 2433	0.165 (0.9967)
Iris OT ($50{\times}50$)	70	37	63/99	$9.3 \cdot 10^{-7}$	0.475, 2.22	126, $[0.05, 0.22]$	7 / 36 / 2430	0.276 (0.9977)

Consequences. Assumption 3 holds on the small instance only. On the two $50{\times}50$ instances the active-set graph has 41 and 37 connected components, $\operatorname{rank} \mathcal{K}_\Omega = 59$ and $63 < p = 99$, and κ is undefined ($s_{\max}^2 / s_{\min}^2 = 9.3$ and 21.9 on the range of $\mathcal{K}_\Omega$); Theorem 3 is therefore not invoked for them. Their Jacobians have $p - \operatorname{rank} \mathcal{K}_\Omega = 40$ and 36 eigenvalues exactly 1 with eigenvectors $(0, \delta y)$, $\mathcal{K}_\Omega^* \delta y = 0$: the directions of non-unique duals, along which T is locally the identity; $p_\alpha(1) = 1$ for every α, so no filter moves them, and Assumption 1 ($\rho(G) < 1$) does not hold at these fixed points. The remaining spectrum has the same structure as on the small instance, and the minimax of Corollary 1 over it (0.165, 0.276), a diagnostic only (on the full state space $q(\alpha) = 1$ for every α), is within 16 T-calls of the full-run grid optima used in Table 1 (27,900 vs. 27,884; 14,986 vs. 14,970). We report these two rows as empirical results with this consistency check and do not claim that Theorem 2 certifies them. Strict complementarity is also weak there ($\delta \approx 10^{-7}$ vs. 10^{-4} on the small instance); the visible consequence, in Figure 8, is that on Quad-OT ($50{\times}50$) the residual plateaus near 10^{-4} until the active set is identified after roughly 15,000 T-calls, so the probes of Algorithm 1 at T-call 60 see a pre-asymptotic iteration and the acceptance test rejects $\hat{\alpha}$ (the "0" in Table 1), whereas on Iris the rule locks $\hat{\alpha} = 0.38$.

D Power Method Jacobian Analysis

We analyze the Jacobian of the power method operator and derive a necessary and sufficient condition for PAPF acceleration in terms of the matrix spectrum.

D.1 Setup and Jacobian analysis

Let $A \in \mathbb{R}^{n \times n}$ be diagonalizable with eigenvalues $\lambda_1, \ldots, \lambda_n$ ordered so that $\lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_n|$, where $\lambda_1 > 0$ is simple, real, and dominant. The power method operator is

$$T(v) = \frac{Av}{\|Av\|}, \tag{20}$$

The unit dominant eigenvectors $\pm v^\star$ ($Av^\star = \lambda_1 v^\star$, $\|v^\star\| = 1$) are fixed points; so is every unit eigenvector of a positive eigenvalue (for $A = \operatorname{diag}(1, \frac{1}{2})$, $\pm e_2$ are fixed points too), but only $\pm v^\star$

are locally attracting. From a starting vector with a nonzero component along the dominant left eigenvector the power method converges to one of $\pm v^\star$, the sign being selected by that component, and we work locally near the chosen $v^\star$. The local rate of the baseline iteration $v^{k+1} = T(v^k)$ is governed by the gap ratio $\rho := |\lambda_2|/\lambda_1 < 1$; we assume $\lambda_2 \neq 0$, so that the angles below are defined.

Proposition 6 (Jacobian of the power method)**.** *Let A be diagonalizable with eigenvalues $\lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_n|$ and right eigenvectors $p_1 = v^\star, p_2, \ldots, p_n$. The Jacobian of the power method operator at $v^\star$ is*

$$G = J_T(v^\star) = \frac{1}{\lambda_1}\big(I - v^\star v^{\star\top}\big)A, \tag{21}$$

and has eigenvalues $0,\ \lambda_2/\lambda_1,\ \ldots,\ \lambda_n/\lambda_1$. In particular, $\rho(G) = |\lambda_2|/\lambda_1 = \rho$.

Proof. Computing the derivative of $T(v) = Av/\|Av\|$ at $v^\star$:

$$J_T(v^\star) = \frac{1}{\|Av^\star\|}\Big(I - \frac{Av^\star(Av^\star)^\top}{\|Av^\star\|^2}\Big)A = \frac{1}{\lambda_1}\big(I - v^\star v^{\star\top}\big)A,$$

since $Av^\star = \lambda_1 v^\star$ and $\|v^\star\| = 1$. To determine the eigenvalues, let $A = PDP^{-1}$ where $P = [v^\star, p_2, \ldots, p_n]$ and $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Write $c_j := v^{\star\top} p_j$ for the inner products between the dominant eigenvector and the remaining right eigenvectors. A direct computation gives

$$P^{-1}GP = \frac{1}{\lambda_1}\begin{pmatrix} 0 & -c_2\lambda_2 & \cdots & -c_n\lambda_n \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix},$$

which is upper triangular with diagonal entries $0,\ \lambda_2/\lambda_1, \ldots, \lambda_n/\lambda_1$. Since similar matrices share eigenvalues, the claim follows. □

Remark 6. The triangular form above is diagonal if and only if $c_j\lambda_j = 0$ for all $j \geq 2$; when all $\lambda_j \neq 0$ this means $c_j = v^{\star\top} p_j = 0$ for all $j \geq 2$, i.e., the dominant eigenvector $v^\star$ is orthogonal to the invariant subspace $\operatorname{span}\{p_2, \ldots, p_n\}$. This is *strictly weaker* than A being normal: it places no orthogonality constraint among the subdominant eigenvectors $p_2, \ldots, p_n$, which can still be highly non-orthogonal to one another. Normality of A ($A^\top A = AA^\top$) is a sufficient condition (it makes the entire eigenbasis orthogonal) but not a necessary one. For general (possibly non-normal) A, G may be non-normal and its eigenvectors need not coincide with those of A (though it can also be diagonal at $v^\star$ for a non-normal A, e.g. $A = \left[\begin{smallmatrix} 1 & b \\ 0 & 0 \end{smallmatrix}\right]$ gives $G = 0$); the *eigenvalues* are always λ_j/λ_1 regardless of the eigenvector geometry.

D.2 Two filtered iterations: on the sphere and in the ambient space

Because T itself normalizes, T^2 maps unit vectors to unit vectors, but the filtered iterate $\mathsf{S}_\alpha(v) = \alpha v + (1-\alpha)T^2(v)$ does not. Two filtered algorithms must therefore be distinguished:

$$\mathsf{S}_\alpha(v) = \alpha v + (1-\alpha)T^2(v) \quad \text{(ambient)}, \qquad \widetilde{\mathsf{S}}_\alpha(v) = N\big(\mathsf{S}_\alpha(v)\big), \quad N(w) = w/\|w\| \quad \text{(sphere)}. \tag{22}$$

All power-method experiments in this paper use the sphere algorithm $\widetilde{\mathsf{S}}_\alpha$: every update, including burn-in steps and probes of Algorithm 1, is followed by renormalization (checked in the code for T^2, relaxed KM, PAPF, the adaptive rule, and TKMA on the four power maps of Table 1; the numerical Jacobian of the implemented T at $v^\star$ agrees with $G = P_\star A/\lambda_1$ to $2 \cdot 10^{-10}$ there). On those four instances the tangent-space and ambient minimizers coincide ($\alpha = 0.378, 0.487, 0.487, 0.490$ on a 10^{-3} grid) because the radial filtered value α stays below the tangent maximum ($0.534, 0.490, 0.490, 0.700$); the example below shows that they can differ, and we reproduced it in code: for $\rho = 0.9$ the sphere algorithm at $\alpha = \rho^2/(1+\rho^2)$ converges superlinearly ($8\ T$-calls to 10^{-12}), whereas without renormalization the same α converges at the radial rate $\alpha = 0.4475$ per step and the ambient optimum $\rho^2/(2+\rho^2)$ at 0.288 (70 vs. 48 T-calls). Both maps fix $v^\star$; since $\mathsf{S}_\alpha(v^\star) = v^\star \neq 0$, N is smooth near $\mathsf{S}_\alpha(v^\star)$ and $\widetilde{\mathsf{S}}_\alpha$ is well defined and C^1 on a neighborhood of $v^\star$ (we make no claim about a global domain of definition). Let $P_\star := I - v^\star v^{\star\top}$ and $H := (v^\star)^\perp = T_{v^\star}\mathbb{S}^{n-1}$. By Proposition 6,

$G = P_\star A/\lambda_1$ maps $\mathbb{R}^n$ into H, $Gv^\star = 0$, and H is G-invariant; the eigenvalues of $G|_H$ are λ_j/λ_1, $j \geq 2$, and the eigenvalue 0 of G with eigenvector $v^\star$ is the *radial* mode. Since $DN(v^\star) = P_\star$,

$$D\widetilde{\mathsf{S}}_\alpha(v^\star) = P_\star\big[\alpha I + (1-\alpha)G^2\big], \qquad D\widetilde{\mathsf{S}}_\alpha(v^\star)\big|_H = \alpha I_H + (1-\alpha)(G|_H)^2, \qquad D\widetilde{\mathsf{S}}_\alpha(v^\star)v^\star = 0, \tag{23}$$

because $(\alpha I + (1-\alpha)G^2)h \in H$ for $h \in H$ and $G^2 v^\star = 0$. Hence the linearized factors of the two algorithms are

$$q_{\mathrm{sph}}(\alpha) = \max_{j\geq 2}\big|\alpha + (1-\alpha)(\lambda_j/\lambda_1)^2\big|, \qquad q_{\mathrm{amb}}(\alpha) = \max\big\{\alpha,\ q_{\mathrm{sph}}(\alpha)\big\}, \tag{24}$$

the ambient algorithm retaining the radial eigenvalue $\alpha + (1-\alpha)\cdot 0^2 = \alpha$ of $D\mathsf{S}_\alpha(v^\star) = \alpha I + (1-\alpha)G^2$. The radial mode can bind: for $A = \mathrm{diag}(1, \rho R(\pi/2))$ the single-pair parameter $\alpha_{\mathrm{pair}} = \rho^2/(1+\rho^2)$ gives $q_{\mathrm{sph}}(\alpha_{\mathrm{pair}}) = 0$ (the tangent pair is annihilated), whereas the ambient algorithm has $q_{\mathrm{amb}}(\alpha_{\mathrm{pair}}) = \alpha_{\mathrm{pair}} > 0$ and its optimum is $\alpha = \rho^2/(2+\rho^2)$ with $q_{\mathrm{amb}} = \rho^2/(2+\rho^2)$. A zero linearized factor does not mean that the nonlinear iteration terminates in one step.

D.3 PAPF ACCELERATION THEOREM

Since the eigenvalues of $G|_H$ are the ratios $\mu_j = \lambda_j/\lambda_1$, $j \geq 2$, the set of dominant modes is $\mathcal{D} = \{j \geq 2 : |\lambda_j| = |\lambda_2|\}$, and the angle of each dominant mode equals the argument of the corresponding eigenvalue of A: $\theta_j = \arg(\lambda_j)$ (since $\lambda_1 > 0$ is real). Theorem 2 therefore specializes as follows.

Theorem 4 (PAPF for the power method). *Let $A \in \mathbb{R}^{n\times n}$ be diagonalizable with $\lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_n|$, $\lambda_1 > 0$, $\lambda_2 \neq 0$, and let $\rho := |\lambda_2|/\lambda_1$. Statements (a)–(c) concern the linearized factor $q_{\mathrm{sph}}(\alpha)$ of the sphere algorithm $\widetilde{\mathsf{S}}_\alpha$ at $v^\star$, compared with $\rho(G)^2 = \rho^2$ for T^2.*

(a) *$q_{\mathrm{sph}}(\alpha) < \rho^2$ for some $\alpha \in (0,1)$ if and only if*

$$\cos\big(2\arg(\lambda_j)\big) < \rho^2 \qquad \text{for every } j \in \mathcal{D}. \tag{25}$$

(b) *If all eigenvalues at the subdominant modulus $|\lambda_2|$ are real, then no $\alpha \in (0,1)$ improves the linearized factor over T^2. In particular, within this family PAPF cannot improve the local factor of the power method for any symmetric (or, more generally, real-spectrum) matrix.*

(c) *If the subdominant eigenvalues form a unique complex conjugate pair $\lambda_{2,3} = r_2\, e^{\pm i\phi}$ with $\phi \in (0,\pi)$ and $\cos(2\phi) < \rho^2$, and all other eigenvalues satisfy $|\lambda_j| \leq r_{\mathrm{sub}} < r_2$, then the per-mode optimal mixing parameter and factor ratio are*

$$\alpha^\star = \frac{\rho^2(\rho^2 - \cos 2\phi)}{|1 - \rho^2 e^{2i\phi}|^2}, \qquad \frac{|\sigma^\star|}{\rho^2} = \frac{|\sin 2\phi|}{|1 - \rho^2 e^{2i\phi}|} < 1, \tag{26}$$

and these are the full-spectrum optima of the sphere algorithm (i.e., $q_{\mathrm{sph}}(\alpha^\star) = |\sigma^\star|$ and $\alpha^\star = \alpha^\star_G$) whenever the subdominant gap condition (31) *of Corollary 2 holds for the tangent spectrum $\{\lambda_j/\lambda_1 : j \geq 4\}$. A simple closed-form sufficient condition is $r_{\mathrm{sub}}^2/\lambda_1^2 < |\sigma^\star|$ together with $\alpha^\star \leq (|\sigma^\star| - r_{\mathrm{sub}}^2/\lambda_1^2)/(1 - r_{\mathrm{sub}}^2/\lambda_1^2)$. As $\rho \to 1^-$ (eigenvalue gap closing), the single-mode factor ratio converges to $|\cos\phi|$.*

(d) *For the ambient algorithm S_α the criterion of (a) is unchanged, but the radial eigenvalue α enters the minimax: $q_{\mathrm{amb}}(\alpha) = \max\{\alpha, q_{\mathrm{sph}}(\alpha)\}$, and the closed form of (c) is the ambient optimum only if additionally $\alpha^\star \leq |\sigma^\star|$.*

Proof. By (23), $q_{\mathrm{sph}}(\alpha) = \rho(\alpha I_H + (1-\alpha)(G|_H)^2)$ and $\rho(G|_H) = \rho(G) = \rho$, so Theorem 2, whose proof uses only the spectrum of the linearization, applies to $G|_H$ with eigenvalues $\mu_j = \lambda_j/\lambda_1$, $j \geq 2$, and gives (a); the local R-linear convergence of $\widetilde{\mathsf{S}}_\alpha$ near $v^\star$ at any rate above $q_{\mathrm{sph}}(\alpha)$ follows from the general local theory applied to $\widetilde{\mathsf{S}}_\alpha$ as a C^1 map on $\mathbb{R}^n$ with $\rho(D\widetilde{\mathsf{S}}_\alpha(v^\star)) = q_{\mathrm{sph}}(\alpha)$. For part (b), if $\lambda_j \in \mathbb{R}$ then $\arg(\lambda_j) \in \{0,\pi\}$, giving $\cos(2\arg(\lambda_j)) = 1 > \rho^2$, so condition (25) fails. Part (c) follows from Corollary 2 applied to $G|_H$ and from Corollary 3. For part (d), $D\mathsf{S}_\alpha(v^\star) = \alpha I + (1-\alpha)G^2$ has the additional eigenvalue α on the radial eigenvector $v^\star$ (Proposition 6), which yields (24); the closed form of (c) minimizes q_{amb} iff the radial value $\alpha^\star$ does not exceed $|\sigma^\star|$. □

Classification of matrices. Theorem 4 yields a simple diagnostic for whether PAPF can improve the local factor of the power method for a given matrix, with $\theta_b := \frac{1}{2}\arccos\rho^2$ the boundary angle of the acceleration region (modes near π are relaxed KM's regime, Proposition 3):

Matrix type	Subdominant spectrum	PAPF outcome
Symmetric / Hermitian	All real	No acceleration
Real spectrum (e.g., triangular)	All real	No acceleration
Non-symmetric, real λ_2	Real at subdominant modulus	No acceleration
Non-sym., complex λ_2, $\phi \notin (\theta_b, \pi - \theta_b)$	outside the region	No improvement
Non-sym., complex λ_2, ϕ just inside	near the boundary	Small improvement
Non-sym., complex λ_2, ϕ near $\pi/2$	ϕ near $\pi/2$	Conditional (see note)

The last row is conditional: every dominant mode must satisfy condition (ii) of Theorem 2, and the single-pair value becomes the full-spectrum factor only under the gap condition of Theorem 4(c). For $A = \mathrm{diag}(1, 0.99R(\pi/2), 0.9899)$ the tangent-space minimax is $\alpha_G^\star = (r-t)/(2+r-t) = 9.9{\cdot}10^{-5}$ with $r = 0.99^2$, $t = 0.9899^2$, and $q^\star/q(0) = 0.9998$: the pair could be annihilated, but the real mode 0.9899 leaves almost no full-spectrum gain.

D.4 Applications and examples

The power method with complex subdominant eigenvalues arises naturally in several settings:

1. **Non-reversible Markov chains.** The stationary distribution of a Markov chain with transition matrix P can be computed via the power method on $P^\top$. The dominant eigenvalue is $\lambda_1 = 1$ (Perron root), and $\rho = |\lambda_2|$ is the subdominant modulus, so the baseline convergence rate is $|\lambda_2|$. For *reversible* chains (satisfying detailed balance), P is self-adjoint with respect to the stationary measure, so all eigenvalues are real and PAPF offers no improvement. For *non-reversible* chains—such as directed random walks on cyclic graphs, asymmetric queueing networks, or MCMC with non-reversible proposals—the subdominant eigenvalues can be complex, and PAPF improves the local factor when $\cos(2\phi) < |\lambda_2|^2$ where $\phi = \arg(\lambda_2)$, i.e., when $\frac{1}{2}\arccos|\lambda_2|^2 < \phi < \pi - \frac{1}{2}\arccos|\lambda_2|^2$. The Euclidean normalization of (20) produces a unit vector; the stationary distribution is obtained by fixing the sign and renormalizing to unit sum, and intermediate iterates are not probability vectors.
2. **PageRank computation.** The Google matrix $M = d\,S + (1-d)\,\frac{1}{n}\mathbf{1}\mathbf{1}^\top$ (with damping factor $d \approx 0.85$ and column-stochastic link matrix S) has $\lambda_1 = 1$ and all other eigenvalues satisfy $|\lambda_j| \le d$. For web graphs with cyclic link structure the subdominant eigenvalues can be complex. When the subdominant pair has modulus $|\lambda_2| = d$, the PAPF condition becomes $\cos(2\phi) < d^2$, i.e., $\frac{1}{2}\arccos(d^2) < \phi < \pi - \frac{1}{2}\arccos(d^2)$ ($21.9^\circ < \phi < 158.1^\circ$ for $d = 0.85$; for $|\lambda_2| < d$, replace d by $|\lambda_2|$); with several dominant modes each must satisfy it.
3. **Non-self-adjoint operators.** Discretizations of convection-diffusion operators, linearized Navier–Stokes equations, and stability operators for dynamical systems yield non-symmetric matrices whose subdominant eigenvalues can be complex. Our experiment for this setting uses a synthetic nonnormal matrix with a prescribed spectrum (App. E), not a PDE discretization.

Section 5.2 tests instances of the first two settings and the synthetic surrogate of the third.

E Reproducibility Details

All experiments use MATLAB R2024a on a standard laptop (Intel Core i9-9880H at 2.3 GHz, 16 GB RAM, macOS); each experiment completes in under 5 minutes. Random seeds are fixed: seed 42 for power method, Markov chain, PageRank, the nonnormal synthetic matrix, and OT experiments; seed 7 for the complex-dominant operator. Convergence tolerance is 10^{-12} on the residual for the synthetic and OT instances, except the small OT instance, which uses $\|z - z^\star\| \le 10^{-14}$, and 10^{-14}

for the power-method family; each method runs until tolerance or until its budget is exhausted: per-instance iteration budgets for the AA, Halpern, and inertial runners and the instance step budget for TKMA (listed in App. G.1); for capped runs the tables report the budget in T-calls, marked †. Empirical rates are estimated by log-linear regression on the last 50–200 residual values. The adaptive search uses $K_{\text{burn}} = 30$, $K_{\text{probe}} = 5$, $\delta = 0.005$, and the acceptance margin $\delta_{\text{acc}} = 0.02$ for every Table 1 run; Section 5.4 and App. G.6 vary these settings deliberately and say so. The tail rates 0.475 on the invariant parabola and 0.867 off it stated in Remark 13 are both measured under the residual $r_k = \|T(z^k) - z^k\|$.

E.1 Instance generation parameters

We list the explicit construction of every operator/instance used in Section 5, so that all results can be regenerated from scratch.

Toy linear ($\mathbb{R}^3$). $T(z) = Az$ with $A = \text{blkdiag}\big(r\,R(\theta),\, \rho_{\text{slow}}\big)$, $r = 0.98$, $\theta = \pi/2$, $\rho_{\text{slow}} = 0.8$, where $R(\theta)$ is the 2×2 rotation by θ. Initial vector $z^0 = (1, 1, 1)^\top/\sqrt{3}$ (deterministic).

Complex-dominant ($\mathbb{R}^{50}$). $A = \text{blkdiag}\big(r_j\,R(\pi/2)\big)_{j=1}^{25}$ with moduli r_j uniformly spaced in $[0.70, 0.99]$ (so all 25 conjugate pairs sit at angle $\pi/2$ with distinct radii). Initial state drawn from a standard Gaussian under `rng(7,'twister')` and normalized.

Generic power method ($\mathbb{R}^{8\times8}$). $A = PD_{\text{real}}P^{-1}$ with

$$D_{\text{real}} = \text{blkdiag}\big(1,\ 0.99\,R(\pi/3),\ \text{diag}(\lambda_4, \ldots, \lambda_8)\big), \qquad \lambda_4, \ldots, \lambda_8 \text{ equispaced in } [0.10, 0.50].$$

Here D_{real} is the real Schur (block-diagonal) form encoding the prescribed eigenvalues $\{1,\ 0.99\,e^{\pm i\pi/3},\ \lambda_4, \ldots, \lambda_8\}$; the complex pair appears as a 2×2 rotation block, not a scalar diagonal entry. The similarity transform P is drawn from a standard Gaussian under `rng(42,'twister')` and resampled until $|\det P| \geq 0.1$ (the 30×30 synthetic instance below uses $\text{cond}(P) \leq 100$). Initial vector: standard Gaussian unit vector under the same seed.

Non-reversible 3-state Markov chain ($n = 3$). Cyclic permutation C on three states ($Ce_1 = e_2$, $Ce_2 = e_3$, $Ce_3 = e_1$), Google-style mixture $P = (1-d)/n \cdot \mathbf{1}\mathbf{1}^\top + d\,C$ with damping $d = 0.98$. We run the power method on $A = P^\top$ (left-eigenvector form); its eigenvalues are $\{1,\ d\,e^{\pm 2\pi i/3}\}$, so $\rho = d = 0.98$ and $\theta = 120°$. Initial state $v_0 = |N(0, I_3)|/\|\cdot\|$ under `rng(42,'twister')`.

Clustered PageRank graph ($n = 15$). A directed ring of three clusters of 5 nodes each: every node in cluster $c \in \{0, 1, 2\}$ has an outgoing edge to every node in cluster $(c + 1) \bmod 3$ (so each node has exactly 5 out-links and the link matrix is column-stochastic with full block structure). Google matrix $M = d\,S + (1 - d)/n \cdot \mathbf{1}\mathbf{1}^\top$ with *damping factor* $d = 0.98$; the dominant non-Perron pair sits at $d\,e^{\pm 2\pi i/3}$, giving $\rho = 0.98$ and $\theta = 120°$. Initial state $v_0 = N(0, I_{15})/\|\cdot\|$ under `rng(42,'twister')`. We use $d = 0.98$ rather than the commonly cited web-search value $d \approx 0.85$ of Appendix D, item 2 in order to push the dominant modulus close to 1 and make the baseline iteration challenging; PAPF accelerates for any d satisfying $\cos(2\theta) < d^2$, which holds for every $d > 0$ at $\theta = 120°$ since $\cos 2\theta = -\frac{1}{2}$.

Nonnormal synthetic matrix with $\theta = \pi/4$ ($\mathbb{R}^{30\times30}$; a surrogate for convection–diffusion-type spectra, not a PDE discretization). $A = PD_{\text{real}}P^{-1}$ with $D_{\text{real}} = \text{blkdiag}\big(1,\ 0.99\,R(\pi/4),\ \text{diag}(\lambda_4, \ldots, \lambda_{30})\big)$ and $\lambda_4, \ldots, \lambda_{30}$ equispaced in $[0.10, 0.60]$. Again D_{real} is a real Schur (block-diagonal) form so that the complex pair $0.99\,e^{\pm i\pi/4}$ is realized as a 2×2 rotation block. P is drawn from $N(0, I)$ under `rng(42,'twister')` and resampled until $\text{cond}(P) \leq 100$.

Quad-OT (small). $n = 6$, $m = 5$; source points $x_i \sim \frac{1}{2}N(0, I_2)$ and target points $y_j \sim \frac{1}{2}N(0, I_2) + (1, 1)$ under `rng(42,'twister')`; cost $C_{ij} = 10^{-3}\|x_i - y_j\|_2^2$ (entries in $[1.6 \cdot 10^{-4}, 7.4 \cdot 10^{-3}]$); uniform marginals $\mathbf{a} = \mathbf{1}/6$, $\mathbf{b} = \mathbf{1}/5$; $\varepsilon = 0.02$; stability stepsizes $\tau = \sigma = 0.99/s_{\max}(\mathcal{K}_r) = 0.3066$ ($a = 0.9939$). Warm start $z^0 = z^\star + 10^{-6}\xi$ with $\xi \sim N(0, I)$

under `rng(7)`, where $z^\star$ is the active-set KKT solution (App. C.4); this is why the instance stops on $\|z - z^\star\| \le 10^{-14}$ rather than on the residual.

Quad-OT (large) and Iris OT. Quad-OT (large): $n = m = 50$ source points $x_i \sim N(0, I_2)$ and target points $y_j \sim N(0, I_2) + (3, 3)$ under `rng(42,'twister')`; squared-Euclidean cost normalized to unit maximum, $C_{ij} = \|x_i - y_j\|_2^2 / \max_{kl} \|x_k - y_l\|_2^2$; uniform marginals $\mathbf{a} = \mathbf{b} = \mathbf{1}/n$; $\varepsilon = 0.02$. Iris OT: all 50 *versicolor* samples (source) and all 50 *virginica* samples (target) of the Fisher Iris dataset (Fisher, 1936; Anderson, 1935) (the converged plan is shown in Fig. 4); each sample is the 4-vector of sepal length, sepal width, petal length, petal width, z-scored across the pooled 100 samples; cost $C_{ij} = \|x_i - y_j\|_2^2 / \max_{kl} \|x_k - y_l\|_2^2$ in standardized units; $\mathbf{a} = \mathbf{b} = \mathbf{1}/n$; $\varepsilon = 0.02$. Both instances use the stability stepsizes $\tau = \sigma = 0.99/s_{\max}(\mathcal{K}_r)$ and the cold start $X^0 = \mathbf{1}/(nm)$, $y^0 = 0$ (the uniform plan, which is feasible).

Adaptive-search hyperparameters. All Table 1 adaptive runs use the defaults $K_{\text{burn}} = 30$, $K_{\text{probe}} = 5$, $\delta = 0.005$, $\delta_{\text{acc}} = 0.02$, with power-method runs additionally renormalizing the iterate to the unit sphere after each step (`renorm=true`).

The nine-instance summary (spectra, best fixed parameters, and T-call counts) is Table 1 in the main text. Its "best fixed α" is the minimax optimum of Corollary 1 over the full computed spectrum for the small OT, power-method, and synthetic instances, and the full-run grid optimum (0.01 step) for the two 50×50 OT instances, whose spectra were computed afterwards and give 0.165 and 0.276 (App. C.4); it is not the single-pair closed form (4), which ignores subdominant modes; the two differ whenever a subdominant mode becomes dominant after filtering (e.g., $\alpha = 0.14$ rather than 0.49 on the toy instance).

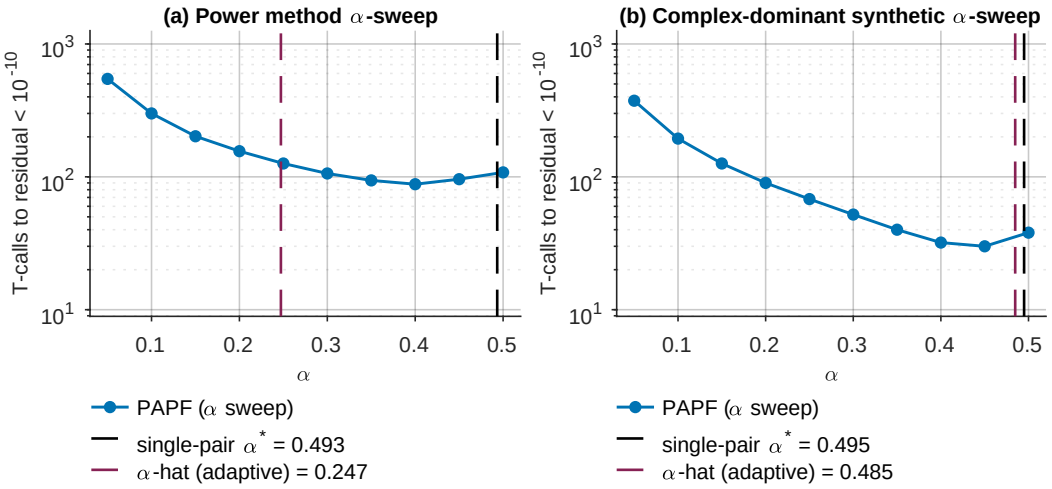

Figure 3: **α-sweep, T-calls to tolerance 10^{-10}.** (a) Power method ($\rho = 0.99$, $\theta = \pi/3$): the empirical minimum sits near $\alpha \approx 0.4$, consistent with the minimax optimum 0.38 of Table 1, between the adaptive selection $\hat{\alpha} = 0.247$ (red dashed) and the single-pair closed form (4), $\alpha = 0.493$ (black dashed); the landscape is shallow on the right half, so the adaptive choice still beats T^2 by nearly $10\times$. (b) Complex-dominant synthetic ($d = 50$, $\theta = \pi/2$, $r_j \in [0.70, 0.99]$): a descent to a minimum near $\alpha \approx 0.45$ (minimax optimum 0.42), where the $\theta = \pi/2$ structure enables near-perfect annihilation of the dominant pair; the adaptive selection lands nearby.

F Proofs of Main Results

Proof of Proposition 1. **Parts 1–2.** Since T is FNE ($\frac{1}{2}$-averaged), the composition rule for averaged operators gives that T^2 is $\frac{2}{3}$-averaged (hence nonexpansive). Then $\mathsf{S}_\alpha = \alpha I + (1-\alpha)T^2$ is a convex combination of the (vacuously) 0-averaged identity and the $\frac{2}{3}$-averaged T^2, so the convex-combination rule yields averagedness constant $\alpha \cdot 0 + (1-\alpha) \cdot \frac{2}{3} = \frac{2(1-\alpha)}{3} \in (0, \frac{2}{3})$. **Part 3:**

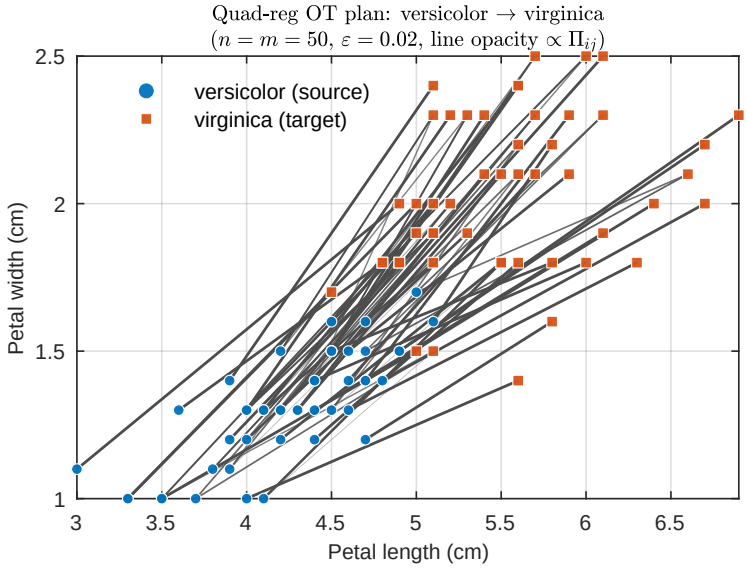

Figure 4: Converged quadratically regularized OT plan (versicolor → virginica), projected onto petal features. Each line connects a source sample x_i to a target sample y_j with $nX_{ij} \geq 0.10$; line opacity and width scale with X_{ij} (the figure's Π_{ij} legend label corresponds to X_{ij} in the main text). The plan concentrates mass on nearby targets in the overlap region; a few extreme virginica samples (upper right) receive only below-threshold mass and therefore have no incoming line.

$\operatorname{Fix} \mathsf{S}_\alpha = \operatorname{Fix} T$. ($\supseteq$) is immediate: $Tx = x$ implies $T^2x = x$ and $\mathsf{S}_\alpha x = x$. ($\subseteq$) If $\mathsf{S}_\alpha x = x$ then $T^2x = x$. Let $y = Tx$; then $Ty = T^2x = x$. The FNE inequality with $u = x, v = y$ gives $\|y - x\|^2 = \|Tx - Ty\|^2 \leq \langle Tx - Ty, x - y\rangle = -\|y - x\|^2$, so $y = x$, i.e. $Tx = x$. □

Proof of Lemma 1. PAPF produces the eigenvalue $\sigma(\alpha) = \alpha + (1-\alpha)\,\rho^2\,e^{2i\theta}$. A direct computation gives

$$|\sigma|^2 = \alpha^2 + 2\alpha(1-\alpha)\,\rho^2\cos 2\theta + (1-\alpha)^2\,\rho^4, \tag{27}$$

which factors as

$$|\sigma|^2 - \rho^4 \;=\; \alpha\Big[\,\alpha\,\underbrace{\big|1-\rho^2 e^{2i\theta}\big|^2}_{=:\,A\,>\,0} + \underbrace{2\rho^2\big(\cos 2\theta - \rho^2\big)}_{=:\,B}\Big]. \tag{28}$$

Part (a). For $\alpha > 0$, $|\sigma|^2 < \rho^4$ iff $\alpha A + B < 0$. If $\cos 2\theta \geq \rho^2$ then $B \geq 0$, so $\alpha A + B > 0$ for all $\alpha > 0$ and PAPF cannot beat T^2. If $\cos 2\theta < \rho^2$ then $B < 0$, and $\alpha A + B < 0$ for $\alpha < -B/A > 0$.

Part (b). From (27), $f(\alpha) := |\sigma(\alpha)|^2$ is a quadratic in α with positive leading coefficient $A > 0$. Setting $f'(\alpha) = 0$ gives $\alpha^\star = -B/(2A) = \rho^2(\rho^2 - \cos 2\theta)/|1-\rho^2 e^{2i\theta}|^2$. Since $\cos 2\theta < \rho^2$, we have $\alpha^\star > 0$. To verify $\alpha^\star < 1$: expanding $|1 - \rho^2 e^{2i\theta}|^2 = 1 - 2\rho^2\cos 2\theta + \rho^4$, the inequality $\rho^2(\rho^2 - \cos 2\theta) < |1 - \rho^2 e^{2i\theta}|^2$ simplifies to $0 < |1 - \rho^2 e^{2i\theta}|^2 - \rho^2(\rho^2 - \cos 2\theta) = 1 - \rho^2\cos 2\theta$, which holds since $\rho^2\cos 2\theta \leq \rho^2 < 1$. From part (a), $|\sigma|^2 < \rho^4$ iff $\alpha < -B/A = 2\alpha^\star$.

Part (c). Substituting $\alpha^\star$ into f: $f(\alpha^\star) = \rho^4 - B^2/(4A) = \rho^4\big(1 - (\rho^2 - \cos 2\theta)^2/|1-\rho^2 e^{2i\theta}|^2\big)$. The numerator simplifies via $|1-\rho^2 e^{2i\theta}|^2 - (\rho^2 - \cos 2\theta)^2 = \sin^2 2\theta$, giving $|\sigma^\star| = \rho^2|\sin 2\theta|/|1-\rho^2 e^{2i\theta}|$. Since $\sin^2 2\theta = |1-\rho^2 e^{2i\theta}|^2 - (\rho^2 - \cos 2\theta)^2$ and $\cos 2\theta \neq \rho^2$ by hypothesis, $|\sin 2\theta| < |1-\rho^2 e^{2i\theta}|$, so the ratio is strictly less than 1. □

Proof of Theorem 2. **(ii) ⇒ (i).** By Lemma 1(b), each dominant mode $j \in \mathcal{D}$ satisfies $|\alpha + (1-\alpha)\mu_j^2| < \rho^2$ for $\alpha < 2\,\alpha^\star(\rho, \theta_j)$. For subdominant modes ($r_j < \rho$), the triangle inequality gives $|\alpha + (1-\alpha)\mu_j^2| \leq \alpha + (1-\alpha)r_j^2 < \rho^2$ for $\alpha < (\rho^2 - r_j^2)/(1 - r_j^2)$. Any α smaller than $\min\big(\min_{j\in\mathcal{D}} 2\,\alpha^\star(\rho, \theta_j),\, \min_{j\notin\mathcal{D}}(\rho^2 - r_j^2)/(1 - r_j^2)\big)$ satisfies $\rho(\alpha I + (1-\alpha)G^2) < \rho^2$.

¬(ii) ⇒ ¬(i). If some $j_0 \in \mathcal{D}$ has $\cos(2\theta_{j_0}) \geq \rho^2$, then Lemma 1(a) gives $|\alpha + (1-\alpha)\mu_{j_0}^2| \geq \rho^2$ for all $\alpha > 0$, so $\rho(\alpha I + (1-\alpha)G^2) \geq \rho^2$. □

Corollary 1 (Optimal mixing parameter and acceleration ratio)**.** *Under the equivalent conditions of Theorem 2, the optimal mixing parameter for the full operator is*

$$\alpha_G^\star \;=\; \arg\min_{\alpha\in(0,1)} \; \max_j \; |\alpha + (1-\alpha)\mu_j^2|, \tag{29}$$

which is a one-dimensional convex problem (Proposition 5); a minimizer exists on the compact interval $[0,1]$*, and since* $\varphi(0) = \rho^2$*,* $\varphi(1) = 1$*, and* $\varphi < \rho^2$ *near* 0^+*, every minimizer lies in* $(0,1)$*. The PAPF acceleration ratio for the full operator is*

$$\frac{\rho\big(\alpha_G^\star I + (1-\alpha_G^\star)G^2\big)}{\rho^2} \;<\; 1. \tag{30}$$

Moreover, PAPF beats T^2 *for all* α *in an interval* $(0, \bar\alpha)$ *with* $\bar\alpha > 0$*.*

Remark 7 (Bounds on $\alpha^\star$)**.** The single-mode optimum satisfies $0 < \alpha^\star(\rho,\theta) \leq \rho^2/(1+\rho^2) \leq \frac{1}{2}$ for all θ with $\cos 2\theta < \rho^2$ (where it is defined), with equality in the first upper bound iff $\theta = \pi/2$, so the optimal mixing parameter is always moderate ($\leq \frac{1}{2}$).

Lemma 3 (Full-spectrum search interval and uniqueness)**.** *Let* $\varphi(\alpha) = \max_j |\alpha + (1-\alpha)\mu_j^2|$ *with* $|\mu_j| \leq \rho < 1$*. Each* $f_j(\alpha) := |\alpha + (1-\alpha)\mu_j^2|^2 = |1-\mu_j^2|^2\alpha^2 + 2(\mathrm{Re}\,\mu_j^2 - |\mu_j|^4)\alpha + |\mu_j|^4$ *is a strictly convex quadratic whose unconstrained minimizer* $\alpha_j = (|\mu_j|^4 - \mathrm{Re}\,\mu_j^2)/|1-\mu_j^2|^2$ *satisfies* $\alpha_j \leq |\mu_j|^2/(1+|\mu_j|^2) \leq \rho^2/(1+\rho^2)$*. Hence every* f_j*, and therefore* φ*, is nondecreasing on* $[\rho^2/(1+\rho^2), 1]$*, and every minimizer of* φ *over* $[0,1]$ *lies in* $[0, \rho^2/(1+\rho^2)] \subset [0, \frac{1}{2}]$*. Moreover* $\varphi^2 = \max_j f_j$ *is strongly convex with modulus* $2\min_j |1-\mu_j^2|^2 > 0$*, so the minimizer* $\alpha_G^\star$ *is unique. These statements concern the exact spectral objective, not the finite-probe objective of Algorithm 1.*

Proof. Expanding $|\alpha(1-\mu_j^2) + \mu_j^2|^2$ gives the quadratic, whose minimizer is α_j; writing $\mu_j = re^{i\theta}$ and $c = \cos 2\theta$, $\alpha_j = r^2(r^2 - c)/(1 - 2r^2c + r^4)$, which is decreasing in c (its derivative in c has the sign of $r^4 - 1 < 0$) and hence at most its value $r^2/(1+r^2)$ at $c = -1$. The maximum of finitely many quadratics with a common strong-convexity modulus is strongly convex with that modulus. □

Proof of Corollary 1. The function $\varphi(\alpha) := \max_j |\alpha + (1-\alpha)\mu_j^2|$ is a pointwise maximum of convex functions of α, hence convex. We show $\varphi(\alpha) < \rho^2$ for all sufficiently small $\alpha > 0$. For each dominant mode $j \in \mathcal{D}$, Lemma 1(b) gives $|\alpha + (1-\alpha)\mu_j^2| < \rho^2$ for $\alpha < 2\,\alpha^\star(\rho,\theta_j)$. For each subdominant mode ($r_j < \rho$), $|\alpha + (1-\alpha)\mu_j^2| \leq \alpha + (1-\alpha)r_j^2 < \rho^2$ for $\alpha < (\rho^2 - r_j^2)/(1-r_j^2)$. Hence $\varphi(\alpha) < \rho^2$ for $\alpha \in (0, \alpha_0)$ with

$$\alpha_0 \;=\; \min\Big(\min_{j\in\mathcal{D}} 2\,\alpha^\star(\rho,\theta_j),\;\; \min_{j\notin\mathcal{D}} \frac{\rho^2 - r_j^2}{1-r_j^2}\Big) \;>\; 0.$$

Since φ is convex with $\varphi(0) = \rho^2$ and $\varphi(\alpha) < \rho^2$ on $(0, \alpha_0)$, the minimizer $\alpha_G^\star \in (0,1)$ achieves $\varphi(\alpha_G^\star) < \rho^2$, establishing (30). The interval of improvement is $\bar\alpha = \sup\{\alpha \in (0,1) : \varphi(\alpha) < \rho^2\} \geq \alpha_0 > 0$. □

Corollary 2 (Unique dominant pair)**.** *Suppose* G *has a unique dominant conjugate pair* $\mu, \bar\mu = \rho\, e^{\pm i\theta}$ *(*$\theta \in (0,\pi)$*) satisfying condition* (ii) *of Theorem 2, with all other eigenvalues satisfying* $|\mu_j| \leq r_{\mathrm{sub}} < \rho$*, and let* $|\sigma^\star| = \rho^2|\sin 2\theta|/|1 - \rho^2 e^{2i\theta}|$ *denote the single-mode optimum from Lemma 1(c). If*

$$\big|\alpha^\star(\rho,\theta) + (1-\alpha^\star(\rho,\theta))\,\mu_j^2\big| \;\leq\; |\sigma^\star| \qquad \textit{for every subdominant eigenvalue } \mu_j, \tag{31}$$

then $\alpha_G^\star = \alpha^\star(\rho,\theta)$ *and the PAPF spectral radius and acceleration ratio equal the single-mode optima:*

$$\rho\big(\alpha_G^\star I + (1-\alpha_G^\star)G^2\big) \;=\; \frac{\rho^2\,|\sin 2\theta|}{|1-\rho^2 e^{2i\theta}|}, \qquad \frac{\rho\big(\alpha_G^\star I + (1-\alpha_G^\star)G^2\big)}{\rho^2} \;=\; \frac{|\sin 2\theta|}{|1-\rho^2 e^{2i\theta}|}.$$

A simple closed-form sufficient condition that implies (31) *is*

$$r_{\mathrm{sub}}^2 < |\sigma^\star| \quad \textit{and} \quad \alpha^\star(\rho,\theta) \;\leq\; \frac{|\sigma^\star| - r_{\mathrm{sub}}^2}{1 - r_{\mathrm{sub}}^2}. \tag{32}$$

Remark 8 (On the gap requirement (31))**.** The looser condition $\alpha^\star(\rho,\theta) < (\rho^2 - r_{\mathrm{sub}}^2)/(1 - r_{\mathrm{sub}}^2)$ only guarantees that the filtered subdominant modes contract *below the unfiltered dominant rate* ρ^2, not below the new accelerated rate $|\sigma^\star| < \rho^2$. When some filtered subdominant mode lands strictly between $|\sigma^\star|$ and ρ^2, it becomes the new spectral bottleneck and the global minimax minimizer $\alpha_G^\star$ shifts away from $\alpha^\star(\rho,\theta)$. Condition (31) (or its closed-form sufficient version (32)) is the correct gap requirement that pins the global minimax to the dominant pair.

Proof of Corollary 2. Under (31), every subdominant filtered eigenvalue satisfies $|\alpha^\star + (1-\alpha^\star)\mu_j^2| \le |\sigma^\star|$, while Lemma 1(c) gives $|\alpha^\star + (1-\alpha^\star)\mu^2| = |\sigma^\star|$ for the dominant pair. Hence $\rho(\alpha^\star I + (1-\alpha^\star)G^2) = |\sigma^\star|$. Since $\alpha^\star$ strictly minimizes the dominant-mode magnitude (Lemma 1(b)), no other α can drive that magnitude—and hence the spectral radius—below $|\sigma^\star|$, so $\alpha_G^\star = \alpha^\star(\rho,\theta)$.

For the closed-form sufficient condition (32): by the triangle inequality applied to a single subdominant eigenvalue, $|\alpha^\star + (1-\alpha^\star)\mu_j^2| \le \alpha^\star + (1-\alpha^\star)r_{\mathrm{sub}}^2$; the right-hand side is $\le |\sigma^\star|$ iff $\alpha^\star \le (|\sigma^\star| - r_{\mathrm{sub}}^2)/(1 - r_{\mathrm{sub}}^2)$, which requires $r_{\mathrm{sub}}^2 < |\sigma^\star|$ to admit a positive $\alpha^\star$. □

Corollary 3 (Limiting spectral-factor ratio as $\rho \to 1$)**.** *For a single mode* $\mu = \rho\, e^{i\theta}$ *with* $\theta \in (0,\pi)$*, the acceleration ratio* (5) *satisfies* $|\sigma^\star|/\rho^2 \to |\cos\theta|$ *as* $\rho \to 1^-$.

Proof of Corollary 3. From (5), $|\sigma^\star|/\rho^2 = |\sin 2\theta|/|1 - \rho^2 e^{2i\theta}|$. As $\rho \to 1^-$, $|1 - \rho^2 e^{2i\theta}| \to |1 - e^{2i\theta}| = 2|\sin\theta|$ for $\theta \neq 0, \pi$. Substituting $|\sin 2\theta| = 2|\sin\theta\cos\theta|$:

$$\frac{|\sigma^\star|}{\rho^2} \;\longrightarrow\; \frac{2|\sin\theta\cos\theta|}{2|\sin\theta|} \;=\; |\cos\theta|. \qquad \square$$

Remark 9 (Per-iteration improvement over T)**.** The weaker cost-asymmetric question—when does PAPF beat the single-evaluation Picard step T?—is essentially trivial since T^2 alone already does so at rate $\rho^2 < \rho$. For completeness: under Assumption 1, $\alpha = 0$ already gives $\rho^2 < \rho$, so by continuity every sufficiently small $\alpha > 0$ improves on the single step T as well; if the dominant eigenvalue is a complex pair with $\cos 2\theta < \rho^2$, then $\alpha^\star(\rho,\theta)$ in (4) also satisfies $|p_{\alpha^\star}(\rho e^{i\theta})| < \rho$, and if it is real $\lambda_\star \in (0,1)$ then $\rho(p_\alpha(G)) < \lambda_\star$ for any $\alpha < \lambda_\star/(1+\lambda_\star)$, both by the modulus-squared identity (3). The substantive question, and the one driving algorithmic and applied content below, is the cost-honest comparison against T^2 resolved by Theorem 2.

Remark 10 (Scope of the cap: nonnegative parameters and FNE geometry)**.** Proposition 2 is a statement about the family $\alpha \in [0,1)$ and about FNE geometry. For FNE T, $T^2 = \frac{1}{3}I + \frac{2}{3}R$ with R nonexpansive, so $\mathsf{S}_\alpha = [1 - \frac{2(1-\alpha)}{3}]I + \frac{2(1-\alpha)}{3}R$ is averaged for every $\alpha \in (-\frac{1}{2}, 1)$, and Proposition 1 extends to that range; over-relaxation $\alpha < 0$ can beat the cap on real modes, e.g. $T = 0.9I$ with $\alpha = -0.4$ gives $q/\rho^2 = 0.734/0.81 \approx 0.906 < \kappa_{\mathrm{FNE}}(0.9) = 0.972$. Likewise convex relaxed KM $\beta I + (1-\beta)T$ is averaged for $\beta \in (-1,1)$ when T is FNE, and Proposition 3 compares only $\beta \in [0,1)$. Finally, FNE is sufficient but not necessary for KM-type global safety: the contraction $T = 0.99R(\pi/2)$ is 0.995-averaged but not FNE, and $\alpha = 0.9801/1.9801$ makes $\mathsf{S}_\alpha \equiv 0$, so global safety and large acceleration can coexist outside the FNE class; we do not pursue these extensions here.

Proof of Proposition 2. Let $\mu = \rho e^{i\theta}$ be a dominant eigenvalue of G; since G is real we may take $\theta \in [0,\pi]$. By the FNE-disk property of Section 2 (valid for FNE with respect to any inner product), $\rho \le \cos\theta$, so $\theta \in [0,\pi/2)$ and $\cos\theta \ge \rho$. Fix $\alpha \in [0,1)$ and let $\sigma(\alpha) = \alpha + (1-\alpha)\mu^2$ be the corresponding eigenvalue of $A_\alpha = \alpha I + (1-\alpha)G^2$, so that $q(\alpha) \ge |\sigma(\alpha)|$.

Step 1 (outside the acceleration region). If $\cos 2\theta \ge \rho^2$, Lemma 1(a) gives $|\sigma(\alpha)| \ge \rho^2$, hence $q(\alpha)/\rho^2 \ge 1 > \kappa_{\mathrm{FNE}}(\rho)$, the last inequality because $4\rho^2 < 1 + 3\rho^2$ for $\rho < 1$.

Step 2 (inside the acceleration region). If $\cos 2\theta < \rho^2$, Lemma 1(c) gives $|\sigma(\alpha)| \ge |\sigma^\star| = \rho^2\, \varphi_\rho(\cos 2\theta)$, where, using $|1 - \rho^2 e^{2i\theta}|^2 = 1 - 2\rho^2\cos 2\theta + \rho^4$ and $\sin^2 2\theta = 1 - \cos^2 2\theta$,

$$\varphi_\rho(c) := \sqrt{\frac{1-c^2}{1 - 2\rho^2 c + \rho^4}}, \qquad c \in (-1, \rho^2).$$

Differentiating the radicand,

$$\frac{d}{dc}\,\frac{1-c^2}{1-2\rho^2c+\rho^4}=\frac{-2c\,(1-2\rho^2c+\rho^4)+2\rho^2(1-c^2)}{(1-2\rho^2c+\rho^4)^2}=\frac{2(\rho^2c-1)(c-\rho^2)}{(1-2\rho^2c+\rho^4)^2},$$

which is positive for $c<\rho^2$ (both factors in the numerator are negative). Hence φ_ρ is increasing on $(-1,\rho^2)$.

Step 3 (the disk constraint). Since $\cos\theta\ge\rho$ and $\theta\in[0,\pi/2)$, $c=\cos 2\theta=2\cos^2\theta-1\ge 2\rho^2-1$, and $2\rho^2-1<\rho^2$ because $\rho<1$; so $\varphi_\rho(c)\ge\varphi_\rho(2\rho^2-1)$. Evaluating, $1-(2\rho^2-1)^2=4\rho^2(1-\rho^2)$ and $1-2\rho^2(2\rho^2-1)+\rho^4=1+2\rho^2-3\rho^4=(1-\rho^2)(1+3\rho^2)$, whence $\varphi_\rho(2\rho^2-1)=\sqrt{4\rho^2/(1+3\rho^2)}=\kappa_{\mathrm{FNE}}(\rho)$. Combining the steps, $q(\alpha)/\rho^2\ge\kappa_{\mathrm{FNE}}(\rho)$ for every $\alpha\in[0,1)$.

Sharpness. Let $T(z)=\rho R(\theta)z$ on $\mathbb{R}^2$ with $\rho=\cos\theta\in(0,1)$. The matrix $2T-I=2\rho R(\theta)-I$ is normal with eigenvalues $2\rho e^{\pm i\theta}-1$, and $|2\rho e^{i\theta}-1|^2=4\rho^2-4\rho\cos\theta+1=1$; hence $\|2T-I\|=1$, i.e., T is FNE for the Euclidean inner product. Its Jacobian has the single conjugate pair $\rho e^{\pm i\theta}$, which lies on the FNE circle, so Lemma 1(c) at $\alpha=\alpha^\star$ gives $q(\alpha^\star)=\rho^2\varphi_\rho(2\rho^2-1)=\rho^2\kappa_{\mathrm{FNE}}(\rho)$. □

Remark 11 (The cap holds for the whole convex degree-2 family)**.** Proposition 2 extends verbatim to every convex filter $p(\lambda)=a+b\lambda+c\lambda^2$ with $a,b,c\ge 0$, $a+b+c=1$: if T is FNE, then $\rho(p(G))\ge\rho^2\kappa_{\mathrm{FNE}}(\rho)$ for every such p, so admitting $b>0$ cannot improve on the $b=0$ family at the level of this bound (for $c>0$ a step costs two T-calls, as for PAPF, so the T-call cap $B(\rho)$ carries over; for $c=0$ the per-mode modulus is at least ρ per T-call, i.e., no gain at all). Let $\mu=\rho e^{i\theta}$ be a dominant eigenvalue; the FNE disk gives $\cos\theta\ge\rho>0$, and $\rho(p(G))\ge|p(\mu)|$ by spectral mapping. As (a,b,c) ranges over the simplex, $p(\mu)$ ranges over the triangle $\Delta=\mathrm{conv}\{1,\mu,\mu^2\}$. If $\theta\neq 0$, then $\mathrm{Im}\,p(\mu)=\rho\sin\theta\,(b+2c\rho\cos\theta)$ vanishes only for $b=c=0$, where $p(\mu)=1$; hence $0\notin\Delta$, and $\min_{z\in\Delta}|z|$ is attained on an edge (from an interior minimizer z, the point $(1-t)z\in\Delta$ would have smaller modulus for small $t>0$). The edge $[1,\mu]$ consists of the relaxed-KM values $\beta+(1-\beta)\mu$, whose modulus is at least $\kappa_{\mathrm{KM}}(\rho,\theta)=\rho$ by (33) because $\cos\theta\ge\rho$; the edge $[\mu,\mu^2]=\mu\,[1,\mu]$ therefore has modulus at least ρ^2; and the edge $[1,\mu^2]$ consists of the PAPF values $\alpha+(1-\alpha)\mu^2$, whose modulus is at least $\rho^2\kappa_{\mathrm{FNE}}(\rho)$ by the proof above. Since $\kappa_{\mathrm{FNE}}(\rho)<1$, the minimum over Δ is at least $\rho^2\kappa_{\mathrm{FNE}}(\rho)$, which proves the claim; for $\theta=0$ the triangle degenerates to the segment $[\rho^2,1]$ and the bound is immediate. The cap is thus a property of FNE geometry and nonnegative coefficients, not of the restriction $b=0$; outside the disk, $b>0$ can help (Proposition 7).

The per-mode optimal moduli of the two filters are

$$\kappa_{\mathrm{KM}}(\rho,\theta)=\begin{cases}\dfrac{\rho\,|\sin\theta|}{|1-\rho e^{i\theta}|}, & \cos\theta<\rho,\\ \rho, & \text{otherwise,}\end{cases}\qquad \kappa_{\mathrm{P}}(\rho,\theta)=\begin{cases}\dfrac{\rho^2\,|\sin 2\theta|}{|1-\rho^2 e^{2i\theta}|}, & \cos 2\theta<\rho^2,\\ \rho^2, & \text{otherwise.}\end{cases}\tag{33}$$

Proof of (33) *and of Proposition 3. Step 1 (relaxed KM single-mode identity).* For $\mu=\rho e^{i\theta}$ and $\beta\in[0,1]$, expanding $|\beta+(1-\beta)\mu|^2=\beta^2+2\beta(1-\beta)\rho\cos\theta+(1-\beta)^2\rho^2$ gives

$$|\beta+(1-\beta)\mu|^2-\rho^2=\beta\big[\beta\,|1-\mu|^2+2\rho(\cos\theta-\rho)\big],\qquad |1-\mu|^2=1-2\rho\cos\theta+\rho^2.$$

Exactly as in Lemma 1, the bracket is negative for small $\beta>0$ iff $\cos\theta<\rho$; in that case the minimizer is $\beta^\star=\rho(\rho-\cos\theta)/|1-\mu|^2\in(0,1)$ (note $\beta^\star<1\iff\rho\cos\theta<1$) and the minimum value is $\rho^2-\rho^2(\cos\theta-\rho)^2/|1-\mu|^2=\rho^2\big(|1-\mu|^2-(\cos\theta-\rho)^2\big)/|1-\mu|^2=\rho^2\sin^2\theta/|1-\mu|^2$, which is κ_{KM}^2; otherwise the minimum over $\beta\in[0,1)$ is ρ^2 at $\beta=0$. The formula for κ_{P} is Lemma 1.

Step 2 (nesting, (i)). For $\theta\in(0,\pi/2]$ we have $\cos\theta\ge 0$, so $\cos\theta<\rho$ implies $\cos 2\theta=2\cos^2\theta-1<2\rho^2-1<\rho^2$. The converse fails on the nonempty band $\theta_b<\theta\le\arccos\rho$, $\theta_b:=\frac{1}{2}\arccos\rho^2$, since $\cos(2\arccos\rho)=2\rho^2-1<\rho^2$.

Step 3 (reduction to g). When both closed forms in (33) apply, factor $1-\rho^2e^{2i\theta}=(1-\rho e^{i\theta})(1+\rho e^{i\theta})$ and use $\sin\theta>0$ on $(0,\pi)$:

$$\kappa_{\mathrm{P}}<\kappa_{\mathrm{KM}}^2\iff\frac{|\sin 2\theta|}{|1-\rho e^{i\theta}|\,|1+\rho e^{i\theta}|}<\frac{\sin^2\theta}{|1-\rho e^{i\theta}|^2}\iff 2|\cos\theta|\,|1-\rho e^{i\theta}|<\sin\theta\,|1+\rho e^{i\theta}|.$$

Squaring (both sides are nonnegative) and writing $c = \cos\theta$, $4c^2(1-2\rho c+\rho^2) < (1-c^2)(1+2\rho c+\rho^2)$, and expanding the difference of the two sides gives exactly $g(\rho,c) > 0$ with $g(\rho,c) = (1+\rho^2)(1-5c^2)+2\rho c+6\rho c^3$.

Step 4 (right half-plane, (ii)). Let $\theta \in (0,\pi/2]$, so $c \in [0,1)$. If $\cos\theta < \rho$, both closed forms apply by Step 2, and polynomial division of $g(\rho,c) - (1-\rho^2)^2$ by $(c-\rho)$ yields the identity

$$g(\rho,c) = (1-\rho^2)^2 + (\rho-c)\,h_\rho(c), \qquad h_\rho(c) := 3\rho - \rho^3 + (5-\rho^2)c - 6\rho c^2.$$

The function h_ρ is concave in c with $h_\rho(0) = \rho(3-\rho^2) > 0$ and $h_\rho(\rho) = 8\rho(1-\rho^2) > 0$, hence $h_\rho > 0$ on $[0,\rho]$ and $g(\rho,c) \geq (1-\rho^2)^2 > 0$ for $0 \leq c \leq \rho$. If instead $\cos\theta \geq \rho$, then $\kappa_{\mathrm{KM}}^2 = \rho^2$, and by Lemma 1(a) $\kappa_{\mathrm{P}} < \rho^2$ iff $\cos 2\theta < \rho^2$, with $\kappa_{\mathrm{P}} = \rho^2$ otherwise. In both cases the strict inequality holds iff $\cos 2\theta < \rho^2$. At $\theta = \pi/2$: $\sin 2\theta = 0$ and $\cos 2\theta = -1 < \rho^2$ give $\kappa_{\mathrm{P}} = 0$, while $|1 - i\rho|^2 = 1+\rho^2$ gives $\kappa_{\mathrm{KM}}^2 = \rho^2/(1+\rho^2)$.

Step 5 (left half-plane, (iii)). Let $\theta \in (\pi/2,\pi)$, so $c \in (-1,0)$ and $\cos\theta < 0 < \rho$: relaxed KM accelerates and its closed form applies. On $(-1,0)$, $\partial_c g = -10(1+\rho^2)c + 2\rho + 18\rho c^2 > 0$, so $g(\rho,\cdot)$ is strictly increasing there; with $g(\rho,-1) = -4(1+\rho^2) - 8\rho < 0 < 1+\rho^2 = g(\rho,0)$ it has a unique zero $c_\times(\rho) \in (-1,0)$. If $\cos 2\theta < \rho^2$, Step 3 shows $\kappa_{\mathrm{P}} < \kappa_{\mathrm{KM}}^2$ iff $g(\rho,c) > 0$ iff $c > c_\times(\rho)$. If $\cos 2\theta \geq \rho^2$, then $\kappa_{\mathrm{P}} = \rho^2 > \kappa_{\mathrm{KM}}^2$, so relaxed KM is better; this is consistent with the criterion because $\cos 2\theta \geq \rho^2$ forces $c \leq -\sqrt{(1+\rho^2)/2} < -1/\sqrt{2}$, while $g(\rho,-1/\sqrt{2}) = -\frac{3}{2}(1+\rho^2) - \sqrt{2}\rho - \frac{3}{\sqrt{2}}\rho < 0$ shows $c_\times(\rho) > -1/\sqrt{2}$. For the limit, $g(1,c) = 6c^3 - 10c^2 + 2c + 2 = 2(3c+1)(c-1)^2$ has the unique zero $-1/3$ in $(-1,0)$, at which it changes sign; since g is jointly continuous, for every $\epsilon > 0$ and ρ close to 1 we have $g(\rho,-1/3-\epsilon) < 0 < g(\rho,-1/3+\epsilon)$, so $c_\times(\rho) \to -1/3$. Finally, as $\theta \to \pi$ the relaxed-KM modulus $\rho\sin\theta/|1-\rho e^{i\theta}| \to 0$ (with $\beta^\star \to \rho/(1+\rho)$, which annihilates the real eigenvalue $-\rho$ exactly), whereas $\cos 2\theta \to 1 > \rho^2$ places PAPF outside its acceleration region. □

Remark 12 (Per-mode versus full-spectrum comparison)**.** Proposition 3 compares the two filters on a single conjugate pair. Because each filter's parameter is selected for the whole instance using the criteria of Table 1, the full-spectrum winner can differ from the dominant-pair winner. For $G = \mathrm{diag}\big(0.99R(\pi/2), -0.98\big)$ the dominant pair sits at $\theta = \pi/2$, PAPF's best case, yet $\min_{\alpha\in[0,1/2]} \max_j |\alpha + (1-\alpha)\mu_j^2| \approx 0.9608$, because the real mode -0.98 is mapped to $\alpha + (1-\alpha)0.9604 \geq 0.9604$ for every $\alpha \geq 0$, whereas $\min_{\beta\in[0,1)} \max_j |\beta + (1-\beta)\mu_j|^2 \approx 0.4950$: convex relaxed KM wins on the full operator. Instance-level outcomes in Section 5 are therefore reported as consistent with the per-mode picture, not as its consequence.

Proposition 7 (Convex annihilation of a dominant pair)**.** *Let $\mu = \rho e^{\pm i\theta}$ with $\rho \in (0,1)$, $\theta \in (0,\pi)$. Among $p(\lambda) = a+b\lambda+c\lambda^2$ with $a+b+c = 1$, the unique polynomial annihilating the eigenvalue pair (if the pair carries nontrivial Jordan chains, $p(G)$ need not vanish on the corresponding generalized eigenspace, and a zero linearized factor does not make a nonlinear iteration terminate in one step) is $p(\lambda) = c(\lambda^2 - 2\rho\cos\theta\,\lambda + \rho^2)$ with $c = |1-\rho e^{i\theta}|^{-2}$, $a = \rho^2 c$, $b = -2\rho\cos\theta\, c$. Hence p is a convex combination iff $\cos\theta \leq 0$, independently of ρ. At $\theta = \pi/2$, $b = 0$ and $p = p_{\alpha^\star}$ with $\alpha^\star = \rho^2/(1+\rho^2)$: PAPF's optimal filter* is *the annihilator. For $\theta < \pi/2$ no convex degree-2 annihilator exists, so exact cancellation is incompatible with the convex-combination structure that preserves averagedness (Proposition 1), and Theorem 2 is the operative statement there; for $\theta \in (\pi/2,\pi)$ a convex annihilator with $b > 0$ exists (future work).*

Proof of Proposition 7. A polynomial of degree at most two vanishing at the distinct non-real points $\mu, \bar\mu$ is $c(\lambda-\mu)(\lambda-\bar\mu) = c(\lambda^2 - 2\rho\cos\theta\,\lambda + \rho^2)$, so $a = \rho^2 c$ and $b = -2\rho\cos\theta\, c$. The normalization $a+b+c = c(\rho^2 - 2\rho\cos\theta + 1) = c\,|1-\rho e^{i\theta}|^2 = 1$ fixes $c = |1-\rho e^{i\theta}|^{-2} > 0$, hence $a > 0$, and $b \geq 0$ iff $\cos\theta \leq 0$. At $\theta = \pi/2$, $b = 0$, $c = 1/(1+\rho^2)$ and $a = \rho^2/(1+\rho^2)$, so $p = p_\alpha$ with $\alpha = \rho^2/(1+\rho^2)$; this equals $\alpha^\star(\rho,\pi/2) = \rho^2(\rho^2+1)/|1+\rho^2|^2$ from (4). □

Proof of Proposition 4. Let U be a ball of radius δ around $z^\star$ on which $T(z) = z^\star + G(z-z^\star)$, and put $u := z^0 - z^\star$. The eigenvalues of A_α are $\alpha + (1-\alpha)\mu_j^2$ with $|\mu_j| < 1$, so $q(\alpha) \leq \alpha + (1-\alpha)\rho(G)^2 < 1$ and $C := \sup_{k\geq 0} \|A_\alpha^k\| < \infty$. If $\|u\| < \delta/(C\max\{1,\|G\|\})$, then by induction every iterate satisfies $z^k - z^\star = A_\alpha^k u$ with $z^k \in U$ and $T(z^k) \in U$ (indeed $\|z^k - z^\star\| \leq C\|u\| < \delta$ and $\|T(z^k) - z^\star\| = \|G(z^k - z^\star)\| \leq \|G\|C\|u\| < \delta$), so the affine representation of T applies at every

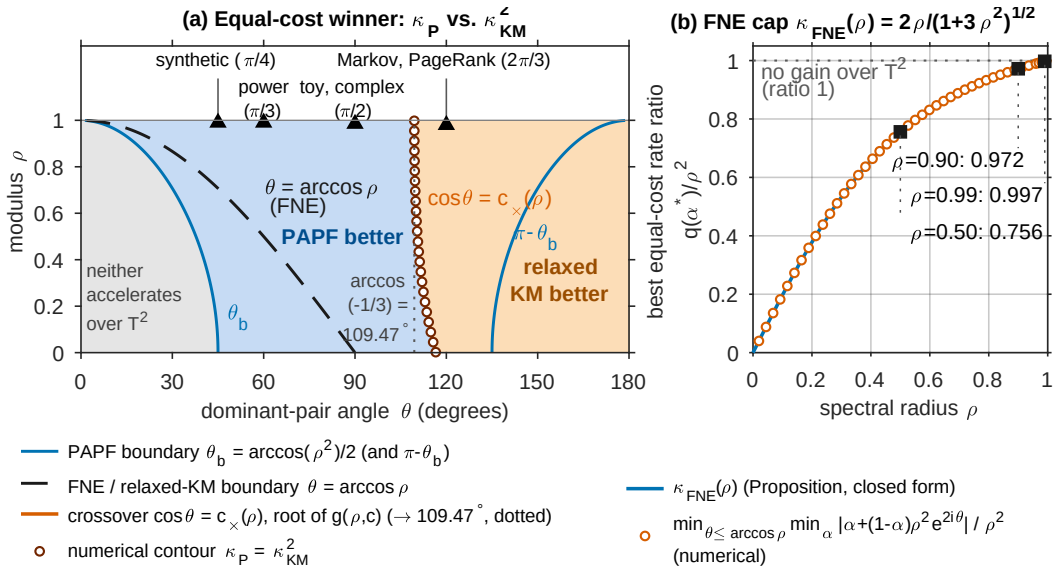

Figure 5: **Left:** equal-cost winner map of Proposition 3 in the (θ, ρ) plane: PAPF is better where $\kappa_{\mathrm{P}} < \kappa_{\mathrm{KM}}^2$, relaxed KM where the reverse holds, and neither accelerates over T^2 below the PAPF boundary $\theta_b = \frac{1}{2}\arccos\rho^2$; the crossover curve $\cos\theta = c_\times(\rho)$ tends to $\arccos(-1/3) \approx 109.5°$ as $\rho \to 1$, and the FNE disk boundary $\theta = \arccos\rho$ is overlaid. **Right:** the FNE cap $\kappa_{\mathrm{FNE}}(\rho) = 2\rho/\sqrt{1+3\rho^2}$ of Proposition 2 coincides with the numerically computed minimum over the FNE-admissible angles of the true optimum $\min_\alpha |\alpha + (1-\alpha)\rho^2 e^{2i\theta}|/\rho^2$.

step. Consequently $r_k = \|T(z^k) - z^k\| = \|(G-I)A_\alpha^k u\|$. Since $\rho(G) < 1$, $G - I$ is invertible, and with $c_1 := \|(G-I)^{-1}\|^{-1}$, $c_2 := \|G - I\|$ we get $c_1\|A_\alpha^k u\| \le r_k \le c_2\|A_\alpha^k u\|$; also $r_0 > 0$ because $u \neq 0$ (as $\Pi_\alpha u \neq 0$). It therefore suffices to show $\|A_\alpha^k u\|^{1/k} \to q := q(\alpha)$.

Upper bound. $\|A_\alpha^k u\| \le \|A_\alpha^k\|\,\|u\|$ and Gelfand's formula $\|A_\alpha^k\|^{1/k} \to \rho(A_\alpha) = q$ give $\limsup_k \|A_\alpha^k u\|^{1/k} \le q$.

Lower bound. Write $u = \Pi_\alpha u + (I - \Pi_\alpha)u$. The subspaces $V_1 := \operatorname{range}\Pi_\alpha$ and $V_2 := \operatorname{range}(I - \Pi_\alpha)$ are A_α-invariant; $A_1 := A_\alpha|_{V_1}$ has all its eigenvalues of modulus exactly q, and $A_2 := A_\alpha|_{V_2}$ has spectral radius $q_2 < q$ (if $V_2 = \{0\}$ the second term below is absent). Because $q > 0$, A_1 is invertible and $\rho(A_1^{-1}) = 1/q$, so by Gelfand's formula, for any $\eta > 0$ and all large k, $\|A_1^{-k}\| \le (1/q + \eta)^k$, whence $\|A_1^k \Pi_\alpha u\| \ge \|\Pi_\alpha u\|/\|A_1^{-k}\| \ge \|\Pi_\alpha u\|\,(1/q+\eta)^{-k}$. Likewise $\|A_2^k(I - \Pi_\alpha)u\| \le C_\eta(q_2 + \eta)^k$. Choosing η so small that $q_2 + \eta < (1/q + \eta)^{-1}$, the triangle inequality gives, for large k,

$$\|A_\alpha^k u\| \;\ge\; \|A_1^k \Pi_\alpha u\| - \|A_2^k(I - \Pi_\alpha)u\| \;\ge\; \tfrac{1}{2}\,\|\Pi_\alpha u\|\,(1/q+\eta)^{-k},$$

so $\liminf_k \|A_\alpha^k u\|^{1/k} \ge (1/q + \eta)^{-1}$; letting $\eta \downarrow 0$ yields $\liminf_k \|A_\alpha^k u\|^{1/k} \ge q$. Hence $\|A_\alpha^k u\|^{1/k} \to q$, and $(r_k/r_0)^{1/k} = r_k^{1/k}\, r_0^{-1/k} \to q \cdot 1$.

For the single-step ratio r_{k+1}/r_k: when the dominant eigenvalue of A_α is real and simple, standard power-iteration arguments give $r_{k+1}/r_k \to q(\alpha)$; when it is a complex-conjugate pair, $A_\alpha^k u$ rotates within the two-dimensional dominant invariant subspace and the single-step ratio can oscillate (it is constant for an isotropic rotation in the Euclidean norm, and may oscillate when the rotation is nonnormal, the norm is not adapted, or other modes are superposed), which is why Algorithm 1 uses the k-step geometric mean. □

Remark 13 (The nonlinear case needs an exclusion beyond the Riesz condition)**.** For smooth nonlinear T, differentiability alone does not yield $z^k - z^\star = A_\alpha^k(z^0 - z^\star) + o(q(\alpha)^k)$, and the conclusion of Proposition 4 can fail even when $\Pi_\alpha(z^0 - z^\star) \neq 0$. Consider $T(x, y) = (ax + y^2,\, by)$ with $0 < b < a < 1$ (an example we owe to an anonymous reviewer). Then $z^\star = 0$, $G = \operatorname{diag}(a, b)$,

and a direct computation gives

$$\mathsf{S}_\alpha(x,y) = \big(\alpha x+(1-\alpha)(a^2x+(a+b^2)y^2),\ h_b\, y\big), \qquad h_b := \alpha+(1-\alpha)b^2 < q(\alpha) = \alpha+(1-\alpha)a^2.$$

The parabola $x = Cy^2$ with $C = -(1-\alpha)(a+b^2)/\big(q(\alpha)-h_b^2\big)$ is invariant under S_α (substitute and compare coefficients of y^2), and on it the iteration reduces to $y \mapsto h_b y$; hence for $z^0 = (Cy_0^2, y_0)$ one gets $(r_k/r_0)^{1/k} \to h_b < q(\alpha)$, although the Riesz projection of z^0 onto the dominant (x-)eigenspace, $(Cy_0^2, 0)$, is nonzero. For $(a, b, \alpha) = (0.9, 0.5, 0.3)$ the two rates are $h_b = 0.475$ and $q(\alpha) = 0.867$, and both are reproduced numerically (App. E). The parabola is the strong stable manifold of S_α at $z^\star$ (tangent to the fast direction). For C^1 maps, under the hypotheses of a strong-stable-manifold theorem for maps (Hirsch et al., 1977; Capiński & Wasieczko-Zając, 2015), the conclusion of the proposition holds for initial points off that manifold, which has positive codimension; because S_α may fail to be invertible when the filter annihilates a mode, we do not state a general nonlinear theorem. For Algorithm 1 this exceptional behavior can only make the residual proxy *underestimate* $q(\alpha)$ along such directions and bias the selection; it does not affect Proposition 1 or Theorem 1.

Proof of Proposition 5. Each $|\alpha + (1-\alpha)\mu_j^2|$ is a convex function of α (it is the modulus of an affine function), and the pointwise maximum of convex functions is convex. □

G Supporting Empirical Results

This appendix reports the auxiliary measurements that complement Section 5: the full per-instance baselines table (§G.1), TKMA in both metrics on both PDHG maps (§G.2), the boundary sweep (§G.3), the OT KKT residuals (§G.4), wall-clock timings (§G.5), and the adaptive-parameter ablation plus a transient-growth probe (§G.6). All experiments share the seeds and instance generators of App. E; raw data are in `codes/results/` (`baselines_table_iclr.csv`, `walltime_iclr.csv`, `ablation.mat`, `transient.mat`).

G.1 Baselines: per-T-call curves and full table

Figure 7 plots per-T-call residuals on four representative instances; Table 3 gives T-call counts and speedups (T^2/method) for all six instances summarized in §5.3.

Baseline configurations. Anderson acceleration is the Walker–Ni Type-II scheme in delta formulation (least squares against finite differences of residuals) with memory $m \in \{5, 10\}$, Tikhonov regularization 10^{-10}, mixing parameter 1 (pure AA), no restarts, and a residual-monotone safeguard: the AA candidate is replaced by the Picard step $T(z^k)$ whenever its residual exceeds $\|T(z^k) - z^k\|$. Halpern iteration is $z^{k+1} = \lambda_k z^0 + (1-\lambda_k)T(z^k)$ with $\lambda_k = 1/(k+2)$. The inertial method is $z^{k+1} = T(z^k + \beta(z^k - z^{k-1}))$ with $\beta = 0.9$ and the same residual-monotone safeguard (fallback to $T(z^k)$). Every T evaluation, including safeguard fallbacks, is counted. Budgets are per instance: the AA, Halpern, and inertial runners stop after 10,000 (small OT), 50,000 (Iris), 10,000 (PageRank), 3,000 (power method), 5,000 (complex-dominant), and 50,000 (large OT) iterations, TKMA after the instance's step budget of App. E; capped runs are marked † and their ratios are upper bounds. The Krylov comparison of Section 5.3 uses the affine variant $x \mapsto Ax + b$ of the complex-dominant instance with the stopping rule $\|b - (I - A)x\| \le 10^{-12}\|b\|$ and one matrix–vector product per T-call: GMRES(full) (Saad & Schultz, 1986) needs 24 products, GMRES(5) 34, GMRES(2) 43, PAPF (best fixed α) 30, relaxed KM 76, and T^2 2582.

On the power-method instance the iterations differ in where they live: T^2, relaxed KM, PAPF (fixed and adaptive, in all phases), and TKMA renormalize after every update (the sphere algorithm of App. D.2); AA and Halpern run in the ambient space (their iterates are not unit vectors, and the reported residual is $\|T(z) - z\|$ with $T(z)$ unit); the accepted iterates of the inertial method are outputs of T and hence unit vectors, while its extrapolated point is off the sphere (T is scale invariant). Without renormalization, relaxed KM and fixed-α PAPF need 246 and 122 T-calls on the power method instead of 245 and 118, and the same counts on the other three power maps.

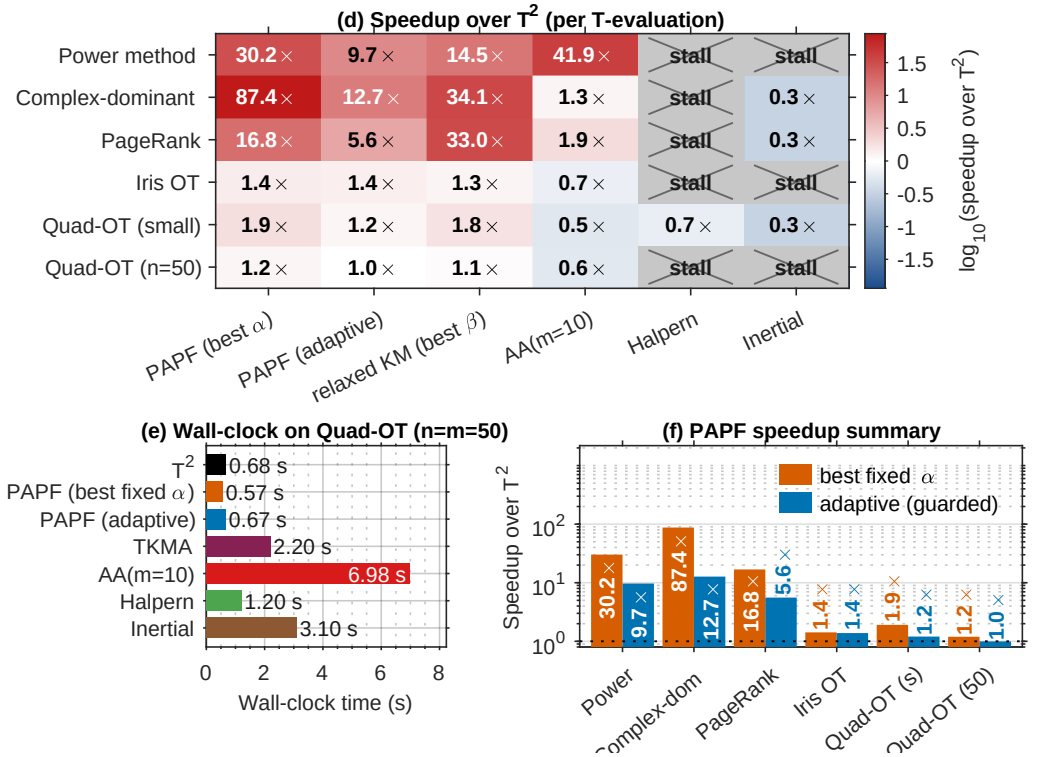

Figure 6: **Empirical overview, second row.** (d) T-call speedup over T^2 on six instances for six methods (log-scale color; "stall" = tolerance not reached). (e) Wall-clock on Quad-OT ($n{=}m{=}50$): PAPF 0.6 s, AA($m{=}10$) 7 s (least-squares cost). (f) PAPF best-fixed vs. adaptive speedup. Same data as Table 3 and Table 7.

Table 3: Per-T-call counts and speedup (T^2/method) under the stopping criterion of the corresponding Table 1 instance (state error $\leq 10^{-14}$ for the small OT instance; residual $\leq 10^{-12}$ for the other OT and synthetic instances; residual $\leq 10^{-14}$ for the power-method family). †: did not reach tolerance within the T-call budget; the ratio then uses the budget and is only an upper bound on the true speedup. Bold: fewest T-calls in the row; underlined: best non-PAPF converged run. Budgets are per instance (App. G.1, "Baseline configurations"); TKMA is the implementation of App. G.2 ($t = \frac{1}{2}$, renormalized on the power maps) with the step budgets of App. E. PAPF (best) uses the fixed parameters and tuning conventions of Table 1; PAPF (ada.) is Algorithm 1 with the acceptance test.

Instance	T^2	AA($m{=}5$)	AA($m{=}10$)	Halpern	Inertial	TKMA	PAPF (best)	PAPF (ada.)
Quad-OT (small)	6516	12.1k (0.54×)	12.2k (0.53×)	10.0k (0.65×)†	19.8k (0.33×)	40.0k (0.16×)†	**3.4k** (1.91×)	5.4k (1.20×)
Iris OT	21264	34.2k (0.62×)	32.3k (0.66×)	50.0k (0.43×)†	147.3k (0.14×)†	100.0k (0.21×)†	**15.0k** (1.42×)	15.4k (1.38×)
PageRank	1616	1213 (1.33×)	829 (1.95×)	10.0k (0.16×)†	4.9k (0.33×)	**56** (28.9×)	96 (16.83×)	290 (5.57×)
Power method	3564	1587 (2.25×)	**85** (41.9×)	3.0k (1.19×)†	7.8k (0.46×)†	6.0k (0.59×)†	118 (30.2×)	368 (9.68×)
Complex-dominant	2798	2555 (1.10×)	2223 (1.26×)	5.0k (0.56×)†	8.4k (0.33×)	158 (17.7×)	**32** (87.4×)	220 (12.72×)
Quad-OT (large)	33356	61.4k (0.54×)	60.0k (0.56×)	50.0k (0.67×)†	145.8k (0.23×)†	100.0k (0.33×)†	**27.9k** (1.20×)	33.5k (1.00×)

G.2 TKMA IN BOTH METRICS ON BOTH PDHG MAPS

Table 4 reports TKMA (He et al., 2025) with all inner products taken either in the Euclidean metric or in the preconditioned metric $M = \left[\begin{smallmatrix} I/\tau & -\mathcal{K}_r^* \\ -\mathcal{K}_r & I/\sigma \end{smallmatrix}\right]$ (positive definite since $\tau\sigma\|\mathcal{K}_r\|^2 = 0.98 < 1$), on the no-extrapolation map T of Section 5.1 and on the extrapolated Chambolle–Pock map T_{cp}, which is FNE in the M-metric. The metric mismatch alone does not explain the stall on T: TKMA stalls in both metrics on all three instances. On T_{cp} it converges in both metrics, near-identically, matching T_{cp}^2 on the large instance and trailing it on the other two.

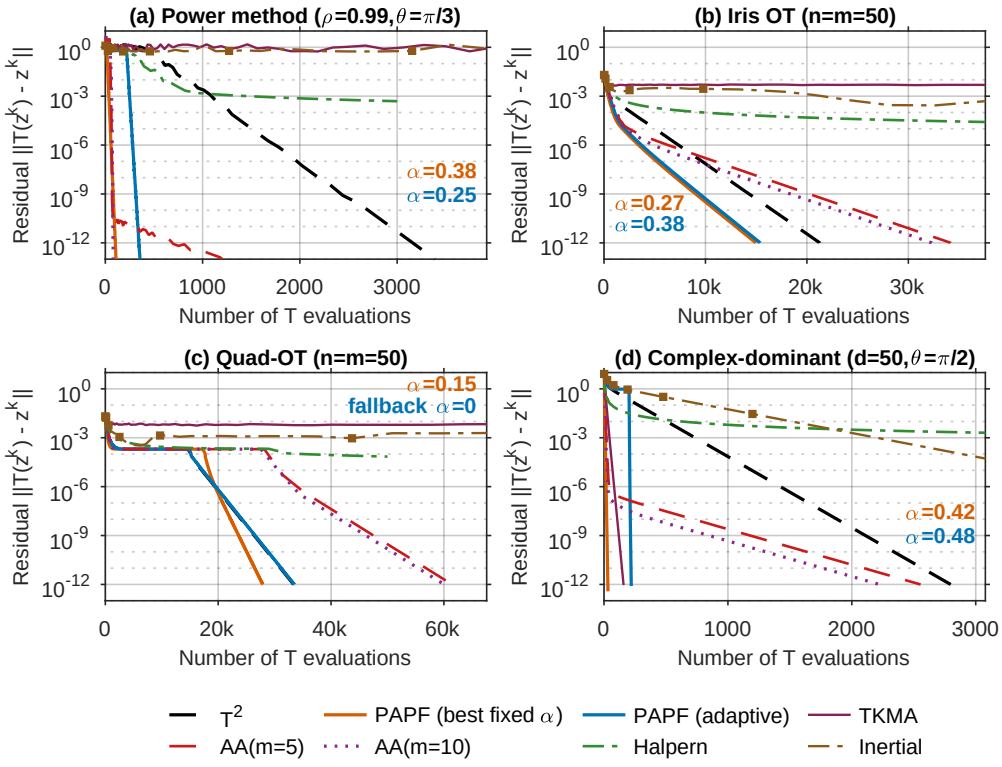

Figure 7: **Baselines comparison.** Per-T-call residual on (a) power method, (b) Iris OT, (c) Quad-OT (n=m=50), (d) complex-dominant. PAPF dominates on PDHG-OT and complex-dominant; AA(m=10) wins the power method but PAPF (best fixed α) wins on complex-dominant; Halpern/Inertial stall on most instances.

The two maps and the TKMA step as implemented. Both maps use $\tau = \sigma = 0.99/\|\mathcal{K}_r\|$. The no-extrapolation map T is the dual-first iteration (9); T_{cp} is the primal-first Chambolle–Pock step $X^+ = [(X - \tau(C + \mathcal{K}_r^* y))/(1 + \tau\varepsilon)]_+$, $y^+ = y + \sigma(\mathcal{K}_r(2X^+ - X) - c_r)$. With the KKT operator $\mathcal{A}(X, y) = (\varepsilon X + C + \mathcal{K}_r^* y + N_{\mathbb{R}_+^{n\times m}}(X),\ c_r - \mathcal{K}_r X)$ we verified $M(z - T_{\mathrm{cp}} z) \in \mathcal{A}(T_{\mathrm{cp}} z)$ at $z^\star$ and at random points of the small instance to $2 \cdot 10^{-15}$, i.e., $T_{\mathrm{cp}} = (I + M^{-1}\mathcal{A})^{-1}$ (M has eigenvalues in $[0.033, 6.49]$ there). TKMA is implemented as in He et al. (2025, eqs. (3.8)–(3.10) of the arXiv v1 PDF, (3.7)–(3.9) of v3; (23)–(25) and (19)–(21) in the HTML renderings): $\bar{z} = T(z)$, $y = T(\bar{z})$, $\theta_k = -\langle \bar{z} - z, \bar{z} - y\rangle / \|\bar{z} - z\|^2$ (Euclidean or M inner product), $w = \bar{z} + \theta_k(\bar{z} - z)$, $z^+ = (1 - t)y + tw$ with $t = \frac{1}{2}$ as in their experiments; two T-calls per step, no clipping of θ_k (set to 0 only if $\|\bar{z} - z\|^2 < 10^{-24}$), no restarts, and renormalization after the step on the power maps only.

Table 4: TKMA T-calls with Euclidean ("Eucl.") or M-metric inner products, on the no-extrapolation map T and on T_{cp}, under each instance's stopping metric ("stall" = plateau at the stated level after the stated budget). Small quad-OT: $\|z - z^\star\| \le 10^{-14}$; large and Iris: $\|T(z) - z\| \le 10^{-12}$.

Instance	Eucl. on T	M on T	T_{cp}^2	Eucl. on T_{cp}	M on T_{cp}
Quad-OT (6×5)	stall (2.0e−1 @ 20k)	stall (3.6e−1 @ 20k)	3214	3744	3716
Quad-OT (50×50)	stall (6.3e−3 @ 50k)	stall (6.2e−3 @ 50k)	22146	22168	22168
Iris OT	stall (4.9e−3 @ 50k)	stall (5.0e−3 @ 50k)	8966	11228	11226

For completeness, on T_{cp} the adaptive rule falls back to $\alpha = 0$ on all three instances, paying exactly the 140-call search overhead (3354, 22,286, and 9106 T-calls vs. 3214, 22,146, and 8966 for T_{cp}^2).

Matched KKT accuracy across the two maps. The counts above stop each map on its own criterion (the state error $\|z - z^\star\| < 10^{-14}$ on the small instance, where both maps share $z^\star$; the map's own residual $< 10^{-12}$ at $n{=}m{=}50$). Since the KKT residuals of App. G.4 depend only on the problem data, they give a common accuracy scale: Table 5 lists the terminal-sustained T-call counts at which the joint residual $\max(R_{\text{stat}}, R_{\text{feas}})$ stays below a level, recorded along the same runs (the T_{cp} records are bitwise identical to those behind Table 4). Stationarity is the binding residual on every converging run. The ranking $T^2_{\text{cp}} <$ PAPF-on-T holds at every common level, with ratios (0.68, 0.79, 0.60 at 10^{-8}) close to the own-criterion ratios at $n{=}m{=}50$ (0.79, 0.60) and wider than on the small instance (0.94): matched accuracy does not weaken the statement that extrapolation is the better algorithm there. The own-criterion stops of the two maps at $n{=}m{=}50$ both correspond to $R_{\text{stat}} \approx 4\text{–}5 \cdot 10^{-10}$, so no run reaches 10^{-10} there. One order changes on the small instance: TKMA on T_{cp} (3744 T-calls on its own criterion) is faster than fixed-α PAPF on T (3416) at matched accuracy (962 vs. 1408 at 10^{-8}; 2746 vs. 3002 at 10^{-12}), while it still trails T^2_{cp} at every level.

Table 5: T-calls from which $\max(R_{\text{stat}}, R_{\text{feas}})$ stays below the stated level (terminal-sustained rule of App. G.4); "own" = the count under the map's own stopping criterion (Tables 1 and 4); "never" = final value above the level.

Instance	Method (map)	own	10^{-6}	10^{-8}	10^{-10}	10^{-12}
Quad-OT ($6{\times}5$)	T^2 (T)	6516	1980	3500	5010	6508
	PAPF $\alpha{=}0.058$ (T)	3416	618	1408	2212	3002
	PAPF adaptive (T)	5408	478	1678	2922	4166
	T^2_{cp}	3214	204	956	1710	2462
	PAPF adaptive (T_{cp})	3354	344	1096	1850	2602
	TKMA (T_{cp})	3744	246	962	1740	2746
Iris OT	T^2 (T)	21264	13388	18042	never	never
	PAPF $\alpha{=}0.27$ (T)	14970	8464	12400	never	never
	PAPF adaptive (T)	15392	8758	12724	never	never
	T^2_{cp}	8966	5132	7460	never	never
	PAPF adaptive (T_{cp})	9106	5272	7600	never	never
	TKMA (T_{cp})	11228	6322	9262	never	never
Quad-OT ($50{\times}50$)	T^2 (T)	33356	25394	30010	never	never
	PAPF $\alpha{=}0.15$ (T)	27884	23112	26022	never	never
	PAPF adaptive (T)	33496	25534	30150	never	never
	T^2_{cp}	22146	18316	20640	never	never
	PAPF adaptive (T_{cp})	22286	18456	20780	never	never
	TKMA (T_{cp})	22168	18560	20674	never	never

G.3 Boundary sweep

Table 6 gives the sweep of Section 5.4: a dominant pair $\rho e^{\pm i\theta}$ with $\rho = 0.99$ plus subdominant real modes 0.6ρ and 0.3ρ, with θ placed outside, on, and inside the boundary $\theta_b = \frac{1}{2}\arccos\rho^2$ of the acceleration region of Fig. 1(a). Rates are empirical tail rates per T-call.

Table 6: Boundary sweep at $\rho = 0.99$ (tolerance 10^{-12}). "Fallback" means the acceptance test rejected the golden-section candidate.

θ/θ_b	regime	T^2 (T-calls)	best fixed α: tail rate vs. T^2	adaptive (T-calls, $\hat{\alpha}$)
0.50	outside	2010	$\alpha = 0$ (no gain exists)	2150, 0 (fallback)
0.95	outside	2072	$\alpha = 0$	2212, 0 (fallback)
1.00	boundary	2076	$\alpha = 0$	2216, 0 (fallback)
1.05	just inside	2082	$\alpha \approx 0.05$: 0.98929 vs. 0.98932	2222, 0 (conservative)
1.60	inside	2124	$\alpha = 0.30$: 0.9870 vs. 0.9895	2264, 0 (conservative miss)
$\theta = \pi/2$	deep inside	2340	$\alpha = 0.24$: 0.7115 vs. 0.9900	**210**, 0.495

At $\rho = 0.9$ the entire T^2 run takes 206–212 T-calls outside the region, so the fixed search cost dominates (adaptive 346–352); at $\theta/\theta_b = 1.6$ the rule accepts $\hat{\alpha} = 0.281$ and still loses, 326 vs. 216

T-calls, because the locked phase saves 30 calls while the search costs 140; only at $\theta = \pi/2$ does the gain barely amortize (208 vs. 226). At $\rho = 0.6$ the instance converges during burn-in and the search never runs.

Break-even rule. In the idealized model in which both trajectories decay geometrically from the common point $\bar{z}$ at their spectral factors (equal prefactors, no transients), with $L := \log(r(\bar{z})/\varepsilon)$, $s_0 := -\log \rho^2$, $s_1 := -\log q(\hat{\alpha})$, two T-calls per update and B search T-calls, T^2 needs $2L/s_0$ calls and the adaptive run $B + 2L/s_1$, so the search pays off iff $B < 2L\,(1/s_0 - 1/s_1)$. With $B = 140$ the required decay is $L > 70/(1/s_0 - 1/s_1)$: at $\rho = 0.9$ a factor improved by 10% ($q = 0.9\rho^2$) needs $L > 44$, a residual reduction by 10^{19}, so the search cannot pay off, whereas at $\rho = 0.99$ the same 10% improvement (a sixfold better rate) pays off after $L > 1.7$. The model ignores the prefactors and transients of Section 4; it is consistent with the sign of every entry of Table 6 and of the $\rho = 0.9$ sweep.

G.4 KKT RESIDUALS FOR PDHG-OT

PDHG iterates are in general infeasible, so a plug-in objective difference $|f(X) - D(y)|$ is not a certificate: it can vanish at infeasible, non-optimal points (for one source and one target with unit mass, $C = 0$ and $\varepsilon = 1$, the point $X = 0, y = 0$ has $f(X) - D(y) = 0$ while the optimal value is $1/2$). For $f(X) = \langle C, X\rangle + \frac{\varepsilon}{2}\|X\|_F^2 + \delta_{X\geq 0}(X)$ the dual is $D(y) = -\langle c_r, y\rangle - \frac{1}{2\varepsilon}\|[-C - \mathcal{K}_r^* y]_+\|_F^2$, and for $X \geq 0$ the Fenchel–Young residual $\mathrm{FY}(X, y) := \frac{\varepsilon}{2}\|X\|_F^2 + \langle C + \mathcal{K}_r^* y, X\rangle + \frac{1}{2\varepsilon}\|[-C - \mathcal{K}_r^* y]_+\|_F^2 \geq 0$ satisfies $f(X) - D(y) = \mathrm{FY}(X, y) - \langle y, \mathcal{K}_r X - c_r\rangle$ (verified numerically at random points). This polynomial expression is a nonnegative gap only on $X \geq 0$ (for $\varepsilon = 1$, $d = C + \mathcal{K}_r^* y = 1$, $X = -0.1$ it equals -0.095), and some iterates have negative entries: fixed-α PAPF carries the warm-start perturbation of the small instance ($\min_{ij} X_{ij}^k = -5.2 \cdot 10^{-8}$, decaying as α^k), and TKMA reaches $-9 \cdot 10^{-3}$, $-1.1 \cdot 10^{-3}$, and $-1.4 \cdot 10^{-3}$ on the three instances and ends with $X < 0$ on the two 50×50 instances. We therefore track, on every real iterate, the two KKT residuals

$$R_{\mathrm{stat}}(X, y) := \left\| X - \tfrac{1}{\varepsilon}[-C - \mathcal{K}_r^* y]_+ \right\|_F, \qquad R_{\mathrm{feas}}(X) := \|\mathcal{K}_r X - c_r\|,$$

which are defined and nonnegative for every real X and vanish simultaneously iff (X, y) satisfies the KKT conditions (the first forces $X \geq 0$, stationarity, and complementarity; the second the marginals). They are residuals, not bounds on the objective gap. Figure 8 shows them for exactly the configuration of Table 1 (fixed $\alpha = 0.058, 0.27, 0.15$; guarded adaptive rule with the default settings; T^2; Euclidean TKMA), on reruns whose recorded residuals and T-call counts are bitwise identical to the Table 1 runs. For a recorded trajectory $z^0, \ldots, z^N$ with cumulative T-call counts t_k we report, for each residual R and threshold η, the *terminal-sustained* count $H_R(\eta) = \min\{t_k : R(z^j) \leq \eta \text{ for all } j = k, \ldots, N\}$ ("never" if $R(z^N) > \eta$). It describes the recorded run only, and isolated early zeros do not count: the uniform cold start is exactly feasible, and on Iris the first PDHG step gives $X^1 = 0$, $y^1 = 0$, where $R_{\mathrm{stat}} = 0$ but $R_{\mathrm{feas}} = 0.199$. The joint count $\min\{t_k : \text{both residuals} \leq \eta \text{ for all } j \geq k\}$ is the larger of the two single-residual counts. At $\eta = 10^{-8}$, $H_{R_{\mathrm{stat}}}$ is 1408 vs. 3500 T-calls for PAPF vs. T^2 on the small instance, 12,400 vs. 18,042 on Iris, and 26,022 vs. 30,010 on the large instance, and $H_{R_{\mathrm{feas}}}$ is 740 vs. 2206, 8978 vs. 14,092, and 23,566 vs. 26,102; since R_{stat} is the later of the two in every run, its counts are also the joint KKT counts. TKMA ends its budget above both thresholds on all three instances. On these runs we measure $R_{\mathrm{stat}}/\|T(z) - z\|$ between 378 and 504 on the 50×50 instances (there $1/(\tau\varepsilon) = 504$), so under the Table 1 stopping rule those runs end at $R_{\mathrm{stat}} \approx 4 \cdot 10^{-10}$; only the small instance reaches 10^{-10} (2212 vs. 5010 for R_{stat}, 1530 vs. 3726 for R_{feas}). On the log axes exact zeros are clipped below the panel.

G.5 WALL-CLOCK MEASUREMENTS

Table 7 reports wall-clock timings on the Quad-OT $n{=}m{=}50$ instance (single thread, Intel Core i9-9880H at 2.3 GHz, MATLAB R2024a, double precision). The cost of one T-call on this instance is $\approx 20\,\mu$s. PAPF's overhead beyond the T-calls (the $\alpha I + (1-\alpha)w_2$ blend) is below the timing noise ($|\mathrm{overhead}| < 0.5\,\mu$s per T-call, i.e., $< 1\,\mu$s per two-call step). Anderson acceleration carries the cost of a regularized least-squares solve on the $m \times d$ history (Tikhonov-regularized normal equations in our implementation) and of the safeguard evaluation at every step (one T-call): $\approx 50\,\mu$s at $m{=}5$ and $\approx 100\,\mu$s at $m{=}10$, so even when AA wins on T-calls (§5.3), the wall-clock advantage erodes or

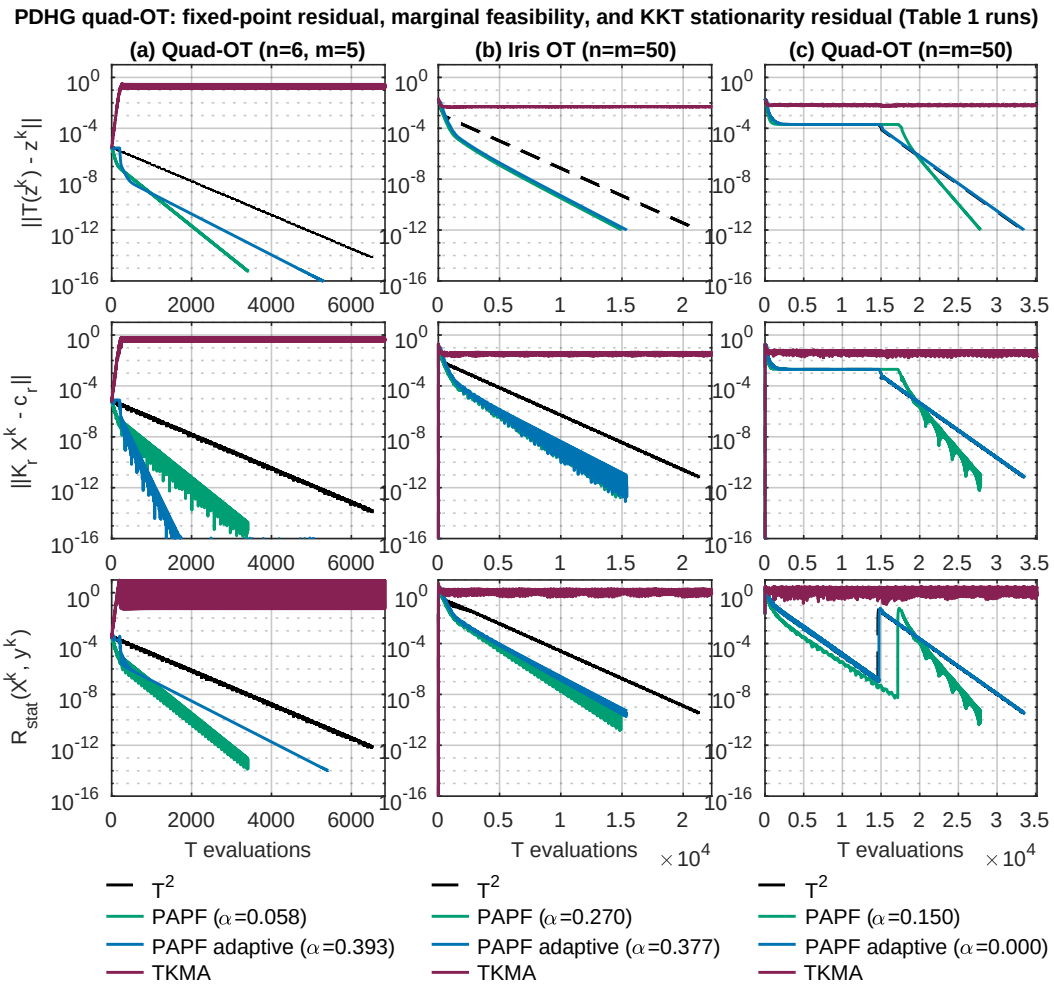

Figure 8: **KKT residuals on PDHG-OT under the Table 1 configuration.** Rows: fixed-point residual $\|T(z^k)-z^k\|$, marginal feasibility $R_{\text{feas}} = \|\mathcal{K}_r X^k - c_r\|$, and stationarity residual $R_{\text{stat}} = \|X^k - [-C - \mathcal{K}_r^* y^k]_+/\varepsilon\|_F$, all defined for every real iterate (exact zeros are clipped below the axes). Columns: Quad-OT (6×5, warm start), Iris OT, Quad-OT (50×50). PAPF's gains transfer to both residuals; TKMA stalls in all three.

reverses on cheap-T problems. Halpern and Inertial carry little per-step overhead but their stalling dominates this comparison. The counts in this table come from the timing harness (median of three single-threaded runs at the Table 1 configuration); TKMA, Halpern, and inertial hit their step budgets.

Wall-clock on the power-method instance. The OT operator above is cheap but not trivial ($\approx 20\,\mu$s per T-call). Table 8 repeats the measurement on the 8×8 power-method instance of Table 1, where a bare T-call costs $0.35\,\mu$s (interpreter floor $0.02\,\mu$s), so per-step overheads dominate. Because a complete run lasts between 0.05 and 10 ms, each timing sample covers 1000 back-to-back complete runs (same machine, single thread, one warm-up, median of five samples; the T-call counts are asserted identical across samples and equal to Tables 1 and 3). AA(m=10) wins on T-calls (85 vs. 118 for fixed PAPF) but spends $7.1\,\mu$s per T-call in its least-squares solve and safeguard, against $0.46\,\mu$s for fixed PAPF and $0.65\,\mu$s for the adaptive rule (burn-in, search, and acceptance included): in wall-clock, fixed PAPF is $11\times$ and the adaptive rule $2.5\times$ faster than AA(m=10), and AA(m=5) is slower than T^2. This is the cheap-T regime in which PAPF's memory-1 structure pays; on expensive operators the T-call count decides, as in Table 7.

Table 7: Wall-clock on Quad-OT $n{=}m{=}50$. "Overhead" = (time per T-call) $-$ (T^2 time per T-call), in μs per T-call (PAPF and TKMA steps use two T-calls; AA, Halpern, and inertial one); positive means costlier than the bare PDHG iteration.

Method	T-calls	Wall (s)	μs / T-call	Overhead μs / T-call
T^2	33 356	0.681	20.41	0.00
PAPF ($\alpha{=}0.15$)	27 884	0.568	20.39	-0.02
PAPF (adaptive)	33 496	0.668	19.96	-0.45
TKMA	100 000	2.205	22.05	1.64
AA ($m{=}5$)	61 499	4.344	70.63	50.22
AA ($m{=}10$)	60 031	6.978	116.23	95.83
Halpern	50 000	1.202	24.05	3.64
Inertial	145 758	3.098	21.26	0.85

Table 8: Wall-clock on the power-method instance ($8{\times}8$, tolerance 10^{-14}; one run = full solve; batched timing as described in the text). "Overhead" = (time per T-call) $-$ (T^2 time per T-call).

Method	T-calls	Wall (ms)	μs / T-call	Overhead μs / T-call
T^2	3 564	1.402	0.39	0.00
Relaxed KM ($\beta{=}0.49$)	245	0.142	0.58	0.18
PAPF ($\alpha{=}0.378$)	118	0.054	0.46	0.07
PAPF (adaptive)	368	0.241	0.65	0.26
TKMA	6 000†	3.203	0.53	0.14
AA ($m{=}5$)	1 587	9.365	5.90	5.51
AA ($m{=}10$)	85	0.606	7.12	6.73
Halpern	3 000†	1.570	0.52	0.13
Inertial	7 823†	3.672	0.47	0.08

G.6 Adaptive-parameter ablation, residual proxy, and transient growth

Call accounting. An instrumented run (a counting wrapper around T; the boundary-sweep instance with $\theta = \theta_b/2$, tolerance 10^{-10}, default settings, guard on) confirms Section 4: 60 burn-in T-calls; 14 probes of exactly 10 T-calls each (2 initial, 10 contractions, 2 acceptance), i.e., 140 search T-calls; and PROBE returns the residual of the iterate after $K_{\text{probe}} - 1 = 4$ updates ($2.8898 \cdot 10^{-2} = r_4$ of the probe sequence $r_0, \dots, r_5$ on that run; r_5 would need one more T-call). If the burn-in reaches the tolerance, the routine returns without probing (four T-calls when $r_0 = 1.3\,\text{tol}$), and $z^0 = z^\star$ returns after one step; no quantity is divided by a residual.

What the probe residual measures. Although the proxy of Section 4 is the root rate $\hat{q}_k(\alpha)$, PROBE compares the *absolute* probe residuals $r_K(\alpha)$ from a shared starting point $\bar{z}$ ($K = K_{\text{probe}} - 1 \geq 1$ filtered updates, i.e., $K_{\text{probe}} \geq 2$; Algorithm 1). Since $r_0 = \|T(\bar{z}) - \bar{z}\| > 0$ is the same for all probes and $x \mapsto (x/r_0)^{1/K}$ is strictly increasing, for every finite K

$$\arg\min_\alpha r_K(\alpha) = \arg\min_\alpha \big(r_K(\alpha)/r_0\big)^{1/K},$$

i.e., the probe ranks parameters by their *empirical* root rate over the probe window. Under the hypotheses of Proposition 4 (locally affine T, nonzero dominant Riesz component) one also has, for each fixed α with $q(\alpha) > 0$, $(r_K(\alpha)/r_0)^{1/K} \to q(\alpha)$ and $\log r_K(\alpha) = K \log q(\alpha) + o(K)$ as $K \to \infty$. No stronger statement is claimed: the prefactor in $r_K(\alpha)$ is not r_0 in general (for $G = \text{diag}(0.9, 0.5)$, $u = (1,1)$, $\alpha = 0$ one has $r_K/0.81^K \to 0.1 \neq r_0 = \sqrt{0.26}$), Jordan blocks add polynomial factors, and pointwise convergence for each α does not imply that the minimizers of the finite-K probe objective converge to $\alpha_G^\star$; the search with $K_{\text{probe}} = 5$ is a heuristic whose failure modes are documented in Section 5.4 and Remark 14. Using r_K rather than a single per-step ratio reduces the sensitivity to any one ratio inside short probe windows; it does not remove finite-window amplitude effects or transients.

A finite-window selection bound. The following elementary bound makes the previous paragraph quantitative for exact-argmin selectors such as the uniform grid of App. G.7 and for the acceptance test. It is a conditional explanation of when short probes can be trusted, not a certificate: its constants are not observable.

Proposition 8 (Finite-window selection bound, affine case)**.** *Let $T(z) = z^\star + G(z - z^\star)$ with $G = V\Lambda V^{-1}$ diagonalizable, eigenvalues λ_j, and $\rho(G) < 1$, and let $\bar{z} \neq z^\star$. Put $v_0 := (G - I)(\bar{z} - z^\star)$, $c := V^{-1}v_0$, and $\kappa := \|V\|\,\|V^{-1}\|$. For a finite candidate set $\mathcal{A} \subset [0, 1)$ let J be the set of indices j such that $|p_\alpha(\lambda_j)| = q(\alpha)$ for some $\alpha \in \mathcal{A}$, and $\gamma := \min_{j\in J} |c_j|/\|c\|$. For $K \geq 1$ let $r_K(\alpha) := \|p_\alpha(G)^K v_0\|$, which is the residual $\|T(z) - z\|$ after K filtered updates $z \leftarrow \alpha z + (1 - \alpha)T(T(z))$ from $\bar{z}$, and let $\hat{\alpha} \in \arg\min_{\alpha\in\mathcal{A}} r_K(\alpha)$. If $\gamma > 0$, then*

$$q(\hat{\alpha}) \;\leq\; (\kappa/\gamma)^{1/K} \min_{\alpha\in\mathcal{A}} q(\alpha). \tag{34}$$

If $\mathcal{A}$ is a grid on $[0, \frac{1}{2}]$ with spacing h containing 0, then $\min_{\mathcal{A}} q \leq \min_{[0,1]} q + (1 + \rho^2)h/2$.

Proof. Since $G-I$ commutes with $p_\alpha(G)$, the residual after K updates from $\bar{z}$ is $(G-I)p_\alpha(G)^K(\bar{z}-z^\star) = p_\alpha(G)^K v_0$; $v_0 \neq 0$ because $G - I$ is invertible. Write $p_\alpha(G)^K v_0 = Vw$ with $w_j = p_\alpha(\lambda_j)^K c_j$. Then $r_K(\alpha) \leq \|V\|\,\|w\| \leq \|V\|\,q(\alpha)^K\|c\|$ for every α, while for an index $j^\star \in J$ attaining $q(\hat{\alpha})$, $r_K(\hat{\alpha}) \geq \|V^{-1}\|^{-1}\|w\| \geq \|V^{-1}\|^{-1}|w_{j^\star}| \geq \|V^{-1}\|^{-1}q(\hat{\alpha})^K\gamma\|c\|$. Combining with $r_K(\hat{\alpha}) \leq r_K(\alpha)$ for a minimizer α of q on $\mathcal{A}$ gives $q(\hat{\alpha})^K \leq (\kappa/\gamma)\min_{\mathcal{A}} q^K$. For the grid statement, each $|p_\alpha(\lambda_j)| = |\alpha(1 - \lambda_j^2) + \lambda_j^2|$ is $(1 + \rho^2)$-Lipschitz in α, hence so is q, and the minimizer of q on $[0, 1]$ lies in $[0, \rho^2/(1 + \rho^2)] \subset [0, \frac{1}{2}]$ (Lemma 3), within $h/2$ of a grid point. □

The bound isolates the three channels through which short probes fail: non-normality (κ), under-excitation of a mode that decides the factor of some candidate (γ), and the window length (K); a spectral loss of at most $1+\eta$ relative to the best candidate is guaranteed once $K \geq \log(\kappa/\gamma)/\log(1+\eta)$. At the default $K = 4$ it is far from a certificate: $\gamma \leq |J|^{-1/2}$ always, so a 10% guarantee at $K = 4$ is impossible whenever three or more modes decide candidate factors, and since $\bar{z}$ is the output of K_{burn} steps of T^2, the coefficient of a mode λ_j carries the factor $\lambda_j^{2K_{\mathrm{burn}}}$, which makes γ tiny for every subdominant mode that decides the factor of some candidate. The bound covers the grid selector (an exact argmin over $\mathcal{A}$) and the acceptance test: with $\mathcal{A} = \{0, \hat{\alpha}\}$ and $r_K(\hat{\alpha}) < (1 - \delta_{\mathrm{acc}})\, r_K(0)$ the same argument gives $q(\hat{\alpha}) \leq \big((1 - \delta_{\mathrm{acc}})\kappa/\gamma\big)^{1/K}\rho^2$, a bound on the harm an accepted candidate can do. It does not cover golden-section search, whose bracketing assumes a unimodal probe objective, it concerns the spectral factor only and not the finite-horizon cost from $\bar{z}$, and it does not extend to Jordan blocks (there r_K/q^K grows polynomially in K and no K-independent κ exists) or to nonlinear maps (Remark 13). Of the instances of Table 1 it applies to the warm-started small OT instance after active-set identification (App. C.4), not to the two 50×50 instances ($\rho(G) = 1$ on the full state) or to the renormalized power maps.

Illustration. On the instance of Remark 14 with the initial amplitude δ of the slow mode varied from 10^{-2} to 10^{-8} ($\kappa = 1$; the slow mode is the only factor-deciding one, $\gamma = 1.2 \cdot 10^{-2}\,\delta$), the uniform grid selects $\hat{\alpha} = 4/26$ at $K = 4$ for every δ (spectral loss 0.3%, bound factor 9.7 to 305) and first selects the spectral optimum $\alpha = 0$ at $K = 9, 10, 12, 14, 15, 17, 19$: the window needed to see the slow mode grows like $1.68\log_{10}(1/\delta)$, matching $\ln 10/\ln\big(q(0)/|p_0(\lambda_{\mathrm{pair}})|\big) = 1.685$, whereas the sufficient window of the bound is 96 to 241; the bound is a worst case, loose by one to two orders of magnitude here, and it is never violated. On the $\theta = \pi/2$ instance of Table 6, after the 30 burn-in steps the two real modes have relative amplitudes $4.8 \cdot 10^{-15}$ and 10^{-32} in the residual (bound factor $3.8 \cdot 10^3$ at $K = 4$ with γ over the factor-deciding modes); the grid's endpoint choice $\hat{\alpha} = 0.5$ has a 29% worse spectral factor than the grid optimum $\alpha = 6/26$ (0.676 vs. 0.523, decided by the real mode 0.6ρ) but reaches the tolerance in 12 T-calls from $\bar{z}$ against 74 for the spectral optimum, because over that horizon only the near-annihilated dominant pair matters. The spectral loss that the bound controls and the finite-horizon cost can thus disagree, and there the selector's choice is the right one for the run.

Global convergence of the retained trajectory. Every probe and the locked phase run S_α with $\alpha \in [0, \frac{1}{2}]$; Proposition 1 covers $\alpha \in (0, 1)$ and $\mathsf{S}_0 = T^2$ is averaged with $\mathrm{Fix}\, T^2 = \mathrm{Fix}\, T$. Probes are discarded, so for FNE T the retained trajectory is a finite burn-in followed by the KM iteration

of one fixed averaged map from $\bar{z}$, and Theorem 1 applies to it whatever $\hat{\alpha}$ is selected; convergence neither requires an accurate $\hat{\alpha}$ nor implies a speedup. In non-FNE settings Algorithm 1 is a local heuristic supported by Theorem 2.

Remark 14 (The acceptance test can accept an asymptotically worse parameter)**.** Let $G = \mathrm{diag}\big(0.99, \frac{1}{2}R(\pi/3)\big)$, which is the Jacobian of an FNE linear map, and $\bar{z} = (10^{-4}, 1, 0)$, $K_{\mathrm{probe}} = 5$. The slow real mode is nonzero but initially under-excited. The spectral factor is minimized at $\alpha = 0$: $q(0) = 0.9801 < q(1/7) = 0.98294$. Yet, with probe residuals measured after K_{probe} filtered updates, $r_5(1/7)/r_5(0) \approx 0.247 < 1 - \delta_{\mathrm{acc}}$ (and ≈ 0.327 with the residual measured inside the last update), so the candidate $\alpha = 1/7$ passes the acceptance test although it worsens the asymptotic factor. The example satisfies the Riesz condition of Proposition 4, so it is not excluded by that result. The test is therefore an empirical safeguard against the most common failure (a candidate that is worse than $\alpha = 0$ over the probe window), not a no-harm guarantee; longer probes or burn-ins do not uniformly remove the issue (Section 5.4).

Non-normality and transient growth. Local linear convergence is asymptotic: for non-normal G with significant pseudospectrum, $\|G^k\|$ can transiently overshoot $\rho(G)^k$ (the classical "hump") before settling, and short residual probes inside the hump can mismeasure $\hat{q}$. A burn-in can reduce some transient effects, but no fixed burn-in length guarantees asymptotic probing: its effect is instance-dependent and need not be monotone. In the unguarded power-method ablation below the tested outcomes follow K_{burn} modulo 3 (the dominant linearized phase repeats every three T^2 steps) rather than the burn-in length: $K_{\mathrm{burn}} = 100$ fails where 99 and 101 succeed. The burn-in length is therefore an empirical tuning choice. The ablation over $(K_{\mathrm{burn}}, K_{\mathrm{probe}}, \delta)$ on a non-normal power-method instance shows the grid points $K_{\mathrm{burn}} \in \{20, 30, 50\}$ recovering 7–17× speedup while $K_{\mathrm{burn}} \in \{10, 100\}$ with the shortest probes return $\hat{\alpha} < 0.01$; as shown below, the tested outcomes follow $K_{\mathrm{burn}} \bmod 3$ (the dominant phase) rather than a burn-in-length window, so we do not recommend a different default (the $(30, 5, 0.005)$ setting is used for every Table 1 result; Section 5.4 and this appendix vary the settings deliberately and say so), and under the perturbations of Figure 10 the selection tracks the per-s single-pair reference.

Memory complexity. PAPF stores only two extra iterates $w_1 = T(z^k)$, $w_2 = T(w_1)$ of dimension n, so the memory overhead is $O(n)$ on top of T's working set; the adaptive scheme additionally caches $\bar{z}$ and a constant number of probe scalars, also $O(n)$.

Figure 9 sweeps $(K_{\mathrm{burn}}, K_{\mathrm{probe}}, \delta) \in \{10, 20, 30, 50, 100\} \times \{3, 5, 10\} \times \{0.001, 0.005, 0.02\}$ without the acceptance test, under exactly the Table 1 configuration of the power-method instance (ρ=0.99, θ=$\pi/3$; residual $\le 10^{-14}$, renormalized; T^2 needs 3564 T-calls) and of the small Quad-OT instance (warm start, $\|z - z^\star\| \le 10^{-14}$ for every method; T^2 needs 6516). On the power method the speedup over T^2 ranges from 1.0× to 16.8× over the 45 cells (δ-averaged heatmap 1.2–15.6×; best cells $(20, 3)$, $(20, 5)$, $(50, 3)$). The failures at $K_{\mathrm{burn}} \in \{10, 100\}$ with $K_{\mathrm{probe}} = 3$ follow the phase of the dominant pair rather than the burn-in length: the dominant linearized phase repeats every three T^2 steps ($2\theta = 2\pi/3$ per step), and the tested probe outcomes show a corresponding dependence on K_{burn} modulo three, an observed pattern rather than an exact periodicity of the nonlinear search; every tested $K_{\mathrm{burn}} \equiv 1 \pmod 3$ (10, 19, 22, 28, 31, 49, 52, 100) returns $\hat{\alpha} < 0.01$ and 2200–3500 T-calls (depending on δ), whereas $K_{\mathrm{burn}} \equiv 2$ (20, 29, 32, 50, 98, 101) and $\equiv 0$ (18, 21, 30, 48, 51, 99, 102) give 230–540 at $\delta = 0.005$; longer probes damp but do not remove the pattern. The Quad-OT panel is flat (0.97–1.25×): the search hyperparameters matter little there. The guarded default $(30, 5, 0.005)$ with $\delta_{\mathrm{acc}} = 0.02$ used for Table 1 is not varied by this ablation. Figure 10 perturbs a separate 8×8 power-method matrix, $A(s) = D + sN_0$ with N_0 strictly upper triangular, $\|N_0\|_F = 1$, $s \in [0, 5]$ (tolerance 10^{-10}; adaptive rule without the acceptance test). Because the complex pair sits in a 2×2 block, the perturbation changes the spectrum as well as the nonnormality: $\rho(s)$ grows from 0.950 to 0.971 and the dominant angle from 1.047 to 1.060, so the references in the figure are recomputed from the true tangent spectrum at each s. The selected $\hat{\alpha}(s) \in [0.466, 0.487]$ tracks the single-pair closed form $\alpha^\star_{\mathrm{pair}}(s) \in [0.466, 0.481]$ rather than the full-spectrum minimax $\alpha^\star_G(s) \in [0.297, 0.309]$, because the subdominant real mode 0.6 is weakly excited within the probe window (the same mechanism as in Remark 14); it still reduces the T-calls from 448–764 for T^2 (growing with s) to 236–242. The right panel reports the tangent-space transient amplification $H_K(s) = \max_{0 \le k \le K} \|M(s)^k\|_2 / \rho(M(s))^k$ with $K = 200$ (unchanged at $K = 2000$), for $M = B(s) = Q^\top G(s) Q$, the tangent Jacobian of T ($G(s) = P_\star(s)A(s)/\lambda_1(s)$, Q an orthonormal basis of $(v^\star(s))^\perp$), and for the PAPF tangent Jacobian $\hat{\alpha}(s)I + (1 - \hat{\alpha}(s))B(s)^2$:

H_K grows from 1.00 at $s = 0$ ($A(0)$ is normal) to 6.9 (T) and 7.5 (PAPF) at $s = 5$, so the sweep does increase transient growth. H_K is an operator-norm quantity, not the amplification along any single trajectory.

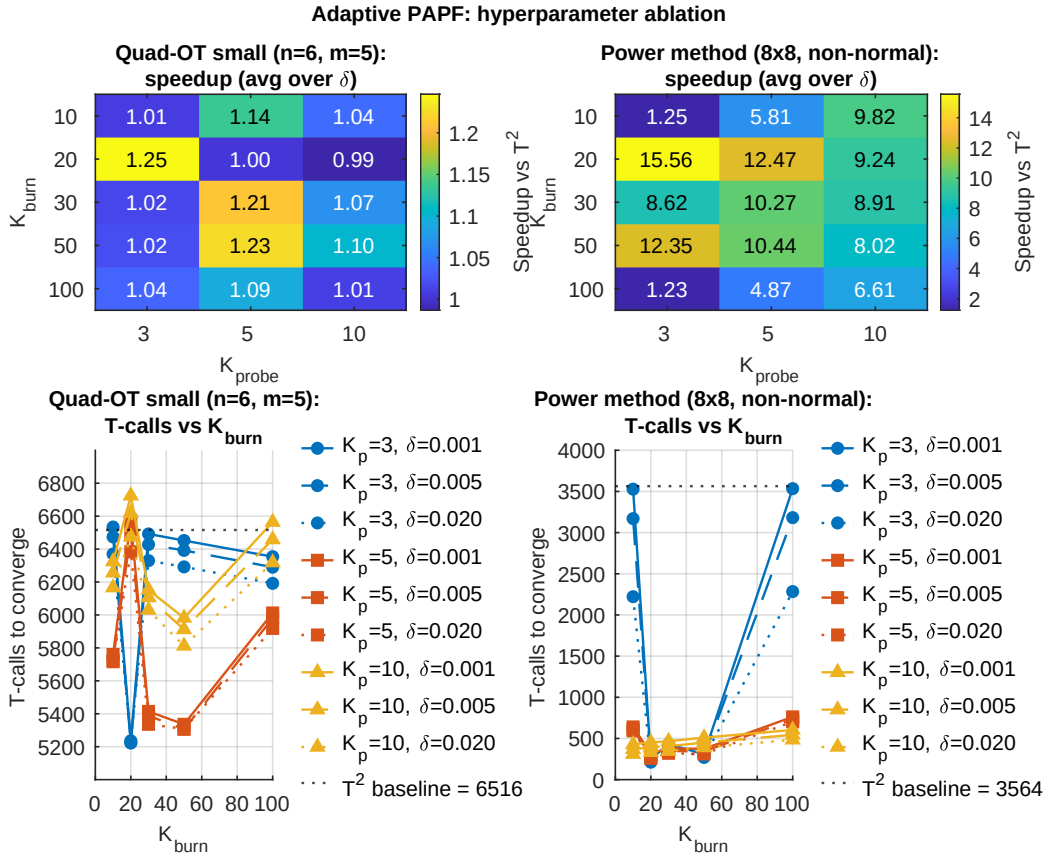

Figure 9: **Ablation over $(K_{\text{burn}}, K_{\text{probe}}, \delta)$ without the acceptance test.** Speedup over T^2 across $5\times3\times3$ configurations under the Table 1 configuration (power method: residual $\le 10^{-14}$, $T^2 = 3564$ T-calls; Quad-OT 6×5: warm start, $\|z - z^\star\| \le 10^{-14}$ for all methods, $T^2 = 6516$). The tested power-method outcomes follow $K_{\text{burn}} \bmod 3$ (the dominant linearized phase repeats every three T^2 steps) rather than K_{burn} monotonically; the Quad-OT panel is flat (0.97–1.25×).

G.7 Equal-budget parameter selectors

To separate the contribution of the golden-section component from that of the probe budget, Table 9 compares Algorithm 1 with two selectors of identical cost (60 burn-in plus 140 search T-calls, verified with a counting wrapper) under the Table 1 protocol: (i) a *uniform-grid* PAPF selector that probes the 14 equispaced values $\alpha_i = i/26$, $i = 0, \ldots, 13$, with the same $K_{\text{probe}} = 5$ probes, takes the smallest probe residual, and applies the same acceptance margin against the $\alpha = 0$ probe (which is one of the grid points); and (ii) an *adaptive relaxed KM* rule with the same structure (60 plain T steps of burn-in, golden-section search over $\beta \in [0, \frac{1}{2}]$ with probes of ten KM steps, two acceptance probes, then the locked KM iteration $\hat\beta z + (1 - \hat\beta)T(z)$). Golden section and the grid agree within 1.3% on eight of the nine instances and make identical acceptance decisions; the exception is the power method, where the five-step probe objective is bimodal (probe residuals 0.55 at $\alpha = 0.23$, 1.38 at 0.42, 0.50 at 0.5): golden section converges to the interior local minimum ($\hat\alpha = 0.247$, 368 T-calls) and the grid to the endpoint ($\hat\alpha = 0.5$, 306), while the best fixed $\alpha = 0.378$ (118) sits where the probe residual is large, so neither short-probe selector finds it. The adaptive KM rule wins on the three instances where relaxed KM is the better filter (Markov chain, PageRank, toy), ties on the large OT instance (all three rules fall back), and loses elsewhere; it falls back on Iris (probe ratio 0.9996)

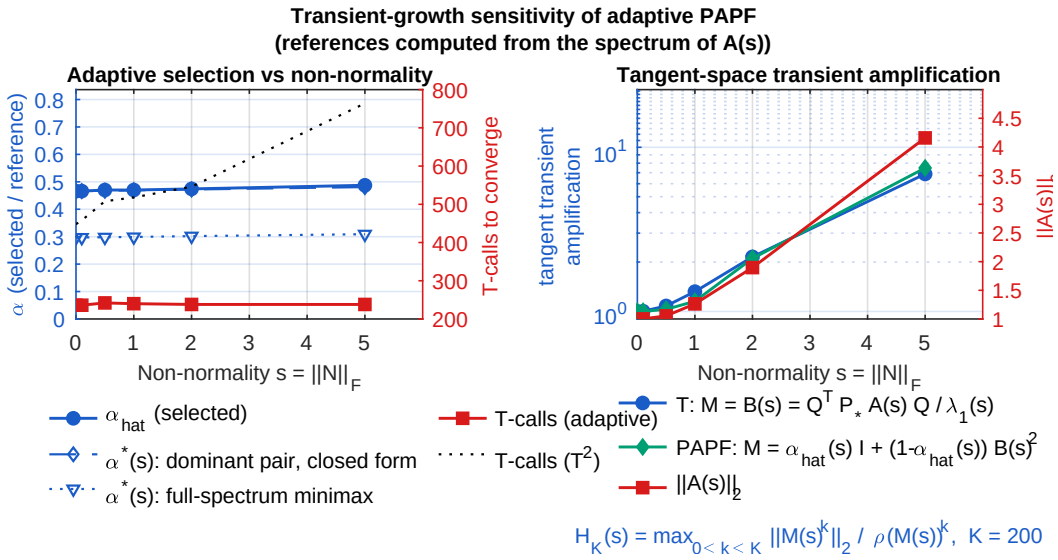

Figure 10: **Transient growth under non-normal perturbation.** $A(s) = D + sN_0$ (N_0 strictly upper triangular, $\|N_0\|_F = 1$), $s \in [0, 5]$, 8×8 power-method instance with $\rho(0) = 0.95$ (adaptive rule without the acceptance test). Left: selected $\hat{\alpha}(s)$ against the single-pair and full-spectrum references computed from the true spectrum of $A(s)$, and T-calls of the adaptive rule and of T^2. Right: tangent-space transient amplification $H_K(s)$ ($K = 200$) of the T and PAPF tangent Jacobians, and $\|A(s)\|_2$.

and never reaches its best fixed count, because the search cost dominates on the cheap instances and a ten-step probe cannot certify the small asymptotic gain of $\beta \approx 0.11$–0.13 on the OT instances. For the record, the Iris acceptance in Table 1 is marginal: its probe ratio is 0.9797 against the threshold 0.98, so $\delta_{\text{acc}} = 0.03$ would turn it into a fallback ($\approx 21{,}400$ T-calls).

Table 9: Equal-budget selectors under the Table 1 protocol: T-calls to tolerance with the selected parameter in parentheses (0 = fallback). Golden-section PAPF is the Table 1 adaptive column; best fixed KM is the Table 1 value.

Instance	T^2	Golden section (PAPF)	Uniform grid (PAPF)	Adaptive KM	Best fixed KM
Quad-OT (6×5)	6516	5408 (0.39)	5340 (0.39)	5640 (0.42)	3625 (0.11)
Quad-OT (50×50)	33356	33496 (0)	33496 (0)	33494 (0)	31008 (0.11)
Iris OT	21264	15392 (0.38)	15454 (0.39)	21403 (0)	16864 (0.13)
Power method	3564	368 (0.25)	306 (0.50)	510 (0.26)	245 (0.49)
Markov chain	1558	288 (0.49)	288 (0.50)	244 (0.49)	46 (0.49)
PageRank	1616	290 (0.49)	292 (0.50)	246 (0.49)	49 (0.49)
Synthetic $\pi/4$	3200	382 (0.50)	382 (0.50)	619 (0.33)	386 (0.48)
Toy linear	1404	326 (0.49)	330 (0.50)	319 (0.49)	157 (0.17)
Complex-dominant	2798	220 (0.49)	222 (0.50)	280 (0.49)	82 (0.50)

G.8 Repeated initializations and similarity transforms

Table 10 reruns T^2, the best fixed PAPF parameter of Table 1, and the complete guarded rule (default settings) from 20 random starts per instance under the Table 1 protocol (`rng(seed)`, seeds 1–20; operator fixed): Gaussian starts, normalized to the unit sphere for the power maps, for the power-method family and the two synthetic operators; the warm-start protocol $z^\star + 10^{-6}\xi$ with a fresh ξ for the small OT instance; and, at $n{=}m{=}50$, the perturbed uniform plan $X^0 \propto (1 + \frac{1}{2}u)$, $u \sim U(0, 1)$ entrywise, rescaled to unit mass, with $y^0 = 0$. For the two prescribed-spectrum instances (power 8×8 and synthetic $\pi/4$) it also redraws the similarity transform P twenty times under the generator's own rule, keeping the start (spectrum fixed, nonnormality varying). A selection is *harmful* if the

adaptive run needs more T-calls than T^2; tail rates are log-linear fits over the last $\min(200, \cdot)$ records of the locked phase, which on the fast instances is shorter than 50 records (10–46). The default starts reproduce Table 1 exactly.

Three observations. (i) On the power-method family, the synthetic operators, and all 40 similarity transforms there are 0 fallbacks, 0 harmful selections, and 0 accepted runs with a worse tail rate than T^2 in 160 runs; $\hat{\alpha}$ varies with the start on the two instances with several subdominant modes (power 0.18–0.50, synthetic $\pi/4$ 0.16–0.50), but the adaptive T-calls stay within 290–424 and 370–592 against 3000–3600 for T^2. (ii) The warm-started small OT instance is the weak spot: every start accepts an $\hat{\alpha} \in [0.34, 0.50]$ far from the spectral optimum 0.058, because at warm-start distance the five-step probe residual is governed by the fast modes, which a large α damps, while the tail rate at $\alpha \approx 0.49$ equals that of T^2; the rule is neutral in the median (1.03×, range 0.98–1.29×; fixed α: 1.89× in every run), 8 of 20 selections are harmful (worst +2.3%), 5 have a strictly worse tail rate (by at most 0.01%), and the Table 1 start (1.20×) is the second most favorable of the 21. (iii) At $n{=}m{=}50$ the perturbed primal starts change the trajectories by only $\approx 3 \cdot 10^{-4}$ over the first 100 T-calls (the first PDHG step moves the primal by 0.02 regardless of X^0), so all 20 large-OT starts give exactly the Table 1 counts and the Iris runs vary by at most 2 T-calls for T^2 and by 15,390–15,558 for the adaptive rule, with the acceptance ratio in $[0.9793, 0.9798]$ against 0.98 in every start. With a random dual start $y^0 \sim N(0, I)$ instead (supplementary rows; $\|y^0\| = 9.0$–11.4 against $\|y^\star\| = 3.51$ and ≈1.00, i.e., 2.6–3.2× and 9.0–11.4×), the third T^2 step clips the primal iterate entirely to zero and it stays zero for the next 86–141 and 72–146 steps (medians 108 and 112), during which the residual sits at the constant $\sigma\|c_r\| = 0.01975$ (to 10^{-14}); the burn-in and the probes fall inside this phase, so the probe residual is the same for every α up to rounding ($3 \cdot 10^{-15}$), the acceptance ratio is exactly 1, and the guard falls back in all 40 runs, at a cost of 140 T-calls (0.35–0.50% and 0.59–0.66% of the T^2 count) and forgoing the median fixed-α gains of 1.18× and 1.26×.

Table 10: Twenty random starts per instance (rows 1–9), twenty random similarity transforms (rows 10–11), and the supplementary random-dual starts at $n{=}m{=}50$ (rows 12–13): median [min, max] T-calls (PAPF fixed: median), median per-run speedups, and counts of fallbacks, harmful selections, and accepted runs with a strictly worse tail rate than T^2.

Instance	T^2	PAPF fixed	Adaptive	Spd. ada / fixed	Fallback	Harmful	Worse tail
Quad-OT (6×5), warm start	6507 [6400, 6614]	3446	6301 [5092, 6642]	1.03 / 1.89	0/20	8/20	5/20
Quad-OT (50×50)	33356 [33356, 33356]	27884	33496 [33496, 33496]	1.00 / 1.20	20/20	20/20	0
Iris OT	21264 [21264, 21266]	14970	15411 [15390, 15558]	1.38 / 1.42	0/20	0	0
Power method	3249 [3078, 3432]	108	306 [292, 424]	10.5 / 29.8	0/20	0	0
Markov chain	1652 [1592, 1760]	96	293 [290, 298]	5.6 / 17.2	0/20	0	0
PageRank	1662 [1532, 1852]	98	294 [286, 304]	5.6 / 16.8	0/20	0	0
Synthetic $\pi/4$	3362 [3160, 3576]	193	413 [376, 452]	8.2 / 17.4	0/20	0	0
Toy linear	1405 [1310, 1434]	156	323 [268, 334]	4.4 / 9.0	0/20	0	0
Complex-dominant	2822 [2630, 2880]	32	220 [220, 222]	12.8 / 88.2	0/20	0	0
Power method, rand. P	3205 [3002, 3420]	106	296 [290, 370]	10.7 / 30.2	0/20	0	0
Synthetic $\pi/4$, rand. P	3223 [2998, 3428]	186	389 [370, 592]	8.3 / 17.4	0/20	0	0
Quad-OT (50×50), rand. y^0	34739 [27828, 39530]	29488	34879 [27968, 39670]	1.00 / 1.18	20/20	20/20	0
Iris OT, rand. y^0	21564 [21154, 23792]	16983	21704 [21294, 23932]	0.99 / 1.26	20/20	20/20	0