SMOOTH OR SEPARABLE? SPARSE BLOCK ACCELERATION FOR ENTROPY-REGULARIZED LINEAR PROGRAMS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study entropy-regularized linear programs with sparse affine constraints and compare two exact dual representations induced by whether a redundant normalization constraint is retained or eliminated. Eliminating it yields a partially separable sum-exp dual with sparse affine factors but unbounded curvature. Our main result is an accelerated randomized block method whose acceleration preserves sparse data-access cost. The key ingredient is a global entropy-specific Hessian–gap bound that remains valid for signed constraint matrices, admits computable block-overlap refinements, and provides deterministic curvature certificates before sampling. This geometry supports square-root importance sampling and an exact lazy implementation in which sparse arithmetic cost is weighted by local curvature rather than by a worst-block factor. We further introduce a gap-dependent curvature dimension that connects the generalized-smooth regime to the classical effective-rank intuition for coordinate methods. Retaining the normalization constraint yields a log-sum-exp dual with globally bounded higher derivatives, enabling gradient-regularized and cubic-regularized Newton methods. Experiments compare the resulting methods in terms of coordinate epochs, arithmetic operations, and wall-clock time, separating the effects of acceleration, touched-data sparsity, and problem-specific structure.

1 INTRODUCTION

Entropy-regularized linear programs (ELPs) with sparse constraints,

$$\min_{x\geq 0}\left\{\langle c,x\rangle+\gamma\sum_j x_j(\log x_j-1) : Ax=b,\ \mathbf{1}^\top x=1\right\},\qquad A\in\mathbb{R}^{m\times n},\ n\gg m, \tag{1}$$

carry a redundant simplex constraint: the affine system already fixes the total mass, in the sense that some $q\in\mathbb{R}^m$ has $A^\top q=\mathbf{1}$ and $\langle q,b\rangle=1$. Discrete optimal transport is the standard example – summing the row marginals makes one of the m equations redundant – and entropic regularization is what made large instances tractable, by turning them into a matrix scaling that alternating projection solves (Sinkhorn and Knopp, 1967; Cuturi, 2013; Altschuler et al., 2017). The matrix is sparse: for transport it is an incidence matrix with two nonzeros per column.

One primal problem, two duals. The redundancy gives *two* exact duals, on opposite sides of a trade. Keeping the simplex equation and dualizing gives the *log-sum-exp* objective; removing the redundant equation first gives the *sum-exp* objective:

$$\phi_{\mathrm{lse}}(\lambda)=-\langle b,\lambda\rangle+\gamma\log\sum_{j=1}^{n}\exp\!\big(\tfrac{a_j^\top\lambda-c_j}{\gamma}\big), \tag{2}$$

$$\phi_{\mathrm{se}}(\lambda)=-\langle b,\lambda\rangle+\gamma\sum_{j=1}^{n}\exp\!\big(\tfrac{a_j^\top\lambda-c_j}{\gamma}\big). \tag{3}$$

The logarithm normalizes the primal response onto the simplex, which bounds every derivative of ϕ_{lse} – Nesterov smoothing (Nesterov, 2005) – but couples all n factors through their sum, so an

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.

accelerated update is not local. Without it ϕ_{se} is a sum of exponentials of sparse affine functions: factor j depends only on the support of column j, so a coordinate update reads only incident data. What ϕ_{se} loses is a global curvature bound; its Hessian grows without limit and it is smooth only in the generalized sense, with curvature controlled by the optimality gap.

Why one wants the separable side, and what blocks it. The reason is spectral: when curvature is spread unevenly across coordinates, square-root importance sampling improves the complexity bound of a full-gradient method by up to $\sqrt{m/d_H}$, with $d_H = \operatorname{tr} H/\|H\|_2$ the effective rank of the Hessian (Allen-Zhu et al., 2016; Nesterov and Stich, 2017; Gasnikov et al., 2015; Tropp, 2015). That is $\sqrt{m}$ only when one eigenvalue dominates, and it is not: on the family we test the factor sits at 4.2 throughout $m \in [100, 1000]$, and on transport between 1.6 and 3.4 (Section I.1). The gain is thus moderate and concerns data access.

Each route then gives up one of the two properties that matter. On ϕ_{lse}, gradient and second-order steps read all n factors or cost $O(m^3)$, and an accelerated coordinate step must re-evaluate the normalizing sum at the extrapolated point, so sparsity is lost. On ϕ_{se} sparsity survives, and acceleration is available in the abstract gap-dependent class: the coordinate method of Lobanov et al. (2024) is not accelerated, and Lobanov (2026), concurrent with this work, accelerates coordinate methods under an (H_0, H_1) condition at the cost of an exact one-dimensional relaxation per update; neither accounts for the arithmetic of an update. The obstruction that remains is concrete: accelerated coordinate methods fix their sampling law and step sizes from curvature constants known *before* a block is drawn (Nesterov, 2012; Lee and Sidford, 2013; Fercoq and Richtárik, 2015; Allen-Zhu et al., 2016; Nesterov and Stich, 2017; Gasnikov et al., 2015), and ϕ_{se} has no finite such constants, while re-estimating curvature at the iterate needs the gap, which touches all n factors and destroys the sparse access one came for. Replacing a constant by a function value is the generalized-smoothness move (Zhang et al., 2020; Crawshaw et al., 2022; Li et al., 2023; Vankov et al., 2025; Gorbunov et al., 2025); what it lacks is the instantiation – constants read off A, b and γ, and an accounting that charges the nonzeros an update touches (Appendix L).

Second-order methods do not separate the two duals. Both are quasi-self-concordant – $M_{\mathrm{qsc}} \le \Delta_A/\gamma$ for ϕ_{lse} and ρ/γ for ϕ_{se} (Proposition 5.2) – so gradient-regularized Newton (GRN) (Doikov, 2026) runs on either; only ϕ_{lse} has the globally bounded third derivative on which cubic-regularized Newton (CRN) (Nesterov and Polyak, 2006) and the accelerated order-2 tensor method (Dvurechensky et al., 2024), with rate $\varepsilon^{-1/3}$, rely. The choice in our title is therefore between two kinds of *step*, a global one costing an $m \times m$ solve or a coordinate one reading only incident data, and since both run on ϕ_{se}, Section 5 compares them on one dual. Transport already uses both (Dvurechensky et al., 2018; Guo et al., 2020; Brauer et al., 2017; Tang et al., 2024; Tang and Qiu, 2024).

Smooth or separable? Smoothness is the property to exploit when the dual dimension is small enough for an $m \times m$ factorization and high accuracy is required: a second-order step costs $O(m^3)$, but its iteration count grows only as $\log \frac{1}{\varepsilon}$, and on our synthetic family this keeps the wall-clock advantage up to m between 400 and 1000. Separability is the property to exploit when m is larger and A is sparse: a coordinate step then costs only the data it reads, and beyond that crossover the coordinate methods are ahead (Table 1). Separability alone does not suffice when the blocks carry structure of their own: on transport the smooth route beats generic coordinate methods, and Sinkhorn, which minimizes exactly over a whole marginal block, is faster than both (Section 6).

The technical core is a curvature bound for ϕ_{se} that holds everywhere and is known before sampling. Writing $F := \phi_{\mathrm{se}} - \phi^\star_{\mathrm{se}}$ for the dual gap and fixing a partition $\mathcal{B}$ of the m dual coordinates into blocks, we prove that every block B admits

$$\nabla^2_{BB}\phi_{\mathrm{se}}(\lambda) \preceq \big(\kappa_B + \nu_B F(\lambda)\big) I_B \qquad \text{for every } \lambda, \tag{4}$$

where ∇^2_{BB} is the diagonal block of the Hessian indexed by B, κ_B the curvature that survives at a solution and ν_B the rate at which it grows with the gap; Theorem 3.1 gives both, and Section 3.2 gives bounds computable from A, b and γ. Neither constant depends on which block is drawn, so two sampling tables are built once, the step sizes follow a deterministic schedule, and the accelerated triple (u_k, v_k, y_k) can be stored so that one sampled block costs $d_B := |B| + \operatorname{nnz}(A_{B:})$ operations and no other coordinate is touched. A block is a set of dual coordinates drawn together, and the partition is an input: singletons are the default, while for transport the natural coarser choice is the

two marginal families, the source rows and the target rows, on which exact block minimization is the Sinkhorn half-step (Section 6).

Contributions. We give an accelerated randomized block coordinate method for ϕ_{se}, so that acceleration and sparse data access hold at once, and we build it on three things specific to the entropy dual rather than to a smoothness abstraction. First, (4) itself: its constants are computable from the data, it survives signed rows without any knowledge of the solution, and its growth coefficient is optimal in its class up to the factor θ. Second, that those constants are fixed *before* the block is drawn, which is what makes the lazy sparse representation exact and the second-moment terms of the accelerated analysis cancel. Third, an arithmetic accounting in which the data a block reads is weighted by that block's own curvature rather than by a worst-block factor; Appendix I.8 exhibits a family on which the two accountings differ polynomially. The leading arithmetic cost is $\widetilde{O}\big(r(\gamma\varepsilon)^{-1/2}\sum_B d_B\sqrt{\gamma\kappa_B}\big)$: every block contributes the data it reads, d_B, weighted by its own curvature, and $\gamma\kappa_B$ carries no γ. Here r bounds the distance from the starting point to a dual solution. Restarting in phases reaches the same guarantee with the sublevel diameter in place of r (Proposition L.1), and the only radius we can compute bounds that diameter, so running without restarts removes phase bookkeeping and not a factor in the rate.

2 Two exact duals

Assume strict primal feasibility and the redundancy certificate $A^\top q = \mathbf{1}$, $\langle q, b\rangle = 1$, and write $x_j(\lambda) = \exp((a_j^\top\lambda - c_j)/\gamma)$, $Z(\lambda) = \sum_j x_j(\lambda)$ and $p_j = x_j/Z$ for the unnormalized and normalized responses; Appendix A states the assumption formally and collects the moment identities of the two duals. Keeping the simplex constraint and dualizing gives (2), with softmax primal response $p(\lambda)$; dropping the redundant equation first gives (3), whose response is the same expression without the normalizing denominator. The two are related by an exact one-dimensional minimization along q. Neither is new: Genevay et al. (2016) state both for transport, the unnormalized one as the dual and the log-sum-exp one as the semi-dual, and run SAG on the first and SGD on the second. Which one an *accelerated block* method wants is open.

Proposition 2.1 (Exact link)**.** *Let $Z(\lambda) = \sum_j x_j(\lambda)$. Then*

$$\min_t \phi_{\mathrm{se}}(\lambda + tq) = \phi_{\mathrm{lse}}(\lambda) + \gamma, \qquad t^\star(\lambda) = -\gamma\log Z(\lambda),$$

and $x(\lambda + t^\star q) = p(\lambda)$. Writing $F_{\mathrm{se}} := \phi_{\mathrm{se}} - \phi_{\mathrm{se}}^\star$ and $F_{\mathrm{lse}} := \phi_{\mathrm{lse}} - \phi_{\mathrm{lse}}^\star$,

$$F_{\mathrm{se}}(\lambda) = F_{\mathrm{lse}}(\lambda) + \gamma\big(Z(\lambda) - 1 - \log Z(\lambda)\big), \tag{5}$$

$$F_{\mathrm{lse}}(\lambda) = \gamma\,\mathrm{KL}(x^\star\|p(\lambda)). \tag{6}$$

Restricted to the line $\lambda + tq$, (3) becomes $-t + \gamma e^{t/\gamma}Z(\lambda)$ up to a constant, and its stationary point gives the claim (Appendix A).

Equation (5) separates two different errors. The LSE gap is exactly the Kullback–Leibler divergence to the optimal primal point; the SE gap is that divergence *plus* a penalty $\gamma(Z - 1 - \log Z)$ for the response carrying the wrong total mass, which vanishes only at $Z(\lambda) = 1$. The two are therefore not interchangeable as accuracy metrics, and an experiment must say which it reports.

3 Curvature of the SE dual

Fix a partition $\mathcal{B}$ of the m dual coordinates. We write $H_{BB}(\lambda) = \nabla^2_{BB}\phi_{\mathrm{se}}(\lambda)$ for the block Hessian, $H^\star_{BB} := H_{BB}(\lambda^\star)$ for its value at a dual solution, $F := \phi_{\mathrm{se}} - \phi^\star_{\mathrm{se}}$ for the dual gap, and $\rho_B = \max_j \|A_{B,j}\|_2$ for the largest norm of a column restricted to B. Dual solutions need not be unique, but $H^\star$ is: every $\lambda^\star$ gives the same primal response $x^\star$, which strict convexity makes unique.

3.1 The Hessian–gap inequality

Theorem 3.1 (Entropy Hessian–gap domination)**.** *For every $\theta > 1$, every block B, and every λ,*

$$H_{BB}(\lambda) \preceq c_\theta H^\star_{BB} + \frac{\theta\rho_B^2}{\gamma^2}F(\lambda)I_B, \qquad c_\theta = \theta\log\tfrac{\theta}{\theta-1}. \tag{7}$$

In particular, for any $\bar{H}_B \geq \|H^\star_{BB}\|_2$ the choice $\kappa_B = 2\log 2\,\bar{H}_B$ and $\nu_B = 2\rho_B^2/\gamma^2$ satisfies (4)*. No sign assumption on A is required.*

Both terms come from a single scalar inequality. Write $x_j(\lambda) = x_j^\star e^{t_j}$ with $t_j = a_j^\top(\lambda - \lambda^\star)/\gamma$. The gap then has the exact Bregman form $F(\lambda) = \gamma \sum_j x_j^\star h(t_j)$, $h(t) = e^t - 1 - t$, while the block Hessian quadratic form at u is $\gamma^{-1}\sum_j x_j^\star e^{t_j}(A_{B,j}^\top u)^2$. The Hessian is weighted by e^t and the gap by $h(t)$, and the elementary inequality $e^t \leq \theta h(t) + c_\theta$ converts one into the other term by term; the gap-dependent piece is exactly the price of that conversion, and $\theta = 2$ gives $c_2 = 2\log 2$. No sign condition on A is needed, because what is bounded is the square $(A_{B,j}^\top u)^2$, and no bounded set, because the right-hand side grows with $F(\lambda)$. The growth coefficient is not improvable: any bound $\nabla^2_{BB}\phi_{\mathrm{se}}(\lambda) \preceq \Psi_B + \nu_B F(\lambda) I_B$ valid for every λ forces $\nu_B \geq \rho_B^2/\gamma^2$, so (7) is optimal in its class up to θ (Proposition B.1, Appendix B).

3.2 Computing the constants

Theorem 3.1 is stated via $H^\star$, which we do not know, but the method needs only a valid upper bound $\bar{H}_B$, and several are available without solving the problem. Universally $\bar{H}_B = \rho_B^2/\gamma$, since $\sum_j x_j^\star = 1$. If the rows are sign-oriented so that $A \geq 0$, then $\sum_j A_{ij} x_j^\star = b_i$ is known exactly and a Cauchy–Schwarz step over the nonzeros of each block column supplies a second bound, so one may take

$$\bar{H}_B = \tfrac{1}{\gamma}\min\Big\{\rho_B^2,\ \max_{i\in B}\big[b_i \max_{j:A_{ij}\neq 0}\omega_{Bj}A_{ij}\big]\Big\}, \quad \omega_{Bj} = |\{i \in B : A_{ij} \neq 0\}|, \tag{8}$$

where the overlap count ω_{Bj} says how many rows of the block a single factor couples. Neither term dominates: for singleton blocks $\omega = 1$ and the second always wins, collapsing to $\bar{H}_i = \|A_{i:}\|_\infty b_i/\gamma$, whereas for a block whose columns concentrate their mass on one row the factor ω_{Bj} can make it the larger by $\omega_{\max} = \max_j \omega_{Bj}$. For signed rows the same idea works from the row mean and the row range alone.

Corollary 3.2 (Range–moment bound for signed rows)**.** *For row i let $\alpha_i^- := \min_j A_{ij}$ and $\alpha_i^+ := \max_j A_{ij}$, counting the implicit zeros of sparse storage. Define*

$$\mu_i := (\alpha_i^- + \alpha_i^+)b_i - \alpha_i^-\alpha_i^+, \qquad \bar{H}_i^{\mathrm{rng}} := \mu_i/\gamma. \tag{9}$$

Then $0 \leq H^\star_{ii} \leq \bar{H}_i^{\mathrm{rng}} \leq \|A_{i:}\|_\infty^2/\gamma$, so $\kappa_i = 2\log 2\,\bar{H}_i^{\mathrm{rng}}$ and $\nu_i = 2\|A_{i:}\|_\infty^2/\gamma^2$ are computable without sign orientation and without knowing $x^\star$. If $\alpha_i^- = 0$ this recovers $\bar{H}_i = \|A_{i:}\|_\infty b_i/\gamma$. ▶

The proof is one line, from $(a - \alpha_i^-)(a - \alpha_i^+) \leq 0$ averaged against $x^\star$. Outside one structural subclass these bounds can be arbitrarily loose; nothing is lost on scaled indicator rows, $A_{ij} \in \{0, \rho_i\}$, where $A_{ij}^2 = \rho_i A_{ij}$ collapses the row's second moment to its mean and feasibility pins that to b_i, giving $H^\star_{ii} = \rho_i b_i/\gamma$ exactly. Incidence matrices, and hence transport, are of this form (Appendix I). Summed over singleton blocks, the two constants define the gap-dependent curvature dimension, which interpolates between the effective rank of $H^\star$ and the spread of the row magnitudes (Section I.1).

4 Accelerated block coordinate descent on the SE dual

The method applies to any convex C^2 function $f : \mathbb{R}^m \to \mathbb{R}$ with an attained minimum satisfying the block gap-growth (BGG) condition

$$0 \preceq \nabla^2_{BB} f(u) \preceq (\kappa_B + \nu_B F(u)) I_B, \qquad F = f - f^\star, \tag{10}$$

globally, with finite $\kappa_B, \nu_B \geq 0$; Theorem 3.1 supplies it for ϕ_{se}. Fix any $u^\star \in \arg\min f$, write $f^\star = f(u^\star)$, let $\mathcal{F}_k$ be the σ-algebra of the first k block draws, and let $U_B : \mathbb{R}^{|B|} \to \mathbb{R}^m$ be the zero-padding injection, so that $U_B \nabla_B f(u)$ is the block gradient viewed in $\mathbb{R}^m$. Blocks with $\kappa_B = \nu_B = 0$ are dropped as exact (Appendix C).

Set $C_0 = \sum_B \sqrt{\kappa_B}$ and $C_1 = \sum_B \sqrt{\nu_B}$: C_0 multiplies the accuracy-dependent term and C_1, which measures how fast curvature inflates, carries no negative power of ε. We assume $C_0 > 0$; otherwise $F \equiv 0$ by a Gronwall argument.

Algorithm 1 Accelerated block step with precomputed curvature bounds

Require: anchor z; $\bar{\Delta} \geq F(z)$; $r \geq \|z - u^\star\|$; failure level δ; target ε; alias tables for $\pi^{(0)}$ and, if $C_1 > 0$, $\pi^{(1)}$.

1: If $C_0 = 0$, $r = 0$, $\bar{\Delta} = 0$, or $\varepsilon \geq \bar{\Delta}$, **return** z.
2: $T_0 \leftarrow r^2/(2\bar{\Delta})$, $u_0 = v_0 \leftarrow z$, $\Pi \leftarrow r^2/\delta$.
3: **for** $k = 0, 1, \ldots$ **do**
4: $\quad \Delta_k \leftarrow \Pi/T_k$ ▷ gap level the potential guarantees
5: $\quad M_k \leftarrow C_0 + \sqrt{5\Delta_k}\, C_1$
6: $\quad$ choose $a_{k+1} > 0$ with $M_k^2 a_{k+1}^2 = T_k + a_{k+1}$; $T_{k+1} \leftarrow T_k + a_{k+1}$
7: $\quad y_k \leftarrow (T_k u_k + a_{k+1} v_k)/T_{k+1}$
8: $\quad$ with probability C_0/M_k draw $B_k \sim \pi^{(0)}$, else draw $B_k \sim \pi^{(1)}$
9: $\quad \sqrt{\bar{L}_{k,B_k}} \leftarrow \sqrt{\kappa_{B_k}} + \sqrt{5\nu_{B_k}\Delta_k}$; $p_{k,B_k} \leftarrow \sqrt{\bar{L}_{k,B_k}}/M_k$
10: $\quad g \leftarrow \nabla_{B_k} f(y_k)$; pick $u_{k+1} \in y_k + \operatorname{range} U_{B_k}$ with $f(u_{k+1}) \leq f(y_k - U_{B_k} g/\bar{L}_{k,B_k})$
11: $\quad v_{k+1} \leftarrow v_k - (a_{k+1}/p_{k,B_k}) U_{B_k} g$
12: $\quad$ **if** $T_{k+1} \geq \Pi/\varepsilon$ **then**
13: $\quad\quad$ **return** u_{k+1}
14: $\quad$ **end if**
15: **end for**

Ideally B is sampled with probability proportional to $\sqrt{\kappa_B + 5\nu_B\Delta_k}$, where Δ_k is the gap level the potential currently guarantees (Lemma 4.1). The separable surrogate $\sqrt{\kappa_B} + \sqrt{5\nu_B\Delta_k}$ loses at most a factor $\sqrt{2}$ and is a mixture of *two fixed distributions* $\pi_B^{(0)} \propto \sqrt{\kappa_B}$ and $\pi_B^{(1)} \propto \sqrt{\nu_B}$, so two alias tables built once in linear time (Vose, 1991) give $O(1)$ expected sampling with no per-iteration scan.

Algorithm 1 realizes the mixture exactly: every block's curvature bound is determined before the draw, and only the sampled block's constant and gradient are evaluated. The constant 5 is a safety margin over $F(y_k) \leq 4\Delta_k$. The explicit step always satisfies line 10, and so does exact block minimization; on the event $\Phi_k \leq \Pi$ (the potential defined below), Lemma 4.1 turns line 10 into the decrease $\|g\|_2^2/(2\bar{L}_{k,B_k})$.

A bound on $F(u_k)$ alone says nothing about the curvature at the extrapolated point y_k; the potential controls both the gap at u_k and the displacement of v_k from a minimizer.

Lemma 4.1 (Pre-sampling block model)**.** *Assume* (10). *Let $T, a, \Pi > 0$, put $T_+ = T + a$ and $\Delta = \Pi/T$, and let $M = C_0 + \sqrt{5\Delta}\, C_1$ satisfy $M^2 a^2 = T_+$. If*

$$TF(u) + \tfrac{1}{2}\|v - u^\star\|_2^2 \leq \Pi, \qquad y := \frac{Tu + av}{T_+}, \tag{11}$$

then $F(y) \leq 4\Delta$. Consequently every block bound $\sqrt{\bar{L}_B} := \sqrt{\kappa_B} + \sqrt{5\nu_B\Delta}$ is valid at y,

$$\nabla_{BB}^2 f(y) \preceq \bar{L}_B I_B, \tag{12}$$

and $y - U_B \nabla_B f(y)/\bar{L}_B$ decreases f by at least $\|\nabla_B f(y)\|_2^2/(2\bar{L}_B)$. ▶

Every quantity in (12) is $\mathcal{F}_k$-measurable at iteration k, so the whole sampling law of line 8 is determined by the potential rather than by the draw.

Theorem 4.2 (One accelerated phase)**.** *Assume* (10) *with $C_0 > 0$, and let the inputs of Algorithm 1 satisfy $\bar{\Delta} \geq F(z)$, $r \geq \|z - u^\star\|_2$ for some minimizer $u^\star$, $\delta \in (0, 1)$ and $0 < \varepsilon < \bar{\Delta}$. Let N be the first index with $T_N \geq \Pi/\varepsilon$. Then $\mathbb{P}(F(u_N) \leq \varepsilon) \geq 1 - \delta$ and*

$$N = \widetilde{O}\left(1 + \frac{rC_0}{\sqrt{\delta\varepsilon}} + \frac{rC_1}{\sqrt{\delta}}\right). \tag{13}$$

▶

The proof runs the potential $\Phi_k = T_k F(u_k) + \frac{1}{2}\|v_k - u^\star\|^2$. On the event $\Phi_k \leq \Pi$, the point y_k lies in $\{F \leq 5\Delta_k\}$, where the block model is valid, and this is known before the draw, so no segment solve or gap evaluation is needed. The quadratic terms then cancel because $a_{k+1}^2/p_{k,B} = T_{k+1} p_{k,B}/\bar{L}_{k,B}$, which is how a_{k+1} is chosen in line 6, and Doob's maximal inequality for the stopped supermartingale turns $\Phi_0 \leq r^2 = \delta\Pi$ into the failure bound (Appendix C). Since the schedule T_k does not depend on the draws, N is deterministic and the expected work is $\sum_{k<N} \mathbb{E}[d_{B_k}]$.

Arithmetic cost. Let d_B be the cost of touching block B and set $S_0 = \sum_B d_B\sqrt{\kappa_B}$, $S_1 = \sum_B d_B\sqrt{\nu_B}$: the same two channels, weighted by the data each block reads. Low effective dimension and low access cost are different advantages, and the pair $(C_\bullet, S_\bullet)$ keeps them apart (Table 3, Appendix C).

Theorem 4.3 (Sparse arithmetic work)**.** *Assume the hypotheses of Theorem 4.2 for ϕ_{se}, with $C_0 > 0$ and $0 < \varepsilon < \bar{\Delta}$. Run $J_{\mathrm{rep}} := \lceil \log(2\bar{\Delta}/\varepsilon)/\log 3 \rceil$ independent copies of Algorithm 1 from the* same anchor z*, each with failure level $1/3$ and target $\varepsilon/2$, and return the point $\widehat{u}$ of smallest objective among z and the outputs. Then $\mathbb{E}F(\widehat{u}) \le \varepsilon$, and with $\chi := \min\{\sqrt{2\bar{\Delta}}/C_0,\, 1/C_1\}$ for $C_1 > 0$ and $\chi := 0$ otherwise,*

$$\mathbb{E}W_{\mathrm{opt}} = \widetilde{O}\left(\frac{S_0}{C_0} + \frac{rS_0}{\sqrt{\varepsilon}} + rS_1 + \chi S_1\right), \tag{14}$$

including phase-end comparisons, best-candidate storage and resets, on top of $W_{\mathrm{init}} = O(\mathrm{nnz}(A) + m + n)$ for anchor projections, constants and alias tables, and again for a full primal response.

The repetitions turn the $1/\sqrt{\delta}$ of (13) into a logarithm; the anchor stays fixed because a better objective value at a new anchor does not imply a smaller distance to $u^\star$. The same repetitions also give $\mathbb{P}(F(\widehat{u}) \le \varepsilon/2) \ge 1 - 3^{-J_{\mathrm{rep}}}$.

Only $rS_0/\sqrt{\varepsilon}$ carries a negative power of ε, while rS_1 scales as $1/\gamma$: on the balanced family of Proposition I.4 the ratio is $\sqrt{m\varepsilon/(\gamma s_{\mathrm{col}} \log 2)}$, so the growth term takes over once $\varepsilon \gtrsim \gamma s_{\mathrm{col}}/m$, and the leading cost quoted in Section 1 is the small-ε face of Theorem 4.3. For signed rows Corollary 3.2 replaces κ_i by the range–moment constant throughout.

4.1 The lazy sparse representation

The representation below is available only because the constants are fixed in advance. Inside one phase, write $v_k = z + \tilde{v}_k$ and $w_k = T_k(u_k - v_k)$; the interpolation point then factors as $y_k = z + \tilde{v}_k + w_k/T_{k+1}$, and the two state vectors obey, with $\xi = a_{k+1}/p_{k,B_k}$,

$$\tilde{v}_{k+1} = \tilde{v}_k - \xi U_{B_k} g, \qquad w_{k+1} = w_k + T_{k+1}\big(\xi U_{B_k} g + \eta_k\big), \qquad \eta_k := u_{k+1} - y_k. \tag{15}$$

The displacement η_k is supported on B_k by the restriction in line 10, and equals $-U_{B_k} g/\bar{L}_{k,B_k}$ for the explicit step, so both updates change only the sampled coordinates whichever admissible u_{k+1} is taken. Maintaining the three projections $\zeta^0 = A^\top z$, $\zeta^v_k = A^\top \tilde{v}_k$ and $\zeta^w_k = A^\top w_k$ lets the sampled gradient be read off as

$$(\nabla_{B_k}\phi_{\mathrm{se}}(y_k))_i + b_i = \sum_{j : A_{ij} \ne 0} A_{ij} \exp\!\big(\tfrac{\zeta^0_j + \zeta^v_{k,j} + \zeta^w_{k,j}/T_{k+1} - c_j}{\gamma}\big), \tag{16}$$

touching only the rows of B_k and the primal factors incident to them. Each projection update is supported on $\mathrm{supp}(A_{B_k:})$, so one sampled block costs $O(d_B)$ with $d_B = |B| + \mathrm{nnz}(A_{B:})$. Phase safeguards are sparse too: the objective difference from the anchor is evaluable over the visited coordinates alone, so accepts, rejects and resets add no hidden $O(m)$ or $O(n)$ rescan (Appendix D).

5 The LSE route

Retaining the normalization constraint destroys separability and buys global smoothness of every order. With $p(\lambda)$ the softmax response,

$$\nabla\phi_{\mathrm{lse}}(\lambda) = Ap(\lambda) - b, \qquad \nabla^2\phi_{\mathrm{lse}}(\lambda) = \tfrac{1}{\gamma} A(\mathrm{Diag}\, p - pp^\top)A^\top,$$

and two things change. The gradient norm *is* the primal infeasibility of the normalized response; $\nabla\phi_{\mathrm{se}} = Ax(\lambda) - b$ is the residual of the unnormalized one, so the SE *coordinate* method has a stopping criterion but no cheap way to maintain it. And the Hessian is a covariance rather than a second moment, so every derivative is bounded on the subspace $\mathcal{H}$ orthogonal to the flat directions of ϕ_{lse}, and each fixed order ℓ is Lipschitz with $O_\ell(\rho^{\ell+1}/\gamma^\ell)$ (Appendix E).

Theorem 5.1 (Global LSE geometry). *Let $\Delta_A := \max_{j,k} \|a_j - a_k\|_2$. Then for every λ and all u, v, w,*

$$\|\nabla^2 \phi_{\mathrm{lse}}(\lambda)\|_2 \le \frac{\Delta_A^2}{4\gamma}, \qquad |\nabla^3 \phi_{\mathrm{lse}}(\lambda)[u,v,w]| \le \frac{\Delta_A^3}{4\gamma^2}\|u\|_2\|v\|_2\|w\|_2,$$
$$|\nabla^3 \phi_{\mathrm{lse}}(\lambda)[u,u,v]| \le \frac{\Delta_A}{\gamma}\big(u^\top \nabla^2 \phi_{\mathrm{lse}}(\lambda) u\big)\|v\|_2. \tag{17}$$

So ϕ_{lse} has $L_{\mathrm{lse}} \le \Delta_A^2/(4\gamma)$, is quasi-self-concordant with $M_{\mathrm{qsc}} \le \Delta_A/\gamma$, and has $M_2 \le \Delta_A^3/(4\gamma^2)$. ▶

Since $\Delta_A \le 2\rho$ with $\rho = \max_j \|a_j\|_2$, the coarser column-norm constants $L_{\mathrm{lse}} \le \rho^2/\gamma$, $M_{\mathrm{qsc}} \le 2\rho/\gamma$ and $M_2 \le 2\rho^3/\gamma^2$ follow and are what the exponent comparisons use. This is Nesterov smoothing (Nesterov, 2005). A single non-accelerated LSE coordinate update can be sparse if the weights and their sum are maintained incrementally; an accelerated one cannot, since the sum is needed at the extrapolated point. The property that decides which dual a second-order method needs is the following.

Proposition 5.2 (The SE dual is quasi-self-concordant). *For every λ and all u, v,*

$$\big|\nabla^3 \phi_{\mathrm{se}}(\lambda)[u,u,v]\big| \le \frac{\max_j |a_j^\top v|}{\gamma}\,\nabla^2 \phi_{\mathrm{se}}(\lambda)[u,u] \le \frac{\rho\|v\|_2}{\gamma}\,\nabla^2 \phi_{\mathrm{se}}(\lambda)[u,u],$$

so $M_{\mathrm{qsc}}(\phi_{\mathrm{se}}) \le \rho/\gamma$, no worse than the column-norm bound $2\rho/\gamma$ for ϕ_{lse}. ▶

Gradient-regularized Newton therefore applies to ϕ_{se} directly, with iterates confined to $z + \operatorname{range} A$ where the SE sublevel sets are bounded (Appendix J); matrix scaling is Doikov's own worked example of a quasi-self-concordant objective of this form (Doikov, 2026). What ϕ_{se} lacks is a global Lipschitz Hessian, so tensor and gradient-norm methods still require ϕ_{lse}. Neither method we run on this dual is ours; Theorem 5.1 contributes the moment representation and the constants that place the dual inside their theory. GRN's shift $\sigma_k\|g_k\|_2 I$ repairs the gauge degeneracy, so quasi-self-concordance alone gives a rate on the whole sublevel set without a cubic subproblem (Doikov, 2026; Doikov and Nesterov, 2024); accelerated alternating minimization (Guminov et al., 2021) needs blocks with a closed form, which general sparse A does not give. Appendices E and L carry the gauge construction, Proposition E.1 and the comparisons at a common tolerance.

Cost as a function of the dual dimension. A coordinate step reads one row at $O(d_i)$, while a second-order step assembles the Hessian in $O(\mathrm{nnz}(A)\omega_{\max} + m^2)$ and factorizes it in $O(m^3)$, with an iteration count almost independent of accuracy and dimension; the same GRN runs on ϕ_{se} (Proposition 5.2) at a fifth more iterations and a third more arithmetic. The fitted exponents of Table 1 and Figure 2 give two crossings: the second-order branch loses its arithmetic advantage between $m = 100$ and 200 but keeps its wall-clock advantage until between $m = 400$ and 1000, because an m^3 factorization runs near peak throughput while a coordinate sweep is a scattered gather over $\mathrm{nnz}(A)$. By Pinsker, a dual gap $\varepsilon = \gamma\varepsilon_{\mathrm{p}}^2/(2\rho^2)$ gives $\mathbb{E}\|A\widehat{x} - b\|_2 \le \varepsilon_{\mathrm{p}}$, in expectation (Appendix F).

6 Experiments

The experiments separate acceleration together with curvature-weighted sampling (ACD against RCD in coordinate epochs), sparse work (a density ablation at fixed m), and problem-specific structure (generic methods against Sinkhorn). Baselines, instance generators, the arithmetic model and the reference optimum are in Appendix G; all curves report the best-so-far relative dual gap $(\phi(\lambda) - \phi^\star)/|\phi^\star_{\mathrm{se}}|$ as medians over five runs. The plotted variant, adaptive ACD, shrinks the conservative gap term using checks on the sampled coordinate, which is not covered by Theorems 4.2–4.3; the fixed-constant schedule is measured separately.

Synthetic instances. On the main sparse instance ($m = 300$, $n = 1200$, $\gamma = 0.05$, five seeds) the adaptive variant reaches relative dual-gap levels $10^{-6}, 10^{-9}, 10^{-12}$ after 80, 145 and 200 epochs. Two baselines bracket it. Plain RCD, taking the envelope step its own analysis prescribes, needs 360, 620 and 900; with the same exact coordinate step our variants take, randomized iterative Bregman projection (Benamou et al., 2015), it needs 185, 320 and 460. Against this baseline, which matches

Table 1: Grouped by step type; GRN (Doikov, 2026) runs on both duals (Proposition 5.2). CRN (Nesterov and Polyak, 2006), accelerated order-2 tensor (Dvurechensky et al., 2024), APDAGD (Dvurechensky et al., 2018), RCD (Lobanov et al., 2024). "ε-dep." is the accuracy exponent of the native guarantee (none for the adaptive variant); α is fitted over $m \in \{100, \dots, 1000\}$ at target 10^{-9}, wall-clock (s) the median over five seeds. A dash marks a target the median seed missed in budget, or too few points to fit.

	step	ε-dep.	α (arith.)	α (wall)	wall-clock	
					m=300	m=1000
GRN on ϕ_{lse}	Newton	$\log \frac{1}{\varepsilon}$	2.92	2.18	0.016	0.37
GRN on ϕ_{se}	Newton	$\log \frac{1}{\varepsilon}$	2.89	2.22	0.019	0.43
CRN	cubic	$\varepsilon^{-1/2}$	3.03	2.13	0.087	1.53
APDAGD	gradient	$\varepsilon^{-1/2}$	0.92	0.69	0.044	0.16
Acc. tensor	tensor	$\varepsilon^{-1/3}$	—	—	0.43	—
RCD	coordinate	ε^{-1}	0.96	0.94	0.044	0.14
RCD, exact coordinate step	coordinate	ε^{-1}	0.96	0.95	0.070	0.22
ACD, analysed schedule (ours)	coordinate	$\varepsilon^{-1/2}$	—	—	—	—
ACD, adaptive (ours, heuristic)	coordinate	none	0.94	0.92	0.037	0.11

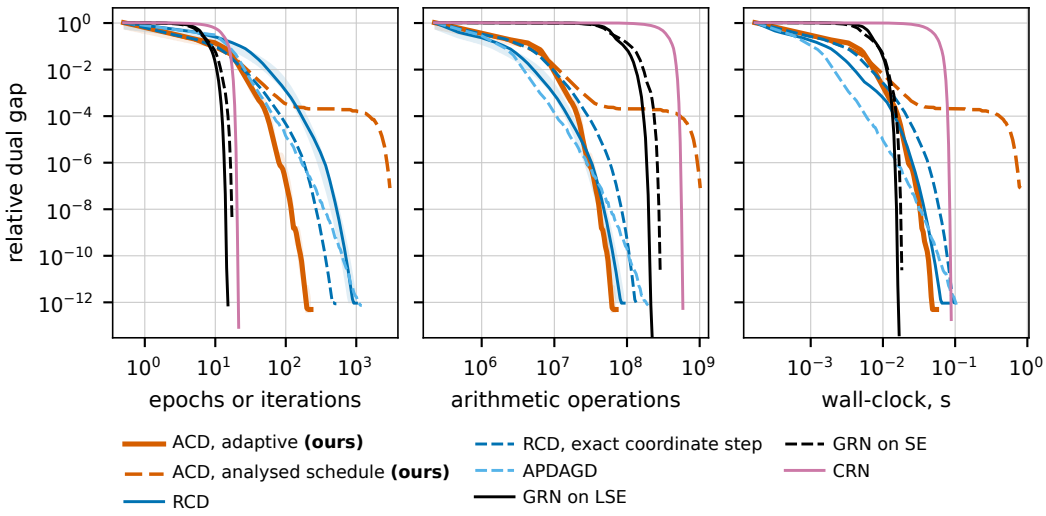

Figure 1: Synthetic sparse entropy-regularized LP instance, $m = 300$, on the three axes the experiments separate; medians over five seeds, with min–max bands on the adaptive and RCD curves. Two ACD curves are shown: the analysed schedule of Theorem 4.2 with its proved constants (dashed) and the adaptive variant that rescales them (solid); Table 7 measures the gap between them. Epoch counts are not work-comparable across the coordinate methods (Appendix G).

the step rule but samples uniformly, the advantage is a factor 2.2. These epoch counts are nearly flat in m (Table 1). On wall-clock at 10^{-9} it is 0.037s against 0.044s and 0.070s: the exact step saves epochs at a matching cost per epoch. We read the comparison on wall-clock rather than on the operation count, because the per-nonzero charge of Appendix G is not calibrated finely enough across the two methods to resolve a margin of this size. GRN sits on the other side of the crossover Section 5 locates: 15 iterations and 0.016s here, against 0.37s at $m = 1000$ where the coordinate methods need 0.11s.

At fixed $m = 1000$, as the row density varies over a factor of 36 in $\text{nnz}(A)$, the epochs to 10^{-9} move only between 125 and 250 while the arithmetic grows from $2.9 \cdot 10^7$ to $6.3 \cdot 10^8$, within a factor 1.8 of proportional to $\text{nnz}(A)$. That is what the $O(d_B)$ block cost predicts: density changes the price of a coordinate step, not how many steps are useful (Figure 3, Appendix G).

The analysed schedule. Run with its proved constants and the fixed anchor of Theorem 4.3, the same method reaches 10^{-6} on the $m = 300$ instance in 2735 epochs but not 10^{-9} within 3000 epochs, and with the envelope step not even 10^{-6}. On the $m = 1000$ instance of Table 7 its termination test is met at none of the thirty settings. At the analysed scale the 10^{-12} tolerance is reached only at $r = 0.5\|\lambda_0 - \lambda^\star_{\mathrm{se}}\|_2$, below the true distance and so outside Theorem 4.2, where the iterate crosses 10^{-6} after 1720 epochs against 345 for RCD and 175 for its exact-step version; at admissible radii it still crosses 10^{-6}, in 205 epochs at $r = 20$, but the schedule cannot recognize the crossing. Confining the rescaling to phase boundaries recovers most of the gap without leaving the certificates: with phases capped at 100 epochs, constants frozen inside each and κ rescaled once per boundary by the slack that phase measured, the method reaches 10^{-9} in 245 epochs at $m = 300$ and 260 at $m = 1000$ against 320 for the like-for-like RCD, with no envelope violation in any run. Two things stay outside the theorems: truncating a phase forfeits its accuracy conclusion, and the rule picking the next κ is retrospective.

Instrumenting the sampled coordinates, the envelope was never violated over thirty settings and five seeds, but it exceeds the curvature actually encountered by 13.6 at $r = 0.5$ and $2.9 \cdot 10^4$ at $r = 20$: the inequality is sound, while the schedule built on it is not competitive at its analysed scale (Section 7).

Optimal transport. The DOTmark experiment (Schrieber et al., 2017) (32×32 images, $m = 2048$, $\gamma = 0.01$) is unfavorable to a generic method and the margin is large. Sinkhorn (Cuturi, 2013) reaches 10^{-6} on all five pairs in 10 to 60 sweeps, the adaptive variant on three (in 1906, 2390 and 2850 epochs) and the analysed schedule and RCD on none; where both finish the factor is 32 to 71 on epochs and 230 to 494 on wall-clock. GRN, CRN and APDAGD finish on all five, so here the non-separable dual is the better route as well. The gap is algorithmic rather than one of data touched: Sinkhorn minimizes exactly over a 1024-dimensional marginal block, and accelerated (Xu and Chen, 2026) and greedy (Altschuler et al., 2017) variants are faster still. That half-step is exact block minimization, which our analysis admits; instantiated there, the analysed schedule reaches 10^{-6} on three pairs in 45, 130 and 750 block passes and within a factor 3.5 on the other two, yet remains 1.9 to 22.9 half-steps behind Sinkhorn, consistent with S_0, which moves by 1.2 to 3.8, never in our favour (Appendices F and G).

7 Discussion and limitations

On the separable side our analysis is much too loose to account for the win of the adaptive variant, which carries no guarantee, and three things limit what has been shown.

Slack. The envelope is three to four orders of magnitude loose at its analysed scale (Section 6), which is why the plotted variant rescales it. Its coefficient on the *global* gap is not improvable (Proposition B.1); the remaining aggregate factor $\sum_i \rho_i/(\sqrt{m}\,\rho)$, five to twelve here, is exactly the price of fixing the constants before the draw (Proposition B.3), so closing it needs a scheme that is not pre-committed per row.

Inputs. The radius r, the gap bound $\bar{\Delta}$ and the partition $\mathcal{B}$ are supplied; too small an r or $\bar{\Delta}$ voids Theorem 4.2 undetectably, and our computable radius is loose, so the runs use an oracle radius. A verified-doubling radius in the spirit of Lobanov (2026) and a budget-driven sampling law are the decisive open items; for $\mathcal{B}$ we give no selection rule, and the only partition we test beyond singletons is the marginal partition of transport.

Numerical range. Unlike ϕ_{lse}, ϕ_{se} is not invariant under a common logit offset, so log-sum-exp recentering requires tracking the removed scale explicitly; our implementation does not, which puts small-γ transport outside double precision without Schmitzer (2019); our theorems assume exact arithmetic, and the $\sqrt{s_{\mathrm{col}}}$ and polynomial separations compare upper bounds under different accounting, not runtimes (Appendices L.2 and L.3).

AI USE STATEMENT

Large language models were used in the following parts of this work. *Writing:* editing and polishing the text and restructuring the LaTeX source. *Mathematical claims:* deriving and checking proofs of the stated results. *Research execution:* implementing the methods and baselines, designing and running the experiments, preparing the figures, and assisting in the interpretation of the results. *Synthetic data:* writing the generators of the synthetic instances described in Appendix G; no data were generated by a model directly. The authors reviewed and verified all statements, proofs, code, and reported results, and take full responsibility for the content of the paper.

REPRODUCIBILITY STATEMENT

Every formal result is stated with its assumptions, and its complete proof is given in the appendices, together with the sparse update and the LSE formulas used by the methods. We do not include code with this submission; instead, Appendix G specifies what is needed to reimplement the experiments: the synthetic instance generator with its distributions and parameters, the DOTmark classes, image pairs, cost normalization and regularization, the seeds, the common starting point and the relative-gap metric, the reference solutions, the inputs supplied to each method (including the oracle radius used by the analysed schedule and the rescaling rule of the adaptive variant), and the arithmetic cost model. Wall-clock times were measured on a single laptop with single-threaded BLAS and are indicative only. Where a reported number aggregates over seeds or image pairs, the text states which aggregate it is.

REFERENCES

Zeyuan Allen-Zhu, Zheng Qu, Peter Richtárik, and Yang Yuan. Even faster accelerated coordinate descent using non-uniform sampling. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *PMLR*, pages 1110–1119, 2016.

Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems*, volume 30, pages 1964–1974, 2017.

Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017. doi: 10.1287/moor.2016.0817.

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015. doi: 10.1137/141000439.

Christoph Brauer, Christian Clason, Dirk Lorenz, and Benedikt Wirth. A Sinkhorn–Newton method for entropic optimal transport. *arXiv preprint arXiv:1710.06635*, 2017.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, Hoboken, NJ, 2 edition, 2006.

Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized SignSGD. In *Advances in Neural Information Processing Systems*, volume 35, pages 9955–9968, 2022.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pages 2292–2300, 2013.

Nikita Doikov. Minimizing quasi-self-concordant functions by gradient regularization of Newton method. *Mathematical Programming*, 219:583–621, 2026. doi: 10.1007/s10107-025-02270-9. First published online in 2025; arXiv:2308.14742.

Nikita Doikov and Yurii Nesterov. Gradient regularization of Newton method with Bregman distances. *Mathematical Programming*, 2024. doi: 10.1007/s10107-023-01943-7.

Radu-Alexandru Dragomir, Adrien B. Taylor, Alexandre d'Aspremont, and Jérôme Bolte. Optimal complexity and certification of Bregman first-order methods. *Mathematical Programming*, 194 (1–2):41–83, 2021. doi: 10.1007/s10107-021-01618-1.

Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *PMLR*, pages 1367–1376, 2018.

Pavel Dvurechensky, Petr Ostroukhov, Alexander Gasnikov, César A. Uribe, and Anastasiya Ivanova. Near-optimal tensor methods for minimizing the gradient norm of convex functions and accelerated primal–dual tensor methods. *Optimization Methods and Software*, 39(5):1068–1103, 2024. doi: 10.1080/10556788.2023.2296443.

Olivier Fercoq and Peter Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015. doi: 10.1137/130949993.

Alexander Gasnikov, Pavel Dvurechensky, and Ilnura Usmanova. About accelerated randomized methods. *arXiv preprint arXiv:1508.02182*, 2015.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pages 3440–3448, 2016.

Eduard Gorbunov, Nazarii Tupitsa, Sayantan Choudhury, Alen Aliakbarpour, Peter Richtárik, and Samuel Horváth. Methods for convex (L_0, L_1)-smooth optimization: Clipping, acceleration, and adaptivity. In *International Conference on Learning Representations*, 2025.

Sergey Guminov, Pavel Dvurechensky, Nazarii Tupitsa, and Alexander Gasnikov. On a combination of alternating minimization and Nesterov's momentum. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *PMLR*, pages 3886–3898, 2021.

Wenshuo Guo, Nhat Ho, and Michael I. Jordan. Fast algorithms for computational optimal transport and Wasserstein barycenter. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *PMLR*, pages 2088–2097, 2020. arXiv:1905.09952.

Filip Hanzely, Peter Richtárik, and Lin Xiao. Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. *Computational Optimization and Applications*, 79(2): 405–440, 2021. doi: 10.1007/s10589-021-00273-8.

Yin Tat Lee and Aaron Sidford. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. In *54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 147–156, 2013. doi: 10.1109/FOCS.2013.24.

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Aleksandr Lobanov. A few accelerated algorithms for convex optimization under (H_0, H_1)-smoothness, 2026.

Aleksandr Lobanov, Alexander Gasnikov, Eduard Gorbunov, and Martin Takáč. Linear convergence rate in convex setup is possible! gradient descent method variants under (L_0, L_1)-smoothness, 2024.

Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018. doi: 10.1137/16M1099546.

Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1): 127–152, 2005. doi: 10.1007/s10107-004-0552-5.

Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012. doi: 10.1137/100802001.

Yurii Nesterov and Boris T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006. doi: 10.1007/s10107-006-0706-8.

Yurii Nesterov and Sebastian U. Stich. Efficiency of the accelerated coordinate descent method on structured optimization problems. *SIAM Journal on Optimization*, 27(1):110–123, 2017. doi: 10.1137/16M1060182.

Yurii Nesterov, Alexander Gasnikov, Sergey Guminov, and Pavel Dvurechensky. Primal–dual accelerated gradient methods with small-dimensional relaxation oracle. *Optimization Methods and Software*, 36(4):773–810, 2021. doi: 10.1080/10556788.2020.1731747.

Bernhard Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM Journal on Scientific Computing*, 41(3):A1443–A1481, 2019. doi: 10.1137/16M1106018.

Jörn Schrieber, Dominic Schuhmacher, and Carsten Gottschlich. DOTmark—a benchmark for discrete optimal transport. *IEEE Access*, 5:271–282, 2017. doi: 10.1109/ACCESS.2016.2639065.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Xun Tang, Michael Shavlovsky, Holakou Rahmanian, Elisa Tardini, Kiran Koshy Thekumparampil, Tesi Xiao, and Lexing Ying. Accelerating Sinkhorn algorithm with sparse Newton iterations. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2401.12253.

Zihao Tang and Yixuan Qiu. Safe and sparse Newton method for entropic-regularized optimal transport. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1–2):1–230, 2015. doi: 10.1561/2200000048.

Daniil Vankov, Anton Rodomanov, Angelia Nedić, Lalitha Sankar, and Sebastian U. Stich. Optimizing (L_0, L_1)-smooth functions by gradient methods. In *International Conference on Learning Representations*, 2025.

Michael D. Vose. A linear algorithm for generating random numbers with a given distribution. *IEEE Transactions on Software Engineering*, 17(9):972–975, 1991. doi: 10.1109/32.92917.

Zeyi Xu and Long Chen. Accelerating Sinkhorn for entropy-regularized optimal transport, 2026.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020.

A Dual derivations and the gap identities

Assumption A.1 (Mass and feasibility)**.** *Some $q \in \mathbb{R}^m$ satisfies $A^\top q = \mathbf{1}$ and $\langle q, b\rangle = 1$, and the affine system admits a strictly positive point: $A\bar{x} = b$ for some $\bar{x} > 0$.*

The first half forces $\mathbf{1}^\top x = 1$ whenever $Ax = b$, so the feasible set is a compact subset of the simplex; the second is what makes the minimizer positive in the Attainment paragraph below. Throughout, the unnormalized and normalized responses are

$$x_j(\lambda) := \exp\Big(\frac{a_j^\top \lambda - c_j}{\gamma}\Big), \qquad Z(\lambda) := \sum_{j=1}^{n} x_j(\lambda), \qquad p_j(\lambda) := \frac{x_j(\lambda)}{Z(\lambda)}. \tag{18}$$

The SE dual follows by minimizing the Lagrangian over $x \geq 0$ coordinatewise; the LSE dual follows by retaining the simplex constraint and minimizing over it. Under $A^\top q = \mathbf{1}$ and $\langle q, b\rangle = 1$,

$$\phi_{\mathrm{se}}(\lambda + tq) = -\langle b, \lambda\rangle - t + \gamma e^{t/\gamma} Z(\lambda),$$

whose stationary point in t proves Proposition 2.1.

Attainment. The feasible set is a closed subset of the simplex, hence compact, and strict feasibility supplies a point with all components positive. The entropy-regularized minimizer $x^\star$ is itself positive: moving from a point with a zero component toward a strictly feasible point produces a one-sided entropy derivative of $-\infty$ at that component, while the derivatives at positive components and of the linear term are finite. The equality-constrained KKT conditions then give a finite $\lambda^\star_{\mathrm{se}}$ with

$$A^\top \lambda^\star_{\mathrm{se}} = c + \gamma \log x^\star, \qquad Ax^\star = b.$$

Both dual optima are attained, though dual minimizers need not be unique.

The two gaps. Proposition 2.1 gives $\phi^\star_{\mathrm{se}} = \phi^\star_{\mathrm{lse}} + \gamma$, and subtracting the two dual functions gives $\phi_{\mathrm{se}} - \phi_{\mathrm{lse}} = \gamma(Z - \log Z)$, which is (5). For (6), since $\mathbf{1}^\top x^\star = 1$,

$$\gamma\,\mathrm{KL}(x^\star \| p(\lambda)) = f_s(x^\star) - \langle b, \lambda\rangle + \gamma \log Z(\lambda) = f_s^\star + \phi_{\mathrm{lse}}(\lambda) = F_{\mathrm{lse}}(\lambda),$$

where $f_s(x) = \langle c, x\rangle + \gamma \sum_j x_j \log x_j$ is the shifted primal objective, equal to the objective of (1) plus γ. The raw SE gap therefore carries a mass-normalization error on top of the error of the normalized response.

Moments of the two duals. The same distinction organizes the derivatives. Let X take the value a_j with probability $p_j(\lambda)$ and put $\mu = \mathbb{E}[X] = Ap(\lambda)$; then

$$\nabla\phi_{\mathrm{se}} = Z\mu - b, \quad \nabla^2\phi_{\mathrm{se}} = \frac{Z}{\gamma}\mathbb{E}[XX^\top], \qquad \nabla\phi_{\mathrm{lse}} = \mu - b, \quad \nabla^2\phi_{\mathrm{lse}} = \frac{1}{\gamma}\,\mathrm{Cov}(X), \tag{19}$$

an unnormalized second moment against a centered one. Everything that separates the two routes follows from that one difference, as Table 2 summarizes.

Table 2: Two exact dual representations of the same primal problem. The last two rows are what Sections 4 and 5 spend the paper measuring.

	Sum-exp (SE)	Log-sum-exp (LSE)
Primal response	unnormalized $x_j(\lambda)$	normalized $p_j(\lambda)$
Nonlinear term	sum of exponentials	log of a sum of exponentials
Hessian	weighted second moment	covariance
Coupling	partially separable	global normalization
Global curvature	unbounded	bounded by the columns of A
Natural update	sparse randomized blocks	regularized Newton steps

B Proof of the Hessian–gap theorem

We first isolate the scalar inequality. For every $\theta > 1$,

$$e^t \le \theta\, h(t) + c_\theta, \qquad h(t) = e^t - 1 - t, \qquad c_\theta = \theta \log \frac{\theta}{\theta - 1}. \tag{20}$$

Indeed $(1-\theta)e^t + \theta(1+t)$ is concave and attains its maximum at $e^t = \theta/(\theta-1)$, where its value is c_θ.

Let $\lambda^\star$ be an SE minimizer and put $t_j = a_j^\top(\lambda - \lambda^\star)/\gamma$. Stationarity gives $x_j(\lambda) = x_j^\star e^{t_j}$ and $Ax^\star = b$, so expanding the dual objective around $\lambda^\star$ yields the exact Bregman identity

$$F(\lambda) = \gamma \sum_j x_j^\star h(t_j). \tag{21}$$

For any $u \in \mathbb{R}^{|B|}$,

$$\begin{aligned} u^\top H_{BB}(\lambda) u &= \frac{1}{\gamma} \sum_j x_j^\star e^{t_j} (A_{B,j}^\top u)^2 \\ &\le \frac{c_\theta}{\gamma} \sum_j x_j^\star (A_{B,j}^\top u)^2 + \frac{\theta}{\gamma} \sum_j x_j^\star h(t_j) (A_{B,j}^\top u)^2 \\ &\le c_\theta\, u^\top H_{BB}^\star u + \frac{\theta \rho_B^2}{\gamma^2} F(\lambda) \|u\|_2^2, \end{aligned}$$

using (20), then $\|A_{B,j}\|_2 \le \rho_B$ and (21). Since u was arbitrary this proves (7), and $\theta = 2$ gives $c_2 = 2\log 2$.

For the universal bound $\bar{H}_B = \rho_B^2/\gamma$, $\|H_{BB}^\star\|_2 \le \gamma^{-1} \sum_j x_j^\star \|A_{B,j}\|_2^2 \le \rho_B^2/\gamma$ because $\sum_j x_j^\star = 1$. For the nonnegative overlap term, Cauchy–Schwarz over the nonzeros of each block column gives $(A_{B,j}^\top u)^2 \le \omega_{Bj} \sum_{i \in B} A_{ij}^2 u_i^2$, hence

$$u^\top H_{BB}^\star u \le \frac{1}{\gamma} \sum_{i \in B} u_i^2 \sum_j x_j^\star \omega_{Bj} A_{ij}^2 \le \frac{1}{\gamma} \max_{i \in B} \Big[b_i \max_{j : A_{ij} \ne 0} (\omega_{Bj} A_{ij}) \Big] \|u\|_2^2,$$

where we used $A \ge 0$ and $\sum_j A_{ij} x_j^\star = b_i$. Both terms bound $\|H_{BB}^\star\|_2$, so their minimum does, which is (8); the singleton formula follows from $\omega_{\{i\}j} = 1$, and the minimum is needed because for $|B| > 1$ the overlap term can exceed ρ_B^2 by up to $\omega_{\max}$.

Proof of Corollary 3.2. For $a \in [\alpha_i^-, \alpha_i^+]$ we have $(a - \alpha_i^-)(a - \alpha_i^+) \le 0$, so $a^2 \le (\alpha_i^- + \alpha_i^+)a - \alpha_i^- \alpha_i^+$. Multiplying by $x_j^\star$ and summing at $a = A_{ij}$ gives $\sum_j x_j^\star A_{ij}^2 \le \mu_i$. If $\alpha_i^- < \alpha_i^+$, put $t = (b_i - \alpha_i^-)/(\alpha_i^+ - \alpha_i^-) \in [0,1]$; then $\mu_i = (1-t)(\alpha_i^-)^2 + t(\alpha_i^+)^2$, so $0 \le \mu_i \le \max\{(\alpha_i^-)^2, (\alpha_i^+)^2\} = \|A_{i:}\|_\infty^2$. For a constant row the same follows from $b_i = \alpha_i^- = \alpha_i^+$ and $\mu_i = (\alpha_i^-)^2$. If $\alpha_i^- = 0$ the numerator is $\alpha_i^+ b_i$.

B.1 Optimality of the growth coefficient

Theorem 3.1 bounds a curvature that grows without limit by a constant $\nu_B = \theta \rho_B^2/\gamma^2$ built from a single entry of the data, which invites the suspicion that the step $A_{B,j}^\top u \le \rho_B$ is wasteful. It is not.

Proposition B.1 (ν_B is optimal up to θ)**.** *Let $\lambda^\star$ be an SE minimizer and let B satisfy $A_{B,:} \neq 0$. If some positive semidefinite Ψ_B and some $\nu_B \geq 0$ satisfy*

$$\nabla^2_{BB}\phi_{\mathrm{se}}(\lambda) \preceq \Psi_B + \nu_B F(\lambda) I_B \qquad \textit{for every } \lambda \in \mathbb{R}^m, \tag{22}$$

then $\nu_B \geq \rho_B^2/\gamma^2$. Moreover, $\theta = 1$ cannot be reached in general: for a scaled indicator row, $\nu_i = \rho_i^2/\gamma^2$ admits no finite κ_i.

The second statement is proved only for indicator rows, and we do not characterize which rows force $\theta > 1$; the first statement holds for every block with $A_{B,:} \neq 0$.

Proof. Choose $j_0 \in \arg\max_j \|A_{B,j}\|_2$ and set $u := A_{B,j_0}/\rho_B$, a unit vector. Along $\lambda_\tau := \lambda^\star + \gamma\tau U_B u$ we have $t_j(\tau) = \tau\beta_j$ with $\beta_j := A_{B,j}^\top u$, so Cauchy–Schwarz gives $\beta_j \leq \|A_{B,j}\|_2 \leq \rho_B$ while $\beta_{j_0} = \rho_B$; thus $\max_j \beta_j = \rho_B$ exactly. Put $J^+ := \{j : \beta_j = \rho_B\} \ni j_0$ and $x^+ := \sum_{J^+} x_j^\star \geq x_{j_0}^\star > 0$. Contracting (22) with u,

$$\frac{1}{\gamma}\sum_j x_j^\star e^{\tau\beta_j}\beta_j^2 \leq u^\top \Psi_B u + \nu_B\,\gamma \sum_j x_j^\star h(\tau\beta_j).$$

Multiply by $e^{-\tau\rho_B}$ and let $\tau \to +\infty$. Every $\beta_j \leq \rho_B$ and the sums are finite, so the left side tends to $x^+\rho_B^2/\gamma$, while $u^\top\Psi_B u\, e^{-\tau\rho_B} \to 0$ and $e^{-\tau\rho_B}\sum_j x_j^\star h(\tau\beta_j) \to x^+$, the $-1-\tau\beta_j$ parts vanishing. Hence $x^+\rho_B^2/\gamma \leq \nu_B\gamma x^+$, which is the claim.

For the second statement take $A_{ij} \in \{0, \rho_i\}$, so that $b_i = \rho_i x_i^+$ with $x_i^+ = \sum_{A_{ij}=\rho_i} x_j^\star$. Along $\lambda_\tau = \lambda^\star + \gamma\tau e_i$ one has $t_j = \tau A_{ij}$, hence $\partial_{ii}^2\phi_{\mathrm{se}} = (\rho_i^2 x_i^+/\gamma)e^{\tau\rho_i}$ and $F = \gamma x_i^+(e^{\tau\rho_i} - 1 - \tau\rho_i)$, so

$$\partial_{ii}^2\phi_{\mathrm{se}}(\lambda_\tau) - \frac{\rho_i^2}{\gamma^2}F(\lambda_\tau) = \frac{\rho_i b_i}{\gamma}(1+\tau\rho_i) \xrightarrow[\tau\to\infty]{} +\infty.$$

□

So the affine envelope of Theorem 3.1 is optimal in its class up to the factor θ, and the exact envelope $\sup\{\partial_{ii}^2\phi_{\mathrm{se}} : F \leq \Delta\} = (\rho_i^2/\gamma^2)\Delta + O(\log\Delta)$ is affine only asymptotically, the correction being $H_{ii}^\star \log\Delta + O(1)$ for scaled indicator rows but $O(1)$ or even negative for signed ones; θ is the price of the affineness that Algorithm 1 needs in order to prebuild its sampling tables. One consequence for Section I.1: $d_A = \sum_i \rho_i^2/\rho^2$ is the far-field dimension of the *per-row* envelopes. Proposition B.1 does not make the aggregate $\sum_i \nu_i$ unimprovable, and Appendix B.2 shows it is loose by $\sum_i \rho_i/(\sqrt{m}\,\rho)$.

B.2 Localized gaps and the aggregate slack

Proposition B.1 closes off any *per-row* improvement. The aggregate is a different matter, and sparsity is the reason: the curvature of row i involves only the columns in $\mathrm{supp}(A_{i:})$, whereas F sums over all n of them.

Proposition B.2 (Localized gap and its budget)**.** *For each i with $\rho_i > 0$ define the* localized gap

$$F_i(\lambda) := \frac{\gamma}{\rho_i^2}\sum_j x_j^\star h(t_j) A_{ij}^2, \qquad \textit{so} \qquad 0 \leq F_i(\lambda) \leq F(\lambda). \tag{23}$$

Then for every $\theta > 1$ and every λ,

$$\partial_{ii}^2\phi_{\mathrm{se}}(\lambda) \leq c_\theta H_{ii}^\star + \frac{\theta\rho_i^2}{\gamma^2}F_i(\lambda), \tag{24}$$

and the localized gaps obey the budget identity, followed by the bound it implies,

$$\sum_i \rho_i^2 F_i(\lambda) = \gamma\sum_j x_j^\star h(t_j)\|a_j\|_2^2 \leq \rho^2 F(\lambda). \tag{25}$$

Proof. For (24), repeat the proof of Theorem 3.1 for the singleton block $\{i\}$ and stop before replacing A_{ij}^2 by ρ_i^2; the surviving sum is exactly $\rho_i^2 F_i/\gamma$. For (25), exchange the two sums: $\sum_i \rho_i^2 F_i = \gamma\sum_j x_j^\star h(t_j)\sum_i A_{ij}^2$, and $\sum_i A_{ij}^2 = \|a_j\|_2^2 \leq \rho^2$, while $\gamma\sum_j x_j^\star h(t_j) = F$. □

The consequence for the quantity the algorithm is actually charged follows from a Cauchy–Schwarz step over (25): with $\nu_i = \theta\rho_i^2/\gamma^2$,

$$\sum_i \sqrt{\nu_i F_i(\lambda)} = \frac{\sqrt{\theta}}{\gamma}\sum_i \rho_i\sqrt{F_i} \le \frac{\sqrt{\theta m}}{\gamma}\,\rho\sqrt{F}, \qquad \text{versus} \qquad \sum_i \sqrt{\nu_i F} = \frac{\sqrt{\theta}}{\gamma}\sqrt{F}\sum_i \rho_i. \tag{26}$$

The ratio is $\sqrt{m}\,\rho/\sum_i \rho_i$. Here ρ is a column 2-norm and ρ_i a row ∞-norm, so with comparable rows this is $\sqrt{s_{\mathrm{col}}/m}$ rather than $1/\sqrt{m}$, where s_{col} is the column sparsity. On the synthetic family of Appendix G ($s_{\mathrm{col}} = 19$) it is 0.196 at $m = 300$, 0.127 at $m = 1000$ and 0.083 at $m = 2000$ – medians over the five seeds of the generator, the spread at $m = 1000$ being 0.106 to 0.129 – a factor of five to twelve in the growth channel. The prediction $\sqrt{s/m}$ gives 0.25, 0.14 and 0.098, which the measured values undershoot by 22, 8 and 15 percent: the same order, not a fixed offset. This is the channel that dominates the early phases.

B.3 The shared-budget form and the cost of pre-commitment

Proposition B.2 is best read as a restatement of the BGG condition in which the rows share one growth budget instead of each owning a coefficient. Writing $w_i := \nu_i F_i(\lambda)$ and $\nu_\bullet := \theta\rho^2/\gamma^2$, (24) and (25) say exactly

$$\partial^2_{ii}\phi_{\mathrm{se}}(\lambda) \le \kappa_i + w_i, \qquad 0 \le w_i \le \nu_i F(\lambda), \qquad \sum_i w_i \le \nu_\bullet F(\lambda), \tag{27}$$

the per-row cap being $F_i \le F$, which holds because $A_{ij}^2 \le \rho_i^2$; and $\sum_i \nu_i = \nu_\bullet d_A$, so the shared budget is smaller than the sum of the individual coefficients by exactly the far-field effective dimension. One consequence is immediate and free: combining (27) with the trace form (35) gives

$$\|\nabla^2\phi_{\mathrm{se}}(\lambda)\|_2 \le \sum_i \kappa_i + \nu_\bullet F(\lambda), \tag{28}$$

which on the synthetic family of Appendix G is a factor $7.8 \cdot 10^3$ below $C(\Delta)^2 := (\sum_B \sqrt{\kappa_B + \nu_B\Delta})^2$, the quantity the sampling normalization forces on the schedule (Appendix C), at $m = 300$ and the $k = 0$ gap level, of which 26.6 comes from the shared budget (close to $d_A = 26.9$, which it approaches as Δ grows) and the remaining 292 from (35).

The rate, however, does not follow, and the reason is worth stating as a limitation of the design rather than of the analysis.

Proposition B.3 (Weight admissible under the cancellation identity)**.** *Fix a gap level Δ and write $B_\star = \nu_\bullet\Delta$ for the growth budget in* (27) *(the symbol W is reserved for arithmetic work). Consider any scheme of the form of Algorithm 1 whose sampling law p and weight a_{k+1} are fixed before the block is drawn and which satisfies the cancellation $p_i \ge a_{k+1}\sqrt{L_i/T_{k+1}}$ for* every *profile admissible under* (27)*, that is every w with $0 \le w_i \le \nu_i\Delta$ and $\sum_i w_i \le B_\star$. Then*

$$a_{k+1} \le \frac{\sqrt{T_{k+1}}}{\sum_i \sqrt{\kappa_i + \nu_i\Delta}}, \tag{29}$$

while a profile-aware normalization may use $\max_w \sum_i \sqrt{\kappa_i + w_i}$ over the same set. The price of pre-commitment, the ratio of the two, satisfies

$$\frac{\sum_i \sqrt{\kappa_i + \nu_i\Delta}}{\sqrt{m\big(\sum_i \kappa_i + B_\star\big)}} \;\le\; \textit{price} \;\le\; \sqrt{d_A}, \tag{30}$$

with equality on the left whenever the capped water-filling level is feasible. In the growth-dominated regime the left-hand side is $\sum_i \rho_i/(\sqrt{m}\,\rho)$, the reciprocal of the aggregate ratio of Proposition B.2.

Proof. Summing $p_i \ge a_{k+1}\sqrt{L_i/T_{k+1}}$ over i with $\sum_i p_i = 1$ gives $a_{k+1} \le \sqrt{T_{k+1}}/\sum_i \sqrt{L_i}$. A pre-committed scheme must respect the worst admissible profile; under the cap that profile is $w_i = \nu_i\Delta$ for every i, which gives (29). For the upper bound in (30), Cauchy–Schwarz over the capped set gives $\max_w \sum_i \sqrt{\kappa_i + w_i} \ge \sum_i \sqrt{\kappa_i + \nu_i\Delta/d_A}$, obtained by the feasible choice $w_i = \nu_i\Delta/d_A$, and the ratio is then at most $\sqrt{d_A}$; for the lower bound, $\max_w \sum_i \sqrt{\kappa_i + w_i} \le \sqrt{m(\sum_i \kappa_i + B_\star)}$ by Cauchy–Schwarz, and this sits in the denominator. □

Measured on the synthetic family at the level $B_\star/\sum_i \kappa_i = 85$, medians over five seeds, the price of (30) is 5.1 at $m = 300$, 7.9 at $m = 1000$ and 11.9 at $m = 2000$, against $\sqrt{d_A} = 5.2$, 8.1 and 12.2: the capped water-filling is feasible here, so the lower bound is attained. Dropping the cap would instead give 17.2, 31.4 and 44.5, close to $\sqrt{m}$; that profile assigns the whole budget to one row, which requires $\rho_i = \rho$ and is not realizable when the columns are sparse.

The slack identified above is therefore not recoverable by a cleverer choice of per-row constants, and the price is exactly the aggregate slack of Proposition B.2 read backwards: what pre-commitment costs is what the localized gap would have saved. Nor have we found a scheme that escapes the price. Committing to the aggregate and verifying per-row admissibility after the draw is the natural candidate, but the deficit incurred on a violating row is weighted by $1/\sqrt{\kappa_i + w_i}$, so an adversarial allocation targets exactly the under-provisioned rows, and the potential of Appendix C carries no slack to absorb it.

The price is not constant: it tends to 1 as $mB_\star/\sum_i \kappa_i \to 0$, that is, once the baseline curvature dominates the growth channel *per row*, a threshold a factor m below the one a careless reading suggests. We tested whether this regime is reachable by raising γ at $m = 300$, and it is not: $B_\star/\sum_i \kappa_i$ falls only from 85 to 5.6 as γ goes from 0.05 to 1, still three orders of magnitude above the $1/m$ that would be needed; the predicted price moves only from 5.1 to 4.8 against $\sqrt{d_A} = 5.2$, and the measured ratio of epochs to RCD *widens* from 5.0 to 9.7, because the non-accelerated method benefits more from the easier problem. We therefore do not claim a regime in which the pre-committed schedule is competitive. On the instances we tested the growth channel dominates M_k at the gap levels the schedule evaluates, which is also why rescaling κ gains less than one power of the rescaling factor: the κ-sweep of Table 8 moves by roughly $\kappa^{1/4}$ per decade, not $\kappa^{1/2}$, because C_1 floors the schedule and is untouched by that rescaling.

C Accelerated phase and work proof

Sampling law versus step length. Line 10 asks only for a point no worse than the explicit step, not for the explicit point itself: $\bar{L}_{k,B}$ must be an *a priori* bound where it sets the sampling law and the weight a_{k+1}, but where it sets the step length any such point will do, so exact block minimization is covered and an inexact iterate is admissible once compared with the explicit step. The restriction $u_{k+1} \in y_k + \operatorname{range} U_{B_k}$ keeps (15) sparse, which is why Theorem 4.3 needs it and not only the rate argument.

Throughout, $F = f - f^\star$ and $0 \preceq \nabla^2_{BB} f(u) \preceq (\kappa_B + \nu_B F(u)) I_B$.

The degenerate case $C_0 = 0$. If $C_0 = 0$ then every $\kappa_B = 0$. For any PSD H and any block vector d,

$$d^\top H d \le \Big(\sum_B \|d_B\|_2 \sqrt{\|H_{BB}\|_2} \Big)^2, \tag{31}$$

by factoring $H = V^\top V$ and applying the triangle inequality to $Vd = \sum_B V U_B d_B$. Hence $\|\nabla^2 f(u)\|_2 \le C_1^2 F(u)$. Along any line, $\psi(t) := F(u^\star + td)$ satisfies $\psi(0) = \psi'(0) = 0$ and $\psi'' \le C_1^2 \|d\|_2^2 \psi$, and the integral form of Gronwall's inequality forces $\psi \equiv 0$. The attained problem is therefore constant, which is why the theorem assumes $C_0 > 0$.

If a single block has $\kappa_B = \nu_B = 0$, its diagonal Hessian block vanishes everywhere; for a PSD matrix a zero diagonal block forces zero cross blocks, so $\nabla_B f$ is constant and vanishes at a minimizer. Dropping that block is exact. The gradient bound below also covers $\kappa_B + \nu_B F(u) = 0$ by applying the descent argument with $M > 0$ and letting $M \downarrow 0$.

Self-validating block descent. Fix u, B and let $M \ge \kappa_B + \nu_B F(u) > 0$. Put $g_B = \nabla_B f(u)$ and, for $g_B \ne 0$, move along $d = -U_B g_B / \|g_B\|$. As long as the objective has not exceeded $f(u)$, the BGG bound gives $d^2 f(u + \tau d)/d\tau^2 \le M$, and a first return to $f(u)$ before $\tau = \|g_B\|/M$ would contradict the resulting quadratic upper model. Evaluating at that endpoint,

$$f\Big(u - \frac{U_B g_B}{M}\Big) \le f(u) - \frac{\|g_B\|_2^2}{2M}, \tag{32}$$

and comparing with $f^\star$,

$$\|\nabla_B f(u)\|_2^2 \le 2F(u)(\kappa_B + \nu_B F(u)). \tag{33}$$

Combining (31) with BGG, on the sublevel $\{F \le \Delta\}$,

$$\|\nabla^2 f\|_2 \le C(\Delta)^2, \qquad C(\Delta) := \sum_B \sqrt{\kappa_B + \nu_B \Delta}. \tag{34}$$

One more application of Cauchy–Schwarz to (31) gives the sharper

$$\|\nabla^2 f\|_2 \le \mathcal{T}(\Delta) := \sum_B (\kappa_B + \nu_B \Delta) \le C(\Delta)^2, \tag{35}$$

since $d^\top H d \le (\sum_B \|d_B\|_2^2)(\sum_B \|H_{BB}\|_2)$. Writing $N_{\mathcal{B}}$ for the number of blocks, to keep it apart from the block size $|B|$: for $N_{\mathcal{B}}$ comparable blocks the two differ by a factor $N_{\mathcal{B}}$, and on the synthetic family of Appendix G the gap is 292 at $m = 300$. The trace form (35) is the sharper of the two and is the statement we record; the schedule below nevertheless uses C, because $M_k = \sum_B \sqrt{\bar{L}_{k,B}}$ is forced by the sampling normalization and already satisfies $M_k^2 \ge \mathcal{T}(5\Delta_k)$ with room to spare. Closing that gap therefore requires a different sampling law, not a different Hessian bound.

Safe interpolation and the constant 5: proof of Lemma 4.1. Suppose $\Phi = TF(u) + \frac{1}{2}\|v - u^\star\|^2 \le \Pi$ and put $\Delta = \Pi/T$. Choose $M \ge C(5\Delta)$ and $a > 0$ with $M^2 a^2 = T^+ := T + a$, and set $y = (Tu + av)/T^+$. Write $\bar{y} = (1 - a/T^+)u + (a/T^+)u^\star$ and $d = (a/T^+)(v - u^\star)$, so that $y = \bar{y} + d$. Convexity gives $F(\bar{y}) \le F(u) \le \Delta$. From $M^2 a^2 = T^+$ we get $a/T^+ = 1/(M\sqrt{T^+})$, so the potential bound $\|v - u^\star\| \le \sqrt{2\Pi}$ yields $\|d\| \le \sqrt{2\Delta}/M$. By summing (33) over blocks and applying (35), $\|\nabla f(\bar{y})\|^2 \le 2\Delta\,\mathcal{T}(\Delta) \le 2\Delta\, C(\Delta)^2$, so $\|\nabla f(\bar{y})\| \le \sqrt{2\Delta}\, C(\Delta) \le \sqrt{2\Delta}\, M$; only $M^2 \ge \mathcal{T}(5\Delta)$ is used.

Now suppose the segment $\bar{y} + \tau d$, $\tau \in [0, 1]$, first reached level 5Δ. Before that crossing (34) bounds the Hessian by M^2, so Taylor's formula gives

$$F(\bar{y} + \tau d) \le \Delta + \tau\sqrt{2\Delta} M \cdot \frac{\sqrt{2\Delta}}{M} + \frac{\tau^2}{2} M^2 \frac{2\Delta}{M^2} = \Delta + 2\tau\Delta + \tau^2\Delta \le 4\Delta,$$

a contradiction. Hence $F(y) \le 4\Delta < 5\Delta$. Any fixed margin above 4 would do.

Conditional accelerated cancellation. At iteration k set $\Delta_k = \Pi/T_k$, $\sqrt{\bar{L}_{k,B}} = \sqrt{\kappa_B} + \sqrt{5\nu_B\Delta_k}$ and $M_k = C_0 + \sqrt{5\Delta_k}\, C_1$. Then $M_k \ge C(5\Delta_k)$ because $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$. On the event $\Phi_k = T_k F(u_k) + \frac{1}{2}\|v_k - u^\star\|^2 \le \Pi$, safe interpolation validates every block model at y_k *before* sampling. The two-table mixture draws exactly $p_{k,B} = \sqrt{\bar{L}_{k,B}}/M_k$. Expanding the momentum norm after the sampled update and combining with (32), the quadratic terms cancel because

$$\frac{a_{k+1}^2}{p_{k,B}} = \frac{T_{k+1} p_{k,B}}{\bar{L}_{k,B}}, \tag{36}$$

which follows from $M_k^2 a_{k+1}^2 = T_{k+1}$ and $p_{k,B} = \sqrt{\bar{L}_{k,B}}/M_k$. In detail, write $g_B := \nabla_B f(y_k)$ and let B be the drawn block, so $v_{k+1} = v_k - (a_{k+1}/p_{k,B}) U_B g_B$ and $\frac{1}{2}\|v_{k+1} - u^\star\|^2 = \frac{1}{2}\|v_k - u^\star\|^2 - \frac{a_{k+1}}{p_{k,B}}\langle g_B, (v_k - u^\star)_B\rangle + \frac{a_{k+1}^2}{2p_{k,B}^2}\|g_B\|^2$. Taking the conditional expectation, the middle term becomes $-a_{k+1}\langle \nabla f(y_k), v_k - u^\star\rangle$ and the last becomes $\sum_B \frac{a_{k+1}^2}{2p_{k,B}}\|g_B\|^2 = T_{k+1}\sum_B \frac{p_{k,B}}{2\bar{L}_{k,B}}\|g_B\|^2$ by (36), which is exactly T_{k+1} times the decrease $\mathbb{E}[F(y_k) - F(u_{k+1}) \mid \mathcal{F}_k]$ guaranteed by (32). What remains is $T_{k+1}F(y_k) - a_{k+1}\langle \nabla f(y_k), v_k - u^\star\rangle \le T_k F(u_k)$, which is the two convexity inequalities $F(y_k) + \langle \nabla f(y_k), u_k - y_k\rangle \le F(u_k)$ and $F(y_k) + \langle \nabla f(y_k), u^\star - y_k\rangle \le 0$ weighted by T_k and a_{k+1} and added, using $T_k(u_k - y_k) + a_{k+1}(v_k - y_k) = 0$. Hence

$$\mathbb{E}[\Phi_{k+1} \mid \mathcal{F}_k] \le \Phi_k \qquad \text{on } \{\Phi_k \le \Pi\}. \tag{37}$$

High-probability phase guarantee: proof of Theorem 4.2. Stop the process at the first crossing of $\Phi_k > \Pi$. The transition that first crosses is covered by (37), so the stopped process is a nonnegative supermartingale. Initially $\Phi_0 = T_0 F(z) + \frac{1}{2}\|z - u^\star\|^2 \le r^2 = \delta\Pi$, so Doob's maximal inequality bounds the probability of any crossing by δ. On the complementary event $T_N F(u_N) \le \Pi$, and the stopping rule $T_N \ge \Pi/\varepsilon$ gives $F(u_N) \le \varepsilon$.

Deterministic growth of the weights. Let $s_k := \sqrt{T_k}$ (not to be confused with the Bregman exponents t_j of Section 3) and $\Gamma := \sqrt{5\Pi}\, C_1$, so that $M_k = C_0 + \Gamma/s_k$. With $a_{k+1} = s_{k+1}^2 - s_k^2$, the defining equation $M_k^2 a_{k+1}^2 = s_{k+1}^2$ becomes

$$s_{k+1} - s_k = \frac{s_{k+1}}{(C_0 + \Gamma/s_k)(s_{k+1} + s_k)}. \tag{38}$$

Since $1/2 \le s_{k+1}/(s_{k+1} + s_k) \le 1$,

$$\frac{1}{2(C_0 + \Gamma/s_k)} \le s_{k+1} - s_k \le \frac{1}{C_0 + \Gamma/s_k}.$$

While $s_k \le \Gamma/C_0$ the lower bound gives $s_{k+1} \ge s_k(1 + 1/(4\Gamma))$, so s_k grows geometrically and reaches Γ/C_0 after $O((1+\Gamma)\log(\Gamma/(C_0 s_0)))$ iterations: for $\Gamma < 1/4$ the factor $1 + 1/(4\Gamma)$ is large and the count is logarithmic, while for large Γ the warm-up is of order Γ up to a logarithm. Thereafter every increment is at least $1/(4C_0)$. Reaching $s_N \ge \sqrt{\Pi/\varepsilon}$ therefore takes

$$N = \widetilde{O}\big(1 + C_0\sqrt{\Pi/\varepsilon} + \Gamma\big) = \widetilde{O}\Big(1 + \frac{rC_0}{\sqrt{\delta\varepsilon}} + \frac{rC_1}{\sqrt{\delta}}\Big),$$

which is (13), and the overshoot obeys

$$s_N \le \sqrt{\Pi/\varepsilon} + 1/C_0. \tag{39}$$

From high probability to expectation. Let $0 < \varepsilon < \bar{\Delta}$. Every repetition uses the original z, the same r and $\bar{\Delta}$, failure level $\delta = 1/3$, target $\varepsilon/2$, and an independent sampling stream; return the smallest-objective point of $\{z, u^{(1)}, \dots, u^{(J_{\mathrm{rep}})}\}$. Each phase succeeds with probability at least $2/3$, so all fail with probability at most $3^{-J_{\mathrm{rep}}}$. Since the return is safeguarded by z,

$$\mathbb{E}F(\widehat{u}) \le \varepsilon/2 + \bar{\Delta}\, 3^{-J_{\mathrm{rep}}} \le \varepsilon, \qquad J_{\mathrm{rep}} = \Big\lceil \frac{\log(2\bar{\Delta}/\varepsilon)}{\log 3} \Big\rceil.$$

No success test and no knowledge of $f^\star$ is needed, only objective comparisons. The same computation gives the high-probability form $\mathbb{P}(F(\widehat{u}) \le \varepsilon/2) \ge 1 - 3^{-J_{\mathrm{rep}}}$. Keeping z fixed is what allows the radius to be reused: a smaller function value at a new anchor does not imply a smaller distance to $u^\star$.

Table 3: The quantities entering the SE complexity bounds, and what each controls.

quantity	meaning	role
κ_B	baseline of the curvature envelope	stiffness near the solution
ν_B	growth of curvature per unit gap	transient inflation far from it
d_B	touched-data cost $\lvert B\rvert + \mathrm{nnz}(A_{B:})$	arithmetic per sampled block
C_0, C_1	unweighted aggregates	number of block queries
S_0, S_1	cost-weighted aggregates	total sparse arithmetic

Weighted sparse work. With $S_0 = \sum_B d_B\sqrt{\kappa_B}$ and $S_1 = \sum_B d_B\sqrt{\nu_B}$, conditional on the history,

$$\mathbb{E}[d_{B_k} \mid \mathcal{F}_k] = \frac{S_0 + \sqrt{5\Delta_k}\, S_1}{C_0 + \sqrt{5\Delta_k}\, C_1} = \frac{S_0 + (\Gamma/s_k)(S_1/C_1)}{C_0 + \Gamma/s_k} \tag{40}$$

when $C_1 > 0$; the first form shows what the identity says, the second is the one the schedule recursion consumes, and they agree because $\sqrt{5\Delta_k} = \sqrt{5\Pi}/s_k = \Gamma/(s_k C_1)$. This is the reason for carrying (C_0, C_1) and (S_0, S_1) separately: the first pair controls how many block queries are made, the second what data those queries read. The lower inequality after (38) bounds the first contribution by $2S_0(s_{k+1} - s_k)$, which telescopes to $2S_0 s_N$. For the second, the *lower* inequality after (38) reads $1/(C_0 s_k + \Gamma) \le 2(s_{k+1} - s_k)/s_k =: 2x_k$, so $\sum_k 1/(C_0 s_k + \Gamma) \le 2\sum_k x_k$; with $x_k \le \vartheta_0 := 1/(C_0 s_0 + \Gamma)$ and $x \le (1+\vartheta_0)\log(1+x)$ on $[0, \vartheta_0]$, and $\sum_k \log(1+x_k) = \log(s_N/s_0)$, the second contribution sums to $O(\sqrt{\Pi}\, S_1(1+\vartheta_0)\log(s_N/s_0))$. The dependence on the initial gap sharpens by keeping both terms in the denominator: for $C_1 > 0$,

$$\sqrt{\Pi}\vartheta_0 = \frac{\sqrt{\Pi}}{C_0 s_0 + \sqrt{5\Pi}\, C_1} \le \min\Big\{ \frac{\sqrt{2\bar{\Delta}}}{\sqrt{\delta}\, C_0},\ \frac{1}{\sqrt{5}\, C_1} \Big\}. \tag{41}$$

Together with (39), at fixed failure level and phase target $\varepsilon/2$,

$$\mathbb{E}W_{\text{inner}} = \widetilde{O}\Big(\frac{S_0}{C_0} + \frac{rS_0}{\sqrt{\varepsilon}} + rS_1 + S_1 \min\{\tfrac{\sqrt{2\bar{\Delta}}}{C_0}, \tfrac{1}{C_1}\}\Big).$$

for the inner work of one phase. Summing over the J_{rep} independent repetitions changes only logarithmic factors, and the phase-boundary, comparison and reset work is absorbed below, so the same expression bounds $\mathbb{E}W_{\text{opt}}$. If $C_1 = 0$ then $S_1 = 0$, the second table is omitted, and the bound reduces to $\widetilde{O}(S_0/C_0 + rS_0/\sqrt{\varepsilon})$ with no division by C_1. Appendix D absorbs comparisons, candidate storage and resets into the accumulated touched work; initialization and any requested dense output are separate passes. This proves Theorem 4.3.

D Exact sparse implementation and phase boundaries

The lazy identities were stated in Section 4.1; we verify them and add the safeguard. From $y_k = (T_k u_k + a_{k+1} v_k)/T_{k+1}$ and $T_{k+1} = T_k + a_{k+1}$,

$$y_k = v_k + \frac{T_k(u_k - v_k)}{T_{k+1}} = z + \tilde{v}_k + \frac{w_k}{T_{k+1}},$$

which is the factorization of y_k used there. For the updates, $v_{k+1} = v_k - \xi U_{B_k} g$ gives the first half of (15) directly, and, writing $u_{k+1} = y_k + \eta_k$ with η_k supported on B_k,

$$w_{k+1} = T_{k+1}(u_{k+1} - v_{k+1}) = T_{k+1}\big(y_k - v_k + \eta_k + \xi U_{B_k} g\big) = w_k + T_{k+1}\big(\xi U_{B_k} g + \eta_k\big),$$

using $T_{k+1}(y_k - v_k) = T_k u_k - T_k v_k = w_k$. The explicit step $\eta_k = -U_{B_k} g / \bar{L}_{k,B_k}$ recovers the coefficient $\xi - \bar{L}_{k,B_k}^{-1}$. The transformed projections follow the same recurrences composed with $A^\top$:

$$\zeta_{k+1}^v = \zeta_k^v - \xi A^\top U_{B_k} g, \qquad \zeta_{k+1}^w = \zeta_k^w + T_{k+1} A^\top \big(\xi U_{B_k} g + \eta_k\big),$$

and each is supported on $\operatorname{supp}(A_{B_k:})$. With row-oriented storage all four updates and (16) cost $O(|B_k| + \operatorname{nnz}(A_{B_k:})) = O(d_{B_k})$; Algorithm 2 collects one such step.

Sparse phase safeguards. Let I be the union of visited dual blocks and $J = \bigcup_{i \in I} \operatorname{supp}(A_{i:})$. The candidate displacement $\hat{h} = \tilde{v}_N + w_N/T_N$ is supported on I and its transform $A^\top \hat{h} = \zeta_N^v + \zeta_N^w/T_N$ on J. Storing $e_j^0 = \exp((\zeta_j^0 - c_j)/\gamma)$ at the phase base,

$$\phi_{\text{se}}(z + \hat{h}) - \phi_{\text{se}}(z) = \gamma \sum_{j \in J} e_j^0 \Big(e^{(A^\top \hat{h})_j/\gamma} - 1\Big) - \sum_{i \in I} b_i \hat{h}_i, \tag{42}$$

because $(A^\top \hat{h})_j = 0$ for $j \notin J$. Timestamp arrays or touched-index lists give exact accept/reject checks and resets in $O(|I| + |J|)$, and $|I| + |J| \le \sum_{k<N} d_{B_k}$, so phase boundaries are absorbed by the cumulative inner work. No $O(m)$ or $O(n)$ restart scan is hidden.

Fixed-anchor preparation and candidate storage. The data ζ^0, e^0, the incidence structure, the row and block constants and the alias tables are prepared once in $O(\operatorname{nnz}(A) + m + n)$ and remain unchanged across the independent repetitions of Theorem 4.3. Each candidate's objective difference from z comes from (42); compare that scalar with the stored best (initialized to zero) and, if smaller, copy only the current visited-coordinate displacement. Freeing or overwriting a previous sparse candidate is charged to its earlier copy, so total copying and resetting is a constant multiple of total visited-set work even when supports differ. Comparisons of nearly equal values need numerically stable summation; the theorem is stated in exact real arithmetic and is not a bit-complexity result.

The sampling law is exactly $p_{k,B} = \frac{C_0}{M_k}\pi_B^{(0)} + \big(1 - \frac{C_0}{M_k}\big)\pi_B^{(1)}$, the second component omitted when $C_1 = 0$, so two static alias tables give $O(1)$ expected sampling after $O(N_{\mathcal{B}})$ preprocessing.

E The LSE high-order branch

E.1 Gauge, attainment and bounded sublevels

Let $\mathcal{G} := \{h : A^\top h \in \operatorname{span}\{\mathbf{1}\}\}$ and $\mathcal{H} := \mathcal{G}^\perp$. If $A^\top h = t\mathbf{1}$, strict feasibility gives $\langle b, h\rangle = t$, so $\phi_{\text{lse}}(\lambda + h) = \phi_{\text{lse}}(\lambda)$; also $A^\top(h - tq) = 0$, which proves $\mathcal{G} = \ker A^\top + \operatorname{span}\{q\}$. For a simplex

Algorithm 2 One exact sparse SE block update

1: Draw B_k by the two-table mixture of Algorithm 1.
2: Evaluate $g = \nabla_{B_k}\phi_{\mathrm{se}}(y_k)$ from incident factors only, via (16).
3: Update $(\tilde{v}_k, w_k)$ and (ζ_k^v, ζ_k^w) by the four sparse recurrences above.
4: Record touched dual rows and primal factors for sparse phase validation and reset.

point p, $\langle h, Ap - b\rangle = t(\mathbf{1}^\top p - 1) = 0$, so $Ap - b \in \mathcal{H}$: an orthonormal parametrization $\lambda = Q\xi$ with columns of Q spanning $\mathcal{H}$ has $\|Q^\top(Ap-b)\|_2 = \|Ap - b\|_2$, and the full Euclidean residual is preserved. A nonorthogonal coordinate fixing need not preserve it.

For coercivity on $\mathcal{H}$, choose strictly feasible $x^{\mathrm{f}} > 0$ with $\mathbf{1}^\top x^{\mathrm{f}} = 1$ and set $\varsigma(h) := \max_j a_j^\top h - \langle b, h\rangle$ for unit $h \in \mathcal{H}$. The second term is the average $\sum_j x_j^{\mathrm{f}} a_j^\top h$ with strictly positive weights, so $\varsigma(h) \geq 0$ with equality only if all $a_j^\top h$ coincide, i.e. $h \in \mathcal{G}$. If $\dim \mathcal{H} > 0$, compactness of the unit sphere gives $\varsigma_* := \min \varsigma > 0$, and for $t \geq 0$, $\phi_{\mathrm{lse}}(th) \geq t\varsigma_* - \max_j c_j$. Hence all sublevels on $\mathcal{H}$ are compact. The positive softmax weights make the Hessian strictly positive on nonzero directions of $\mathcal{H}$, so the minimizer is unique. If $\mathcal{H} = \{0\}$ the residual at the sole gauge representative vanishes and no iterations are needed.

For transport with two complete marginal families, $\lambda = (\alpha, \beta)$ and $a_{ij}^\top \lambda = \alpha_i + \beta_j$, so $\mathcal{G} = \mathrm{span}\{(\mathbf{1}, 0), (0, \mathbf{1})\}$: both marginal-constant directions must be removed, for example by $\mathbf{1}^\top\alpha = \mathbf{1}^\top\beta = 0$. Removing only one is insufficient.

E.2 Derivative bounds and second-order rates

Section 5 states the constants (17) and the two second-order branches; this appendix derives them and records the gauge details. Let X take the value a_j with probability $p_j(\lambda)$ and put $\mu = \mathbb{E}[X]$. Differentiating the softmax weights in a direction u gives $Dp_j(\lambda)[u] = \gamma^{-1} p_j \langle a_j - \mu, u\rangle$, so with $Y_u := \langle X - \mu, u\rangle$,

$$\nabla\phi_{\mathrm{lse}}(\lambda) = \mu - b, \qquad \nabla^2\phi_{\mathrm{lse}}(\lambda) = \tfrac{1}{\gamma}\,\mathrm{Cov}(X), \qquad \nabla^3\phi_{\mathrm{lse}}(\lambda)[u,v,w] = \tfrac{1}{\gamma^2}\mathbb{E}[Y_u Y_v Y_w], \quad (43)$$

the two terms from differentiating μ in the third identity vanishing because $\mathbb{E}[Y_u] = 0$. The same identities survive an orthogonal restriction to $\mathcal{H}$.

Proof of Proposition 5.2. Differentiating ϕ_{se} three times gives $\nabla^3\phi_{\mathrm{se}}(\lambda)[u,u,v] = \gamma^{-2}\sum_j x_j(\lambda)(a_j^\top u)^2(a_j^\top v)$, while $\nabla^2\phi_{\mathrm{se}}(\lambda)[u,u] = \gamma^{-1}\sum_j x_j(\lambda)(a_j^\top u)^2$. Every term of the first sum shares the nonnegative weight $x_j(a_j^\top u)^2$ of the second, so pulling out $\max_j |a_j^\top v|$ and using $|a_j^\top v| \leq \rho\|v\|_2$ gives the claim. The bound is attained in the limit where one column carries the whole response. Note that it needs no sign assumption on A and no restriction of λ, unlike the global Hessian bound, which does not exist for ϕ_{se}.

Proof of Theorem 5.1. Every bound follows from one observation: for a unit u the scalar variable $u^\top X$ is supported in an interval of length at most $\Delta_A = \max_{j,k}\|a_j - a_k\|_2$, and a variable supported in an interval of length d has variance at most $d^2/4$. Hence $u^\top\nabla^2\phi_{\mathrm{lse}}u = \gamma^{-1}\,\mathrm{Var}(u^\top X) \leq \Delta_A^2/(4\gamma)$, which is the first bound. For the second, $|Y_v| \leq \Delta_A\|v\|_2$ pointwise, so (43) gives $|\nabla^3\phi_{\mathrm{lse}}[u,u,v]| \leq \gamma^{-2}\Delta_A\|v\|_2\mathbb{E}[Y_u^2] = \gamma^{-1}\Delta_A\|v\|_2(u^\top\nabla^2\phi_{\mathrm{lse}}u)$, which is quasi-self-concordance with $M_{\mathrm{qsc}} = \Delta_A/\gamma$. For the third, take u, v, w unit, bound $|Y_w| \leq \Delta_A$ and apply Cauchy–Schwarz with the variance bound to get $\mathbb{E}|Y_uY_v| \leq \Delta_A^2/4$; homogeneity extends the result to arbitrary arguments. Since $\Delta_A \leq 2\rho$ for $\rho = \max_j\|a_j\|_2 = \|A\|_{1\to 2}$, the column-norm constants $L_{\mathrm{lse}} \leq \rho^2/\gamma$, $M_{\mathrm{qsc}} \leq 2\rho/\gamma$ and $M_2 \leq 2\rho^3/\gamma^2$ follow, and the exponent comparisons below are stated in those. More generally each fixed order ℓ has a derivative Lipschitz bound $O_\ell(\rho^{\ell+1}/\gamma^\ell)$ from the bounded-support cumulant formula.

Gradient regularization. The quasi-self-concordance bound controls relative rather than additive Hessian variation, $e^{-M_{\mathrm{qsc}}\|h\|_2}H(\lambda) \preceq H(\lambda + h) \preceq e^{M_{\mathrm{qsc}}\|h\|_2}H(\lambda)$, which is what licenses a regularizer proportional to the current gradient norm: far from a solution the shift limits the step, and at feasibility it vanishes. Writing $f = \phi_{\mathrm{lse}}|_{\mathcal{H}}$, $g_k = \nabla f(\lambda_k)$ and $\varrho_k = \|g_k\|_2$ (not the supplied radius

r of Section 4), the last quantity is exactly the primal residual $\|Ap(\lambda_k) - b\|_2$ by (43). The constant choice $\sigma_k = M_{\mathrm{qsc}}$ is admissible; if the constant is not known tightly, Doikov (2026) double σ_k until $\langle g_{k+1}, -h_k\rangle \geq \varrho_{k+1}^2/(2\sigma_k\varrho_k)$, a test in quantities already computed at the trial point, which costs at most one extra trial on average with the usual warm start.

Proposition E.1 (GRN on the LSE dual)**.** *Let $\mathcal{L}_0 = \{\lambda \in \mathcal{H} : f(\lambda) \leq f(\lambda_0)\}$ and $D \geq \sup_{u,v\in\mathcal{L}_0} \|u - v\|_2$, finite by the coercivity above. If $M_{\mathrm{qsc}} > 0$ and $\sigma_k \in (0, 2M_{\mathrm{qsc}}]$ satisfies the progress test, then for $k \geq 1$*

$$f(\lambda_k) - f^\star \leq e^{-k/(8M_{\mathrm{qsc}}D)}\big(f(\lambda_0) - f^\star\big) + e^{-k/4}\varrho_0 D, \tag{44}$$

so an LSE dual gap at most ε is reached in $O\big((1 + M_{\mathrm{qsc}}D)\log((f(\lambda_0) - f^\star + \varrho_0 D)/\varepsilon)\big)$ iterations, with no strong-convexity assumption. With $M_{\mathrm{qsc}} \leq \Delta_A/\gamma$ this is the instantiation used in Section 5.

Equation (44) is the Euclidean smooth specialization of Doikov (2026, Theorem 3.3) applied on $\mathcal{H}$, whose initial sublevel set has diameter D; Theorem 5.1 supplies the quasi-self-concordance parameter. The stated iteration count follows by making each term at most $\varepsilon/2$, which needs $k \geq 8M_{\mathrm{qsc}}D\log(2(f(\lambda_0) - f^\star)/\varepsilon)$ and $k \geq 4\log(2\varrho_0 D/\varepsilon)$ respectively. Near a solution the local analysis of Doikov (2026, Section 5) adds a quadratic phase in the corresponding normalized gradient measure. Unlike CRN below, GRN needs no valid Lipschitz-Hessian constant, but it does need the progress test; CRN needs the constant and no search, at the price of a cubic subproblem instead of one shifted linear system.

Cubic regularization. From $\lambda_0 \in \mathcal{H}$, let $D \geq \sup_{\lambda\in\mathcal{L}_0} \|\lambda - \lambda^\star_{\mathrm{lse}}\|_2$ on the initial sublevel $\mathcal{L}_0$. The cubic step

$$h_k \in \arg\min_{h\in\mathcal{H}}\Big\{\langle g_k, h\rangle + \tfrac{1}{2}\langle H_k h, h\rangle + \tfrac{M_2}{6}\|h\|_2^3\Big\} \tag{45}$$

is monotone and satisfies $F_{\mathrm{lse}}(\lambda_k) = O(M_2 D^3/k^2)$ (Nesterov and Polyak, 2006); Algorithm 3 states the iteration. Global smoothness gives $\|Ap(\lambda_k) - b\|_2^2 \leq 2L_{\mathrm{lse}}F_{\mathrm{lse}}(\lambda_k)$, hence $k_{\mathrm{CRN}}(\varepsilon_{\mathrm{p}}) = O(1 + \rho^{5/2}D^{3/2}/(\gamma^{3/2}\varepsilon_{\mathrm{p}}))$.

Accelerated primal–dual tensor method, order 2. For zero initial iterates on $\mathcal{H}$ and $R \geq \|\lambda^\star_{\mathrm{lse}}\|_2$, apply Dvurechensky et al. (2024, Theorem 5.4) to the simplex primal and its LSE dual. With the sign convention reversed and weights reindexed as a_{i+1}, the returned primal point is the weighted average

$$\widehat{p}_k = \frac{1}{A_k}\sum_{i=0}^{k-1} a_{i+1}p(\lambda_{i+1}), \qquad A_k = \sum_{i=0}^{k-1} a_{i+1}, \tag{46}$$

and for a regularization parameter equal to a fixed multiple strictly greater than M_2,

$$\|A\widehat{p}_k - b\|_2 = O(M_2R^2/k^3), \qquad \big|f_s(\widehat{p}_k) + \phi_{\mathrm{lse}}(\lambda_k)\big| \leq O(M_2R^3/k^3). \tag{47}$$

Hence $k_{\mathrm{accT}}(\varepsilon_{\mathrm{p}}) = O(1 + \rho R^{2/3}/(\gamma^{2/3}\varepsilon_{\mathrm{p}}^{1/3}))$. This is the reason the method averages: the estimate is primal–dual and direct, rather than a dual-gap rate converted to feasibility at the cost of a square root. A diameter or sublevel radius may replace R only if it actually bounds the distance from the stipulated zero start.

Translation to a nonzero start. Let $\lambda_0 \in \mathcal{H}$ and $R \geq \|\lambda^\star_{\mathrm{lse}} - \lambda_0\|_2$. Writing $v = \lambda - \lambda_0$, $\widetilde{c} = c - A^\top\lambda_0$, $\widetilde{f}_s(p) = f_s(p) - \langle\lambda_0, Ap\rangle$ and $\widetilde{\phi}(v) = \phi_{\mathrm{lse}}(\lambda_0 + v) + \langle b, \lambda_0\rangle$ leaves the primal optimizer and responses unchanged, and the zero-start theorem gives

$$\big|f_s(\widehat{p}_k) + \phi_{\mathrm{lse}}(\lambda_k) - \langle\lambda_0, A\widehat{p}_k - b\rangle\big| \leq O(M_2R^3/k^3). \tag{48}$$

The uncentered gap is bounded only after adding $\|\lambda_0\|_2\|A\widehat{p}_k - b\|_2$, which is why the starting point cannot be suppressed in a direct primal–dual guarantee.

Near-optimal gradient-norm method. With $R \geq \|\lambda_0 - \lambda^\star_{\mathrm{lse}}\|_2$, the initial-distance specialization of Dvurechensky et al. (2024) gives $k_{\mathrm{nearT}}(\varepsilon_{\mathrm{p}}) = \widetilde{O}(1 + (M_2R^2/\varepsilon_{\mathrm{p}})^{2/7}) = \widetilde{O}(1 + \rho^{6/7}R^{4/7}/(\gamma^{4/7}\varepsilon_{\mathrm{p}}^{2/7}))$. This is a theoretical benchmark, not an implementation in this paper; its line searches and regularizations are part of the logarithmic oracle overhead.

Table 4: Second-order LSE regimes at the common target $\|Ap - b\|_2 \le \varepsilon_{\mathrm{p}}$. All methods start at zero on $\mathcal{H}$. D is a sublevel radius, R a solution-norm bound. Bounds are understood with an additive constant and exact local subproblem solves.

Method	primal output	iterations
CRN	$p(\lambda_k)$	$O(\rho^{5/2}D^{3/2}/(\gamma^{3/2}\varepsilon_{\mathrm{p}}))$
accelerated tensor	$\widehat{p}_k$ in (46)	$O(\rho R^{2/3}/(\gamma^{2/3}\varepsilon_{\mathrm{p}}^{1/3}))$
near-optimal 2nd order	response at returned iterate	$\widetilde{O}(\rho^{6/7}R^{4/7}/(\gamma^{4/7}\varepsilon_{\mathrm{p}}^{2/7}))$

Algorithm 3 Cubic-regularized Newton on the LSE dual

Require: $\lambda_0 \in \mathcal{H}$; $\lambda_0 = 0$ for the comparison in Table 4.

1: **for** $k = 0, \dots, N-1$ **do**
2: $\quad p_k \leftarrow \mathrm{softmax}((A^\top \lambda_k - c)/\gamma)$, $g_k \leftarrow Ap_k - b$
3: $\quad H_k \leftarrow \gamma^{-1} A(\mathrm{Diag}\, p_k - p_k p_k^\top) A^\top$
4: $\quad$ solve (45) over $h_k \in \mathcal{H}$; $\lambda_{k+1} \leftarrow \lambda_k + h_k$
5: **end for**

Preparing an explicit orthonormal gauge basis, when used, is additional preprocessing, and the complexity discussion counts high-order oracle steps and their dense algebra, not unit-cost Hessian or cubic solves. The quasi-self-concordance constant $M_{\mathrm{qsc}} \le 2\rho/\gamma$ used by GRN follows from the same third-derivative formula: bounding one factor by $2\rho\|v\|_2$ and the remaining quadratic form by $\gamma\,\nabla^2\phi_{\mathrm{lse}}[u,u]$ gives $|\nabla^3\phi_{\mathrm{lse}}[u,u,v]| \le (2\rho\|v\|_2/\gamma)\,\nabla^2\phi_{\mathrm{lse}}[u,u]$.

F Primal reconstruction, certificates and transport

For SE, normalize $\widehat{x} = x(\lambda)/\|x(\lambda)\|_1$. Writing $\mathrm{gKL}(u\|v) = \sum_j \big(u_j \log(u_j/v_j) - u_j + v_j\big)$ for the generalized (unnormalized) Kullback–Leibler divergence, the exact identity $F_{\mathrm{se}}(\lambda) = \gamma\,\mathrm{gKL}(x^\star\|x(\lambda))$ gives

$$\mathrm{KL}(x^\star\|\widehat{x}) \le F_{\mathrm{se}}(\lambda)/\gamma, \qquad \|\widehat{x} - x^\star\|_1 \le \sqrt{2F_{\mathrm{se}}(\lambda)/\gamma}, \tag{49}$$

the second by Pinsker's inequality (Cover and Thomas, 2006). The stronger statement is $F_{\mathrm{lse}}(\lambda) = \gamma\,\mathrm{KL}(x^\star\|\widehat{x})$ together with $F_{\mathrm{se}} = F_{\mathrm{lse}} + \gamma(Z - 1 - \log Z)$: the two gaps quantify different errors even at the same normalized primal point.

F.1 A radius-based computable certificate

Suppose $R_{\mathrm{abs}} \ge \|\lambda^\star_{\mathrm{lse}}\|_2$ is supplied for the orthogonal gauge representative. For any simplex point w and any dual λ, define

$$\mathcal{C}_{R_{\mathrm{abs}}}(w,\lambda) := f_s(w) + \phi_{\mathrm{lse}}(\lambda) + R_{\mathrm{abs}}\|Aw - b\|_2. \tag{50}$$

Weak duality at $\lambda^\star_{\mathrm{lse}}$ gives $f_s(w) - f_s^\star \ge \langle \lambda^\star_{\mathrm{lse}}, Aw - b\rangle \ge -R_{\mathrm{abs}}\|Aw - b\|_2$; adding $F_{\mathrm{lse}}(\lambda) = \phi_{\mathrm{lse}}(\lambda) + f_s^\star$ and using $F_{\mathrm{lse}} \ge 0$ yields

$$0 \le F_{\mathrm{lse}}(\lambda) \le \mathcal{C}_{R_{\mathrm{abs}}}(w,\lambda), \qquad -R_{\mathrm{abs}}\|Aw - b\|_2 \le f_s(w) - f_s^\star \le \mathcal{C}_{R_{\mathrm{abs}}}(w,\lambda). \tag{51}$$

Checking $\|Aw - b\|_2 \le \varepsilon_{\mathrm{p}}$ and $\max\{\mathcal{C}_{R_{\mathrm{abs}}}(w,\lambda), R_{\mathrm{abs}}\|Aw - b\|_2\} \le \varepsilon_{\mathrm{f}}$ therefore certifies both the feasibility tolerance and $|f_s(w) - f_s^\star| \le \varepsilon_{\mathrm{f}}$. Taking $w = p(\lambda)$, the identity $f_s(p(\lambda)) + \phi_{\mathrm{lse}}(\lambda) = \langle \lambda, Ap(\lambda) - b\rangle$ makes (50) particularly cheap. This is not a radius-free certificate: a known R_{abs} is an extra input. The softmax, the residual and the value (50) are formed in one final $O(\mathrm{nnz}(A) + m + n)$ pass using log-sum-exp evaluation. A known feasible $x^{\mathrm{f}} > 0$ also supplies an initial SE gap bound without $f^\star$, $\bar{\Delta} = f_s(x^{\mathrm{f}}) + \phi_{\mathrm{se}}(z) \ge F_{\mathrm{se}}(z)$, since $\phi^\star_{\mathrm{se}} + f_s^\star = 0$.

F.2 Comparison at a common residual tolerance

Since $\|Av\|_2 \le \rho\|v\|_1$ and $\mathbb{E}\|A\widehat{x} - b\|_2 \le \rho\sqrt{2\mathbb{E}F_{\mathrm{se}}(\widehat{u})/\gamma}$ by Jensen, running Theorem 4.3 at $\varepsilon = \gamma\varepsilon_{\mathrm{p}}^2/(2\rho^2)$ gives $\mathbb{E}\|A\widehat{x} - b\|_2 \le \varepsilon_{\mathrm{p}}$ at leading accuracy-dependent cost

$$\widetilde{O}\left(\frac{r\rho}{\gamma\varepsilon_{\mathrm{p}}}\sum_B d_B\sqrt{\gamma\kappa_B}\right). \tag{52}$$

For a single phase the same implication holds on its success event. Let $K_{\mathrm{dense}} := \mathrm{nnz}(A)\omega_{\max} + m^2 + m^3 + n$ with $\omega_{\max} = \max_j \|a_j\|_0$ denote the leading dense local algebra plus factor access. Table 5 collects the resulting leading costs.

Table 5: Leading accuracy-dependent arithmetic at residual tolerance ε_{p}. SE also carries the initialization, transient and output terms of Theorem 4.3; dense-method setup is separate. The radii r, R, D belong to different algorithms and are not equated. No universal runtime ordering follows from the powers of ε_{p} alone.

Method	Guarantee	Leading work
SE fixed anchor	expectation	$\widetilde{O}\left(\frac{r\rho}{\gamma\varepsilon_{\mathrm{p}}} \sum_B d_B \sqrt{\gamma\kappa_B}\right)$
LSE CRN	deterministic	$\widetilde{O}\left(K_{\mathrm{dense}} \frac{\rho^{5/2} D^{3/2}}{\gamma^{3/2}\varepsilon_{\mathrm{p}}}\right)$
LSE acc. tensor	deterministic	$\widetilde{O}\left(K_{\mathrm{dense}} \frac{\rho R^{2/3}}{\gamma^{2/3}\varepsilon_{\mathrm{p}}^{1/3}}\right)$
LSE near-optimal	deterministic	$\widetilde{O}\left(K_{\mathrm{dense}} \frac{\rho^{6/7} R^{4/7}}{\gamma^{4/7}\varepsilon_{\mathrm{p}}^{2/7}}\right)$

F.3 Transport specialization

For uniform $d \times d$ transport, singleton and two-marginal block partitions show the difference between query count and arithmetic. The two-block partition lowers the aggregate curvature query factor by d, but each block costs d times more to touch, so the leading weighted work is unchanged. This is the structural reason specialized Sinkhorn updates can dominate generic methods despite favourable block-query counts.

Measured. We instantiate Algorithm 1 on this partition, the one run in the paper that is not singleton. The blocks are the two marginal families, so a column meets exactly one row of each block, $\rho_B = 1$, and (10) holds with $\kappa_B = 2 \log 2 \max_{i \in B} b_i/\gamma$ and $\nu_B = 2/\gamma^2$. Exact minimization over a block is the Sinkhorn half-step in the log domain, and it cannot have a larger value than the explicit step $y - U_B \nabla_B f(y)/\bar{L}_B$, so the admissibility test of Algorithm 1 is met at no cost and the potential argument applies verbatim. A block pass and a Sinkhorn half-step each read the cost matrix once, so half-steps are the implementation-independent axis and a Sinkhorn sweep is two of them; the budget is 4000 passes, what the singleton runs receive in epochs.

Table 6 reports it. The partition is what makes the analysed schedule usable on transport at all: with $\kappa = 1$ it reaches 10^{-6} on three of the five pairs and stops at 1.9 to $3.5 \cdot 10^{-6}$ on the other two, one to two orders below the singleton analysed schedule taking the same exact step, which stalls between $1.5 \cdot 10^{-5}$ and $1.5 \cdot 10^{-4}$ (with the envelope step instead, between $2.1 \cdot 10^{-3}$ and $4.3 \cdot 10^{-2}$). The *rescaled* singleton variant also reaches 10^{-6} on three pairs, in 1906, 2390 and 2850 epochs, but not the same three; on MicroscopyImages, where both finish, the certified block schedule needs 130 passes against its 2850 epochs. What neither does is overtake the method whose step it borrows. Rescaling the gap term by the measured $F_{\mathrm{se}}(\lambda_0)/\bar{\Delta}$ (between $2.1 \cdot 10^{-2}$ and $2.3 \cdot 10^{-1}$, and no more covered by the theorems than the singleton adaptation is) rescues the two pairs that miss but costs CauchyDensity its own, so neither schedule dominates.

Two checks on the implementation, one of which discriminates and one of which does not. The step the analysis turns on is the cancellation identity $M_k^2 a_{k+1}^2 = T_{k+1}$, which is what makes the second moment of the momentum update cancel against the block descent term. Comparing the two sides along a trajectory started at λ_0, they agree to $1.3 \cdot 10^{-15}$ or better on every pair, and dropping the $1/p_{k,B}$ from the momentum update moves the disagreement to $7.5 \cdot 10^{-1}$; that test has teeth. The potential $\Phi_k = T_k F_{\mathrm{se}}(u_k) + \frac{1}{2}\|v_k - \lambda_{\mathrm{se}}^\star\|_2^2$ of Theorem 4.2 also never increases in conditional expectation, taken exactly over both blocks rather than sampled, over 300 steps on each pair – but that test does *not* discriminate, and the reason is worth reporting: $\bar{L}_{k,B}$ exceeds $\|\nabla^2_{BB}\phi_{\mathrm{se}}(y_k)\|_2$ here by a factor $4.5 \cdot 10^5$ to $8.0 \cdot 10^5$, so the inequality has room to absorb a wrong constant. Scaling ν_B down by 10^4 leaves a twelvefold margin and produces no violation. That is the same looseness Section 6 measures, seen from the inside. Realized Φ_k is free to rise; the statement is a supermartingale, not a

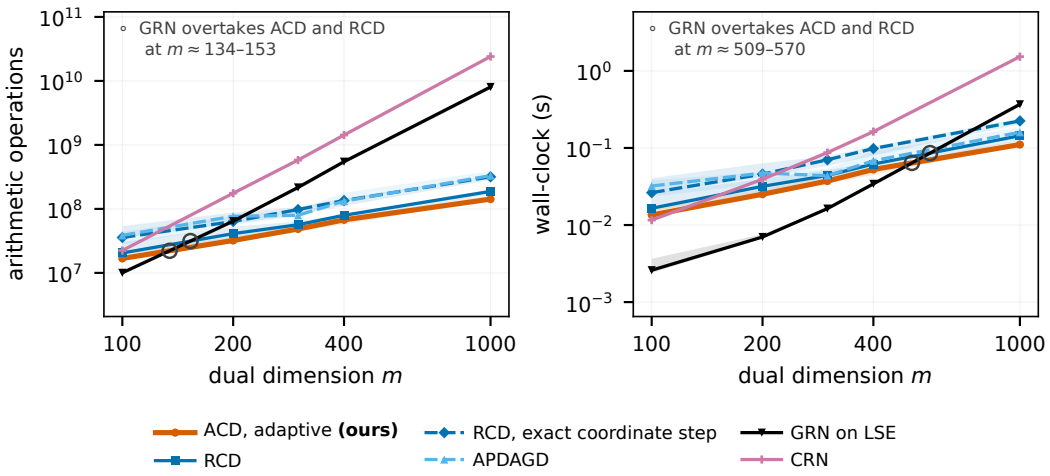

Figure 2: Cost to relative dual gap 10^{-9} against dual dimension, synthetic family ($n = 4m$, $\bar{d} = 75$, $\gamma = 0.05$); medians over five seeds with min–max bands, both axes logarithmic. Second-order curves rise with slope ≈ 3 and coordinate curves with slope ≈ 1, so each second-order/coordinate pair crosses once. Open circles mark where GRN crosses ACD and RCD, located by log-log interpolation between the sampled dimensions. Different pairs cross at different m: CRN is already the more expensive of the two branches in arithmetic at $m = 100$, so its arithmetic crossing lies below the range plotted.

monotone decrease. The reference point must be the SE minimizer, which is the LSE one shifted by $-\gamma \log Z$ along the gauge direction q (Proposition 2.1); on the Shapes pair the unshifted point has $\phi_{\mathrm{se}} = 2.1 \cdot 10^6$ against the optimal 0.101, so the shift is not cosmetic.

Table 6: The two-marginal block partition on DOTmark, 4000 passes. Both block columns are ours: "analysed" is Algorithm 1 with its own constants; "rescaled" divides the gap term by the measured $\bar{\Delta}/F_{\mathrm{se}}(\lambda_0)$ and is not certified. A parenthesized entry is the best relative dual gap reached instead of the target. Half-steps are the comparable axis: a block pass and a Sinkhorn half-step each read the cost matrix once. "Slack" is the smallest ratio $\bar{L}_{k,B}/\|\nabla^2_{BB}\phi_{\mathrm{se}}(y_k)\|_2$ over 300 steps; the potential of Theorem 4.2 decreased in conditional expectation at every one of them, which at this slack is not a test of the constants. One run per pair, at seed 0. Sinkhorn sweeps are logged on this run's finer grid, so they differ by one logging step from the counts quoted in Section 6 (35 against 40 here, 45 against 50 on GRFmoderate).

	block passes to 10^{-6}		Sinkhorn	half-steps		
Pair	analysed	rescaled	sweeps	vs. Sinkhorn	S_0 ratio	slack
ClassicImages	$(3.5 \cdot 10^{-6})$	2750	60	22.9	1.24	$8.0 \cdot 10^5$
Shapes	45	45	10	2.2	1.60	$5.3 \cdot 10^5$
CauchyDensity	750	$(3.0 \cdot 10^{-6})$	50	7.5	3.77	$4.5 \cdot 10^5$
GRFmoderate	$(1.9 \cdot 10^{-6})$	1695	45	18.8	1.34	$6.8 \cdot 10^5$
MicroscopyImages	130	160	35	1.9	3.49	$6.4 \cdot 10^5$

G Experimental protocol

Instances. Synthetic sparse entropy-regularized LPs are drawn with column nonzero counts $\max(1, \mathrm{Bin}(m, s_{\mathrm{col}}/m))$, where the target mean row degree $\bar{d}$ fixes the column sparsity $s_{\mathrm{col}} = \mathrm{round}(\bar{d}\, m/n)$, entries $\mathrm{Unif}(0.5, 1.5)$, and columns rescaled so that $A^\top q = \mathbf{1}$ for $q = \mathbf{1}/m$; a planted feasible $x^{\mathrm{f}} \sim \mathrm{Dir}(\mathbf{1})$ gives $b = Ax^{\mathrm{f}}$, and $c \sim \mathrm{Unif}(0, 1)$. We use $n = 4m$, $\bar{d} = 75$, $\gamma = 0.05$, seeds 0–4; the main instance is $m = 300$ and the density ablation fixes $m = 1000$ with $\bar{d} \in \{10, 25, 75, 150, 300\}$. DOTmark (Schrieber et al., 2017) uses 32×32 images ($m = 2048$, $n = 2^{20}$) with squared Euclidean cost normalized to maximum 1 and $\gamma = 0.01$, on pairs $(1, 2)$ of ClassicImages, Shapes, CauchyDensity, GRFmoderate and MicroscopyImages.

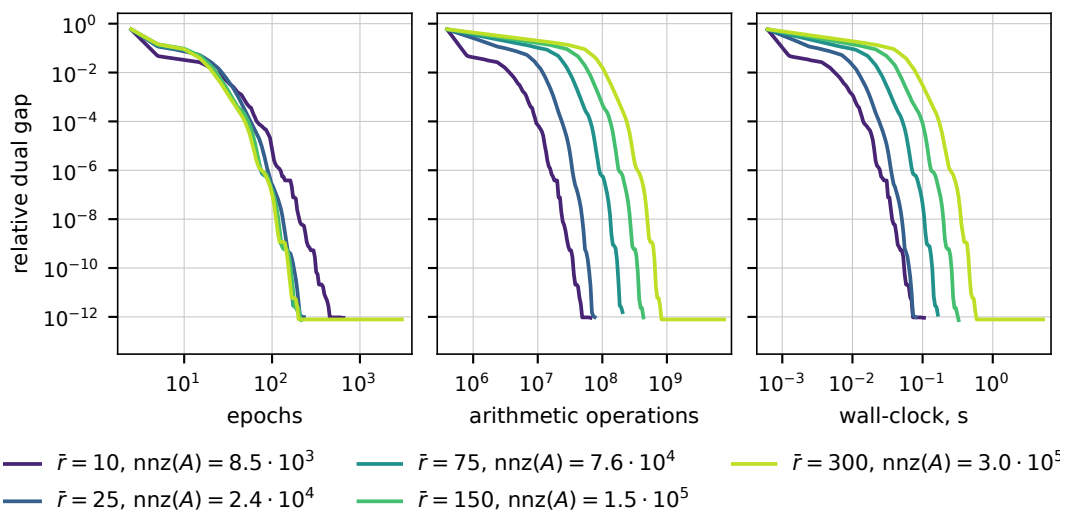

Figure 3: Density ablation at fixed dual dimension. Coordinate epochs are almost density independent, while arithmetic and wall-clock costs track the number of touched nonzeros.

Common start and metric. LSE methods start at $\lambda_0 = 0$; SE methods start at $\lambda_0 = t^\star(0)q = -\gamma \log Z(0)q$, which by Proposition 2.1 gives $x(\lambda_0) = p(0)$ and hence a common initial dual gap for both families. Every curve reports the best-so-far relative dual gap $(\phi(\lambda) - \phi^\star)/|\phi^\star_{\mathrm{se}}|$, with each method measured against its own dual. Because $F_{\mathrm{se}} \geq F_{\mathrm{lse}}$ by (5), this holds the SE methods to the stricter of the two criteria. Medians are over five seeds (or five image pairs). Min–max bands are drawn for the adaptive ACD and RCD curves on the epoch and arithmetic axes; the wall-clock axis and the DOTmark figure carry none. The reference optimum comes from L-BFGS-B on the LSE dual followed by a GRN polish, reaching an SE stationarity residual with median $2 \cdot 10^{-14}$ and worst case $1.7 \cdot 10^{-12}$ across all runs; the 10^{-12} level is therefore at the resolution floor and should not be read as a converged digit.

Methods and inputs. ACD uses $\delta = 1/3$ and $\bar{\Delta} = \phi_{\mathrm{se}}(\lambda_0) + f_s(x^{\mathrm{f}})$ from the planted point. The radius is taken as $r = 2\|\lambda_0 - \lambda^\star_{\mathrm{se}}\|_2$ *from the reference solution*: the runs therefore use an oracle radius, not the computable bound of Proposition J.1, and Section 7 should be read with that in mind. The adaptive variant rescales the baseline κ by backtracking on validity at the sampled coordinates: per epoch, divide by 4 if no violation occurred and the minimum observed ratio was at least 4, and multiply by 4 (capped at the analysed value) after any violation. This rule has no guarantee; its violation counts are reported in Table 7. RCD is Algorithm 4 of Lobanov et al. (2024) with uniform sampling and row-wise constants. GRN, CRN, APDAGD and Sinkhorn use their theoretical constants with adaptive backtracking, with Sinkhorn run in the log domain.

Arithmetic model. A coordinate update is charged per nonzero of the sampled row: 5 operations for ACD (three to form the gradient and the local curvature, two to update the lazy state) and 4 for RCD, plus 3 per nonzero per inner Newton iteration when the exact block step is used, capped at 24 iterations. These per-nonzero constants are the same order but are not calibrated against each other: counting elementary operations in the two loop bodies gives roughly 15 for an ACD coordinate against 6 for an RCD one, so the model favours ACD by about a factor two on this axis. It is adequate for growth exponents, which is what we fit from it, and not for margins near unity; where such a margin matters we quote wall-clock. A matrix–vector product with A costs 1 per nonzero; a Cholesky factorization costs $m^3/3$, a symmetric eigendecomposition m^3, and the sparse Hessian assembly $\sum_j \|a_j\|_0^2 + m^2$; objective evaluations at logging points are charged as incurred. Wall-clock is indicative only: all runs were executed on one laptop, single process and single-threaded BLAS, so the timings describe that one machine, and fitted wall-clock exponents should be read to one decimal.

G.1 THE ANALYSED SCHEDULE AGAINST ITS ADAPTIVE RESCALING

Table 7 reports the sweep described in Section 6, on the $m = 1000$ synthetic instance at target 10^{-6} with a budget of 3000 epochs, against 345 epochs for RCD and 175 for RCD with the exact coordinate step. The final column is the smallest observed ratio between the bound $\kappa_B + \nu_B F$ and the curvature actually encountered at the sampled block, so it measures how much slack the analysis carries on these instances.

Table 7: The analysed schedule (ours) against rescalings of its baseline, $m = 1000$ synthetic instance; medians over five seeds, target 10^{-6} relative, budget 3000 epochs; RCD needs 345 and RCD with an exact coordinate step 175. "scale" multiplies κ_B inside $\bar{L}_{k,B}$, so 1 is Theorem 3.1 as proved and a scale of 10^{-4} shrinks $\sqrt{\kappa_B}$, and with it the accuracy channel of M_k, by 10^{-2}; the growth channel C_1 is left untouched throughout; "epochs" is when the *iterate* first reaches the target, by the same rule applied to RCD; "10^{-12}" counts seeds reaching the stopping tolerance within budget; the schedule's own termination test $T_k \geq \Pi/\varepsilon$ is met at none of the thirty settings, so every epoch count in this sweep is an iterate crossing and never a completed phase; "slack" is the smallest observed ratio of envelope to realized curvature. The full 5×6 sweep over radius and scale is Table 8.

$r/\|\lambda_0 - \lambda^\star_{\mathrm{se}}\|_2$	scale	epochs	10^{-12}	violations	slack
0.5	1	1720	5/5	0	13.6
1	1	—	0/5	0	77.2
1	10^{-2}	280	5/5	0	2.9
1	10^{-4}	125	5/5	0	1.9
2	1	—	0/5	0	$1.5 \cdot 10^3$
5	1	—	0/5	0	$1.1 \cdot 10^4$
20	1	205	0/5	0	$2.9 \cdot 10^4$

Three readings. The inequality holds: at the analysed scale no violation was recorded anywhere in the sweep, on any seed. It is loose, and how loose depends entirely on the radius: the slack runs from 13.6 at $r = 0.5$ to $2.9 \cdot 10^4$ at $r = 20$, because a larger r inflates Π and hence the gap level at which the envelope is evaluated. And the looseness is what the schedule pays: at the analysed scale the 10^{-12} tolerance is reached only at $r = 0.5$, no setting completes a phase, and competitiveness with RCD needs κ shrunk by 10^2 to 10^4, at which point the method is no longer the one analysed in Section 4.

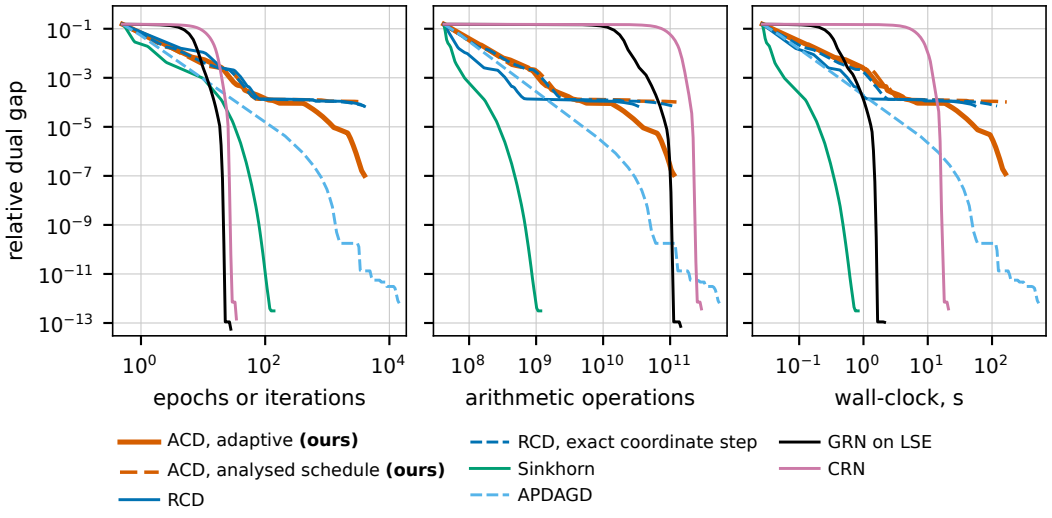

Figure 4: DOTmark 32×32 transport ($m = 2048$, $n = 2^{20}$, $\gamma = 0.01$), five image pairs. Sinkhorn dominates on every axis, by minimizing exactly over each 1024-dimensional marginal block in closed form: where both methods finish, by 32 to 71 on epochs and 230 to 494 on wall-clock.

Provenance of the reported numbers. Every figure and table in the paper is read off one set of runs under the protocol above, Figures 1–4 included. Two quantities are not. The fitted growth exponents of Table 1 and the two-block query estimate quoted in Section 6 are post-processing of the recorded per-iteration traces rather than outputs of a run. And the deterministic checks of the algebraic identities are not optimization runs at all; Appendix K states what they verify.

G.2 The full radius-by-scale sweep

Table 8 is the sweep abbreviated in Table 7. Reading down a column shows the effect of rescaling the analysed baseline at a fixed radius; reading across a row shows that the slack grows with the supplied radius. The epoch count does not. The schedule's own termination test is met at none of the thirty settings, so every entry records the first time the *iterate* passes the target under a schedule that would not have stopped there; the 205 at $r = 20$, scale 1, is that reading and not a faster method. No violation of the envelope was recorded at any of the thirty settings, on any seed.

Table 8: Epochs to relative dual gap 10^{-6} for the analysed schedule (ours), $m = 1000$, medians over five seeds, budget 3000 epochs; "—" means at least one seed did not reach the target, and RCD needs 345. The row under "scale $= 1$" gives the smallest observed ratio of envelope to realized curvature.

scale	$r/\|\cdot\| = 0.5$	1	2	5	20
1	1720	—	—	—	205
slack	13.6	77.2	$1.5{\cdot}10^3$	$1.1{\cdot}10^4$	$2.9{\cdot}10^4$
10^{-1}	530	880	1430	2545	—
10^{-2}	230	280	425	710	355
10^{-3}	150	155	190	195	310
10^{-4}	120	125	135	150	210
10^{-5}	110	115	115	140	205

G.3 A known-solution generator

This is a proposed protocol, not a claim about completed runs. Construct a sparse A with $A^\top q = \mathbf{1}$: one concrete nonnegative construction splits the rows into two nonempty families and makes each column sum to one within each family, using random nonempty supports and normalized positive weights, so that $q = (\mathbf{1}, 0)$ works. For a signed variant, add signed zero-sum perturbations within a family's support, using at least two selected rows so that mixed signs do not change the column sums. The support law, weight distribution, perturbation law and seeds must be fixed and reported. Then choose $x^\star > 0$ with $\mathbf{1}^\top x^\star = 1$, a finite $\lambda^\star_{\mathrm{se}}$, and set

$$b := Ax^\star, \qquad c := A^\top \lambda^\star_{\mathrm{se}} - \gamma \log x^\star. \tag{53}$$

Then $q^\top b = 1$, strict feasibility holds, and $x(\lambda^\star_{\mathrm{se}}) = x^\star$, so by stationarity and convexity this is an exact SE minimizer, and its projection onto $\mathcal{H}$ is an LSE minimizer. Any chosen start z then has the exactly valid radius $r = \|z - \lambda^\star_{\mathrm{se}}\|_2$, and a zero LSE start has $R = \|P_{\mathcal{H}} \lambda^\star_{\mathrm{se}}\|_2$. This supplies a reference solution without running a numerical optimizer.

G.4 Metrics and numerical accuracy

Report the full residual $\|Ap(\lambda) - b\|_2$ and the common normalized gap $F_{\mathrm{lse}}(\lambda)$ for every method. With a known optimum,

$$F_{\mathrm{lse}}(\lambda) = \gamma\, \mathrm{KL}(x^\star \| p(\lambda)), \qquad F_{\mathrm{se}}(\lambda) = \gamma \sum_j x^\star_j h(t_j), \quad t_j = a_j^\top (\lambda - \lambda^\star_{\mathrm{se}})/\gamma.$$

Use log-sum-exp for normalization, compensated or higher-precision summation where cancellation matters, and a series or `expm1`-based evaluation of $h(t)$ near zero. To support a reported 10^{-12} gap, verify that the reference error is well below that threshold rather than estimating it by subtracting nearly equal objective values. The figures divide the gap by $|\phi^\star_{\mathrm{se}}|$, a per-instance constant shared by all methods; the ratio $F_{\mathrm{lse}}(\lambda)/F_{\mathrm{lse}}(\lambda_0)$ at a common start is a normalization-invariant alternative.

Where the implementation still differs from the analysed method. Every run but the two-marginal transport experiment of Appendix F is the $|B| = 1$ specialization of Theorem 4.3, and no run chooses a partition: the one we test is the one transport hands us. Each analysed run keeps the anchor fixed and drives a single phase to the final target, which is the fixed-anchor repetition of Theorem 4.3, and RCD is reported both with the envelope step of its own analysis and with the same exact coordinate solve our variants take, so that epoch counts are comparable. Two deviations remain, both in favour of the analysed method: the runs report a single phase, which is the high-probability guarantee of Theorem 4.2 rather than the expectation guarantee of Theorem 4.3, and the phase loop is bounded by the epoch budget rather than by J_{rep}.

Validity tests and violations of the adaptive rule. The adaptive variant checks the envelope at each sampled coordinate and rescales κ from that. On the $m = 300$ instance it performed 312 107 such tests over five seeds and recorded *no* violation; on $m = 1000$, 1 010 637 tests and again none. The scale it settles on is $\kappa = 1.53 \cdot 10^{-5}$, reached by dividing by 4 once per epoch for eight epochs, and the smallest observed ratio of envelope to realized curvature falls with it, from $2.0 \cdot 10^4$ at $\kappa = 1$ to 73 after five epochs. This is evidence that the rule is not reckless on these instances; it is not a guarantee, because the test only ever sees the sampled coordinate.

G.5 Fixed-constant versus adaptive runs

A direct validation must compare the unchanged fixed-constant wrapper of Theorem 4.3, the adaptive heuristic and RCD on identical instances and starts, reporting the radius and gap inputs, sampling laws, stopping rules, restart behavior, and all preparation, phase-boundary and output work. Changing a sampled block's constant and then reusing that same sample under new sampling weights needs its own conditional-expectation analysis and is not covered by a proof that fixes every constant beforehand. For each adaptive run, log the number of validity tests, violations, rejections and restarts, the violation magnitudes and the recovery rule. A convenient observable is the sampled descent residual $f(y - U_B \nabla_B f(y)/\bar{L}_B) - f(y) + \|\nabla_B f(y)\|_2^2/(2\bar{L}_B)$, together with the floating-point tolerance used to call it positive; passing this local test does not certify all blocks or the global accelerated guarantee. Further useful ablations compare signed against nonnegative rows, uniform against curvature-weighted sampling, and different block partitions, all measured in total arithmetic rather than query counts.

H Smoothness models and their precise implications

All statements here assume a convex C^2 function on $\mathbb{R}^m$ with an attained minimum and $F = f - f^\star$; the Hessian conditions are pointwise global statements. A first-order condition posed only on a neighbourhood also requires its radius to be tracked, and the two formulations are not identified without that step.

H.1 Three distinct conditions

For a convex C^2 function, the following are genuinely different:

$$\partial_{ii}^2 f(u) \le L_{0,i} + L_{1,i}|\partial_i f(u)|, \qquad \text{coordinate-gradient control,} \tag{54}$$

$$\partial_{ii}^2 f(u) \le L_{0,i} + L_{1,i}\|\nabla f(u)\|_2, \qquad \text{full-gradient control of a diagonal,} \tag{55}$$

$$\partial_{ii}^2 f(u) \le \kappa_i + \nu_i F(u), \qquad \text{gap-dependent control (BGG).} \tag{56}$$

Condition (54) implies (55) with the same constants, but the converse fails even for an SE dual, as the example below shows. The *global* (L_0, L_1) model bounds $\|\nabla^2 f\|_2$ rather than only its diagonal. Our accelerated theorem uses (56) and does not silently assume (54) for signed matrices.

H.2 From coordinate-gradient control to a gap bound

Lemma H.1 (A coordinate self-bound)**.** *Under* (54)*, for every u and every coordinate i,*

$$|\partial_i f(u)|^2 \le 2F(u)\big(L_{0,i} + L_{1,i}|\partial_i f(u)|\big), \tag{57}$$

$$\partial_{ii}^2 f(u) \le 2L_{0,i} + 2L_{1,i}^2 F(u). \tag{58}$$

Proof. Write $g = \partial_i f(u)$ and $M = L_{0,i} + L_{1,i}|g|$. For $g \neq 0$, $M > 0$, move along $u - t\,\mathrm{sign}(g)e_i$. Until the directional derivative first reaches zero its magnitude is at most $|g|$ by convexity, so the second derivative is at most M; a zero cannot occur before $t = |g|/M$, since changing the initial derivative $-|g|$ to zero at rate at most M takes at least that distance. The quadratic upper model is therefore valid up to $|g|/M$, giving $f(u - ge_i/M) \leq f(u) - g^2/(2M)$, and comparison with $f^\star$ proves (57). For $g = 0$ it is immediate; if $g \neq 0$ and $M = 0$ then $L_{0,i} = L_{1,i} = 0$ and f is affine with nonzero slope on that line, contradicting attainment.

If $L_{1,i} = 0$, (58) is direct. Otherwise $(M - L_{0,i})^2 = L_{1,i}^2 g^2 \leq 2L_{1,i}^2 F(u)M$, and for $M > 0$ expanding and dividing by M gives $M \leq 2(L_{0,i} + L_{1,i}^2 F(u)) - L_{0,i}^2/M \leq 2(L_{0,i} + L_{1,i}^2 F(u))$. Since $\partial_{ii}^2 f(u) \leq M$ the claim follows; $M = 0$ is immediate from Hessian nonnegativity. □

The proof uses one-dimensional monotonicity of a *partial* derivative; replacing $|\partial_i f|$ by the full gradient norm inside it is not justified, since the other components can change along the coordinate line. Conversely, positive semidefiniteness and the trace bound show that (54) implies a global norm model with $L_0 = \sum_i L_{0,i}$ and $L_1 = (\sum_i L_{1,i}^2)^{1/2}$, that is, aggregated constants rather than the original per-coordinate ones. Applying Theorem 4.2 through (58) yields, at fixed failure level, $\widetilde{O}(1 + r\sum_i \sqrt{L_{0,i}}/\sqrt{\varepsilon} + r\sum_i L_{1,i})$ queries. That is a specialization, not a better L_1 exponent.

H.3 A LOCAL UPPER MODEL FROM THE BGG CONDITION

Lemma H.2 (Explicit local radius)**.** *Suppose* $0 \leq \partial_{ii}^2 f(u) \leq \kappa_i + \nu_i F(u)$ *everywhere. If* $\nu_i > 0$ *and* $|t| \leq 1/(4\sqrt{\nu_i})$*, then*

$$f(u + te_i) \leq f(u) + t\,\partial_i f(u) + (\kappa_i + \nu_i F(u))t^2. \tag{59}$$

For $\nu_i = 0$ *the same holds for every* t *with the sharper coefficient* $\kappa_i/2$*.*

Proof. The block descent self-bound of Appendix C gives $|\partial_i f(u)|^2 \leq 2F(u)(\kappa_i + \nu_i F(u))$. For $\nu_i > 0$ let $\psi(t) = F(u + te_i) + \kappa_i/\nu_i$; then $\psi \geq 0$, $\psi'' \leq \nu_i \psi$ and $|\psi'(0)| \leq \sqrt{2\nu_i}\,\psi(0)$. Integral comparison with the solution of $y'' = \nu_i y$ on either side of zero gives $\psi(t) \leq \psi(0)(\cosh(\sqrt{\nu_i}|t|) + \sqrt{2}\sinh(\sqrt{\nu_i}|t|)) < 2\psi(0)$ for $0 < |t| \leq 1/(4\sqrt{\nu_i})$ (non-strict if $\psi(0) = 0$). Integrating $\partial_{ii}^2 f(u + te_i) \leq 2(\kappa_i + \nu_i F(u))$ proves the model, and since the constant is $\cosh\frac{1}{4} + \sqrt{2}\sinh\frac{1}{4} = 1.389$ the factor 2 could be replaced by 1.39; $\nu_i = 0$ is the global one-dimensional smoothness inequality. □

This makes the comparison with coordinate gap-dependent assumptions explicit: only absolute curvature and radius factors change, and a coordinate condition is still not identified with a full-gradient one.

H.4 SE CONSTANTS AND SIGNED ROWS

For $A \geq 0$, using $A_{ij}^2 \leq \rho_i A_{ij}$,

$$\partial_{ii}^2 \phi_{\mathrm{se}} = \gamma^{-1} \sum_j A_{ij}^2 x_j \leq \frac{\rho_i}{\gamma}(b_i + \partial_i \phi_{\mathrm{se}}) \leq \frac{\rho_i b_i}{\gamma} + \frac{\rho_i}{\gamma}|\partial_i \phi_{\mathrm{se}}|, \tag{60}$$

so $L_{0,i} = \rho_i b_i/\gamma$ and $L_{1,i} = \rho_i/\gamma$: the baseline combines the target constraint value with the largest row coefficient, and $L_{1,i}^{-1}$ is the coordinate displacement over which factor weights change substantially. The same holds for a nonpositive row after flipping signs, but not for a row with both signs. For arbitrary signs the mass identity $q^\top \nabla \phi_{\mathrm{se}} = Z - 1$ gives instead the valid global bound

$$\|\nabla^2 \phi_{\mathrm{se}}\|_2 \leq \gamma^{-1} \sum_j x_j \|a_j\|_2^2 \leq \frac{\rho^2}{\gamma} Z \leq \frac{\rho^2}{\gamma} + \frac{\rho^2 \|q\|_2}{\gamma} \|\nabla \phi_{\mathrm{se}}\|_2. \tag{61}$$

These sign-free constants can be loose and do not imply (54).

To see that the gap is real, take $\gamma = 1$, $c = 0$, $A = \left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$, $b = (1, 0)^\top$, $q = (1, 0)^\top$. The strictly feasible primal point is $(1/2, 1/2)$ and the SE dual is $f(u, v) = e^{u+v} + e^{u-v} - u$ with minimizer $(-\log 2, 0)$. At $v = 0$ we have $\partial_v f = 0$ while $\partial_{vv}^2 f = 2e^u$, so no finite constants satisfy (54) in that coordinate, yet (61) still holds. The sign-free theorem of Section 3 therefore covers objectives that the strong coordinate model excludes.

I Computable geometry and the balanced example

I.1 Effective dimension and its crossover

Summing the singleton envelopes of Theorem 3.1 at $\theta = 2$ turns (κ, ν) into one gap-dependent dimension $d_{\mathrm{curv}}(\Delta)$, interpolating from the effective rank $\operatorname{tr} H^\star / \|H^\star\|_2$ at $\Delta = 0$ to the row spread $d_A = \sum_i \rho_i^2/\rho^2$ as $\Delta \to \infty$, with $\rho_i = \|A_{i:}\|_\infty$ and $\rho = \max_j \|a_j\|_2$. It is what decides square-root sampling for a fixed Hessian, and measuring both ends calibrates the motivation for coordinate methods: on our family the effective rank grows in proportion to m, so the predicted gain is a constant 4.2 rather than $\sqrt{m}$, and on transport it is 1.6 to 3.4 – the instances with the least spectral asymmetry to exploit, which is part of why Section 6 loses to Sinkhorn there. Appendix I carries the proof, the crossover, the balanced family and the full measurement.

I.2 The gap-dependent dimension, and what it measures

The interpolation is exact: $d_{\mathrm{curv}}(\Delta)$ equals $\operatorname{tr} H^\star / \|H^\star\|_2$ at $\Delta = 0$, tends to d_A as $\Delta \to \infty$, exchanges the two at $\Delta_{\mathrm{cross}} = \gamma^2 \log 2 \, \|H^\star\|_2/\rho^2$, and has both endpoints in $[1, m]$. The matching statement about the method is that $\nu_B \equiv 0$ gives $C_1 = 0$ and $N = \widetilde{O}(1 + rC_0/\sqrt{\delta\varepsilon})$, exactly the accelerated coordinate bound of Allen-Zhu et al. (2016); Nesterov and Stich (2017): generalized smoothness recovers the classical rate when its growth channel vanishes.

Measured directly on the generated instances, on the synthetic family d_H grows in proportion to m, from 5.7 at $m = 100$ to 57.9 at $m = 1000$, so $\sqrt{m/d_H}$ is a constant 4.2 across the sweep; it moves with density instead, from 2.4 at $\bar{d} = 10$ to 8.3 at $\bar{d} = 300$. On DOTmark $d_A = m/2$ exactly, since an incidence matrix has $\rho_i = 1$ and $\rho = \sqrt{2}$, while d_H ranges over 175 to 821, giving 1.6 to 3.4. Δ_{cross} is $2 \cdot 10^{-4}$ to $3 \cdot 10^{-3}$ on the synthetic family, far below the gaps the runs traverse, so the method spends its time at the d_A end.

I.3 The gap-dependent dimension

With $\rho_i = \|A_{i:}\|_\infty$, $\rho = \max_j \|a_j\|_2$ and $\theta = 2$ in Theorem 3.1, the singleton envelopes on $\{F \le \Delta\}$ are

$$\ell_i(\Delta) = 2 \log 2 \, H^\star_{ii} + 2\Delta\rho_i^2/\gamma^2, \qquad \ell_{\mathrm{full}}(\Delta) = 2 \log 2 \, \|H^\star\|_2 + 2\Delta\rho^2/\gamma^2. \tag{62}$$

Proposition I.1 (Effective dimension and its crossover)**.** *Assume $H^\star \neq 0$ and define*

$$d_{\mathrm{curv}}(\Delta) := \frac{\sum_i \ell_i(\Delta)}{\ell_{\mathrm{full}}(\Delta)}, \qquad d_H := \frac{\operatorname{tr} H^\star}{\|H^\star\|_2}, \qquad d_A := \frac{\sum_i \rho_i^2}{\rho^2}. \tag{63}$$

With $\Delta_{\mathrm{cross}} := \gamma^2 \log 2 \, \|H^\star\|_2/\rho^2$,

$$d_{\mathrm{curv}}(\Delta) = \frac{\Delta_{\mathrm{cross}}}{\Delta_{\mathrm{cross}} + \Delta} \, d_H + \frac{\Delta}{\Delta_{\mathrm{cross}} + \Delta} \, d_A. \tag{64}$$

Both d_H and d_A lie in $[1, m]$, so $d_{\mathrm{curv}}(0) = d_H$ and $d_{\mathrm{curv}}(\Delta) \to d_A$ as $\Delta \to \infty$.

The identity follows by summing (62). Above Δ_{cross} the relevant dimension is d_A, which asks how the largest row coefficients are spread across factors; below it d_H takes over. A balanced incidence family on which $d_H = d_A = m/s$ collapses the interpolation is given in Proposition I.4.

I.4 Remarks on the two dimensions

First, the range claimed in Proposition I.1. For a nonzero PSD matrix, $\|H^\star\|_2 \le \operatorname{tr} H^\star \le m\|H^\star\|_2$, giving $d_H \in [1, m]$; and since $\rho^2 = \max_j \|a_j\|_2^2 \ge \max_{i,j} A_{ij}^2$ while $\sum_i \rho_i^2 = \sum_i \max_j A_{ij}^2 \ge \max_j \|a_j\|_2^2$, we get $\rho^2 \le \sum_i \rho_i^2 \le m\rho^2$ and hence $d_A \in [1, m]$. A convex combination of the two then lies in $[1, m]$ as well. Three further points deferred from Section I.1. First, d_A is not invariant under a change of dual coordinates, whereas d_H is invariant under orthogonal ones (a non-orthogonal rescaling changes $\operatorname{tr} H^\star / \|H^\star\|_2$ as well): the quantity $\sum_i \rho_i^2$ depends on the basis in which the rows are expressed, so both the coordinate system and the data-access costs matter when reading d_{curv}. Second, the two terms of ℓ_i scale differently in the regularization, as γ^{-1} for the baseline and γ^{-2} for

the growth coefficient; since $H^\star$ also depends on γ, individual entries need not be monotone in γ and the crossover level Δ_{cross} moves with it. Third, (64) is stated for singleton coordinates on purpose: block operator norms do not sum to a trace, so while Theorem 3.1 and the algorithm remain valid for any partition, the literal effective-rank identity does not survive blocking. The cost-weighted ratio below is the quantity that does generalize.

Corollary I.2 (Exact constants for scaled indicator rows)**.** *If $A_{ij} \in \{0, \rho_i\}$ with $\rho_i > 0$, then $A_{ij}^2 = \rho_i A_{ij}$, so $H^\star_{ii} = \rho_i \sum_j A_{ij} x^\star_j/\gamma = \rho_i b_i/\gamma$ and $\kappa_i = (2\log 2/\gamma)\rho_i b_i$ uses the exact diagonal of $H^\star$ with no knowledge of $x^\star$. Incidence matrices, and hence optimal transport, are of this form.*

I.5 Cost-weighted work ratio

For a convex quadratic with Hessian H the coordinate curvatures are $L_i = H_{ii}$ and the full-gradient curvature is $L = \|H\|_2$, so square-root importance sampling (Allen-Zhu et al., 2016) compares $\sum_i \sqrt{H_{ii}}$ against $m\sqrt{\|H\|_2}$:

$$\frac{\sum_i \sqrt{H_{ii}}}{m\sqrt{\|H\|_2}} \le \sqrt{\frac{d_H}{m}}, \qquad d_H := \frac{\operatorname{tr} H}{\|H\|_2}, \tag{65}$$

which is the classical effective-rank criterion that Section I.1 generalizes. With $d_i = 1 + \mathrm{nnz}(A_{i:})$ and $d_{\mathrm{full}} = \sum_i d_i$ as the nominal full-pass cost, define

$$\mathcal{R}_{\mathrm{work}}(\Delta) := \frac{\sum_i d_i \sqrt{\ell_i(\Delta)}}{d_{\mathrm{full}}\sqrt{\ell_{\mathrm{full}}(\Delta)}}. \tag{66}$$

For equal d_i, (65) gives $\mathcal{R}_{\mathrm{work}}(\Delta) \le \sqrt{d_{\mathrm{curv}}(\Delta)/m}$; for heterogeneous costs it also measures whether the stiff coordinates are the expensive ones. Replacing $H^\star_{ii}$ and $\|H^\star\|_2$ by chosen upper bounds gives the *implemented* comparison, and a ratio of loose bounds is not an intrinsic dimension.

Proposition I.3 (Accuracy and growth contributions)**.** *Choose valid singleton constants (κ_i, ν_i) and full-block constants $(\kappa_{\mathrm{f}}, \nu_{\mathrm{f}})$ with $\kappa_{\mathrm{f}}, \nu_{\mathrm{f}} > 0$, and set $G_0 := S_0/(d_{\mathrm{full}}\sqrt{\kappa_{\mathrm{f}}})$ and $G_1 := S_1/(d_{\mathrm{full}}\sqrt{\nu_{\mathrm{f}}})$. For $r, \varepsilon > 0$ and the same start, radius, gap bound, target and confidence, the ratio of the two leading work expressions of Theorem 4.3 and its one-block specialization is*

$$\frac{rS_0/\sqrt{\varepsilon} + rS_1}{rd_{\mathrm{full}}(\sqrt{\kappa_{\mathrm{f}}}/\sqrt{\varepsilon} + \sqrt{\nu_{\mathrm{f}}})} = \omega_\varepsilon G_0 + (1-\omega_\varepsilon)G_1, \qquad \omega_\varepsilon := \frac{\sqrt{\kappa_{\mathrm{f}}}/\sqrt{\varepsilon}}{\sqrt{\kappa_{\mathrm{f}}}/\sqrt{\varepsilon} + \sqrt{\nu_{\mathrm{f}}}}. \tag{67}$$

If all growth coefficients vanish, only G_0 remains.

The identity is algebraic; that these expressions occur in valid bounds follows by applying Theorem 4.3 to the two partitions. Logarithms, initialization, output and additive sampling terms stay in the full bounds and their ratio is not claimed to equal (67). This is neither a runtime lower bound nor a claim of optimality among sampling laws.

I.6 The balanced incidence family

Proposition I.4 (Balanced incidence family)**.** *Let $A \in \{0,1\}^{m\times n}$ have exactly s_{col} ones in every column, let $b = (s_{\mathrm{col}}/m)\mathbf{1}$, and assume strict feasibility. Then $q = \mathbf{1}/s_{\mathrm{col}}$ certifies the redundant mass equation, $H^\star_{ii} = s_{\mathrm{col}}/(m\gamma)$, $\|H^\star\|_2 = s^2_{\mathrm{col}}/(m\gamma)$, and*

$$d_H = d_A = d_{\mathrm{curv}}(\Delta) = m/s_{\mathrm{col}} \qquad (\Delta \ge 0), \tag{68}$$

with the bounds of (8) *exact. If row degrees are equal, the cost-weighted ratio $\mathcal{R}_{\mathrm{work}}$ of* (66) *equals $1/\sqrt{s_{\mathrm{col}}}$ at every gap level, so a factor coupling s_{col} constraints separates the coordinate and one-block* leading bounds *by $\sqrt{s_{\mathrm{col}}}$; when $s_{\mathrm{col}} = O(1)$, sparsity alone delivers no large spectral gain. This compares upper bounds for one method under two partitions, not a lower bound for any competitor.*

For a binary column with s_{col} ones, $\mathbf{1}^\top a_j = s_{\mathrm{col}}$ and $\|a_j\|_2^2 = s_{\mathrm{col}}$, so $q = \mathbf{1}/s_{\mathrm{col}}$ satisfies $A^\top q = \mathbf{1}$ and $q^\top b = 1$; strict feasibility makes every row nonempty. Binary entries give $H^\star_{ii} = \gamma^{-1}\sum_j A_{ij}x^\star_j = b_i/\gamma = s_{\mathrm{col}}/(m\gamma)$, and

$$H^\star\mathbf{1} = \gamma^{-1}\sum_j x^\star_j a_j(a_j^\top \mathbf{1}) = (s_{\mathrm{col}}/\gamma)b = (s^2_{\mathrm{col}}/(m\gamma))\mathbf{1}.$$

Every row sum of the symmetric nonnegative matrix $H^\star$ equals $s_{\mathrm{col}}^2/(m\gamma)$, which is therefore an eigenvalue and bounds every eigenvalue in absolute value, so $\|H^\star\|_2 = s_{\mathrm{col}}^2/(m\gamma)$. The trace is s_{col}/γ, giving $d_H = m/s_{\mathrm{col}}$; since $\rho_i = 1$ and $\rho^2 = s_{\mathrm{col}}$, also $d_A = m/s_{\mathrm{col}}$. The singleton overlap bound equals $b_i/\gamma = H^\star_{ii}$, and for the full block $\omega_{Bj} = s_{\mathrm{col}}$ gives $s_{\mathrm{col}} \max_i b_i/\gamma = s_{\mathrm{col}}^2/(m\gamma) = \|H^\star\|_2$: both baselines are available without spectral information about the unknown optimizer. For every gap level $\ell_{\mathrm{full}}(\Delta) = s_{\mathrm{col}}\, \ell_i(\Delta)$, and with equal row degrees $d_i = d$, $d_{\mathrm{full}} = md$, substitution gives $\mathcal{R}_{\mathrm{work}} = 1/\sqrt{s_{\mathrm{col}}}$. The implemented constants are

$$\kappa_i = 2\log 2\, s_{\mathrm{col}}/(m\gamma), \quad \nu_i = 2/\gamma^2, \qquad \kappa_{\mathrm{f}} = 2\log 2\, s_{\mathrm{col}}^2/(m\gamma), \quad \nu_{\mathrm{f}} = 2s_{\mathrm{col}}/\gamma^2,$$

so $G_0 = G_1 = 1/\sqrt{s_{\mathrm{col}}}$ in Proposition I.3. Initialization and finite phase overhead do not scale by this factor and remain in Theorem 4.3.

I.7 A fully specified construction

Let $n = m$, index rows and columns modulo m, and set

$$A_{ij} = \mathbf{1}\{(i-j) \bmod m \in \{0, \dots, s_{\mathrm{col}} - 1\}\}. \tag{69}$$

Every row and column has s_{col} ones and $x^\star = \mathbf{1}/m$ gives $b = (s_{\mathrm{col}}/m)\mathbf{1}$. For any fixed finite $\lambda^\star$, choosing $c = A^\top \lambda^\star - \gamma \log x^\star$ makes the KKT equations hold at that optimizer. Repeating each column equally often gives larger n with the same formulas after splitting the primal mass. Taking $\gcd(m, s) = 1$ and $s < m$ gives a full-rank square construction when that is useful, though full rank is not needed for Proposition I.4.

I.8 A polynomial gap between the two accountings

Let $t \geq 2$ be an integer, $m = 2$, $n = t^2 + 1$ and $\gamma = 1$. The first t^2 columns of A equal e_1 and the last equals e_2. Set

$$b = (t^{-4}, 1 - t^{-4})^\top, \qquad x_j^\star = t^{-6}\ (j \leq t^2), \qquad x_n^\star = 1 - t^{-4}, \qquad c = -\log x^\star.$$

Then $q = (1, 1)^\top$ certifies total mass, $x^\star > 0$ is feasible, and $\lambda^\star = 0$ is an SE minimizer. From the common start $z = (1, 1)^\top$ the supplied radius is $r = \sqrt{2}$ and $F(z) = e - 2$, both independent of t. The exact singleton constants and costs are $\kappa_i = 2\log 2\, b_i$, $\nu_i = 2$, $d_1 = t^2 + 1$, $d_2 = 2$, so $C_0 = \Theta(1)$, $C_1 = \Theta(1)$, $S_0 = \Theta(1)$ and $S_1 = \Theta(t^2)$, whereas $(\max_i d_i)C_0 = \Theta(t^2)$. At target $\varepsilon = t^{-6}$, Theorem 4.3 including initialization and output gives $\mathbb{E}W_{\mathrm{total}} = \widetilde{O}(t^3)$, while multiplying the query bound by the worst block cost gives only $\widetilde{O}(t^5)$; all transient and preparation terms are $\widetilde{O}(t^2)$ and do not conceal the difference. This separates two *accounting rules for the same method*; it is not a lower bound on another algorithm, and the example has simple structure that a specialized solver could exploit.

I.9 Tightness of the subclass

The range–moment bound need not be tight for signed rows. Append a mass row to $(-1, 0, 1)$ and set $x^\star = (t, 1 - 2t, t)$ with $0 < t < 1/2$. For the signed row $b_i = 0$ and $H^\star_{ii} = 2t/\gamma$ while $\bar{H}_i^{\mathrm{rng}} = 1/\gamma$, so the ratio $1/(2t)$ is unbounded. Choosing $c = -\gamma \log x^\star$ makes $\lambda^\star = 0$ an attained SE minimizer and all assumptions hold. This is not a counterexample to the work bound; it shows why effective rank alone cannot guarantee a comparable curvature sum in the implementation. More generally, if $H^\star_{ii} > 0$ and $\bar{H}_i \leq \varkappa H^\star_{ii}$ for all active coordinates with $\varkappa \geq 1$, then $2\log 2\, \bar{H}_i + \nu_i \Delta \leq \varkappa\, \ell_i(\Delta)$, so against the full envelope the weighted numerator loses at most $\sqrt{\varkappa}$. The indicator-row corollary has $\varkappa = 1$; no uniform finite $\varkappa$ holds for arbitrary signed data.

J Initial gap and distance bounds

Obtaining r and $\bar{\Delta}$. Neither input is free. A strictly feasible $\bar{x} > 0$ gives $\bar{\Delta} = \phi_{\mathrm{se}}(z) + P(\bar{x})$ by weak duality, P being the objective of (1). The radius is harder: Proposition J.1 derives one from $\bar{x}$ and a lower bound on the smallest nonzero singular value of A, but it is conservative (Section 7), and obtaining both inputs is preprocessing not counted in W_{init}. Let $P(x) = \langle c, x\rangle + \gamma \sum_j x_j(\log x_j - 1)$ be the primal objective of (1). Producing a feasible point is additional information and the statement below does not assume it comes for free.

Proposition J.1 (Bounds from strict feasibility)**.** *Suppose $\bar{x} > 0$ with $A\bar{x} = b$ and $\mathbf{1}^\top \bar{x} = 1$ is supplied. Define*

$$B_z := \phi_{\mathrm{se}}(z) + P(\bar{x}), \qquad x_{\min} := \min_j \bar{x}_j > 0, \qquad \bar{t}_j := 1 + \frac{B_z}{\gamma \bar{x}_j}.$$

Then $B_z \geq F_{\mathrm{se}}(z) \geq 0$, so $\bar{\Delta} = B_z$ is valid. Let $0 < \widehat{\sigma} \leq \sigma^+_{\min}(A)$ be a verified lower bound on the smallest nonzero singular value. There exists an SE minimizer $\lambda^\star$ with the same orthogonal projection onto $\ker A^\top$ as z and

$$\|z - \lambda^\star\|_2 \leq r_{\mathrm{cert}} := \frac{2\gamma}{\widehat{\sigma}} \|\bar{t}\|_2 \leq \frac{2\gamma\sqrt{n}}{\widehat{\sigma}} \Big(1 + \frac{B_z}{\gamma x_{\min}}\Big). \tag{70}$$

The first bound also controls the diameter of the initial SE sublevel restricted to that gauge slice. If $B_z = 0$, return z.

Proof. For $t_j(\lambda) = (a_j^\top \lambda - c_j)/\gamma - \log \bar{x}_j$, feasibility of $\bar{x}$ gives the exact identity

$$\phi_{\mathrm{se}}(\lambda) + P(\bar{x}) = \gamma \sum_j \bar{x}_j \, h(t_j(\lambda)), \tag{71}$$

and weak duality gives $B_z \geq F_{\mathrm{se}}(z)$. On $\{\phi_{\mathrm{se}}(\lambda) \leq \phi_{\mathrm{se}}(z)\}$ each nonnegative summand is at most B_z. We have $|t| \leq 1 + h(t)$: for $t < 0$ from $h(t) \geq -t - 1$, and for $t \geq 0$ from $h(t) \geq t^2/2$ and $\sqrt{2h} \leq 1 + h$. Hence $|t_j(\lambda)| \leq \bar{t}_j$ on the sublevel. The dual is invariant along $\ker A^\top$ because $b = A\bar{x}$ is orthogonal to that kernel, so shift an attained minimizer inside the kernel to match the kernel component of z. Then $z - \lambda^\star \in \operatorname{range} A$, where the singular-value bound applies, and both endpoints lie in the initial sublevel:

$$\widehat{\sigma}\|z - \lambda^\star\|_2 \leq \|A^\top(z - \lambda^\star)\|_2 = \gamma \|t(z) - t(\lambda^\star)\|_2 \leq 2\gamma \|\bar{t}\|_2.$$

The same argument applies to any two sublevel points in the chosen gauge, and the second inequality follows from $\bar{x}_j \geq x_{\min}$. □

This SE gauge removes only $\ker A^\top$; it is *not* the larger LSE invariance space $\mathcal{G}$, since ϕ_{se} is not invariant in the mass direction q. No gauge projection is needed in the sparse iterations; the theorem only needs a valid initial distance to some minimizer.

A spectral bound for the cyclic family. For the square cyclic matrix (69), the Fourier singular values are s_{col} and $|\sin(\pi s_{\mathrm{col}} k/m)/\sin(\pi k/m)|$ for $k = 1, \ldots, m-1$. If $1 \leq s_{\mathrm{col}} < m$ and $\gcd(m, s_{\mathrm{col}}) = 1$, then $s_{\mathrm{col}} k \not\equiv 0$, every numerator is at least $\sin(\pi/m)$ in magnitude, and every denominator is at most one, so $\widehat{\sigma} = \sin(\pi/m)$ is valid. The uniform feasible point has $x_{\min} = 1/m$, giving

$$r_{\mathrm{cert}} \leq \frac{2\gamma\sqrt{m}}{\sin(\pi/m)} \Big(1 + \frac{mB_z}{\gamma}\Big).$$

This shows computability without a dense SVD; it is not a sharp practical radius, and on this family it exceeds the true distance by a factor that grows with m; see Section 7. The known-solution protocol of Appendix G instead supplies the exact distance to the constructed minimizer.

Preprocessing. Given $\bar{x}$ and $\widehat{\sigma}$, forming these bounds costs one data pass and vector operations. Finding $\bar{x}$ and verifying $\widehat{\sigma}$ can be considerably more expensive; denote that work by W_{geometry} and add it to $W_{\mathrm{init}} + W_{\mathrm{opt}} + W_{\mathrm{output}}$ when the inputs are not provided. A small feasibility margin $x_{\min}$ or a small spectral gap makes (70) very conservative: a finite bound is not a guarantee of a well-conditioned instance.

K FIXED-SUPPORT VALIDATION

This specifies *additional validation*, not completed optimization benchmarks. The generator below is deterministic and what it feeds are numerical checks of the algebraic identities; no new convergence curves are produced.

Fix one binary matrix (69), a $\gamma > 0$ and an index j_0. For $0 \le \vartheta < 1$ set

$$x^\star(\vartheta) = (1-\vartheta)\mathbf{1}/m + \vartheta e_{j_0}, \qquad b(\vartheta) = Ax^\star(\vartheta), \qquad c(\vartheta) = -\gamma \log x^\star(\vartheta). \tag{72}$$

The matrix and its support never change, strict positivity holds, and $\lambda^\star = 0$ is an exact SE minimizer throughout. Use a common nonoptimal start, for example $z = \gamma e_{i_0}$ on a nonempty row, for both the singleton and one-block methods; the supplied radius $r = \|z\|_2 = \gamma$ is then valid for every ϑ. A zero start would already solve the generated problem and must not be used.

For every ϑ the coordinate constant is exact by Corollary I.2, and at $\vartheta = 0$ the balanced formulas apply. As $\vartheta \uparrow 1$,

$$H^\star(\vartheta) \longrightarrow a_{j_0} a_{j_0}^\top/\gamma, \qquad d_H(\vartheta) \longrightarrow 1, \qquad d_A = m/s \text{ for all } \vartheta,$$

so the protocol sweeps d_H across its whole range while holding sparsity, access costs and the radius fixed. Do not use $\vartheta = 1$, which violates strict positivity. A computable full-block constant is $\bar{H}_{\mathrm{f}}(\vartheta) = s \max_i b_i(\vartheta)/\gamma$; away from balance it need not equal $\|H^\star(\vartheta)\|_2$.

Record d_H, d_A, Δ_{cross}, the exact and computed curvature sums, $\bar{H}_{\mathrm{f}}/\|H^\star\|_2$, and the weighted work factors. For actual optimization runs also record the starting gap, radius input, common absolute target and residual, seeds, phase comparisons and all data-access work. Spectral diagnostics are computed offline, their cost reported separately, and must not be silently handed to an otherwise bound-only algorithm. Note that the protocol changes b, c and in general the initial gap, so those changes should be reported rather than claiming the spectrum is the only thing that moved. A separate ablation should vary the association between row cost d_i and baseline curvature at fixed aggregate sparsity, reporting $S_0 = \sum_i d_i \sqrt{\kappa_i}$ alongside $(\max_i d_i) \sum_i \sqrt{\kappa_i}$.

L Resource comparison and a sharper L_1 target

L.1 Comparing native guarantees

In Table 9, R is a distance to a solution, $\widetilde{R}$ a sublevel radius in the cited coordinate analysis, and r a supplied upper bound used in our schedule. Our row describes expectation after fixed-anchor repetition, not a deterministic guarantee.

The cost of a frozen constant. The comparison the introduction states in one line is this. The method already takes an input $\bar{\Delta} \ge F(z)$, and the interpolation argument of Appendix C confines the accelerated iterate to $\{F \le 10\bar{\Delta}/\delta\}$, so $\kappa_B + 10\nu_B\bar{\Delta}/\delta$ is a legitimate coordinate Lipschitz constant, fixed before the first iteration and requiring no schedule of ours. Feeding it to Allen-Zhu et al. (2016) gives $N = \widetilde{O}\big(r(C_0 + \sqrt{\bar{\Delta}}C_1)/\sqrt{\varepsilon}\big)$, against $N = \widetilde{O}\big(1 + rC_0/\sqrt{\delta\varepsilon} + rC_1/\sqrt{\delta}\big)$ in Theorem 4.2. The accuracy channel $rC_0/\sqrt{\varepsilon}$ is identical; the growth channel is charged $\sqrt{\bar{\Delta}/\varepsilon}$ times more by the frozen constant. Since $\bar{\Delta}$ is an input that a user can only over-estimate, and the runs of Section 6 show the envelope exceeding the curvature actually met by up to $2.9 \cdot 10^4$, that factor is paid on the instances we run and not only in the bound.

Restarts, and what the continuous schedule does not buy. A single frozen constant is not the sharpest alternative; restarts are. Run the method in phases $p = 0, 1, \ldots$, each targeting half the gap of the last, and inside phase p freeze the constants at that phase's certified level $\kappa_B + 10\nu_B\Delta_p/\delta$ with $\Delta_p = 2^{-p}\bar{\Delta}$, which is the level the same interpolation argument supplies at the phase's own gap bound. Lemma 4.1 applies verbatim with M frozen: it only needs $M^2 \ge \mathcal{T}(10\Delta_p/\delta)$ on the phase's sublevel set, which the phase target guarantees. Summing the per-phase counts of Theorem 4.2 over the geometric schedule gives

Proposition L.1 (Restarted ACD)**.** *Let D bound the diameter of $\{F \le \bar{\Delta}\}$ in the gauge slice. Restarted ACD with per-phase frozen constants reaches $\mathbb{E}F \le \varepsilon$ in $\widetilde{O}\big(DC_0/\sqrt{\varepsilon} + DC_1\big)$ block queries.*

This is (13) with r replaced by D, and both channels match. The comparison therefore turns entirely on whether a radius smaller than the sublevel diameter is available, and here we must be plain: the only radius we can compute, Proposition J.1, bounds exactly that diameter. On the inputs a user

can actually supply, the continuous schedule does not improve on restarts; what it removes is the phase bookkeeping – re-validating the anchor, recomputing the sampling tables and discarding the momentum at every boundary – not a factor in the rate. We therefore present it as restart-free rather than as faster, and leave open whether an instance family with a computable $r \ll D$ exists. The geometric schedule is out of reach at the budgets of Section 6: with the proved constants, certifying the first phase target needs $T_k = 4\Pi/\bar{\Delta}$, and 3000 epochs cover a quarter of that distance at $m = 300$ and a sixth at $m = 1000$, so the runs tagged `ACD:restart` complete a single phase and carry no comparison between the two schedules.

Table 9: Native gap-accuracy bounds and the extra resources each method consumes. Logarithms and absolute constants are suppressed. A segment query is not a unit-cost partial derivative.

Method	Native update count	Information and auxiliary work
Ordinary ACD (Allen-Zhu et al., 2016)	$\widetilde{O}(R\sum_i \sqrt{L_i}/\sqrt{\varepsilon})$	Fixed coordinate smoothness; sparse accelerated state when the problem supports it.
Lobanov (2026, Theorem 8)	$\widetilde{O}\big(\widetilde{R}(\sum_i \sqrt{H_{0,i}}/\sqrt{\varepsilon} + \sum_i \sqrt{H_{1,i}})\big)$	Coordinate gap model; known $f^\star$; segment relaxation. Phase-wise sampling and verified doubling available.
AGMsDR (Nesterov et al., 2021; Vankov et al., 2025)	$\widetilde{O}(R\sqrt{L_0/\varepsilon} + (L_1 R)^{2/3})$	Global gradient-dependent model; full gradients and function differences; segment minimization costs extra oracle calls.
This paper	$\widetilde{O}(1 + rC_0/\sqrt{\varepsilon} + rC_1)$	BGG constants, supplied r, $\bar{\Delta}$; one sampled block gradient; no segment solve and no current-gap evaluation inside a phase.

Lobanov's extensions retain a known $f^\star$ while removing the sublevel radius from the inputs by verified doubling, and inexact relaxation and phase-wise sampling belong in any fair comparison (Lobanov, 2026). Neither fixing the parameters beforehand nor non-uniform square-root weights distinguishes our method; the claim concerns the missing auxiliary solve and the sparse implementation of the fixed-constant schedule, not a better native exponent. The bridge between the Hessian condition and a local model is Lemma H.2.

The full-gradient result of Vankov et al. (2025) motivates a sharper coordinate L_1 dependence, but its proof cannot be imported by replacing a full gradient with an unbiased coordinate estimator: its monotonicity, segment optimality and accumulated-descent arguments would all have to survive that replacement. Function-value and directional-derivative costs on generic affine sums should be counted by touched factors rather than hidden behind an update count.

Relation to accelerated alternating minimization. Guminov et al. (2021) give an accelerated alternating-minimization method with a k^{-2} rate, which on the entropic transport dual is accelerated block coordinate ascent over the two marginal families. It is the natural point of comparison for this paper on the LSE side, and the comparison is instructive precisely because the two methods buy different things. Their per-block step is an exact minimization, which is affordable only when the block subproblem has closed form; ours is a fixed-curvature step, which needs no subproblem solver at all but pays the envelope. Their iteration touches an entire marginal, so the per-iteration cost is a full pass over the data and no sparsity claim is made; ours touches d_B nonzeros. And their rate is stated for the normalized dual, where the Hessian is bounded, so the difficulty our Theorem 3.1 addresses does not arise. A fair head-to-head would need a family carrying both a block structure with solvable subproblems *and* enough sparsity for the access cost to matter; transport has the first and not usefully the second, and our synthetic family the reverse. We therefore report the structural comparison that transport does allow — Sinkhorn against a generic sparse method, Section 6 — and do not claim a comparison we have not run.

Relation to relative smoothness. The SE objective is relatively smooth with respect to a suitable Bregman reference, and the Bregman literature (Bauschke et al., 2017; Lu et al., 2018; Hanzely et al., 2021) supplies methods that avoid a global Lipschitz constant by a different route. Two things separate

the present analysis. First, accelerated Bregman methods require a triangle-scaling or symmetry condition on the reference function, and their optimal rates are constrained by the lower bounds of Dragomir et al. (2021); we instead keep the Euclidean geometry and move the difficulty into a gap-dependent constant. Second, and this is the binding objection, a Bregman proximal step on the SE dual does not decompose over sparse blocks in a way that preserves the exact lazy representation of Section 4.1, which is the property the arithmetic bound rests on. A proper comparison of the two routes on entropy duals is worth doing and is not attempted here.

L.2 A proved aggregation inequality and an unproved extension

Lemma L.2 (Gradient-sensitive square-root aggregation)**.** *For $L_{0,i}, L_{1,i} \geq 0$, set $\mathcal{A}_0 = \sum_i \sqrt{L_{0,i}}$ and $\mathcal{B}_1 = \sum_i L_{1,i}^{2/3}$. Every $g \in \mathbb{R}^m$ satisfies*

$$\Big(\sum_i \sqrt{L_{0,i} + L_{1,i}|g_i|}\Big)^2 \leq 2\mathcal{A}_0^2 + 2\mathcal{B}_1^{3/2}\|g\|_2. \tag{73}$$

Proof. By Hölder with exponents $4/3$ and 4, $\sum_i \sqrt{L_{1,i}|g_i|} \leq (\sum_i L_{1,i}^{2/3})^{3/4}(\sum_i |g_i|^2)^{1/4}$. Apply $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ termwise, then $(a+b)^2 \leq 2a^2 + 2b^2$. □

An unproved target. Under (54), (73) suggests seeking a method with query complexity $\widetilde{O}(1 + r\mathcal{A}_0/\sqrt{\varepsilon} + r^{2/3}\mathcal{B}_1)$. For equal $L_{1,i} = L_1$ its growth contribution would be $m(L_1 r)^{2/3}$ instead of the $mL_1 r$ that the established gap-based reduction gives, an improvement in the large-$L_1 r$ regime only. A plausible sufficient ingredient would be a one-phase bound

$$\mathbb{E}F(u_N) \leq K\frac{r^2\mathcal{A}_0^2}{N^2} + K\frac{r^2\mathcal{B}_1^3}{N^3}\Delta, \qquad F(z) \leq \Delta, \tag{74}$$

with universal K, an admissible sampling and acceptance scheme, and an appropriate radius invariant. Equation (74) is a target, not a lemma of this paper, and is not established for Algorithm 1; even if proved for one phase it would still need a separate radius argument before moving-anchor restarts. Computing every $|g_i|$ to set probabilities would cost a full gradient, so the aggregation lemma by itself provides neither a sparse sampling implementation nor the required stochastic monotonicity. The proved query bound remains Theorem 4.2 and its specialization through Lemma H.1.

L.3 Limitations in detail

Looseness of the supplied radius. Proposition J.1 bounds the distance to the optimum by $\gamma\sqrt{n}\,\widehat{\sigma}^{-1}(1 + B_z/(\gamma x_{\min}))$ with $x_{\min} = \min_j \bar{x}_j$. On the balanced family of Proposition I.4 this grows polynomially in m while the true distance stays $O(\gamma)$, so feeding the bound to Algorithm 1 would inflate every phase length; the runs therefore supply an oracle radius, and Table 7 shows the slack growing with r.

The dynamic range of the SE iteration. With $e^{O(\|A^\top(\lambda-\lambda^\star)\|_\infty/\gamma)}$ as the range of the unnormalized factors, overflow sets in once $\|A^\top(\lambda-\lambda^\star)\|_\infty/\gamma \gtrsim 710$, a displacement of about 7 at $\gamma = 10^{-2}$ for unit entries and much less for the entries of size $\approx m/s$ in our family – long before the iterate is inaccurate in any meaningful sense. The LSE dual is immune because $p(\lambda)$ is shift invariant and can be formed in the log domain; that asymmetry, not the rate, is what keeps small-γ transport off the SE route.

Scope of the separations. The $\sqrt{s_{\mathrm{col}}}$ gain of Proposition I.4 and the polynomial gain of Appendix I.8 both hold between *our own* upper bounds under two accountings of the same method. They are not runtime claims, and they are not lower bounds against any competitor; Appendix L states what a fair comparison of native guarantees would need instead.