STOCHASTIC GRADIENT DESCENT FOR SEMI-DISCRETE OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Semi-discrete optimal transport maps a continuous probability distribution onto finitely many target points with prescribed masses. We study stochastic gradient descent (SGD) for its Kantorovich dual energy, using source samples to estimate the masses of the associated Laguerre cells. For quadratic cost, a compact convex source domain, and a positive bounded Hölder continuous density, we derive an explicit lower bound on the smallest positive eigenvalue of the dual Hessian from a transport variance inequality. The bound is linear in the smallest cell mass. The main difficulty is that cells can become empty during SGD, making this eigenvalue bound vanish. We prove an explicit restricted strong-convexity bound on an entire bounded region of potentials, including empty-cell configurations. Clipping and centering keep every iterate in this region. Together with an explicit smoothness bound, decreasing steps and weighted averaging give an $O(1/T)$ expected dual-energy error, with separate initialization and sampling terms. We also introduce MC-guided SGD, which uses sampled spectra to choose practical steps and can restart the guaranteed algorithm when needed. Experiments up to dimension 128 and 5000 targets show lower errors than the global schedule in many settings, with competitive accuracy against projected SGD and a measurable spectral-estimation overhead.

1 INTRODUCTION

Optimal transport (OT) redistributes one probability distribution into another at minimum cost. Its geometry supports image comparison, shape matching, domain adaptation, and generative modeling (Peyré & Cuturi, 2019). In *semi-discrete* OT, a continuous source is matched to finitely many target sites with prescribed masses. For quadratic cost, one scalar potential per target determines a cell of source points sent to that target. Sampling these cells estimates the gradient of a convex energy, making stochastic gradient descent (SGD) a natural solver even when explicit cell construction is expensive.

The main obstacle in the convergence analysis is that cells can become empty during optimization. A positive Hessian eigenvalue bound at the solution does not provide the same bound along stochastic iterates. We prove an explicit inequality controlling the gradient toward the solution throughout a bounded region of potentials, even at empty cells. Keeping the iterates in this region then supplies the condition needed for an $O(1/T)$ convergence guarantee.

Geometric and learning background. Prescribed-mass power diagrams (Aurenhammer et al., 1998) and the variational formulation of Gu et al. (2016) connect semi-discrete OT to convex geometry and discrete Monge–Ampère equations. This structure supports multiscale and geometric optimization algorithms (Mérigot, 2011; Lévy, 2015). Gangbo & McCann (1996, Example 1.6) proposed following the mass-mismatch vector field and conjectured convergence; the quantitative concavity theory of Kitagawa–Mérigot–Thibert (KMT) answers this gradient-flow conjecture under their geometric assumptions (Kitagawa et al., 2019).

OT also underlies Wasserstein GANs and their gradient-penalty variants (Arjovsky et al., 2017; Gulrajani et al., 2017). AE-OT uses Monte Carlo cell masses to fit a latent-space transport map after autoencoder training (An et al., 2020). In flow matching, OT-CFM forms training pairs by minibatch OT (Tong et al., 2024), convex neural potentials provide another coupling construction (Kornilov

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.

et al., 2024), and SD-FM fits a semi-discrete coupling to the full dataset (Mousavi-Hosseini et al., 2026). These applications motivate efficient solvers with quantitative guidance for their learning rates.

Stochastic optimization and the choice of step size. Genevay et al. (2016) study stochastic optimization of the semi-dual objective. More recently, projected stochastic gradient descent (PSGD) has been analyzed for unregularized semi-discrete OT, including $O(1/T)$ guarantees for averaged potentials and dual-energy errors (Genans et al., 2025b). Decreasing entropic regularization averaged gradient (DRAG) obtains these rates while reducing entropic regularization (Genans et al., 2025a). These results motivate a concrete question: *which properties of the source distribution and target points control the constants used to choose an SGD step size?* Other results control cell-mass discrepancies (Mousavi-Hosseini et al., 2026) or statistical map-estimation error (Pooladian et al., 2023). Here T denotes optimization steps for a fixed transport problem; Appendix C compares the error measures.

Our proof has two geometric steps. First, the transport variance inequality of Letrouit & Mérigot (2026) gives a Hessian lower bound proportional to the smallest cell mass. Second, we integrate this bound along segments from the optimizer and account separately for empty cells. This proves restricted strong convexity (RSC) on the whole bounded region used by SGD (Section 4.3). The second step is essential for our $O(1/T)$ analysis: the RSC inequality must hold at every iterate, not just near the optimizer. Our estimate makes the dependence on source geometry, target masses, and the allowed potential range explicit.

Contributions. For quadratic cost on a compact convex source domain with a positive bounded Hölder continuous density, we establish the following results.

- **Explicit eigenvalue bounds.** We prove $\lambda_2(\nabla^2 E(h)) \geq p_{\min}(h)/A$, where $p_{\min}(h)$ is the smallest cell mass and A is an explicit geometric constant. The dependence on the smallest mass is linear, compared with the cubic dependence in KMT's bound. Together with a smoothness bound, this gives an upper bound on the Hessian condition number (Propositions 4.1 and 4.2).
- **A bound valid throughout optimization.** We prove an explicit RSC inequality on a bounded set of potentials, including configurations with empty cells (Theorem 4.5). This is the key step connecting the Hessian estimate to the stochastic iteration: clipping and centering keep every iterate within the set where the inequality holds.
- **SGD with explicit parameters.** Using these bounds, decreasing steps and weighted averaging give $O(1/T)$ expected dual-energy error, with explicit initialization and batch-size dependence (Theorem 3.3).
- **MC-guided practical steps.** Algorithm 2 uses sampled eigenvalues to guide SGD and permits a restart with the proved schedule. Experiments up to $d = 128, k = 5000$ show substantially lower errors than the global schedule in many settings and competitive accuracy against PSGD, while accounting for estimation cost.

Section 2 introduces the transport problem and dual energy. Section 3 states the algorithm and convergence results; Section 4 derives the required geometric constants. Section 5 reports experiments. Full proofs and experimental protocols are in the appendices.

2 Semi-discrete optimal transport

Consider a probability measure $\mu = f(x)\,dx$ on a compact convex set $\Omega \subset \mathbb{R}^d$ with nonempty interior. This means that μ has density f with respect to Lebesgue measure. Let $\nu = \sum_{i=1}^k \nu_i \delta_{y_i}$ be a probability measure supported on k distinct target points $y_i \in \mathbb{R}^d$, where δ_{y_i} is a unit point mass. Semi-discrete optimal transport asks for a map that sends the prescribed amount of source mass to each target at minimum cost:

$$\min_{T:\Omega\to\{y_1,\dots,y_k\}} \int_\Omega \tfrac{1}{2}\|x - T(x)\|^2\, d\mu(x), \qquad \mu(T^{-1}(\{y_i\})) = \nu_i \quad \text{for every } i. \tag{1}$$

We use the following assumptions throughout the analysis.

Assumption 2.1. The source density satisfies $0 < f_{\min} \le f(x) \le f_{\max} < \infty$ on Ω and $\int_\Omega f(x)\,dx = 1$. It is Hölder continuous: there are $L_f < \infty$ and $\gamma \in (0, 1]$ such that $|f(x) - f(x')| \le L_f \|x - x'\|^\gamma$ for all $x, x' \in \Omega$. The target weights satisfy $\nu_i \ge \nu_{\min} > 0$ and $\sum_i \nu_i = 1$, with $k \ge 2$ distinct target points.

The quadratic-cost transport theorem gives an optimal map, unique μ-almost everywhere, because μ is absolutely continuous (Villani, 2009).

The dual energy and its cells. Kantorovich duality reduces (1) to a finite-dimensional optimization problem. After a change of variables in the dual potentials, there is one scalar h_i per target, and we minimize

$$E(h) = \int_\Omega \phi_h(x)\,d\mu(x) - \sum_i \nu_i h_i, \qquad \phi_h(x) = \max_i \{\langle x, y_i\rangle + h_i\}. \tag{2}$$

The function ϕ_h is the upper envelope of k affine functions. The part where plane i is highest projects onto the *Laguerre cell*

$$W_i(h) = \{x \in \Omega : \langle x, y_i\rangle + h_i \ge \langle x, y_j\rangle + h_j \text{ for every } j\}, \qquad p_i(h) = \mu(W_i(h)).$$

Thus $p_i(h)$ is the mass currently assigned to target y_i. A tie means that two or more planes attain the same maximum. Shared cell boundaries lie in $(d-1)$-dimensional hyperplanes and have zero μ measure. Differentiating the energy gives

$$\nabla E(h) = p(h) - \nu. \tag{3}$$

A gradient step raises the potential of a cell with insufficient mass and lowers that of a cell with excess mass. At a minimizer h^*, $p_i(h^*) = \nu_i$, and the optimal map is simply $T(x) = y_i$ for $x \in W_i(h^*)$, almost everywhere.

Adding the same number to every potential changes neither the cells nor E. We therefore choose the centered representative

$$H_0 = \{h \in \mathbb{R}^k : \mathbf{1}^\top h = 0\}, \qquad \Pi_{H_0} h = h - \tfrac{1}{k}(\mathbf{1}^\top h)\mathbf{1},$$

where $\mathbf{1} = (1, \dots, 1)^\top$ and Π_{H_0} is the orthogonal projection onto H_0. Under Assumption 2.1, the centered minimizer $h^* = \arg\min_{h \in H_0} E(h)$ exists and is unique (Gu et al., 2016; Kitagawa et al., 2019). Appendix A.1 gives the dual change of variables and differentiation details; see also Genans et al. (2025b).

The Hessian as a graph Laplacian. When all cells have positive mass, the Hessian is the Jacobian of p:

$$L(h) = \nabla^2 E(h) = Dp(h), \qquad L_{ij} = -w_{ij}\ (i \ne j), \qquad L_{ii} = \sum_{j \ne i} w_{ij}, \tag{4}$$

where

$$w_{ij} = \frac{1}{\|y_i - y_j\|} \int_{W_i(h) \cap W_j(h)} f(x)\,d\mathcal{H}^{d-1}(x).$$

Here $\mathcal{H}^{d-1}$ measures interface area (length in two dimensions). Cells sharing an interface of positive area are connected by an edge; otherwise $w_{ij} = 0$. This is the weighted Laguerre graph, and $L(h)$ is its graph Laplacian. Its eigenvalues determine how rapidly the gradient changes with h. The target weights enter E linearly, so its Hessian at a fixed h does not depend on ν.

A two-dimensional example. Figure 1 uses the uniform source on $\Omega = [0, 1]^2$, sites $y_1 = (0.2, 0.2)$, $y_2 = (0.8, 0.2)$, $y_3 = (0.5, 0.8)$, and masses $\nu = (0.2, 0.3, 0.5)$. Here $\int \phi_h\,d\mu$ is the signed volume under the envelope. Raising plane i changes this integral at rate $p_i(h)$, while the linear term contributes $-\nu_i$. Minimizing E balances the cell areas with the prescribed masses.

3 Stochastic gradient descent and convergence

We estimate the cell masses using fresh source samples and update the potentials by SGD. We first state the conditions needed by the algorithm, then give its convergence guarantees. Section 4 proves these conditions for semi-discrete OT and provides explicit constants.

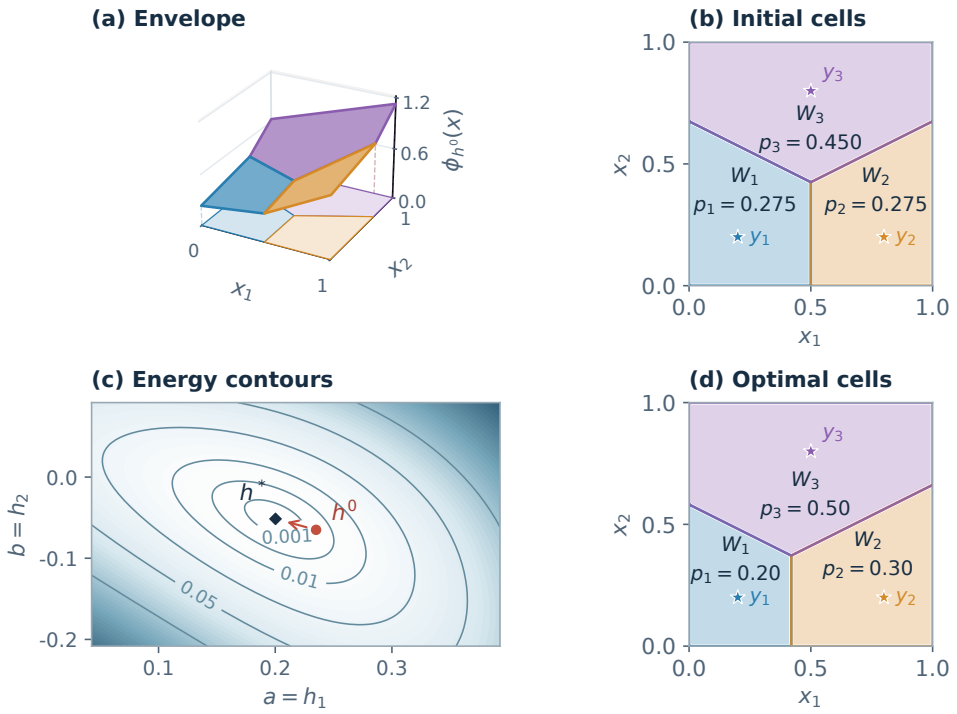


Figure 1: Geometry of the dual energy. (a) The envelope and (b) cells at centered Voronoi potentials $h^0 = (0.235, -0.065, -0.170)$; stars mark target sites. (c) Energy contours on H_0, parameterized by $h = (a, b, -a - b)$; the arrow connects the displayed configurations. (d) At the minimum, cell masses equal ν.

3.1 Conditions and algorithm

Besides convexity, the analysis uses two inequalities:

(A1) **Smoothness.** For all $h, g \in H_0$, $\|\nabla E(h) - \nabla E(g)\| \le \beta\|h - g\|$. The constant β limits the size of a stable gradient step.

(A2) **Restricted strong convexity (RSC).** On the region $\mathcal{R}_C = \{h \in H_0 : \max_i h_i - \min_i h_i \le 2C\}$,

$$\langle \nabla E(h), h - h^* \rangle \ge a\|h - h^*\|^2, \qquad a > 0. \tag{5}$$

This inequality measures the gradient component toward the fixed minimizer. Ordinary strong convexity implies it, but is not required. We use the RSC convention of Genans et al. (2025b, Lemma 4.2).

For $h_0 \in H_0$, choose $C \ge \max\{\|h^*\|_\infty, \|h_0\|_\infty\}$. Proposition 4.1 gives a bound $R_0 \ge \|h^*\|_\infty$ using only the source domain and target locations, so h^* need not be known. Theorem 4.5 supplies a for this choice of C; we take $0 < a \le \beta$.

After each step, clip coordinates to $[-C, C]$ and subtract their mean. With $P_{[-C,C]^k}$ denoting coordinatewise clipping, the update map $\mathcal{T} = \Pi_{H_0} P_{[-C,C]^k}$ satisfies

$$\|\mathcal{T}(z) - h^*\| \le \|z - h^*\|, \qquad \max_i \mathcal{T}(z)_i - \min_i \mathcal{T}(z)_i \le 2C. \tag{6}$$

Both projections fix h^*, and subtracting the mean preserves the coordinate range. Thus every iterate stays in the region where (A2) holds. A useful initialization is $h_0 = \Pi_{H_0}(-\|y_i\|^2/2)_{i=1}^k$: its cells are ordinary Voronoi cells intersected with Ω.

Algorithm 1 uses step sizes $\eta_t = 2/[a(t + s)]$, where $s = 4\beta/a$ ensures $\eta_0 = 1/(2\beta)$. Later steps decrease as $2/(at)$. It averages the iterates with weights $w_t = t + s - 1$, assigning more weight to

Algorithm 1 RSC-SGD with decreasing steps and weighted averaging

Require: Source sampler, target points and weights; $h_0 \in H_0$; C, a, β as above; batch size $B \geq 1$; iterations $T \geq 1$

$s \leftarrow 4\beta/a$; $S \leftarrow 0$; $\bar{h} \leftarrow h_0$
for $t = 0, \dots, T-1$ **do**
  $w \leftarrow t + s - 1$; $S \leftarrow S + w$; $\bar{h} \leftarrow \bar{h} + (w/S)(h_t - \bar{h})$
  Draw $X_{t,1}, \dots, X_{t,B} \sim \mu$ independently
  Assign each sample to its cell: $I_{t,b} \in \arg\max_i(\langle X_{t,b}, y_i \rangle + h_{t,i})$
  $G_t \leftarrow B^{-1} \sum_{b=1}^{B} e_{I_{t,b}} - \nu$
  $\eta_t \leftarrow 2/[a(t+s)]$; $h_{t+1} \leftarrow \Pi_{H_0} \operatorname{clip}(h_t - \eta_t G_t, -C, C)$
end for
return $\bar{h}$ (the weighted average of $h_0, \dots, h_{T-1}$)

later estimates. This choice follows the weighted SGD analysis of Lacoste-Julien et al. (2012); Stich (2019); Appendix A.5 derives the recursion under (A1)–(A2).

Here e_i is the ith coordinate vector. Each update requires $O(Bkd)$ arithmetic; averaging adds $O(k)$ storage.

3.2 Gradient estimate and convergence rates

The basic statistical property is particularly simple: each source sample belongs to exactly one cell, except on a set of probability zero.

Lemma 3.1 (Unbiased gradient and bounded variance)**.** *Let $\mathcal{F}_t$ denote the history before drawing the fresh batch. The gradient estimate in Algorithm 1 satisfies*

$$\mathbb{E}[G_t \mid \mathcal{F}_t] = \nabla E(h_t), \qquad \mathbb{E}[\|G_t - \nabla E(h_t)\|^2 \mid \mathcal{F}_t] = \frac{1 - \|p(h_t)\|^2}{B} \leq \frac{1}{B}.$$

The variance bound does not depend on the number of targets or dimension.

Proof. The assignment vector $e_{I_{t,b}}$ has mean $p(h_t)$ and squared norm one. Independence of the B samples divides its variance by B. □

For comparison, smooth convex optimization already gives a $T^{-1/2}$ guarantee without (A2).

Theorem 3.2 (Constant steps and uniform averaging)**.** *Assume (A1), and let $U_0 \geq \mathbb{E}\|h_0 - h^*\|^2$. Use the same gradient and clipping/centering update with a constant $0 < \eta \leq 1/\beta$, and return $\widetilde{h}_T = T^{-1} \sum_{t=0}^{T-1} h_t$. Then*

$$\mathbb{E}[E(\widetilde{h}_T) - E(h^*)] \leq \frac{U_0}{\eta T} + \frac{\eta}{B}.$$

For $U_0 > 0$, choosing $\eta = \min\{1/\beta, \sqrt{BU_0/T}\}$ gives $O(\beta U_0/T + \sqrt{U_0/(BT)})$.

Using (A2) improves the rate to T^{-1} with decreasing steps and weighted averaging.

Theorem 3.3 (Decreasing steps and weighted averaging)**.** *Assume (A1)–(A2), and let $U_0 \geq \mathbb{E}\|h_0 - h^*\|^2$. Algorithm 1, with any $s \geq 4\beta/a$, satisfies*

$$\mathbb{E}[E(\bar{h}_T) - E(h^*)] \leq \frac{2a(s-1)(s-2)U_0}{T(T+2s-3)} + \frac{8}{aB(T+2s-3)}. \tag{7}$$

In particular, $s = 4\beta/a$ gives $O(\beta^2 U_0/(aT^2) + 1/(aBT))$.

The first term records the effect of initialization; the second records sampling noise. The bound $U_0 = k(C + R_0)^2$ is always available and is used only in the analysis. No separate warmup is needed. For fixed geometry and batch size, the expected energy error is $O(1/T)$. The total number of source samples is BT. Although the gradient variance is dimension independent, a, β, and U_0 depend on the problem geometry.

Proof outline. Expanding the squared distance to h^* gives a recursion with three terms: contraction, energy decrease, and sampling noise. With the stated steps and weights, consecutive distance terms cancel when summed. Convexity then bounds the energy of the weighted average. Appendix A.5 gives this recursion and proves the rate.

4 Explicit estimates for the SGD constants

We now verify (A1)–(A2). The upper eigenvalue bound gives smoothness. A variance inequality gives a lower eigenvalue bound when cells have positive mass. A further geometric argument yields the RSC inequality on the entire region visited by the algorithm.

4.1 Smoothness and a bound on the optimal potentials

Proposition 4.1 (Smoothness and potential bounds)**.** *Under Assumption 2.1, let $R_X = \sup_{x\in\Omega}\|x\|$, $D_Y = \max_{i,j}\|y_i - y_j\|$, and $R_0 = R_X D_Y$. Then $\|h^*\|_\infty \le R_0$. Let A_Ω be the largest $(d-1)$-dimensional area of the intersection of Ω with a hyperplane, with $A_\Omega = 1$ for $d = 1$. The energy satisfies (A1) with*

$$\beta = 2 f_{\max} A_\Omega \max_i \sum_{j\neq i} \|y_i - y_j\|^{-1}. \tag{8}$$

Moreover, $|p_i(h) - p_i(g)| \le \beta\|h - g\|_\infty$ for every i, h, g.

A hyperplane section is a line segment in two dimensions or a planar slice in three dimensions. The proof first bounds the mass gained or lost by each cell when the potentials change. Since $p - \nu = \nabla E$, Lemma A.1 converts this componentwise bound into the Euclidean smoothness bound (8), using symmetry of the Hessian of a smoothed energy. The argument includes empty cells and changes in cell adjacency (Appendix A.1). Equation (8) is a sufficient upper bound; uniform boxes admit the sharper bounds in Proposition A.2.

4.2 A lower bound from the transport variance inequality

The next result is the main estimate. It makes the dependence on cell masses explicit and supplies the lower bound used in the SGD analysis.

Proposition 4.2 (Lower bound on the Hessian eigenvalues)**.** *Under Assumption 2.1, fix a target center $y_0 \in \mathbb{R}^d$ and set*

$$D_X = \operatorname{diam}\Omega, \qquad R_Y = \max_i \|y_i - y_0\|, \qquad A = e\frac{f_{\max}}{f_{\min}} D_X R_Y. \tag{9}$$

At every h with $p_{\min}(h) := \min_i p_i(h) > 0$,

$$L(h) \succeq A^{-1}\{\operatorname{diag}(p(h)) - p(h)p(h)^\top\}, \qquad L(h)|_{H_0} \succeq \frac{p_{\min}(h)}{A} I. \tag{10}$$

In particular, $\lambda_2(L(h^*)) \ge \nu_{\min}/A$. Here $\succeq$ compares quadratic forms, and λ_2 is the smallest positive eigenvalue; $L(h)\mathbf{1} = 0$ accounts for the zero eigenvalue.

Proof. Specializing Letrouit & Mérigot (2026, Theorem 2.1) gives

$$\operatorname{Var}_\mu(\phi_g - \phi_h) \le A\langle g - h, p(g) - p(h)\rangle, \qquad h, g \in \mathbb{R}^k, \tag{11}$$

where $\operatorname{Var}_\mu(u) = \min_c \int_\Omega (u - c)^2 \, d\mu$. Set $g = h + tv$ and divide by t^2. The envelope difference quotient converges almost everywhere to v_i on cell i and is bounded by $\|v\|_\infty$. Differentiability of p at positive-cell configurations (Kitagawa et al., 2019, Theorem 1.3) and dominated convergence give

$$v^\top\{\operatorname{diag}(p(h)) - p(h)p(h)^\top\}v = \operatorname{Var}_{p(h)}(v) \le A v^\top L(h) v.$$

For $v \in H_0$, $\operatorname{Var}_{p(h)}(v) \ge p_{\min}(h)\min_c \sum_i (v_i - c)^2 = p_{\min}(h)\|v\|^2$. At h^*, $p(h^*) = \nu$. These inequalities prove all the claims; Appendix A.2 details the specialization and limit. □

Corollary 4.3 (Strong convexity and condition number)**.** *Under Assumption 2.1, on any convex region $\mathcal{C} \subset H_0$ where $p_i(h) \geq \varepsilon > 0$ for every i, E is strongly convex with constant $\alpha = \varepsilon/A$. At the optimum, the condition number of the Hessian restricted to H_0 is the ratio of its largest to smallest eigenvalue and satisfies*

$$\kappa_* := \frac{\lambda_{\max}(L(h^*))}{\lambda_2(L(h^*))} \leq \frac{\beta A}{\nu_{\min}}.$$

Proof. Integrating (10) along segments in $\mathcal{C}$ proves strong convexity. The ratio bound follows from $\lambda_{\max}(L(h^*)) \leq \beta$ (Proposition 4.1) and $\lambda_2(L(h^*)) \geq \nu_{\min}/A$ (Proposition 4.2). □

Remark 4.4 (Mass dependence and step sizes)**.** The lower bound ε/A is linear in the smallest cell mass, compared with $C_{\mathrm{KMT}}\varepsilon^3$ in KMT's Theorem 5.1. The geometric constants differ; KMT also treats other costs. If $A \leq A_0$ and $\nu_i \geq c_0/k$ for fixed $A_0, c_0 > 0$, then $\lambda_2(L(h^*)) \geq c_0/(A_0 k)$. A large condition number means that steps stable in directions with large eigenvalues can progress slowly in directions with small eigenvalues; both eigenvalues therefore matter when choosing a learning rate.

4.3 Verifying restricted strong convexity along the iterations

The optimum eigenvalue bound alone is not sufficient for SGD: an iterate may have an empty cell and a zero eigenvalue on H_0. The following theorem completes the link between the Hessian estimate and the SGD convergence guarantee. It proves condition (A2) throughout the region preserved by the algorithm, without requiring every intermediate cell to have positive mass.

Theorem 4.5 (Explicit RSC constant)**.** *Under Assumption 2.1, for every $D > 0$ and $h \in H_0$ with $\operatorname{osc}(h) := \max_i h_i - \min_i h_i \leq D$,*

$$\langle \nabla E(h), h - h^* \rangle \geq a_D \|h - h^*\|^2, \qquad a_D = \frac{\nu_{\min}}{A(d+1)f_{\max}/f_{\min} + R_0 + D}. \tag{12}$$

Consequently, $a = \min\{a_{2C}, \beta\}$ verifies (A2) for Algorithm 1, including iterates with empty cells.

The proof integrates (10) along segments from h^* and handles empty cells by raising their affine functions until they touch the envelope. It is given in Appendix A.3. The constant a_D controls the gradient relative to the optimum; it is not a lower bound on every Hessian eigenvalue in this region. Enlarging the allowed potential range D makes this guarantee weaker.

Together, Propositions 4.1–4.2 and Theorem 4.5 provide explicit inputs for the convergence results of Section 3. These bounds can be conservative; Section 5 compares them with exact and sampled Hessian eigenvalues.

5 Experimental evaluation

We test geometric constants and compare theoretical and MC-guided SGD with matched averaging. All accuracy targets are nonuniform. One-dimensional tests use analytic population solutions; higher-dimensional tests estimate cell-mass errors on independent samples. Appendix B gives reference calculations and experimental protocols.

Costs beyond the quadratic case. An exact change of potentials extends our analysis to forward Bregman costs: affine score slopes become $z_i = \nabla\phi(y_i)$ and the geometric constants are recomputed for these slopes (Appendix B.1). This covers the forward KL and Itakura–Saito (IS) tests below on compact positive-coordinate domains.

5.1 How conservative are the constants?

We use $k = 12$ equally spaced sites, target mass ratio four, and uniform sources on $[0, 1]$ for quadratic cost and $[0.2, 1]$ for KL and IS. Table 1 compares the RSC constant a, the optimum lower bound $\alpha_{\mathrm{LM},*} = \nu_{\min}/A$, and global smoothness β with exact eigenvalues $\alpha_* = \lambda_2(L(h^*))$ and $\beta_* = \lambda_{\max}(L(h^*))$. For quadratic cost, the lower bound 0.0268 is about 30 times below $\alpha_* = 0.818$; the RSC constant is smaller still. The path bound reduces β from 113.6 in (8) to 48. The step-size ratio $\beta/a = 7548$ and optimum condition number $\kappa_* = 57.7$ concern different validity regions.

Problem	a	$\alpha_{\mathrm{LM},*}$	β	α_*	β_*	β/a	κ_*
Quadratic ($k = 2$)	0.07	0.294	4	4	4	57.2	1
Quadratic ($k = 12$)	6.36×10^{-3}	0.0268	48	0.818	47.2	7.55×10^{3}	57.7
Forward KL ($k = 12$)	3.75×10^{-3}	0.0179	70	0.699	57.7	1.87×10^{4}	82.5
Itakura-Saito ($k = 12$)	1.41×10^{-3}	6.72×10^{-3}	65.3	0.333	49.6	4.62×10^{4}	149

Table 1: RSC constant, the lower eigenvalue bound at the optimum, sharper global smoothness, and exact eigenvalues at the optimum. All target weights are positive and nonuniform. Exact spectra are diagnostic references, not inputs to Algorithm 2.

5.2 Monte Carlo guidance for practical step sizes

The gap between global bounds and local eigenvalues motivates Algorithm 2. For each target pair, count samples whose top-two scores belong to that pair and differ by less than δ. Dividing by $2N\delta$ estimates its interface weight; the resulting Laplacian estimates both extreme eigenvalues (Appendix B.2).

At selected iterations, use independent samples and two bandwidths $r \in \{\delta/2, \delta\}$. Set $\widetilde{\beta}$ to twice the larger upper eigenvalue estimate. If both graphs are connected and their spectral solves succeed, use half the smaller gap estimate for $\widetilde{a}$; otherwise retain the theoretical a. This can provide larger practical steps while retaining the same clip-and-center update and averaging rule. A full-fallback policy rejects an unusable upper estimate and may also reject an unresolved gap.

Algorithm 2 MC-guided SGD with a fallback to Algorithm 1

Require: Algorithm 1 inputs; refresh times $\mathcal{J}$; MC budget N; fallback policy
$s \leftarrow 4\beta/a$; $S \leftarrow 0$; $\bar{h} \leftarrow h_0$; $(\widetilde{a}, \widetilde{\beta}) \leftarrow (a, \beta)$
for $t = 0, \dots, T-1$ **do**
  if $t \in \mathcal{J}$ **then**
    Estimate $(\widehat{\alpha}_r, \widehat{\beta}_r)$ at h_t for both bandwidths
    if the fallback policy rejects the estimates **then**
      return Algorithm 1 restarted at h_t for $T - t$ steps
    end if
    $\widetilde{\beta} \leftarrow 2 \max_r \widehat{\beta}_r$
    $\widetilde{a} \leftarrow \frac{1}{2} \min_r \widehat{\alpha}_r$ if both gaps are usable, else a; $\widetilde{a} \leftarrow \min\{\widetilde{a}, \widetilde{\beta}\}$
  end if
  Use Algorithm 1's average and update, with $w_t = t + s - 1$ and $\eta_t = 2/(\widetilde{a}t + 4\widetilde{\beta})$
end for
return Weighted average

A full fallback recomputes the valid region constant and resets both the step counter and average. Theorem 3.3 then guarantees $O(1/N_{\mathrm{tail}})$ error for the remaining N_{tail} steps (Corollary A.4). MC-guided steps are assessed empirically: finite-sample spectra and safety factors alone do not certify an entire optimization path. Replacing only $\widetilde{a}$ by a is a partial fallback, not a recovery of that guarantee.

5.3 Accuracy and cost of MC-guided SGD

We compare Algorithms 1–2 with PSGD at scales $0.5, 5, \sqrt{Bk}$ multiplying $(t+1)^{-3/4}$. Quadratic tests use uniform sources on $[0,1]^d$, nonuniform target mass ratio four, $B = 64$, and 8192 updates. Three paired seeds share initialization, training samples, and averaging weights within each problem. Figure 2 holds $k = 1000$ fixed across three dimensions; Appendix B.2 specifies the sampling and timing protocols and gives the corresponding time comparison.

At $d = 10, 50, 128$, MC-guided final L^1 mass errors are $0.0504, 0.0497, 0.0490$, versus $0.443, 0.752, 0.864$ for Algorithm 1 and $0.156, 0.0756, 0.0656$ for default PSGD. Most large-target spectral graphs are disconnected; the improvement comes mainly from using a smaller empirical estimate of the largest eigenvalue. With half-suffix averaging, MC-guided and well-scaled PSGD are closer. At $k = 5000, d = 128$, MC-guided error is 0.114, versus 0.942 and 0.171, but total time

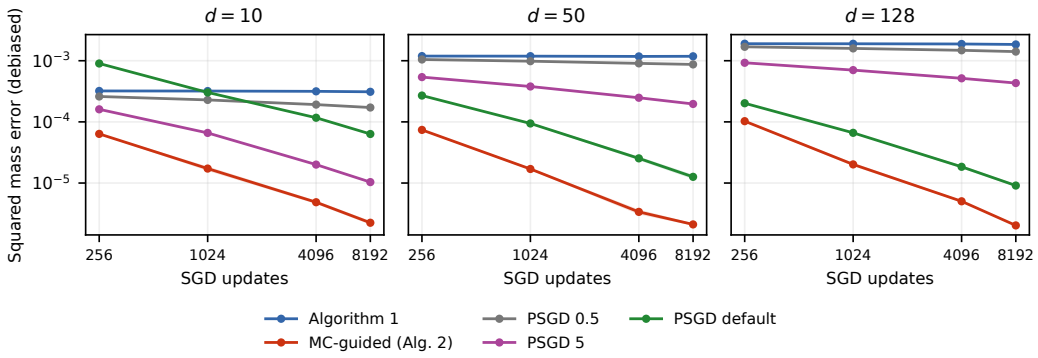


Figure 2: Debiased squared cell-mass error versus SGD updates at $k = 1000$ ($B = 64$). Three paired seeds; bands show one training standard error; common weighted averages. MC calibration samples are additional; total-time curves appear in Appendix B.2.

is 11.8 s versus 6.2 s for default PSGD. MC estimation costs 0.12–0.36 s at $k = 1000$ and 2.3–5.0 s at $k = 5000$; Appendix B.2 specifies timing and additional sampling budgets.

Thus MC guidance often improves finite-budget accuracy, while the explicit theory supplies a guaranteed fallback. It does not uniformly outperform PSGD or Algorithm 1: one-dimensional controls favor the sharper analytic bounds.

6 Conclusion and Limitations

We give explicit smoothness and restricted strong-convexity constants for SGD applied to the semi-discrete OT dual energy. The main estimate follows from a transport variance inequality: the smallest positive Hessian eigenvalue is bounded below linearly in the smallest cell mass. These estimates support an $O(1/T)$ guarantee for decreasing steps and weighted averaging, with explicit initialization and sampling terms.

MC-guided SGD translates the observed gap between global constants and local spectra into practical step sizes. It achieves lower errors in many multidimensional tests, with a full restart option covered by the theory. The assumptions require a compact convex source domain and a positive bounded Hölder continuous density. MC-guided phases are not covered by the global rate; unreliable gap estimates and the cost of calibration at large k motivate cheaper estimators and validated local schedules.

REFERENCES

Dongsheng An, Yang Guo, Na Lei, Zhongxuan Luo, Shing-Tung Yau, and Xianfeng Gu. AE-OT: A new generative model based on extended semi-discrete optimal transport. In *International Conference on Learning Representations*, 2020. URL https://iclr.cc/virtual_2020/poster_HkldyTNYwH.html.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 2017. URL https://proceedings.mlr.press/v70/arjovsky17a.html.

Franz Aurenhammer, Friedrich Hoffmann, and Boris Aronov. Minkowski-type theorems and least-squares clustering. *Algorithmica*, 20(1):61–76, 1998. doi: 10.1007/PL00009187. URL https://link.springer.com/article/10.1007/PL00009187.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013. URL https://papers.nips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

Alex Delalande and Quentin Mérigot. Quantitative stability of optimal transport maps under variations of the target measure. *Duke Mathematical Journal*, 172(17):3321–3357, 2023. doi: 10.1215/00127094-2022-0106. URL https://doi.org/10.1215/00127094-2022-0106.

Wilfrid Gangbo and Robert J. McCann. The geometry of optimal transportation. *Acta Mathematica*, 177:113–161, 1996. URL https://www.math.toronto.edu/mccann/assignments/477/GangboMcCann96.pdf.

Ferdinand Genans, Antoine Godichon-Baggioni, François-Xavier Vialard, and Olivier Wintenberger. Decreasing entropic regularization averaged gradient for semi-discrete optimal transport. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. doi: 10.52202/085713-4915. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d7efa12e98f5e0dd8b4f48cd60b4e3aa-Abstract-Conference.html.

Ferdinand Genans, Antoine Godichon-Baggioni, François-Xavier Vialard, and Olivier Wintenberger. Stochastic optimization in semi-discrete optimal transport: Convergence analysis and minimax rate. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. doi: 10.52202/085713-1716. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/49f7b16c2799e4f12feb1d037a39fd37-Abstract-Conference.html.

Aude Genevay. *Entropy-Regularized Optimal Transport for Machine Learning*. PhD thesis, PSL Research University, 2019. URL https://audeg.github.io/publications/these_aude.pdf.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/2a27b8144ac02f67687f76782a3b5d8f-Abstract.html.

Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of Sinkhorn divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1574–1583. PMLR, 2019. URL https://proceedings.mlr.press/v89/genevay19a.html.

Xianfeng Gu, Feng Luo, Jian Sun, and Shing-Tung Yau. Variational principles for Minkowski type problems, discrete optimal transport, and discrete Monge–Ampère equations. *Asian Journal of Mathematics*, 20(2):383–398, 2016. doi: 10.4310/AJM.2016.v20.n2.a7. URL https://intlpress.com/site/pub/files/_fulltext/journals/ajm/2016/0020/0002/AJM-2016-0020-0002-a007.pdf.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1704.00028.

Jun Kitagawa, Quentin Mérigot, and Boris Thibert. Convergence of a Newton algorithm for semi-discrete optimal transport. *Journal of the European Mathematical Society*, 21(9):2603–2651, 2019. doi: 10.4171/JEMS/889. URL https://ems.press/journals/jems/articles/16184.

Nikita Kornilov, Petr Mokrov, Alexander Gasnikov, and Alexander Korotin. Optimal flow matching: Learning straight trajectories in just one step. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bc8f76d9caadd48f77025b1c889d2e2d-Abstract-Conference.html.

Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method. arXiv:1212.2002, 2012. URL https://arxiv.org/abs/1212.2002.

Cyril Letrouit and Quentin Mérigot. Gluing methods for quantitative stability of optimal transport maps. arXiv:2411.04908v3, 2026. URL https://arxiv.org/abs/2411.04908v3. Version 3, 16 July 2026; to appear in Annales scientifiques de lÉcole normale supérieure.

Bruno Lévy. A numerical algorithm for L_2 semi-discrete optimal transport in 3D. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1693–1715, 2015. doi: 10.1051/m2an/2015055. URL https://www.numdam.org/item/M2AN_2015__49_6_1693_0/.

Quentin Mérigot. A multiscale approach to optimal transport. *Computer Graphics Forum*, 30(5):1583–1592, 2011. doi: 10.1111/j.1467-8659.2011.02032.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8659.2011.02032.x.

Alireza Mousavi-Hosseini, Stephen Y. Zhang, Michal Klein, and Marco Cuturi. Flow matching with semidiscrete couplings. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/9a263e235f6d1521d13a8531c7974951-Abstract-Conference.html.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009. doi: 10.1137/070704277. URL https://doi.org/10.1137/070704277.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. doi: 10.1561/2200000073. URL https://www.nowpublishers.com/article/Details/MAL-073.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. doi: 10.1137/0330046. URL https://epubs.siam.org/doi/10.1137/0330046.

Aram-Alexandre Pooladian, Vincent Divol, and Jonathan Niles-Weed. Minimax estimation of discontinuous optimal transport maps: The semi-discrete case. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28128–28150. PMLR, 2023. URL https://proceedings.mlr.press/v202/pooladian23b.html.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 1571–1578. Omnipress, 2012. URL https://icml.cc/2012/papers/261.pdf.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 71–79. PMLR, 2013. URL https://proceedings.mlr.press/v28/shamir13.html.

Sebastian U. Stich. Unified optimal analysis of the (Stochastic) gradient method. arXiv:1907.04232, 2019. URL https://arxiv.org/abs/1907.04232.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=CD9Snc73AW.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009. ISBN 978-3-540-71049-3. doi: 10.1007/978-3-540-71050-9. URL https://link.springer.com/book/10.1007/978-3-540-71050-9.

A COMPLETE PROOFS

A.1 DUAL FORMULATION, GRADIENT, AND SMOOTHNESS

Dual formulation. Kantorovich duality expresses this problem using one scalar potential g_i per target, giving the finite-dimensional *semi-dual*

$$\max_{g\in\mathbb{R}^k}\left\{\sum_i \nu_i g_i + \int_\Omega \min_i\big(c(x,y_i)-g_i\big)\,d\mu(x)\right\}.$$

Writing $g_i = h_i + \frac{1}{2}\|y_i\|^2$ turns this maximization into minimization of the convex energy E in (2), up to a constant independent of h.

Gradient and Hessian. For distinct sites the equality of two affine scores defines a hyperplane, which has zero μ measure. At the unique winning index $I_h(x)$, $\partial_{h_i}\phi_h(x) = \mathbf{1}_{\{I_h(x)=i\}}$. The corresponding difference quotients are bounded by one, so dominated convergence gives $\nabla E(h) = p(h) - \nu$ for every h. The same argument gives continuity of p. Replacing ν by another fixed mass vector only changes a linear term, hence leaves the Hessian unchanged wherever it exists.

At positive-cell configurations the Hessian $L(h) = D_h p(h)$ is a weighted graph Laplacian (Gu et al., 2016; Kitagawa et al., 2019):

$$L_{ij} = -w_{ij}\quad (i\neq j),\qquad L_{ii} = \sum_{j\neq i} w_{ij},\qquad w_{ij} = \int_{W_i(h)\cap W_j(h)} \frac{f(x)}{\|y_i - y_j\|}\,d\mathcal{H}^{d-1}(x).$$

Here $\mathcal{H}^{d-1}$ is interface area (counting measure for $d=1$). The weight measures the mass transferred across an interface per unit change in the relative potentials; interfaces of zero area contribute zero.

Lemma A.1 (From componentwise bounds to Euclidean smoothness)**.** *Let $F\in C^1(\mathbb{R}^k)$ and $K\geq 0$. Suppose that, for every i and all $h,g\in\mathbb{R}^k$,*

$$|\partial_i F(h) - \partial_i F(g)| \leq K\|h-g\|_\infty. \tag{13}$$

Then $\|\nabla F(h) - \nabla F(g)\|_2 \leq K\|h-g\|_2$ for all h,g.

Proof. Choose a nonnegative smooth compactly supported kernel ζ with integral one, and set $\zeta_\epsilon(u) = \epsilon^{-k}\zeta(u/\epsilon)$, $F_\epsilon = F * \zeta_\epsilon$, and $q_\epsilon = \nabla F_\epsilon$. Differentiation of the convolution gives $q_\epsilon = (\nabla F) * \zeta_\epsilon$. Averaging (13) against this kernel shows that $|(q_\epsilon)_i(h) - (q_\epsilon)_i(g)| \leq K\|h-g\|_\infty$. For the smooth symmetric Hessian $H_\epsilon(h) = D^2F_\epsilon(h)$, differentiating in any direction v therefore gives $|(H_\epsilon(h)v)_i| \leq K\|v\|_\infty$. Taking the supremum over $\|v\|_\infty\leq 1$ yields

$$\sum_j |(H_\epsilon(h))_{ij}| \leq K \quad \text{for every } i.$$

Write $\|H\|_{\mathrm{op}} = \sup_{\|v\|_2=1}\|Hv\|_2$ for the Euclidean operator norm. For a symmetric matrix this is the largest absolute eigenvalue. Gershgorin's circle theorem bounds each absolute eigenvalue by the largest absolute row sum, so

$$\|H_\epsilon(h)\|_{\mathrm{op}} = \max_j |\lambda_j(H_\epsilon(h))| \leq \max_i \sum_j |(H_\epsilon(h))_{ij}| \leq K.$$

Integrating H_ϵ along the segment from g to h gives $\|q_\epsilon(h) - q_\epsilon(g)\|_2 \leq K\|h-g\|_2$. Since ∇F is continuous, $q_\epsilon \to \nabla F$ locally uniformly. Passing to the limit proves the claim for every pair h,g. □

Proof of Proposition 4.1. The largest hyperplane-section area is

$$A_\Omega = \sup_{\|n\|=1,\, b\in\mathbb{R}} \mathcal{H}^{d-1}(\Omega\cap\{x : n^\top x = b\}),$$

with $A_\Omega = 1$ for $d=1$. Let K denote the value of β in (8). We first prove the asserted cell-mass stability bound, then the smoothness and potential bounds.

Let $\delta = \|h-g\|_\infty$. If a point leaves cell i, one of its score inequalities changes sign after a threshold displacement of at most 2δ. For each competitor j the lost points lie in a one-sided slab of width at most $2\delta/\|y_i - y_j\|$. The section-area bound and a union bound control the lost mass. Interchanging h, g gives the same bound on the gained mass. The absolute difference of the masses is at most the larger of the lost and gained masses, giving

$$|p_i(h) - p_i(g)| \le 2f_{\max} A_\Omega \delta \sum_{j\neq i} \|y_i - y_j\|^{-1} \le K\delta.$$

For a general vector-valued map this componentwise bound only gives a Euclidean Lipschitz constant $\sqrt{k}K$. Here $\nabla E = p - \nu$ and $E \in C^1(\mathbb{R}^k)$ by the preceding differentiation argument. Lemma A.1, applied to $F = E$, therefore gives the sharper constant K in (8). The slab bound and the lemma hold for all potentials, including those with empty cells or changes in cell adjacency. No extension of the positive-cell Hessian formula is required.

If every closed cell is nonempty, selecting a point in $W_j(h)$ gives $h_i - h_j \le R_X\|y_i - y_j\| \le R_0$. Interchanging i, j and centering yields $\|h\|_\infty \le R_0$. In particular this holds at h^*. Existence of a balanced potential for positive target masses follows from the semi-discrete dual theory (Gu et al., 2016; Kitagawa et al., 2019). The positive Hessian lower bound proved next, together with convexity, ensures uniqueness on H_0. □

The following proposition provides the sharper analytic smoothness bounds used to choose β in Table 1 and in Algorithm 1 for the uniform-cube experiments.

Proposition A.2 (Global smoothness on uniform boxes)**.** *Let the source be uniform on an axis-aligned cube of side length ℓ, with distinct score slopes z_i. A global smoothness constant is*

$$\beta_{\mathrm{box}} = \frac{2}{\ell} \max_i \sum_{j\neq i} \frac{1}{\|z_i - z_j\|_\infty}. \tag{14}$$

In one dimension, one can also take

$$\beta_{\mathrm{path}} = \frac{4}{\ell \min_{i\neq j} |z_i - z_j|}. \tag{15}$$

The minimum of these bounds and (8) *remains valid globally. For a forward Bregman cost, $z_i = \nabla\phi(y_i)$.*

Proof. Write $\delta = \|h-g\|_\infty$. For each pair i, j, choose a coordinate a with $|(z_i - z_j)_a| = \|z_i - z_j\|_\infty$. A score threshold changes by at most 2δ. After fixing all other coordinates, the corresponding one-sided slab has length at most $2\delta/\|z_i - z_j\|_\infty$ in coordinate a. Fubini's theorem and the density ℓ^{-d} therefore bound its mass by $2\delta/(\ell\|z_i - z_j\|_\infty)$. The same lost/gained-mass argument as above gives

$$|p_i(h) - p_i(g)| \le \frac{2\delta}{\ell} \sum_{j\neq i} \|z_i - z_j\|_\infty^{-1}.$$

Lemma A.1 proves (14).

For (15), let the source interval be $[l, u]$, with $u - l = \ell$, and order the distinct slopes as $z_1 < \cdots < z_k$. Set $m = \min_{i\neq j} |z_i - z_j| > 0$. The closed cell of i is $[L_i(h), R_i(h)]$ when $L_i(h) \le R_i(h)$ and is empty otherwise, where

$$L_i(h) = \max\left(\{l\} \cup \left\{\frac{h_j - h_i}{z_i - z_j} : j < i\right\}\right), \qquad R_i(h) = \min\left(\{u\} \cup \left\{\frac{h_i - h_j}{z_j - z_i} : j > i\right\}\right).$$

Both endpoints are $2/m$-Lipschitz in the infinity norm: each affine threshold has this bound, and finite maxima and minima preserve it. Since $p_i(h) = [R_i(h) - L_i(h)]_+/\ell$, where $[r]_+ = \max\{r, 0\}$,

$$|p_i(h) - p_i(g)| \le \frac{4}{\ell m}\|h - g\|_\infty.$$

Another application of Lemma A.1 proves the global bound, including empty and zero-length cells. □

A.2 Proof of Proposition 4.2

Specializing the variance inequality. Theorem 2.1 of Letrouit & Mérigot (2026) uses a density ρ satisfying $m_\rho\sigma \le \rho \le M_\rho\sigma$, where σ is a log-concave probability density. To apply it, set $\rho = f$ and $\sigma = \mathbf{1}_\Omega/|\Omega|$, where $|\Omega|$ denotes the d-dimensional Lebesgue volume of Ω (area for $d = 2$). Convexity of Ω makes σ log-concave. The choices $m_\rho = f_{\min}|\Omega|$ and $M_\rho = f_{\max}|\Omega|$ satisfy the comparison assumptions, with $M_\rho/m_\rho = f_{\max}/f_{\min}$. Translation of all sites by $-y_0$ adds an h-independent linear function of x to the envelope and does not change its cells or envelope differences. The target radius in the cited theorem is then R_Y. On the finite target set, put $\psi_0(y_i - y_0) = -h_i$ and $\psi_1(y_i - y_0) = -g_i$; both functions are continuous on that set. Their convex conjugates are $\phi_h(x) - x^\top y_0$ and $\phi_g(x) - x^\top y_0$, and their gradient pushforwards have masses $p(h)$ and $p(g)$. Substitution into its inequality (2.2) gives (11) with $A = e(f_{\max}/f_{\min})D_X R_Y$. No optimality or positive-cell condition is needed for this step.

Differentiating the inequality. Set $g = h + tv$, with $t > 0$, in (11) and divide by t^2 to obtain

$$\mathrm{Var}_\mu\left(\frac{\phi_{h+tv} - \phi_h}{t}\right) \le A\langle v, \frac{p(h+tv) - p(h)}{t}\rangle.$$

At a positive-cell configuration, Kitagawa et al. (2019, Theorem 1.3) makes p differentiable under Assumption 2.1. The envelope difference quotient converges almost everywhere to $v_{I_h(x)}$ and is bounded by $\|v\|_\infty$. Dominated convergence therefore gives $\mathrm{Var}_{p(h)}(v) \le Av^\top L(h)v$. For $v \in H_0$,

$$\mathrm{Var}_p(v) = \min_c \sum_i p_i(v_i - c)^2 \ge p_{\min} \min_c \sum_i (v_i - c)^2 = p_{\min}\|v\|^2.$$

This proves both the matrix estimate and the Euclidean spectral bound.

From eigenvalue bounds to strong convexity. On a convex region $\mathcal{C} \subset H_0$ with $p_i(h) \ge \varepsilon > 0$ for all $h \in \mathcal{C}$, Proposition 4.2 implies

$$E(g) \ge E(h) + \langle \nabla E(h), g - h\rangle + \frac{\varepsilon}{2A}\|g - h\|^2, \qquad h, g \in \mathcal{C}. \tag{16}$$

Indeed, setting $v = g - h$ and integrating along the segment in $\mathcal{C}$ gives

$$E(g) - E(h) - \langle \nabla E(h), v\rangle = \int_0^1 (1-t)v^\top L(h+tv)v\, dt \ge \frac{\varepsilon}{2A}\|v\|^2.$$

The next proof handles bounded regions without a positive current-mass floor.

A.3 Proof of Theorem 4.5

Let $\mathcal{A} = \{h \in H_0 : W_i(h) \ne \varnothing \text{ for every } i\}$, where the cells are closed and may have zero mass. This set is convex: cell witnesses and potential coordinates satisfy jointly linear score inequalities, and their domain is convex. It is closed by compactness of Ω, and bounded by the preceding R_0 estimate.

For $z \in \mathcal{A}$, choose $x_i \in W_i(z)$. Every score inequality shows that

$$(1-s)W_i(h^*) + sx_i \subseteq W_i((1-s)h^* + sz), \qquad 0 \le s < 1.$$

Volume scaling and density bounds imply $p_i((1-s)h^* + sz) \ge (f_{\min}/f_{\max})(1-s)^d\nu_i$. Let $u = z - h^*$. The segment has positive cell masses before its endpoint, so Proposition 4.2 and integration give

$$\langle \nabla E(z), u\rangle = \int_0^1 u^\top L(h^* + su)u\, ds \ge \frac{f_{\min}\nu_{\min}}{f_{\max}A(d+1)}\|u\|^2.$$

The endpoint follows by continuity; global smoothness also justifies absolute continuity of the gradient along the segment.

For arbitrary $h \in H_0$ with $\mathrm{osc}(h) \le D$, define

$$\bar{h}_i = \min_{x\in\Omega}\{\phi_h(x) - x^\top y_i\}, \qquad q_i = \bar{h}_i - h_i \ge 0, \qquad z = \Pi_{H_0}\bar{h}.$$

All minima are attained. Each raised affine piece lies below the old envelope and each old affine piece lies below its raised version, so $\phi_{\tilde{h}} = \phi_h$. Every raised piece contacts the envelope, hence $z \in \mathcal{A}$. Distinct slopes and absolute continuity imply $p(z) = p(h)$. If $q_i > 0$, the original affine piece is strictly below the envelope everywhere; therefore $p_i(h) = 0$ and $p(h)^\top q = 0$.

Writing $h - h^* = u - v$ with $u = z - h^*$ and $v = \Pi_{H_0} q$, and using the zero sum of the gradient, gives

$$\langle \nabla E(h), h - h^* \rangle = \langle \nabla E(z), u \rangle + \nu^\top q \geq a_0 \|u\|^2 + \nu^\top q, \qquad a_0 = \frac{f_{\min} \nu_{\min}}{f_{\max} A(d+1)}.$$

For every $x \in \Omega$, $\phi_h(x) - x^\top y_i - h_i = \max_j \{x^\top (y_j - y_i) + h_j - h_i\} \leq R_X D_Y + D$. Thus $0 \leq q_i \leq R_0 + D =: M$, and $\nu^\top q \geq (\nu_{\min}/M)\|q\|^2 \geq b_0 \|v\|^2$, where $b_0 = \nu_{\min}/M$. The weighted Cauchy–Schwarz inequality gives

$$\|u - v\|^2 \leq (a_0^{-1} + b_0^{-1})(a_0 \|u\|^2 + b_0 \|v\|^2).$$

The reciprocal of $a_0^{-1} + b_0^{-1}$ equals a_D, proving the theorem.

Corollary A.3 (Residual-based error bounds)**.** *On the region of Theorem 4.5,*

$$\|h - h^*\| \leq \frac{\|p(h) - \nu\|}{a_D}, \qquad E(h) - E(h^*) \leq \frac{\beta}{2a_D^2} \|p(h) - \nu\|^2.$$

Proof. Cauchy–Schwarz and (12) give the distance bound; $E(h) - E(h^*) \leq \beta \|h - h^*\|^2/2$ gives the objective bound. □

These bounds concern true masses. With an independent batch of size N, Hoeffding's inequality gives $\|\widehat{p} - p\|_\infty \leq \sqrt{\log(2k/\delta)/(2N)}$ with probability at least $1 - \delta$. Add the corresponding ℓ_2 error before applying the corollary; repeated checks require a confidence budget across checks.

A.4 Proof of Theorem 3.2

Let $\mathcal{G}_t = \langle \nabla E(h_t), h_t - h^* \rangle$ and $r_t = \mathbb{E}\|h_t - h^*\|^2$. Convex smoothness gives $\|\nabla E(h_t)\|^2 \leq \beta \mathcal{G}_t$, and convexity gives $\mathcal{G}_t \geq E(h_t) - E(h^*)$. Lemma 3.1 and (6) therefore imply, for $0 < \eta \leq 1/\beta$,

$$r_{t+1} \leq r_t - (2\eta - \beta\eta^2)\mathbb{E}\mathcal{G}_t + \eta^2/B \leq r_t - \eta \mathbb{E}[E(h_t) - E(h^*)] + \eta^2/B.$$

Summing, using $r_T \geq 0$, and applying convexity to the uniform average gives $U_0/(\eta T) + \eta/B$. For the stated choice of η, this is at most $\beta U_0/T + 2\sqrt{U_0/(BT)}$.

A.5 Proof of the weighted objective bound

The map $\mathcal{T}$ fixes h^* and is nonexpansive relative to h^*, by the same two projection arguments as in (6). The composition need not be the Euclidean projection onto $[-C, C]^k \cap H_0$; the two properties in (6) suffice. All queries lie in $\mathcal{R}_C = \{h \in H_0 : \mathrm{osc}(h) \leq 2C\}$, a convex set on which the RSC bound holds, including at empty cells. The weighted output also belongs to $\mathcal{R}_C$.

Conditional descent. Set $d_t = h_t - h^*$, $\mathcal{G}_t = \langle \nabla E(h_t), d_t \rangle$, and $\delta_t = E(h_t) - E(h^*)$. For the independent batch,

$$\mathbb{E}[\|G_t\|^2 \mid \mathcal{F}_t] = \|\nabla E(h_t)\|^2 + \mathbb{E}[\|G_t - \nabla E(h_t)\|^2 \mid \mathcal{F}_t] \leq \|\nabla E(h_t)\|^2 + 1/B.$$

Consequently, distance expansion before applying $\mathcal{T}$ yields

$$\mathbb{E}[\|d_{t+1}\|^2 \mid \mathcal{F}_t] \leq \|d_t\|^2 - 2\eta_t \mathcal{G}_t + \eta_t^2 \|\nabla E(h_t)\|^2 + \eta_t^2/B.$$

Since E is convex and globally β-smooth with $\nabla E(h^*) = 0$, co-coercivity gives $\|\nabla E(h_t)\|^2 \leq \beta \mathcal{G}_t$. For $\eta_t \leq 1/(2\beta)$,

$$2\mathcal{G}_t - \eta_t \|\nabla E(h_t)\|^2 \geq \tfrac{3}{2}\mathcal{G}_t \geq a\|d_t\|^2 + \tfrac{1}{2}\delta_t.$$

The last inequality combines $\mathcal{G}_t \geq a\|d_t\|^2$ and $\mathcal{G}_t \geq \delta_t$; no two-point strong-convexity assumption is used. Taking expectations gives

$$r_{t+1} \leq (1 - a\eta_t) r_t - \tfrac{1}{2}\eta_t \Delta_t + \eta_t^2/B, \qquad r_t = \mathbb{E}\|h_t - h^*\|^2, \quad \Delta_t = \mathbb{E}[E(h_t) - E(h^*)]. \tag{17}$$

Connection to the classical stochastic recursion. Equation (17) is the recursion of Stich (2019, Section 4, equation (8)), with their coefficients $(a, b, c, d) = (a, 1/2, 1/B, 2\beta)$ and their objective sequence set to Δ_t. Their Assumption 3 uses the stronger one-point condition

$$\langle \nabla E(h), h - h^* \rangle \geq E(h) - E(h^*) + \frac{\mu}{2}\|h - h^*\|^2.$$

Our RSC inequality does not by itself give this condition with $\mu = a$. The preceding derivation supplies the recursion directly from convexity, RSC, smoothness, and a distance-nonincreasing update map. The telescoping below follows the decreasing-step argument of their Lemma 7, using $w_t = t + s - 1$ in place of $t + s$ to cancel consecutive distance terms exactly. We retain the initialization term and give the constants for this choice of weights.

Exact telescoping and initialization. For $s \geq 4\beta/a$, $\eta_t = 2/[a(t+s)] \leq 1/(2\beta)$. Let $u = t + s$. Rearranging (17) gives

$$\Delta_t \leq a(u-2)r_t - aur_{t+1} + \frac{4}{aBu}.$$

Multiplication by $w_t = u - 1$ and summation over $t = 0, \ldots, T-1$ give

$$\begin{aligned}\sum_{t=0}^{T-1} w_t \Delta_t + a(T+s-2)(T+s-1)r_T \\ &\leq a(s-1)(s-2)r_0 + \frac{4}{aB}\sum_{t=0}^{T-1}\frac{t+s-1}{t+s} \\ &\leq a(s-1)(s-2)U_0 + \frac{4T}{aB}.\end{aligned}$$

Here $s \geq 4$, so all displayed coefficients are nonnegative. Drop the nonnegative terminal term on the left, divide by $S_T = \sum_t w_t = T(T+2s-3)/2$, and apply convexity to the weighted potential. This proves (7) for every $T \geq 1$, including $T = 1$, where the returned point is h_0. No independence between iterates is required. The batch variance bound, not a bound of $1/B$ on the entire gradient second moment, supplies the noise term. The weights refer to query points before each update; shifting them to post-update points requires a corresponding reindexing.

A.6 A GUARANTEED RESTART AFTER MC-GUIDED STEPS

Corollary A.4 (Full fallback to the proved schedule)**.** *Under Assumption 2.1, let a preliminary algorithm start in $\mathcal{R}_C$, with $C \geq R_0$, and apply the clip-and-center map $\mathcal{T} = \Pi_{H_0} P_{[-C,C]^k}$ after each update. Its step sizes may depend on independent MC spectral estimates. At any finite stopping time τ, condition on its history $\mathcal{F}_\tau$ and define*

$$C_\tau = \max\{C, \|h_\tau\|_\infty\}, \qquad a_\tau = \min\{a_{2C_\tau}, \beta\}, \qquad s_\tau = 4\beta/a_\tau, \qquad U_\tau = k(C_\tau + R_0)^2.$$

Restart Algorithm 1 at h_τ, with these constants, fresh source batches, and a new average that excludes the preliminary iterates. For every $N_{\mathrm{tail}} \geq 1$, its output satisfies

$$\mathbb{E}[E(\bar{h}^{\mathrm{tail}}_{N_{\mathrm{tail}}}) - E(h^*) \mid \mathcal{F}_\tau] \leq \frac{2a_\tau(s_\tau - 1)(s_\tau - 2)U_\tau}{N_{\mathrm{tail}}(N_{\mathrm{tail}} + 2s_\tau - 3)} + \frac{8}{a_\tau B(N_{\mathrm{tail}} + 2s_\tau - 3)}.$$

Proof. Every preliminary iterate belongs to H_0 and has oscillation at most $2C$. Hence $\|h_\tau\|_\infty \leq 2C$ and $C \leq C_\tau \leq 2C$. The potential bound gives $\|h^*\|_\infty \leq R_0 \leq C_\tau$, so $\|h_\tau - h^*\|^2 \leq U_\tau$. Conditionally on $\mathcal{F}_\tau$, all restart parameters are fixed. Proposition 4.1 and Theorem 4.5 verify (A1)–(A2) for the new clip-and-center region. Fresh batches retain the unbiased gradient and variance bound. Theorem 3.3, applied conditionally to the restarted sequence, yields the displayed estimate. □

The guarantee counts only the restarted steps; it does not assert an $O(1/T)$ rate for an arbitrary MC prefix or for an average mixing both phases. The radius adjustment is necessary because centering preserves the clipped coordinate range but may place a coordinate outside $[-C, C]$. Only the region-dependent a and averaging shift need updating; the global β remains valid. Partial fallback of the gap estimate alone retains an empirical upper estimate and does not invoke this corollary.

A.7 Literature estimates and the scope of strong convexity

Existence and uniqueness versus quantitative growth. A unique minimizer does not imply strong convexity, even for a smooth strictly convex function: $F(t) = t^4$ has a unique minimizer at zero but $F(t)/t^2 \to 0$ as $t \to 0$. For our energy, translation invariance excludes strong convexity on $\mathbb{R}^k$. On H_0, global strong convexity with constant $m > 0$ would imply

$$m\|h - h^*\|^2 \le \langle \nabla E(h), h - h^* \rangle \le \sqrt{2}\|h - h^*\|,$$

since $p(h)$ and ν are probability vectors. This contradicts the unboundedness of H_0. Even bounded centered regions can fail: for $\mu = \mathrm{Unif}[0,1]$, $y_1 = 0, y_2 = 1$ and $\nu = (q, 1-q)$ with $0 < q < 1$, direct integration gives

$$E(t, -t) = \begin{cases} 1/2 - 2qt, & t \le 0, \\ 1/2 - 2qt + 2t^2, & 0 \le t \le 1/2, \\ 2(1-q)t, & t \ge 1/2. \end{cases}$$

The unique centered minimizer is $t^* = q/2$, but the empty-cell tails are affine. Neither positive target weights nor centering prevents this.

The quantitative KMT estimate. Write Φ for KMT's concave functional and $E_Y = \mathbf{1}^\perp$. Kitagawa et al. (2019, Theorem 5.1) prove

$$-v^\top D^2\Phi(\psi)v \ge C_{\mathrm{KMT}}\varepsilon^3\|v\|^2, \qquad v \in E_Y, \quad \psi \in \mathcal{K}^\varepsilon,$$

where $\mathcal{K}^\varepsilon$ requires every current cell mass to exceed ε. Equation (5.46) in arXiv:1603.05579v2 specifies the explicit constant, in their notation,

$$C_{\mathrm{KMT}} = \frac{\varepsilon_{\mathrm{tw}}^2}{2C_{\exp}^{4(d-1)}C_\nabla^3[\mathcal{H}_g^{d-1}(\partial X)]^2\|f\|_\infty^2 C_{\mathrm{PW}}^3}.$$

Here f denotes the source density (written ρ in KMT). Their Section 5 derives the bound through the weighted graph Laplacian, Cheeger inequality, and source Poincaré–Wirtinger inequality. In the quadratic, convex-source setting, we obtain linear dependence on the mass floor from the transport variance inequality. KMT's estimate also covers costs satisfying their more general geometric assumptions.

Variance versus Euclidean curvature. Delalande & Mérigot (2023, Theorem 2.1) control differences of conjugate potentials in a target-weighted variance, with constants depending on dimension, density bounds, and potential ranges. The source-space inequality of Letrouit & Mérigot (2026, Theorem 2.1), stated in (11), yields $\mathrm{Var}_{p(h)}(v)$ after differentiation. Appendix A.2 converts this variance to Euclidean curvature using a current-mass floor. At h^*, the identity $p(h^*) = \nu$ supplies that floor; elsewhere, the regional argument in Appendix A.3 handles empty cells.

B Experimental protocols and supporting results

This appendix specifies the reference calculations behind Table 1, the MC spectral estimator used by Algorithm 2, and the matched comparisons supporting Section 5.

B.1 Reference solutions and implementation controls

Data and analytic constants. On $[l, u]$, sites are $y_i = l + (u-l)(i - \frac{1}{2})/k$; Table 1 also includes $k = 2$ on $[0, 1]$. In higher dimensions, sites are independent uniform draws from the source cube. Weights are a random permutation of k equally spaced values from one to four, normalized to sum to one (geometry seed 1729). For the score slopes z_i, we center the target geometry at $z_0 = k^{-1}\sum_i z_i$ and use $2R_Y$ as an upper bound on its diameter when computing R_0. We set $C = \max\{R_0, \|h_0\|_\infty\}$, $a = \min\{a_{2C}, \beta\}$, and choose the smallest applicable smoothness bound in Proposition A.2; the general section-area bound uses the containing ball of radius $\sqrt{d}(u-l)/2$.

Exact one-dimensional reference. For a uniform source on $[l, u]$, sorted sites, and positive masses, let $x_i = l + (u - l)\sum_{j \le i} \nu_j$. For increasing affine slopes z_i, optimal heights satisfy $h^*_{i+1} = h^*_i + x_i(z_i - z_{i+1})$, followed by centering. For arbitrary heights, intersecting the interval with all pairwise score inequalities gives exact cell masses and exact integrals of the envelope. This reference covers quadratic and forward Bregman costs. At the optimum the Hessian is a path Laplacian with $w_i = [(u - l)(z_{i+1} - z_i)]^{-1}$, evaluated in FP64.

KL and IS score coordinates. For a forward Bregman cost, $c(x, y_i) = \phi(x) - x^\top z_i + b_i$, where $z_i = \nabla\phi(y_i)$ and $b_i = y_i^\top z_i - \phi(y_i)$. Setting $h_i = g_i - b_i$ makes $g_i - c(x, y_i) = x^\top z_i + h_i - \phi(x)$. Thus the dual energy differs from the affine-score energy only by a constant, and its derivatives and geometric bounds use the slopes z_i. The KL and IS tests use $\phi(x) = x \log x$ and $\phi(x) = -\log x$, respectively, on $[0.2, 1]$; centering and clipping act on h.

B.2 MC-GUIDED SGD AND TIMING COMPARISON

Matched comparisons. The quadratic tests use uniform sources on $[0, 1]^d$, target mass ratio four, geometry seed 1729, and three paired training seeds (0,1,2). Each method starts at zero cost potentials and uses the same training samples, $B = 64$, and 8192 updates. PSGD retains its native cost-coordinate box $[-C_g, C_g]^k$, where $C_g = \max_{x\in[0,1]^d, i} \frac{1}{2}\|x - y_i\|^2$, no training-time recentering, and steps $c/(t + 1)^{3/4}$ for $c \in \{0.5, 5, \sqrt{Bk}\}$. All methods average pre-update queries with weights $w_t = t + 4\beta/a - 1$; half-suffix checks average the second half of each run. Beyond the three dimensions in Figure 2, controls use $k = 12$ at $d = 1$, $k = 100$ at $d \in \{1, 2, 10, 50, 128\}$, $k = 1000$ at $d = 2$, and $k = 5000$ at $d \in \{10, 128\}$.

MC calibration. For scores $s_i(x) = x^\top y_i + h_i$, let $C_{ij}(\delta)$ count samples whose unordered top-two pair is $\{i, j\}$ and whose score gap is below δ. Set $\widehat{w}_{ij} = C_{ij}(\delta)/(2N_{\mathrm{MC}}\delta)$ and form the Laplacian with off-diagonals $-\widehat{w}_{ij}$ and zero row sums. At a regular interface, the coarea formula gives $\Pr(\{I_1, I_2\} = \{i, j\}, s_{I_1} - s_{I_2} < \delta)/(2\delta) \to w_{ij}$; the factor two counts both sides, and no additional distance normalization is needed. At $t = 0, 256, 1024, 4096$, an independent 4096-sample pilot sets δ to the 15th percentile of the score gap. A further $N_{\mathrm{MC}} = \max\{32768, 64k\}$ samples estimate both δ and $\delta/2$ spectra. Algorithm 2 uses safety factors two and the absolute iteration index, without resetting it at refreshes. Edge storage and spectral solves are sparse for $k > 128$; score evaluation costs $O(N_{\mathrm{MC}}kd)$. A failed upper-eigenvalue solve may use twice the estimated maximum degree.

All reported upper estimates are finite and positive, so no full restart is triggered. In the Figure 2 cases and both $k = 5000$ cases, all 12 refreshes per setting reject the estimated gap and retain a, while still using the empirical upper estimate. This partial fallback is distinct from the guaranteed restart of Corollary A.4. Finite samples and bandwidth bias prevent interpreting these spectra as proved pathwise bounds.

Evaluation and time. Independent shared validation streams use $N_{\mathrm{val}} = 65536$ samples at intermediate checkpoints and 524288 at the endpoint. For cell counts n_i, the plotted unbiased squared residual is

$$\widehat{\|p - \nu\|_2^2} = \frac{\sum_i n_i(n_i - 1)}{N_{\mathrm{val}}(N_{\mathrm{val}} - 1)} - 2\sum_i \frac{n_i\nu_i}{N_{\mathrm{val}}} + \|\nu\|_2^2.$$

Negative estimates are retained; bands show one training standard error conditional on the shared validation stream. The L^1 values in the text use $\|\widehat{p} - \nu\|_1$. Each trajectory is timed separately in FP64 on an RTX A6000, with TF32 disabled and synchronized timing. Figure 3 includes setup, training, MC sampling, and spectral solves; validation is excluded. The common averaging shift requires global-constant setup for every method, although native PSGD alone would not need it (0.34–0.38 s at $k = 5000$). MC guidance adds $4(4096 + N_{\mathrm{MC}})$ source draws to the 524288 training draws; equal update budgets therefore do not mean equal total sampling cost.

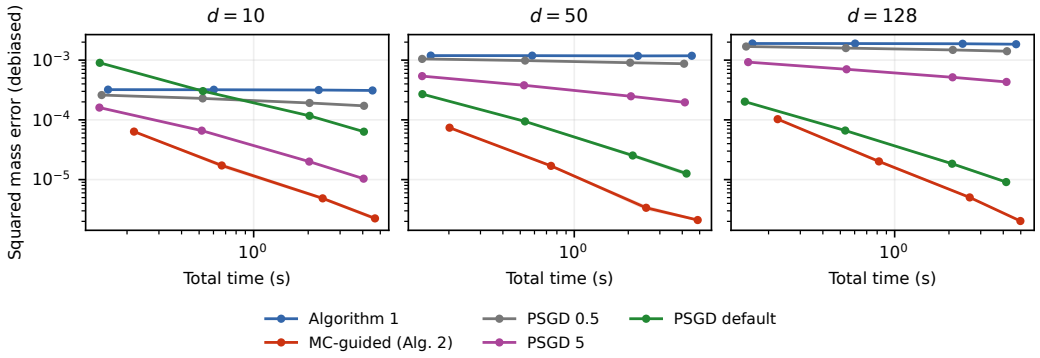


Figure 3: Figure 2 plotted against total time, including MC calibration and excluding validation. Three paired seeds; bands show one training standard error.

C Detailed related work

Hessian eigenvalue bounds. The introduction describes the geometric formulation and solvers. Appendix A.7 gives the quantitative KMT estimate and distinguishes the two transport variance inequalities. KMT's Laplacian/Cheeger argument supports a Newton line search that preserves positive cell masses. Our RSC bound on the bounded potential set also covers empty cells. Genans' local-to-regional argument (Genans et al., 2025b, Lemma 4.2) likewise uses convexity to extend local growth; our bound makes the source, mass, and region dependence explicit.

Stochastic semi-discrete optimization. Entropic regularization softens assignments and enables Sinkhorn scaling on discrete inputs (Cuturi, 2013). Genevay et al. (2016, Section 4) formulate the semi-dual as an expectation and analyze averaged SGD, reporting an $O(T^{-1/2})$ rate in the regularized setting. SD-FM (Mousavi-Hosseini et al., 2026, Theorem 2) covers zero regularization under a surface-area condition: at a random iterate, expected χ^2 marginal error is $O(T^{-1/2})$ and the excess-cost upper bound for the induced, generally marginal-inexact coupling is $O(T^{-1/4})$. These are different metrics from our averaged dual-objective gap. Genans et al. (2025b, Lemma 4.2, Theorem 4.5, and Corollary 4.6) establish RSC on the bounded potential set and $O(1/T)$ averaged squared-potential and objective rates under additional regularity. DRAG (Genans et al., 2025a) decreases entropic regularization while averaging, obtaining $O(1/T)$ potential and unregularized objective rates in its stated equal-mass setting. We quantify geometric constants for nonuniform masses and use them to specify steps and initialization costs.

Statistical sample complexity. Genevay et al. (2019) and Genevay (2019, Chapter 3) analyze the error from replacing distributions by finite samples. Their $O(n^{-1/2})$ Sinkhorn-divergence bound has a constant that deteriorates as regularization vanishes; the dimension-dependent empirical-OT rates discussed there concern this statistical problem. Pooladian et al. (2023) obtain a minimax $O(n^{-1/2})$ expected squared L^2 map error with an entropic estimator in the semi-discrete two-sample setting. Such dimension-independent exponents concern estimation from data. Our optimization bounds concern a fixed population transport problem, with geometry-dependent constants.

Classical SGD and averaging. Our proof uses distance recursions from stochastic convex optimization (Nemirovski et al., 2009), trajectory averaging (Polyak & Juditsky, 1992), and suffix or increasing-weight schemes (Rakhlin et al., 2012; Lacoste-Julien et al., 2012; Shamir & Zhang, 2013). Retaining the objective-decrease term (Stich, 2019) gives the initialization and noise terms in Theorem 3.3.