Preprint - Under Review 1–36, Feb 9 - 2026

Fast and Large-Scale Unbalanced Optimal Transport via its Semi-Dual and Adaptive Gradient Methods

Ferdinand Genans GENANS.FERDINAND@GMAIL.COM
LPSM, Sorbonne Université, Paris, France

Abstract

Unbalanced Optimal Transport (UOT) has emerged as a robust relaxation of standard Optimal Transport, particularly effective for handling outliers and mass variations. However, scalable algorithms for UOT, specifically those based on Gradient Descent (SGD), remain largely underexplored. In this work, we address this gap by analyzing the semi-dual formulation of Entropic UOT and demonstrating its suitability for adaptive gradient methods. While the semi-dual is a standard tool for large-scale balanced OT, its geometry in the unbalanced setting appears ill-conditioned under standard analysis. Specifically, worst-case bounds on the marginal penalties using χ^2 divergence suggest a condition number scaling with n/ε, implying poor scalability. In contrast, we show that the local condition number actually scales as $\mathcal{O}(1/\varepsilon)$, effectively removing the ill-conditioned dependence on n. Exploiting this property, we prove that SGD methods adapt to this local curvature, achieving a convergence rate of $\mathcal{O}(n/\varepsilon T)$ in the stochastic and online regimes, making it suitable for large-scale and semi-discrete applications. Finally, for the full batch discrete setting, we derive a nearly tight upper bound on local smoothness depending solely on the gradient. Using it to adapt step sizes, we propose a modified Adaptive Nesterov Accelerated Gradient (ANAG) method on the semi-dual functional and prove that it achieves a local complexity of $\mathcal{O}(n^2\sqrt{1/\varepsilon}\ln(1/\delta))$.

Keywords: Unbalanced Optimal Transport, Entropic Regularization, Convex Optimization, Gradient Descent

1. Introduction

Optimal Transport (OT) has firmly established itself as a fundamental tool in machine learning and statistics (Peyré et al., 2019), offering a geometrically meaningful way to compare probability distributions. Its applications span a vast landscape, including domain adaptation (Courty et al., 2014), generative modeling (An et al., 2019; Li et al., 2023), and biological data analysis (Schiebinger et al., 2019). The widespread adoption of OT is largely attributable to the computational breakthrough of entropic regularization (Cuturi, 2013), which enabled the use of efficient Sinkhorn-type scaling algorithms.

Building upon this foundation, Unbalanced OT (UOT) has emerged as a flexible generalization specifically designed to handle scenarios involving outliers, mass variations, or partial matching. By relaxing the strict mass conservation constraints of standard OT using φ-divergences, UOT accommodates datasets with varying total masses (Liero et al., 2018; Chizat et al., 2018b).

Despite the success of UOT, its algorithmic landscape remains heavily skewed toward generalized Sinkhorn methods (Chizat et al., 2018a; Séjourné et al., 2023). This is in stark contrast to balanced OT, where semi-dual formulations have become the standard for large-scale (Genevay et al., 2016; Seguy et al., 2017) and semi-discrete applications (Kitagawa et al., 2016). Surprisingly, semi-dual based algorithms have neither been widely proposed nor analyzed for the UOT problem. The notable exceptions are the works of Vacher and Vialard (2023), where the statistical properties

of unregularized continuous UOT are analyzed via its semi-dual formulation, and Choi et al. (2023), who successfully used SGD on the unregularized semi-dual for generative model training. However, both the optimization geometry of the regularized semi-dual and the theoretical convergence guarantees of SGD schemes in this setting remain largely unexplored.

This gap likely stems from the theoretical "stiffness" of the unbalanced formulation. Unlike the Sinkhorn algorithm, which sees its computational complexity improve from $\mathcal{O}(n^2/\varepsilon^2)$ to $\mathcal{O}(n^2/\varepsilon)$ in the unbalanced setting, gradient-based methods on the semi-dual appear to suffer from poor conditioning. Furthermore, the standard use of Kullback-Leibler (KL) marginal penalties is often taken for granted, as it is a prerequisite for the coordinate-descent updates of Sinkhorn. This hegemony has likely overshadowed the exploration of alternative divergences for Entropic UOT. In this work, we demonstrate that employing the χ^2 divergence on the target measure is actually a key factor in mitigating ill-conditioning, rendering the problem tractable for first-order methods.

Contributions. We challenge the perspective that UOT is ill-suited for gradient methods by providing a comprehensive analysis of the Entropic UOT semi-dual geometry. Our contributions are threefold:

• **Theoretical Analysis of the Entropic Semi-Dual Geometry:** We provide a thorough analysis of the Entropic UOT semi-dual. Our key observation is that the local condition number at the optimizer scales strictly as $\mathcal{O}(1/\varepsilon)$, effectively removing the dependence on the problem size n. Furthermore, we identify key properties enabling gradient schemes to adapt to this favorable geometry, specifically establishing the generalized self-concordance, global smoothness bounds, and (L_0, L_1)-type smoothness of the semi-dual.

• **Stochastic Regime:** We analyze the Projected Averaged SGD (PASGD) for the UOT semi-dual. We prove that PASGD naturally adapts to the benign local geometry, achieving a convergence rate of $\mathcal{O}(n/\varepsilon T)$. This yields a lightweight solver suitable for massive datasets.

• **Deterministic Regime:** For the full-batch discrete setting, we leverage a tight, data-dependent upper bound on the local smoothness. We utilize this bound to design an Adaptive Nesterov Accelerated Gradient (NAG) method. Unlike standard acceleration, which relies on conservative global constants, our method adjusts its step size dynamically, achieving a local complexity of $\mathcal{O}(n^2\sqrt{1/\varepsilon}\ln(1/\delta))$.

2. Background on Unbalanced Optimal Transport

Let $(\mathcal{X}, d_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y})$ be Polish spaces. We consider finite nonnegative measures $\mu \in \mathcal{M}_+(\mathcal{X})$ and $\nu \in \mathcal{M}_+(\mathcal{Y})$, and a continuous non-negative cost $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. For a coupling $\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})$, we denote its marginals by $\pi_1 \in \mathcal{M}_+(\mathcal{X})$ and $\pi_2 \in \mathcal{M}_+(\mathcal{Y})$.

Unbalanced OT with entropic regularization. Unbalanced optimal transport (UOT) compares μ and ν while allowing mass variation: instead of enforcing $\pi_1 = \mu$ and $\pi_2 = \nu$, it penalizes marginal mismatch. With entropic regularization $\varepsilon > 0$ and penalty weights $\rho_1, \rho_2 > 0$:

$$\mathrm{UOT}^{\rho_1,\rho_2}_{\varepsilon,c}(\mu,\nu) := \min_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int_{\mathcal{X}\times\mathcal{Y}} c\mathrm{d}\pi + \varepsilon\,\mathrm{KL}(\pi \mid \mu\otimes\nu) + \rho_1\, D_1(\pi_1 \mid \mu) + \rho_2\, D_2(\pi_2 \mid \nu)\,. \tag{1}$$

The marginal penalties D_1 and D_2 are chosen as Csiszár (f-)divergences (Csiszár, 1967; Ali and Silvey, 1966), which are of the form

$$D_\varphi(\alpha \mid \beta) := \int \varphi\left(\frac{d\alpha}{d\beta}\right) d\beta, \quad \text{for } \alpha \ll \beta, \quad \text{otherwise} : +\infty.$$

for $\varphi : [0, \infty) \to \mathbb{R} \cup \{+\infty\}$ convex with $\varphi(1) = 0$. The usual choices in OT are (i) KL, with $\varphi_{\mathrm{KL}}(t) = t \ln t - t + 1$ and $D_{\varphi_{\mathrm{KL}}} = \mathrm{KL}$, and (ii) the quadratic (Pearson χ^2) divergence, with $\varphi_{\chi^2}(t) = \frac{1}{2}(t-1)^2$ and $D_{\varphi_{\chi^2}} = D_{\chi^2}$.
The joint term $\mathrm{KL}(\pi \mid \mu \otimes \nu)$ is the standard entropic regularization, which makes the objective smoother and enables efficient iterative solvers (Sinkhorn-type schemes in OT (Cuturi, 2013) and in UOT (Chizat et al., 2018a)). Finally, $\varepsilon \to 0$ recovers the unregularized UOT objective, while $\rho_1, \rho_2 \to +\infty$ formally enforces balanced marginals and recovers (entropic) OT.

Entropic dual. Entropic UOT admits a dual formulation over measurable potentials $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$:

$$\mathrm{UOT}_{\varepsilon,c}^{\rho_1,\rho_2}(\mu,\nu) = \sup_{f,g} \left\{ -\varepsilon \int \exp\left(\tfrac{f+g-c}{\varepsilon}\right) \mathrm{d}\mu\mathrm{d}\nu - \rho_1 \int_{\mathcal{X}} \varphi_1^c\left(-\tfrac{f}{\rho_1}\right) \mathrm{d}\mu - \rho_2 \int_{\mathcal{Y}} \varphi_2^c\left(-\tfrac{g}{\rho_2}\right) \mathrm{d}\nu \right\},$$

up to additive constants independent of (f, g). Here $\varphi^c : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is the convex conjugate $\varphi^c(s) = \sup_{t \geq 0}\{st - \varphi(t)\}$. In particular:

$$\begin{aligned} \varphi_{\mathrm{KL}}^c(s) &= e^s - 1, \qquad s \in \mathbb{R}, \\ \varphi_{\chi^2}^c(s) &= s + \tfrac{1}{2}s^2 \quad \text{if } s \geq -1\,, \quad \text{otherwise } -\tfrac{1}{2}\,. \end{aligned}$$

Dual optimizers and induced coupling. The dual potentials explicitly parameterize the optimal coupling: if $(f^\star, g^\star)$ maximizes the dual, then $\pi^\star \ll \mu \otimes \nu$ and

$$\frac{d\pi^\star}{d(\mu \otimes \nu)}(x, y) = \exp\left(\frac{f^\star(x) + g^\star(y) - c(x, y)}{\varepsilon}\right). \tag{2}$$

3. Entropic UOT Semi-Dual: Derivation and Properties

Having established the primal and dual formulations, we now specialize to the, at least, semi-discrete paradigm. In this setting, the target measure ν is discrete (or has been discretized), while the source measure μ remains abstract (continuous or discrete). The choice of the target measure being discrete, rather than the source one, is arbitrary here.

Setting 1 *We assume ν is a discrete positive measure supported on n points $\{y_1, \ldots, y_n\} \subset \mathcal{Y}$:*

$$\nu = \sum_{j=1}^{n} \beta_j \delta_{y_j}, \quad \textit{with } \beta_j > 0.$$

To ensure stability, we assume the weights are balanced relative to the resolution n. Specifically, there exist constants $b/B \approx 1$ such that for all j, $b/n \leq \beta_j \leq B/n$.

3.1. Semi-Dual Formulation and Gradient

The semi-dual functional is derived by explicitly solving the maximization of the dual objective with respect to the source potential f. This reduces the problem to an unconstrained optimization over the target potential vector $\mathbf{g} = (g_1, \ldots, g_n) \in \mathbb{R}^n$. Before introducing the semi-dual functional, we define the following auxiliary quantities for any $x \in \mathcal{X}$:

$$B_j(x, \mathbf{g}) := \beta_j \exp\left(\tfrac{g_j - c(x, y_j)}{\varepsilon}\right), Z(x, \mathbf{g}) := \sum_{j=1}^{n} B_j(x, \mathbf{g})\,, \quad w_j(x, \mathbf{g}) := \tfrac{B_j(x, \mathbf{g})}{Z(x, \mathbf{g})}.$$

We maintain a fixed target penalty $D_2 = D_{\chi^2}$ but consider two standard options for the source penalty D_1:
• **(KL-source)** $D_1 = \mathrm{KL}$, allowing for a closed-form elimination of f.
• **(χ^2-source)** $D_1 = D_{\chi^2}$, which involves the Lambert W function.
To make the presentation easier, in the core of the paper, we fix $D_1 = \mathrm{KL}$, since it has a slightly easier form to present and derive the theorems. However, all the results here are the same for $D_1 = \chi^2$, up to constants. The proofs for the case $D_1 = \chi^2$ are also in the appendix.

Proposition 2 (Semi-Dual Objective and Gradient) *(Proof in Appendix B.) With $\alpha := \frac{\varepsilon}{\varepsilon+\rho_1}$, the semi-dual objective $\mathcal{J} : \mathbb{R}^n \to \mathbb{R}$ is*

$$\mathcal{J}(\mathbf{g}) = (\rho_1 + \varepsilon) \int_{\mathcal{X}} Z(x, \mathbf{g})^{\alpha} \, d\mu(x) + \sum_{j=1}^{n} \beta_j \left(\frac{g_j^2}{2\rho_2} - g_j \right). \tag{3}$$

Its gradient with respect to the k-th component is given by:

$$\nabla_k \mathcal{J}(\mathbf{g}) = \int_{\mathcal{X}} Z(x, \mathbf{g})^{\alpha} \, w_k(x, \mathbf{g}) \, d\mu(x) + \tfrac{\beta_k}{\rho_2} g_k - \beta_k. \tag{4}$$

Remark: While we allow different divergences for the source measure, fixing the χ^2 divergence for the target marginal is compulsory to have a data independent strong convexity. For instance, this is not (always) the case when $D_2 = \mathrm{KL}$, we refer to Section 3.3 for more details.

3.2. The First Order Condition Keystone

The efficiency of gradient-based algorithms is governed by the geometry of the objective function. While the strong convexity of $\mathcal{J}$ is evident from the target χ^2 divergence term, the smoothness is more subtle. We show here that by analyzing the smoothness locally, we can derive a much tighter conditioning bound. To do so, we identify the *transport* part of the objective, which is the key component of our analysis:

$$\mathcal{J}_{\mathrm{trans}}(\mathbf{g}) = (\rho_1 + \varepsilon) \int_{\mathcal{X}} Z(x, \mathbf{g})^{\alpha} \, d\mu(x); \quad [\nabla \mathcal{J}_{\mathrm{trans}}(\mathbf{g})]_k = \int_{\mathcal{X}} Z(x, \mathbf{g})^{\alpha} \, w_k(x, \mathbf{g}) \, d\mu(x) \, .$$

From it, we state our key smoothness result, which holds for both source divergences.

Theorem 3 (Smoothness Bound via Gradient Transport) *(Proof in Appendix C.) For all $\mathbf{g} \in \mathbb{R}^n$, the operator norm of the Hessian satisfies:*

$$\|\nabla^2 \mathcal{J}(\mathbf{g})\|_{\mathrm{op}} \leq \frac{1}{\varepsilon} \|\nabla \mathcal{J}_{\mathrm{trans}}(\mathbf{g})\|_{\infty} + \frac{\beta_{\max}}{\rho_2}. \tag{5}$$

This theorem is our *keystone*: it transforms the problem of bounding curvature into the problem of bounding the gradient. This link will be further leveraged in the next section.

The next proposition is straightforward to derive yet fundamental; it establishes that the optimal potentials are naturally confined to a region where the geometry is well-behaved.

Proposition 4 (First-Order Optimality) *At the global minimizer $\mathbf{g}^\star$, the condition $\nabla \mathcal{J}(\mathbf{g}^\star) = 0$ implies $[\nabla \mathcal{J}_{\text{trans}}(\mathbf{g}^\star)]_k = \beta_k(1 - g_k^\star/\rho_2)$. Since $\nabla \mathcal{J}_{\text{trans}} \geq 0$, the optimizer satisfies the automatic box constraint:*

$$g_k^\star \leq \rho_2, \quad \forall k \in \{1, \ldots, n\}. \tag{6}$$

Building on this, we define the feasible region $\mathcal{K}$, which will serve as the focal point for our subsequent analysis and in our algorithms, where we will use $\delta = 0.1$ by default.

$$\mathcal{K}_\delta := \{\mathbf{g} \in \mathbb{R}^n \mid g_k \leq \rho_2 + \delta, \forall k\},$$

where $\delta > 0$ is a small margin ensuring $\mathbf{g}^\star \in \text{int}(\mathcal{K})$. By focusing on $\mathcal{K}$, we can transition from point-wise optimality to a more global understanding of the objective's geometry. In particular, in the following, we leverage this set to derive dimension-independent bounds on the curvature.

3.3. Global and Local Curvature

Lemma 5 (Uniform Gradient Bound and Smoothness) *(Proof in Appendix D.) On $\mathcal{K}$, the L_1-norm of the transport gradient is uniformly bounded: $\|\nabla \mathcal{J}_{\text{trans}}(\mathbf{g})\|_1 \leq C_{\text{bound}}$, where:*

$$C_{\text{bound}}^{\text{KL}} := \mu(\mathcal{X})\|\nu\|_1^\alpha \exp\left(\frac{\rho_2 + \delta}{\rho_1 + \varepsilon}\right) \tag{7}$$

Consequently, the Hessian is bounded on $\mathcal{K}$, and $\mathcal{J}$ is L-smooth with $L = \mathcal{O}(1/\varepsilon)$.

This result demonstrates that the objective remains "flat" enough for stable optimization even as the resolution $n \to \infty$, when we are in $\mathcal{K}$. Notably, the bound is independent of n since $\|\nu\|_1 \leq B$. However, it grows exponentially with the margin δ, illustrating that $\mathcal{J}$ is not globally smooth; its curvature is only controlled near the optimizer. Finally, we establish the conditioning of the problem, where the condition number for an L-smooth and γ-strongly convex objective is defined as

$$\kappa := \frac{L}{\gamma}.$$

Corollary 6 (Local Conditioning) *The objective $\mathcal{J}$ is $\frac{\beta_{\min}}{\rho_2}$-strongly convex on $\mathbb{R}^n$. At the optimizer $\mathbf{g}^\star$, the local condition number κ satisfies:*

$$\kappa(\nabla^2 \mathcal{J}(\mathbf{g}^\star)) \leq \frac{\beta_{\max}}{\beta_{\min}}\left(1 + \frac{\max_k\{\rho_2 - g_k^\star\}}{\varepsilon}\right). \tag{8}$$

This result highlights a crucial disconnect between the global worst-case analysis derived in Lemma 5 and the local reality of the problem. Even on $\mathcal{K}$, a global bound on L is of the form $\mathcal{O}(1/\varepsilon)$, leading to a condition number of $\mathcal{O}(N\rho_2/\varepsilon)$. In contrast, Corollary 6 shows that locally, the conditioning depends only on the regularization ratio ρ_2/ε and on the mass balance ratio $\beta_{\max}/\beta_{\min}$, which we assume is ≈ 1. Crucially, this local condition number is independent of n. This observation motivates the use of first-order methods that can adapt to the local curvature and converge at a rate independent of the problem size n.

The key to unlocking these local acceleration guarantees lies in our proof of generalized self-concordance for the semi-dual. While this property has been successfully applied to logistic regression Bach (2010, 2014) and discrete optimal transport Sun and Tran-Dinh (2019), establishing it in

our context allows us to strictly control the change of the Hessian locally. We utilize this property in both our semi-discrete and discrete settings to derive rates that depend on the favorable local geometry rather than global worst-case bounds.

Proposition 7 (Generalized self-concordance) *(Proof in Appendix E.) The semi-dual $\mathcal{J}$ is generalized self-concordant. That is, for $M = \frac{2+3\alpha}{\varepsilon}$ for* KL *source, and $M = \frac{6}{\varepsilon}$ for χ^2 , we have for any $\mathbf{g} \in \mathbb{R}^n$ and any direction $\mathbf{h} \in \mathbb{R}^n$:*

$$\left|\nabla^3 \mathcal{J}(\mathbf{g})[\mathbf{h}, \mathbf{h}, \mathbf{h}]\right| \leq M \|\mathbf{h}\|_\infty \langle \mathbf{h}, \nabla^2 \mathcal{J}(\mathbf{g})\mathbf{h} \rangle.$$

The Necessity of the Target χ^2 Divergence. As highlighted in Remark 2, the χ^2 target penalty offers a decisive geometric advantage over the standard KL-KL formulation. In the KL-KL setting, the diagonal terms of the dual Hessian scale with e^{-g/ρ_2}. Consequently, the curvature becomes highly anisotropic and vanishes exponentially as $g \to \infty$ (a phenomenon linked to mass destruction). This lack of uniform strong convexity severely complicates the analysis of accelerated algorithms. In stark contrast, the χ^2 penalty ensures a constant curvature of $1/\rho_2$ and guarantees global strong convexity. While the KL-KL geometry remains manageable in the batch discrete setting (discussed in Section 4.2), the χ^2 formulation provides the structural stability required for our results.

4. Adaptive Gradient Descent on the Semi-Dual

4.1. Large-Scale and Semi-Discrete Settings

We first address the setting where the source measure μ is continuous, or discrete with a cardinality large enough to make full-batch processing impossible. In this regime, Stochastic Gradient Descent (SGD) is the natural algorithmic choice. We demonstrate that a simple Projected Averaged SGD (PASGD) scheme achieves efficient convergence, escapes the worst-case analysis, and leverages the local smoothness of the semi-dual.

Unbiased Gradient Estimator. Assume $\mu(\mathcal{X}) < \infty$ and let $X_1, \dots, X_{m_b} \overset{\text{i.i.d.}}{\sim} \mu/\mu(\mathcal{X})$ be a batch of samples drawn from the normalized source measure. Using the gradient formulation (4), we define the stochastic gradient estimator $\widehat{\nabla}\mathcal{J}(\mathbf{g})$ component-wise for $k \in \{1, \dots, n\}$:

$$\widehat{\nabla}_k \mathcal{J}(\mathbf{g}) := \frac{\mu(\mathcal{X})}{m_b} \sum_{i=1}^{m_b} Z(X_i, \mathbf{g})^\alpha \, w_k(X_i, \mathbf{g}) + \frac{\beta_k}{\rho_2} g_k - \beta_k. \tag{9}$$

By the linearity of expectation and the identity $\mathbb{E}_{X \sim \mu/\mu(\mathcal{X})}[h(X)] = \frac{1}{\mu(\mathcal{X})} \int h \, d\mu$, it is immediate that this estimator is unbiased: $\mathbb{E}[\widehat{\nabla}\mathcal{J}(\mathbf{g})] = \nabla \mathcal{J}(\mathbf{g})$.

Complexity. Computing this estimator for the full vector $\mathbf{g} \in \mathbb{R}^n$ requires only $\mathcal{O}(m_b \cdot n)$ operations. This linear complexity in n makes the approach scalable to large target supports.

The Online Regime. A significant advantage of this stochastic formulation is its direct applicability to the *online* setting Hazan et al. (2016). Unlike batch methods that require repeated passes over a fixed dataset, our approach naturally handles streaming data where samples X_t arrive sequentially. In this regime, the algorithm effectively minimizes the population risk (the integral against μ) directly, rather than the empirical risk. This is particularly valuable for semi-discrete tasks, as in generative modeling (Li et al., 2023; Choi et al., 2023) where storing the full history of samples is memory-prohibitive. The constant memory footprint, storing only the potential $\mathbf{g} \in \mathbb{R}^n$, remains independent of the number of samples processed.

Adaptivity and Convergence. We propose to solve the semi-dual problem using Projected Averaged SGD (PASGD). While standard SGD can suffer from oscillation in ill-conditioned settings, averaging the iterates (Polyak-Ruppert averaging) is known to statistically adapt to the local curvature, achieving optimal asymptotic rates.

We define the update rule at step t with step size η_t as:

$$\mathbf{g}_{t+1} = \Pi_{\mathcal{K}}\left(\mathbf{g}_t - \eta_t \widehat{\nabla}\mathcal{J}(\mathbf{g}_t)\right), \quad \bar{\mathbf{g}}_T = \frac{1}{T}\sum_{t=1}^{T}\mathbf{g}_t. \tag{10}$$

Here, $\Pi_{\mathcal{K}}$ denotes the projection onto the feasible set $\mathcal{K} = \{\mathbf{g} \mid g_k \leq \rho_2 + \delta\}$, which ensures the iterates remain in the region where the gradient variance and Hessian are well-behaved (as established in Section 3).

Before stating the main convergence result, we give a simple lemma, important for the analysis of SGD schemes.

Proposition 8 (Variance Bound of Mini-Batch Gradient) *(Proof in Appendix D.1.) Let $\widehat{\nabla}\mathcal{J}(\mathbf{g})$ be the mini-batch gradient estimator computed with batch size $m_b \geq 1$, as defined in Eq. (9). For any $\mathbf{g} \in \mathcal{K}$, using the uniform bound C_{bound} from Lemma 5, the variance is bounded by:*

$$\mathbb{E}\left[\|\widehat{\nabla}\mathcal{J}(\mathbf{g}) - \nabla\mathcal{J}(\mathbf{g})\|_2^2\right] \leq \frac{4C_{\text{bound}}^2}{m_b}. \tag{11}$$

Theorem 9 (Convergence of PASGD) *(Proof in Appendix G.) Let the step sizes be chosen as $\eta_t = Ct^{-\gamma}$ with $\gamma \in (1/2, 1)$. Under Setting 1 and the projection onto $\mathcal{K}$, the averaged iterate $\bar{\mathbf{g}}_T$ converges to the optimum $\mathbf{g}^\star$ in objective value with an expected error of:*

$$\mathbb{E}\left[\mathcal{J}(\bar{\mathbf{g}}_T) - \mathcal{J}(\mathbf{g}^\star)\right] = \mathcal{O}\left(\frac{n\rho_2^2}{\varepsilon T}\right).$$

We give here a sketch of proof, to illustrate where and how we are able to leverage the local condtionning near the optimum. *Sketch of Proof.* Classical results Polyak and Juditsky (1992); Gadat and Panloup (2017) using global strong convexity and uniform gradient error bound from Lemma 5 assure, with $\mathbf{H} = \nabla^2\mathcal{J}(\mathbf{g}^\star)$ and $\mathbf{\Sigma}$ the noise covariance at the optimum:

$$\mathbb{E}[\|\bar{\mathbf{g}}_T - \mathbf{g}^\star\|^2] \leq \frac{\mathrm{Tr}(\mathbf{H}^{-1}\mathbf{\Sigma}\mathbf{H}^{-1})}{Tm_b} + o(1/T).$$

From the global strong convexity of $\mathcal{J}$, this term scales as $\mathcal{O}(\rho_2^2 n^2/T)$. To handle the non-global smoothness, we split the objective gap and will use the generalized self-concordance property. Locally, $\mathcal{J}$ is L-smooth with $L \propto 1/n\varepsilon$. From the generalized self-concordance, we have (see Corollary 13): $L(\mathbf{g}_T) \leq \exp(\frac{2+3\alpha}{\varepsilon}\|\mathbf{g}_T - \mathbf{g}^\star\|)L(\mathbf{g}^\star)$. Therefore

$$\mathbb{E}[\mathcal{J}(\bar{\mathbf{g}}_T) - \mathcal{J}^\star] \leq \frac{L(\mathbf{g}^\star)(2+3\alpha)}{2}\mathbb{E}[\|\bar{\mathbf{g}}_T - \mathbf{g}^\star\|^2] + \frac{C_1}{\varepsilon}\mathbb{P}(\|\bar{\mathbf{g}}_T - \mathbf{g}^\star\| \geq \varepsilon).$$

Due to the concentration of the average SGD, with $\gamma_t \propto t^{-b}, b \in (1/2, 1)$, which gives high order moments for both the SGD and ASGD schemes, $\mathbb{P}(\|\bar{\mathbf{g}}_T - \mathbf{g}^\star\| \geq \varepsilon)$ is negligible, which concludes. □

Remark: Observe that for this proof of adaptivity, we need $\gamma_t \propto t^{-b}$ with $b < 1$. However, we refer to Appendix I, for a motivation of the study of other SGD schemes and learning rate, that could lead to an even better complexity, by exploiting even more the good local conditioning.

4.1.1. NUMERICAL EXPERIMENTS: SEMI-DISCRETE SETTING

We validate our theoretical findings on a synthetic 10-dimensional mixture of Gaussians with 4 modes. The target ν is discrete with $n = 2,000$ uniformly weighted points. To evaluate convergence, we compute a high-precision ground truth $\mathbf{g}^\star$ using the deterministic Adaptive NAG solver (Section 4.2) run to machine precision, with $n' = 10,000$ points for the source measure. All results report the average of 20 independent runs; variance was negligible and is omitted for clarity.

Efficiency of PASGD. We reaffirm here that scaling the learning rate by the inverse of the strong convexity yields optimal results. Figure 1 compares standard SGD against Projected Averaged SGD (PASGD) using a step size decay of $\eta_t = C\frac{n}{\rho_2}(t + 1/\varepsilon)^{-2/3}$ for varying scales C. Here, $\varepsilon = 0.01$. For $b = 2/3$, the choice $C = 1$, which corresponds to the natural scale dictated by the strong convexity of $\mathcal{J}$, achieves the best performance.

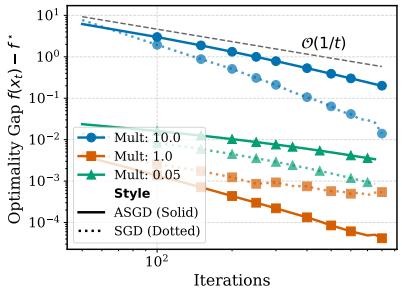

Figure 1: **PASGD vs. SGD.** Convergence of the objective gap on a semi-discrete UOT problem ($n = 2000$), with $\eta_t = C\frac{n}{\rho_2}(t + 1/\varepsilon)^{-2/3}$, $C \in \{0.05, 1, 10\}$. PASGD confirms the $\mathcal{O}(1/T)$ rate and shows superior performance compared to SGD.

Impact of Regularization. Figure 2 examines the sensitivity of the algorithm to the regularization parameter ε. While optimal theoretical bounds suggest the possibility of eliminating the dependency on ε, we observe that practical performance retains some sensitivity. Specifically, decreasing ε impacts the convergence of the objective function, which scales with roughly $1/\varepsilon$. However, importantly, the squared distance to the optimum $\|\bar{\mathbf{g}}_t - \mathbf{g}^\star\|^2$ exhibits a much milder dependence on ε, demonstrating significant robustness in parameter space.

Large-Scale Application: Color Transfer. To demonstrate scalability, we replicate the color transfer task of Kemertas et al. (2025) on high-resolution images (n=512^2 and n=1024^2 pixels). We employ our PASGD solver on a single modern GPU using a batch size of $m_b = 32$ and a robust, non-tuned learning rate schedule $\eta_t = \frac{n\rho_2}{\varepsilon^{-1}+t^{2/3}}$. We set $\varepsilon = 5 \cdot 10^{-3}$. Our solver exhibits great efficiency gains compared to the state-of-the-art: while Kemertas et al. (2025) reports a runtime of ≈ 10 hours for the 1024^2 task (using `PyKeops` for memory management), our method converges in just 30 minutes (and ≈ 2 minutes for 512^2). This strictly linear $\mathcal{O}(n)$ complexity, both in memory and compute, establishes Entropic UOT via PASGD as a scalable alternative to $\mathcal{O}(n^2)$ per iteration OT baselines for large-scale tasks.

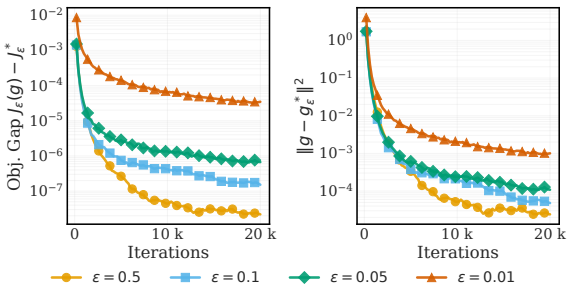

Figure 2: **Effect of ε.** Convergence profiles for varying entropic regularization levels. The objective gap (Left) reflects a practical dependence on ε, whereas the parameter error $\|\bar{\mathbf{g}}_t - \mathbf{g}^\star\|^2$ (Right) demonstrates higher robustness to the regularization parameter.

(*a*) Source (*b*) Target (*c*) $\rho = 0.1$ (*d*) $\rho = 1$ (*e*) $\rho = 10$

Figure 3: **High-Resolution Color Transfer** (1024×1024)**.** We transport the source color distribution (a) to the target geometry (b). The parameter ρ controls the fidelity of the mass transfer. At $\rho = 0.1$ (c), the relaxation allows for partial matching. At $\rho = 10$ (e), the penalty enforces nearly balanced transport.

4.2. Discrete - Full Batch setting

We now transition to the full-batch setting, which represents the most common scenario for practitioners where measures are discrete or have been pre-discretized.

Setting 10 (Full Batch Setting) *We assume μ and ν are discrete positive measures supported on n_1 and n_2 points, respectively: $\mu = \sum_{j=1}^{n_1} \alpha_j \delta_{x_j}$, $\nu = \sum_{j=1}^{n_2} \beta_j \delta_{y_j}$.*

To ensure numerical stability, we assume the weights of ν are balanced relative to the resolution n_2: there exist constants $b, B \approx 1$ such that $b/n_2 \leq \beta_j \leq B/n_2$ for all j.

Data-Dependent Smoothness. Standard acceleration schemes rely on a global Lipschitz constant L to determine step sizes. In the UOT semi-dual, however, the global L is prohibitively large, while the local curvature near the optimum is up to n times better. We link here Entropic UOT to the recent line of work on adaptive gradient methods for (L_0, L_1)-smooth functions (Zhang et al., 2019), where local smoothness scales with the gradient norm. We derive a specialized form of this property that holds explicitly along gradient descent trajectories.

Proposition 11 (Asymmetric L_0–L_1 smoothness along a gradient step) *Consider the segment $g(s) := g_t - s\lambda_t \nabla \mathcal{J}(g_t)$, $s \in [0,1]$, with step size*

$$\lambda_t := \frac{1}{L(g_t)}, \qquad L(g_t) := \frac{C}{\varepsilon}\|\nabla \mathcal{J}_{\text{trans}}(g_t)\|_\infty + \frac{C\beta_{\max}}{\rho_2} + \frac{C}{e\varepsilon}\|\nabla \mathcal{J}(g_t)\|_\infty,$$

where $C = 6e$ for χ^2-source and $C = (2+3\alpha)e$ for KL-source. Then for any two points g_1, g_2 on the segment $\{g(s) : s \in [0,1]\}$,

$$\|\nabla \mathcal{J}(g_1) - \nabla \mathcal{J}(g_2)\| \;\le\; L(g_t)\,\|g_1 - g_2\|.$$

This proposition is pivotal. It shows that, rather than relying on a crude worst-case global constant, we can use a local smoothness estimate $L(g_t)$ that is directly controlled by the computed gradient. As the algorithm converges and $\nabla \mathcal{J}(g_t) \to 0$, this bound tightens naturally toward $L(g^\star) \approx \beta_{\max}\big(1 + \rho_2/\varepsilon\big)$, which in turn allows the effective step size to increase. Leveraging this local geometry, we propose an Adaptive Nesterov Accelerated Gradient (ANAG) method. In our setting, each ANAG iteration has complexity $\mathcal{O}(n_1 n_2)$.

While ANAG is structurally related to the heuristic adaptive schemes of Malitsky and Mishchenko (2019), which estimate curvature via finite differences, our method uses the analytical upper bound $L(g_t)$ derived from the problem structure; combined with our (L_0, L_1)-type smoothness control and the projection set, this enables convergence guarantees. Finally, to ensure that all smoothness arguments remain valid, we include a simple safeguard restart whenever the extrapolated point leaves the region $\mathcal{K}_1$, where $\mathcal{J}$ has controlled smoothness. This restart is expected to be rare in practice, and often absent, since the optimizer lies in $\mathcal{K}_0$; it is introduced primarily to simplify the analysis.

Algorithm 1 Smoothness Adaptive NAG (with safeguard restarts)

Input data: $\mu = \sum_{i=1}^{n_1} \alpha_i \delta_{x_i}, \nu = \sum_{j=1}^{n_2} \beta_j \delta_{y_j}$
UOT parameters: $\varepsilon, \rho_1, \rho_2 > 0$
Sets: $\mathcal{K} := \{g \in \mathbb{R}^n : g_i \le \rho_2 + 0.1\}$, $\mathcal{K}_1 := \{g \in \mathbb{R}^n : g_i \le \rho_2 + 1\}$
$y_0 \leftarrow g_0 = (0, \dots, 0) \in \mathbb{R}^n$
for $t = 0$ **to** $T-1$ **do**
  $L_t \leftarrow \frac{C}{\varepsilon}\|\nabla \mathcal{J}_{\text{trans}}(y_t)\|_\infty + \frac{C\beta_{\max}}{\rho_2} + \frac{C}{e\varepsilon}\|\nabla \mathcal{J}(y_t)\|_\infty$
  $\theta_t \leftarrow \frac{\sqrt{L_t} - \sqrt{\beta_{\min}/\rho_2}}{\sqrt{L_t} + \sqrt{\beta_{\min}/\rho_2}}$
  $g_{t+1} \leftarrow \Pi_{\mathcal{K}, \cdot, \infty}[y_t - \frac{1}{L_t}\nabla \mathcal{J}(y_t)]$
  $y_{t+1} \leftarrow g_{t+1} + \theta_t(g_{t+1} - g_t)$
  Restart: if $y_{t+1} \notin \mathcal{K}_1$ **then** $y_{t+1} \leftarrow g_t$
end for
Output: g_T

Theorem 12 (Adaptive NAG Convergence Rate) *(Proof in Appendix H.) Let R be the number of restarts. Then, the iterates generated by Algorithm 1 satisfy*

$$\mathcal{J}(g_{T+1}) - \mathcal{J}^\star \le 2^R \left(\mathcal{J}(g_0) - \mathcal{J}^\star + \frac{\beta_{\min}}{2\rho_2}\|g_0 - g^\star\|^2\right) \prod_{t=0}^{T} \left(1 - \sqrt{\frac{\beta_{\min}}{\rho_2 L_t}}\right), \tag{12}$$

Furthermore, the algorithm ensures $y_t \in K_1$ for all t, so using C_{bound} from Lemma 5, we have $L_t \le \bar{L} = \mathcal{O}(C_{bound}/\varepsilon)$ for all t. This implies the following rates:

1. ***Global Rate:*** *We have at least the contraction rate $1 - \mathcal{O}(\sqrt{\varepsilon/\beta_{\min}\rho_2})$ for both the objective gap and gradient norm.*

2. ***Local Rate:*** *Since* $L_t \leq 2\beta_{\max}\left(\frac{1}{\varepsilon}+\frac{1}{\rho_2}\right)+2\|\nabla\mathcal{J}(g_{T+1})\|$, *assuming* $\beta_{\min}/\beta_{\max}\simeq 1$ *and substituting the contraction rate of the gradient norm into* (12) *shows that, locally, we have a contraction rate of* $1-\mathcal{O}(\sqrt{\varepsilon/\rho_2})$.

Therefore, in our setting, we have a local total complexity of $\mathcal{O}(n_1 n_2\sqrt{\rho_2/\varepsilon})\ln(1/\delta_{\text{acc}})$ for δ_{acc}-accuracy. This result formally confirms our geometric intuition: while the initial convergence may depend on the problem size, the adaptive solver rapidly transitions to a regime where the complexity is governed solely by the local condition number $\mathcal{O}(1/\varepsilon)$, yielding a highly scalable discrete solver.

Experiment: Scale Invariance and Acceleration. We validate Theorem 12 using synthetic measures supported on n points in $[0,1]^{10}$ ($\beta_i = 1/n, \varepsilon = 10^{-2}, \rho_{1,2} = 10$). We compare ANAG against Adaptive GD and two baselines: (i) fixed (conservative) learning rate GD and (ii) fixed learning rate NAG. Figure 4 highlights the **scale invariance** of ANAG: trajectories for $n \in \{1500, 3000, 4500\}$ overlap, confirming that the local condition number is independent of problem size n. Furthermore, the results isolate the benefits of adaptivity: Adaptive GD significantly outperforms Conservative NAG, demonstrating that exploiting local smoothness ($L(g_t) \ll L_{\text{global}}$) is more critical than blind acceleration. ANAG yields the fastest rates by combining both advantages.

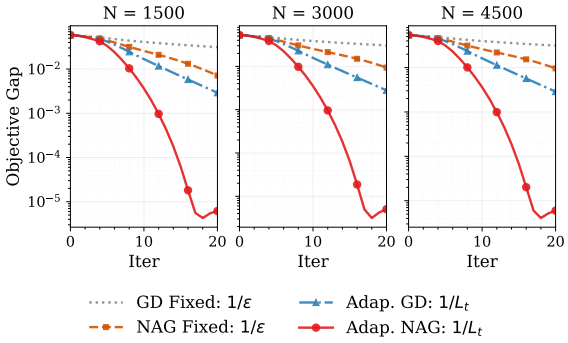

Figure 4: **Scale Invariance and Adaptive Acceleration.** Convergence on random measures with varying support sizes n ($\varepsilon = 0.01, \rho = 10$). We compare ANAG against Adaptive GD and Conservative NAG (fixed step $1/L_{\text{global}}$). The overlap of ANAG curves confirms the dimension-independent local complexity, while the superiority of adaptive schemes highlights the benefit of local step sizes.

Remark: ANAG for the KL-KL divergences case. Our restriction to the χ^2 target penalty simplifies the analysis by fixing the strong convexity parameter. However, for the $D_1 = D_2 = \text{KL}$ case, the strong convexity depends on the diagonal of the Hessian, given by the vector $\frac{1}{\rho_2}e^{-\mathbf{g}/\rho_2}\odot\boldsymbol{\beta}$. Assuming the optimal potentials are bounded, one could adapt Algorithm 1 to update the momentum parameter θ_t using a data-dependent strong convexity estimate $\mu_t \approx \min_j\left(\frac{\beta_j}{\rho_2}e^{-(g_t)_j/\rho_2}\right)$.

Comparison to the literature

UOT algorithms. The primary solvers for Entropic UOT are generalized Sinkhorn algorithms (Chizat et al., 2018a) and their translation-invariant variants (Séjourné et al., 2022). In the unbal-

anced setting, Sinkhorn iterations enjoy an enhanced complexity of $\mathcal{O}(n^2/\varepsilon)$ (Pham et al., 2020), improving upon the $\mathcal{O}(n^2/\varepsilon^2)$ scaling of balanced OT (Dvurechensky et al., 2018). Beyond Sinkhorn, Nguyen et al. (2023) proposed a Gradient Extrapolation Method (GEM) for L_2-regularized UOT, which produces sparse transport plans. While GEM achieves linear convergence, its condition number κ depends on the input measures and scales as $\mathcal{O}(n)$, which is prohibitive for large-scale problems. Alternatively, Chapel et al. (2021) introduced a Majorization-Minimization scheme with $\mathcal{O}(n^2)$ per-iteration cost, though it currently lacks global convergence rates. In the continuous domain, neural network approximations exist (Gazdieva et al., 2024), but these methods are computationally heavy and lack convex optimization guarantees.

Table 1: Complexity of UOT algorithms to achieve δ-accuracy, given discrete measures of size n. Our methods provide the first accelerated rates for discrete UOT and the first rigorous rates for the semi-discrete regime.

Algorithm	**Regime**	**Regularization**	**Complexity / Rate**
Gen. Sinkhorn (Pham et al., 2020)	Discrete	Entropic	$\mathcal{O}(n^2 \ln(1/\delta)/\varepsilon)$
GEM (Nguyen et al., 2023)	Discrete	L_2	$\mathcal{O}(\kappa(n) n^2 \ln(1/\delta))$
MM (Chapel et al., 2021)	Discrete	L_2, Entropic	N/A
Neural UOT (Eyring et al., 2023)	Continuous	Entropic	Heuristic
Adaptive NAG (Ours)	**Discrete**	Entropic	$\tilde{\mathcal{O}}(n^2 \ln(1/\delta)/\sqrt{\varepsilon})$ (Local)
PASGD (Ours)	**Semi-Disc.**	Entropic	$\mathcal{O}(n/\varepsilon T)$

Adaptive Gradient Methods. The strategy of adapting step sizes via local smoothness estimation was pioneered by Malitsky and Mishchenko (2019) and recently extended to (L_0, L_1)-smooth functions (Gorbunov et al., 2025; Vankov et al., 2025). While these general frameworks often require complex algorithmic adjustments to handle potentially unbounded curvature, our analysis exploits the specific geometry of the UOT semi-dual. We show that the smoothness is locally bounded, enabling us to prove the convergence of a simpler Adaptive NAG scheme with tighter complexity guarantees than generic (L_0, L_1) approaches.

5. Conclusion

In this work, we demonstrate that the geometry of the Entropic UOT semi-dual is naturally suited for adaptive first-order methods. In the semi-discrete regime, our stochastic scheme provides the first rigorous convergence guarantees for UOT, matching the $\mathcal{O}(n^2/\varepsilon)$ efficiency of Sinkhorn while enabling strictly linear scalability. In the full-batch discrete setting, our Adaptive NAG achieves a superior local accelerated complexity of $\tilde{\mathcal{O}}(n^2/\sqrt{\varepsilon})$ and a global worst-case of $\tilde{\mathcal{O}}(n^{2.5}/\sqrt{\varepsilon})$.

Looking forward, **exploiting the key geometric properties derived in this work suggests** further potential in bridging UOT with generalized (L_0, L_1)-smoothness frameworks to design tailored solvers. Furthermore, promising directions to improve the global worst-case complexity include (i) AdaGrad-type algorithms (Duchi et al., 2011) in the semi-discrete setting to further leverage global conditioning and (ii) the integration of decreasing regularization schedules (Schmitzer, 2019) in the discrete setting. We hypothesize that leveraging the strong convexity of the semi-dual in this manner could eliminate the current initialization overhead, potentially securing a global $\tilde{\mathcal{O}}(n^2/\sqrt{\varepsilon})$ convergence rate.

References

Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1): 131–142, 1966.

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. *ICLR 2020*, 2019.

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4: 384–414, 2010.

Francis Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *The Journal of Machine Learning Research*, 15(1):595–627, 2014.

Nikhil Bansal and Anupam Gupta. Potential-function proofs for gradient methods. *Theory of Computing*, 15(1):1–32, 2019.

Bernard Bercu and Jérémie Bigot. Asymptotic distribution and convergence rates of stochastic algorithms for entropic optimal transportation between probability measures. *The Annals of Statistics*, 49(2):968–987, 2021.

Laetitia Chapel, Rémi Flamary, Haoran Wu, Cédric Févotte, and Gilles Gasso. Unbalanced optimal transport through non-negative penalized linear regression. *Advances in Neural Information Processing Systems*, 34:23270–23282, 2021.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of computation*, 87(314):2563–2609, 2018a.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and kantorovich formulations. *Journal of Functional Analysis*, 274(11): 3090–3123, 2018b.

Jaemoo Choi, Jaewoong Choi, and Myungjoo Kang. Generative modeling through the semi-dual formulation of unbalanced optimal transport. *Advances in Neural Information Processing Systems*, 36:42433–42455, 2023.

Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I 14*, pages 274–289. Springer, 2014.

Imre Csiszár. On information-type measure of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances In Neural Information Processing Systems*, 26, 2013.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

P. Dvurechensky, A. Gasnikov, and A. Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. In *International Conference On Machine Learning*, pages 1367–1376, 2018.

Luca Eyring, Dominik Klein, Théo Uscidda, Giovanni Palla, Niki Kilbertus, Zeynep Akata, and Fabian Theis. Unbalancedness in neural monge maps improves unpaired domain translation. *arXiv preprint arXiv:2311.15100*, 2023.

Sébastien Gadat and Fabien Panloup. Optimal non-asymptotic bound of the ruppert-polyak averaging without strong convexity. *arXiv preprint arXiv:1709.03342*, 2017.

Milena Gazdieva, Arip Asadulaev, Evgeny Burnaev, and Alexander Korotin. Light unbalanced optimal transport. *Advances in Neural Information Processing Systems*, 37:93907–93938, 2024.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. *Advances in neural information processing systems*, 29, 2016.

Antoine Godichon-Baggioni. Lp and almost sure rates of convergence of averaged stochastic gradient algorithms: locally strongly convex objective. *ESAIM: Probability and Statistics*, 23:841–873, 2019.

Eduard Gorbunov, Nazarii Tupitsa, Sayantan Choudhury, Alen Aliev, Peter Richtárik, Samuel Horváth, and Martin Takáč. Methods for convex (l_0, l_1)-smooth optimization: Clipping, acceleration, and adaptivity. In *The Thirteenth International Conference on Learning Representations*, 2025.

Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Mete Kemertas, Amir-massoud Farahmand, and Allan Douglas Jepson. A truncated newton method for optimal transport. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jun Kitagawa, Quentin Mérigot, and Boris Thibert. A newton algorithm for semi-discrete optimal transport. *Journal of the European Mathematical Society*, 21, 03 2016. doi: 10.4171/JEMS/889.

Zezeng Li, Shenghao Li, Zhanpeng Wang, Na Lei, Zhongxuan Luo, and David Xianfeng Gu. Dpm-ot: a new diffusion probabilistic model based on optimal transport. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 22624–22633, 2023.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, 211 (3):969–1117, 2018.

Yura Malitsky and Konstantin Mishchenko. Adaptive gradient descent without descent. *arXiv preprint arXiv:1910.09529*, 2019.

Yu Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1):381–404, 2015.

Quang Minh Nguyen, Hoang H Nguyen, Yi Zhou, and Lam M Nguyen. On unbalanced optimal transport: Gradient methods, sparsity and approximation error. *Journal of Machine Learning Research*, 24(384):1–41, 2023.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Khiem Pham, Khang Le, Nhat Ho, Tung Pham, and Hung Bui. On unbalanced optimal transport: An analysis of sinkhorn algorithm. In *International Conference on Machine Learning*, pages 7673–7682. PMLR, 2020.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

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

Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-scale optimal transport and mapping estimation. *arXiv preprint arXiv:1711.02283*, 2017.

Thibault Séjourné, François-Xavier Vialard, and Gabriel Peyré. Faster unbalanced optimal transport: Translation invariant sinkhorn and 1-d frank-wolfe. In *International Conference on Artificial Intelligence and Statistics*, pages 4995–5021. PMLR, 2022.

Thibault Séjourné, Gabriel Peyré, and François-Xavier Vialard. Unbalanced optimal transport, from theory to numerics. *Handbook of Numerical Analysis*, 24:407–471, 2023.

Sebastian U Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232*, 2019.

Tianxiao Sun and Quoc Tran-Dinh. Generalized self-concordant functions: a recipe for newton-type methods. *Mathematical Programming*, 178(1):145–213, 2019.

Adrien Vacher and François-Xavier Vialard. Semi-dual unbalanced quadratic optimal transport: fast statistical rates and convergent algorithm. In *International Conference on Machine Learning*, pages 34734–34758. PMLR, 2023.

Daniil Vankov, Anton Rodomanov, Angelia Nedich, Lalitha Sankar, and Sebastian U Stich. Optimizing (l_0, l_1)-smooth functions by gradient methods. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.

Table of Contents

Appendix A. Nomenclature

Table 2 summarizes the specific mathematical notations and operators used throughout the paper.

Table 2: Nomenclature and List of Symbols

Symbol	Description
	MEASURE THEORY & OPERATORS
$\mathcal{M}_+(\mathcal{X})$	Space of finite non-negative measures on $\mathcal{X}$.
$\alpha \ll \beta$	Absolute continuity: measure α is dominated by β.
δ_x	Dirac mass at location x.
$f \lesssim f$	Inequality up to a positive constant: $f(\cdot) \leq C \cdot g(\cdot)$ for some universal $C > 0$.
$\langle \cdot, \cdot \rangle$	Standard Euclidean inner product.
$\odot$	Hadamard (element-wise) product.
$\Pi_{\mathcal{K}}$	Euclidean projection onto the set $\mathcal{K}$.
	UNBALANCED OT & DIVERGENCES
$D_\varphi(\cdot \mid \cdot)$	Csiszár φ-divergence defined by convex generator φ.
φ^c	Convex conjugate (Legendre-Fenchel transform) of φ.
ρ_1, ρ_2	Marginal penalty weights for source and target, respectively.
α	Scaling exponent for the KL-source case: $\alpha = \frac{\varepsilon}{\varepsilon+\rho_1}$.
$W(\cdot)$	Lambert W function (inverse of $z \mapsto ze^z$).
	SEMI-DUAL GEOMETRY
$B_j(x, \mathbf{g})$	Gibbs kernel term: $\beta_j \exp((g_j - c(x, y_j))/\varepsilon)$.
$Z(x, \mathbf{g})$	Partition function: $\sum_{j=1}^n B_j(x, \mathbf{g})$.
$w_k(x, \mathbf{g})$	Normalized transport weights: $B_k(x, \mathbf{g})/Z(x, \mathbf{g})$.
$U(x, \mathbf{g})$	Scaled partition function (χ^2-source): $\frac{\rho_1}{\varepsilon} e^{\rho_1/\varepsilon} Z(x, \mathbf{g})$.
$\mathcal{J}_{\text{trans}}(\mathbf{g})$	Transport component of the semi-dual functional.
	OPTIMIZATION & ALGORITHMS
$\mathcal{K}_\delta$	Effective domain (feasible set): $\{\mathbf{g} \in \mathbb{R}^n \mid g_j \leq \rho_2 + \delta\}$.
$L(\mathbf{g})$	Local smoothness upper bound at $\mathbf{g}$.
$\widehat{\nabla}\mathcal{J}$	Stochastic gradient estimator (batch size m_b).

Appendix B. Derivation of the semi-dual functionals.

Summary of the semi-dual objectives, gradients, and recovery of the dual potentials. Assume $\nu = \sum_{j=1}^{n} \beta_j \delta_{y_j}$ with $\beta_j > 0$ and fix the target marginal penalty $D_2 = D_{\chi^2}$ with weight $\rho_2 > 0$. Define, for $x \in \mathcal{X}$ and $g \in \mathcal{R}^n$,

$$B_j(x,g) := \beta_j \exp\Big(\tfrac{g_j - c(x,y_j)}{\varepsilon}\Big), \qquad Z(x,g) := \sum_{j=1}^{n} B_j(x,g), \qquad w_j(x,g) := \frac{B_j(x,g)}{Z(x,g)}.$$

Then any dual maximizer satisfies $g_j^\star \le \rho_2$ for all j, and the dual problem reduces to the minimization of a strictly convex semi-dual functional $\mathcal{J}(g)$ of the form

$$\mathcal{J}(g) = \mathcal{J}_{\mathrm{trans}}(g) \; + \; \sum_{j=1}^{n} \beta_j \Big(\tfrac{g_j^2}{2\rho_2} - g_j\Big),$$

where the transport term $\mathcal{J}_{\mathrm{trans}}$ depends on the choice of the source divergence D_1:
KL-source ($D_1 = \mathrm{KL}$). Let $\alpha := \frac{\varepsilon}{\rho_1 + \varepsilon}$. Then

$$\mathcal{J}_{\mathrm{trans}}(g) = (\rho_1 + \varepsilon) \int_{\mathcal{X}} Z(x,g)^\alpha \, \mathrm{d}\mu(x), \tag{13}$$

and the gradient is

$$\nabla_k \mathcal{J}(g) = \int_{\mathcal{X}} Z(x,g)^\alpha \, w_k(x,g) \, \mathrm{d}\mu(x) \; + \; \beta_k \Big(\tfrac{g_k}{\rho_2} - 1\Big). \tag{14}$$

Moreover, for any fixed g, the maximizer in the eliminated variable f is unique and given in closed form by

$$f^\star(x;g) = -\frac{\rho_1 \varepsilon}{\rho_1 + \varepsilon} \ln Z(x,g), \qquad x \in \mathcal{X}. \tag{15}$$

χ^2-source ($D_1 = D_{\chi^2}$). Define

$$U(x,g) := \frac{\rho_1}{\varepsilon} e^{\rho_1/\varepsilon} Z(x,g), \qquad W(\cdot) \text{ the Lambert function.}$$

Then

$$\mathcal{J}_{\mathrm{trans}}(g) = \frac{\varepsilon^2}{\rho_1} \int_{\mathcal{X}} \Big(W(U(x,g)) + \tfrac{1}{2} W(U(x,g))^2 \Big) \, \mathrm{d}\mu(x), \tag{16}$$

and the gradient is

$$\nabla_k \mathcal{J}(g) = \int_{\mathcal{X}} \frac{\varepsilon}{\rho_1} W(U(x,g)) \, w_k(x,g) \, \mathrm{d}\mu(x) \; + \; \beta_k \Big(\tfrac{g_k}{\rho_2} - 1\Big). \tag{17}$$

Moreover, for any fixed g, the maximizer in the eliminated variable f is unique and given by

$$f^\star(x;g) = \rho_1 - \varepsilon \, W(U(x,g)), \qquad x \in \mathcal{X}. \tag{18}$$

Finally, if $g^\star$ minimizes $\mathcal{J}$, then $(f^\star(\cdot;g^\star), g^\star)$ is a maximizer of the original dual (19). The associated optimal coupling is recovered from the dual potentials by

$$\frac{\mathrm{d}\pi^\star}{\mathrm{d}(\mu \otimes \nu)}(x, y_j) = \exp\left(\frac{f^\star(x;g^\star) + g_j^\star - c(x,y_j)}{\varepsilon} \right).$$

Proof We consider entropic unbalanced OT with discrete target $\nu = \sum_{j=1}^{n} \beta_j \delta_{y_j}$, $\beta_j > 0$. Starting from the dual (up to additive constants independent of $(f, \mathbf{g})$),

$$\sup_{f,\mathbf{g}} \left\{ -\varepsilon \int_{\mathcal{X}} \sum_{j=1}^{n} \beta_j \exp\Big(\tfrac{f(x)+g_j-c(x,y_j)}{\varepsilon}\Big) \mathrm{d}\mu(x) - \rho_1 \int_{\mathcal{X}} \varphi_1^c\Big(-\tfrac{f(x)}{\rho_1}\Big) \mathrm{d}\mu(x) - \rho_2 \sum_{j=1}^{n} \beta_j \, \varphi_{\chi^2}^c\Big(-\tfrac{g_j}{\rho_2}\Big) \right\}. \tag{19}$$

Pearson χ^2 generator and conjugate. The Pearson χ^2 divergence between measures $\pi \ll \nu$ is generated by

$$\varphi_{\chi^2}(t) = \tfrac{1}{2}(t-1)^2, \qquad t \geq 0,$$

so that $D_{\chi^2}(\pi \| \nu) = \int \varphi_{\chi^2}\big(\frac{d\pi}{d\nu}\big) \mathrm{d}\nu$. Its convex conjugate over $t \geq 0$ is

$$\varphi_{\chi^2}^c(s) = \sup_{t \geq 0} \{st - \tfrac{1}{2}(t-1)^2\} = \begin{cases} s + \frac{1}{2}s^2, & s \geq -1, \\ -\frac{1}{2}, & s < -1. \end{cases} \tag{20}$$

Softmax quantities. For any $x \in \mathcal{X}$ and $\mathbf{g} \in \mathbb{R}^n$, define

$$B_j(x, \mathbf{g}) := \beta_j \exp\Big(\tfrac{g_j - c(x,y_j)}{\varepsilon}\Big), \qquad Z(x, \mathbf{g}) := \sum_{j=1}^{n} B_j(x, \mathbf{g}), \qquad w_j(x, \mathbf{g}) := \frac{B_j(x, \mathbf{g})}{Z(x, \mathbf{g})}. \tag{21}$$

Then

$$\sum_{j=1}^{n} \beta_j \exp\Big(\tfrac{f(x)+g_j-c(x,y_j)}{\varepsilon}\Big) = e^{f(x)/\varepsilon} Z(x, \mathbf{g}).$$

Let $s_j = -g_j/\rho_2$. Using (20),

$$-\rho_2 \, \varphi_{\chi^2}^c(s_j) = \begin{cases} g_j - \frac{g_j^2}{2\rho_2}, & g_j \leq \rho_2 \ (s_j \geq -1), \\ \frac{\rho_2}{2}, & g_j > \rho_2 \ (s_j < -1). \end{cases} \tag{22}$$

In our dual objective, the only other dependence on g_j is through the entropic term $-\varepsilon \int e^{f/\varepsilon} B_j(x, \mathbf{g}) \mathrm{d}\mu(x)$, which is strictly decreasing in g_j. Therefore, increasing g_j above ρ_2 can only decrease the objective, hence at any maximizer one has

$$g_j^\star \leq \rho_2 \quad \text{for all } j. \tag{23}$$

As a consequence, we may use the quadratic branch in (22) and write

$$-\rho_2 \sum_{j=1}^{n} \beta_j \, \varphi_{\chi^2}^c\Big(-\tfrac{g_j}{\rho_2}\Big) = \sum_{j=1}^{n} \beta_j \Big(g_j - \tfrac{g_j^2}{2\rho_2}\Big). \tag{24}$$

Plugging (21) and (24) into (19) yields

$$\sup_{f,\mathbf{g}} \left\{ \int_{\mathcal{X}} \Big[-\varepsilon \, e^{f(x)/\varepsilon} Z(x, \mathbf{g}) - \rho_1 \, \varphi_1^c\Big(-\tfrac{f(x)}{\rho_1}\Big) \Big] \mathrm{d}\mu(x) + \sum_{j=1}^{n} \beta_j \Big(g_j - \tfrac{g_j^2}{2\rho_2}\Big) \right\}. \tag{25}$$

For fixed $\mathbf{g}$, the maximization over f is separable in x.

Case A: KL-source ($\varphi_1 = \mathrm{KL}$). Here $\varphi^c_{\mathrm{KL}}(s) = e^s - 1$, hence

$$-\rho_1 \, \varphi^c_{\mathrm{KL}}\Big(-\tfrac{f}{\rho_1}\Big) = -\rho_1\big(e^{-f/\rho_1} - 1\big) = -\rho_1 e^{-f/\rho_1} + \rho_1.$$

Fix x and abbreviate $Z = Z(x, \mathbf{g})$. We maximize over $f \in \mathbb{R}$:

$$\Phi_x(f) := -\varepsilon Z e^{f/\varepsilon} - \rho_1 e^{-f/\rho_1} + \rho_1. \tag{26}$$

The first-order condition gives

$$0 = \Phi'_x(f) = -Z e^{f/\varepsilon} + e^{-f/\rho_1} \quad \Longleftrightarrow \quad e^{-f/\rho_1} = Z e^{f/\varepsilon}.$$

Let $\alpha := \frac{\varepsilon}{\rho_1 + \varepsilon}$. The unique maximizer is

$$f^\star(x; \mathbf{g}) = -\frac{\rho_1 \varepsilon}{\rho_1 + \varepsilon} \ln Z(x, \mathbf{g}) \quad \text{and} \quad e^{-f^\star/\rho_1} = Z^\alpha, \;\; e^{f^\star/\varepsilon} = Z^{\alpha - 1}. \tag{27}$$

Plugging into (26) yields the exact pointwise optimum value

$$\sup_f \Phi_x(f) = \rho_1 - (\rho_1 + \varepsilon)\, Z(x, \mathbf{g})^\alpha. \tag{28}$$

Therefore the dual reduces to

$$\sup_{\mathbf{g}} \left\{ \rho_1 \, \mu(\mathcal{X}) - (\rho_1 + \varepsilon) \int_{\mathcal{X}} Z(x, \mathbf{g})^\alpha \mathrm{d}\mu(x) + \sum_{j=1}^n \beta_j \Big(g_j - \tfrac{g_j^2}{2\rho_2} \Big) \right\}. \tag{29}$$

Equivalently, maximizing (29) is the same as minimizing the strictly convex objective

$$\mathcal{J}_{\mathrm{KL}}(\mathbf{g}) := (\rho_1 + \varepsilon) \int_{\mathcal{X}} Z(x, \mathbf{g})^\alpha \mathrm{d}\mu(x) + \sum_{j=1}^n \beta_j \Big(\tfrac{g_j^2}{2\rho_2} - g_j \Big), \tag{30}$$

and the optimal dual value equals $\rho_1 \mu(\mathcal{X}) - \inf_{\mathbf{g}} \mathcal{J}_{\mathrm{KL}}(\mathbf{g})$.

Case B: χ^2-source ($\varphi_1 = \chi^2$). Using (20) with $s = -f/\rho_1$, the quadratic branch (valid when $f \le \rho_1$) gives

$$-\rho_1 \, \varphi^c_{\chi^2}\Big(-\tfrac{f}{\rho_1}\Big) = -\rho_1 \Big(-\tfrac{f}{\rho_1} + \tfrac{1}{2}\tfrac{f^2}{\rho_1^2} \Big) = f - \tfrac{f^2}{2\rho_1}.$$

If $f > \rho_1$, the penalty becomes constant $-\rho_1(-\frac{1}{2}) = \rho_1/2$ while the entropic term $-\varepsilon Z e^{f/\varepsilon}$ strictly decreases with f, so at any maximizer we have $f^\star(x) \le \rho_1$. Fix x and write $Z = Z(x, \mathbf{g})$. We maximize over $f \le \rho_1$:

$$\Phi_x(f) := -\varepsilon Z e^{f/\varepsilon} + f - \tfrac{f^2}{2\rho_1}. \tag{31}$$

The first-order condition is

$$0 = \Phi'_x(f) = -Z e^{f/\varepsilon} + 1 - \tfrac{f}{\rho_1} \quad \Longleftrightarrow \quad 1 - \tfrac{f}{\rho_1} = Z e^{f/\varepsilon}. \tag{32}$$

Let $u := 1 - \frac{f}{\rho_1} \ge 0$. Then (32) becomes

$$u = Z \exp\Big(\tfrac{\rho_1}{\varepsilon}(1 - u) \Big) = Z e^{\rho_1/\varepsilon} \, e^{-(\rho_1/\varepsilon)u}.$$

Equivalently,

$$\left(\tfrac{\rho_1}{\varepsilon}u\right)\exp\left(\tfrac{\rho_1}{\varepsilon}u\right) = \frac{\rho_1}{\varepsilon}Ze^{\rho_1/\varepsilon}.$$

Define

$$U(x,\mathbf{g}) := \frac{\rho_1}{\varepsilon}Z(x,\mathbf{g})e^{\rho_1/\varepsilon}, \qquad W(\cdot) \text{ the Lambert function.} \tag{33}$$

Then the unique maximizer is

$$u(x,\mathbf{g}) = \frac{\varepsilon}{\rho_1}W\big(U(x,\mathbf{g})\big), \qquad f^\star(x;\mathbf{g}) = \rho_1 - \varepsilon\, W\big(U(x,\mathbf{g})\big) \le \rho_1. \tag{34}$$

To evaluate the optimum value, note from (32) that $Ze^{f^\star/\varepsilon} = u$. Plugging $f^\star = \rho_1(1-u)$ into (31) gives

$$\sup_f \Phi_x(f) = -\varepsilon u + \rho_1(1-u) - \tfrac{\rho_1}{2}(1-u)^2 = \tfrac{\rho_1}{2} - \varepsilon u - \tfrac{\rho_1}{2}u^2.$$

Using (34) with $u = \frac{\varepsilon}{\rho_1}w$ and $w := W(U)$, we obtain the exact value

$$\sup_f \Phi_x(f) = \frac{\rho_1}{2} - \frac{\varepsilon^2}{\rho_1}\Big(w + \tfrac{1}{2}w^2\Big), \qquad w = W\big(U(x,\mathbf{g})\big). \tag{35}$$

Therefore the dual reduces to

$$\sup_{\mathbf{g}}\left\{\frac{\rho_1}{2}\,\mu(\mathcal{X}) - \frac{\varepsilon^2}{\rho_1}\int_{\mathcal{X}}\Big(W(U) + \tfrac{1}{2}W(U)^2\Big)\mathrm{d}\mu + \sum_{j=1}^n \beta_j\Big(g_j - \tfrac{g_j^2}{2\rho_2}\Big)\right\}. \tag{36}$$

Equivalently, maximizing (36) is the same as minimizing the strictly convex objective

$$\mathcal{J}_{\chi^2}(\mathbf{g}) := \frac{\varepsilon^2}{\rho_1}\int_{\mathcal{X}}\Big(W(U(x,\mathbf{g})) + \tfrac{1}{2}W(U(x,\mathbf{g}))^2\Big)\mathrm{d}\mu(x) + \sum_{j=1}^n \beta_j\Big(\tfrac{g_j^2}{2\rho_2} - g_j\Big), \tag{37}$$

and the optimal dual value equals $\frac{\rho_1}{2}\mu(\mathcal{X}) - \inf_{\mathbf{g}}\mathcal{J}_{\chi^2}(\mathbf{g})$. ■

Gradients of the semi-dual objectives

We now differentiate $\mathcal{J}(\mathbf{g})$ in the two cases. First note that, from (21),

$$\frac{\partial Z(x,\mathbf{g})}{\partial g_k} = \frac{1}{\varepsilon}B_k(x,\mathbf{g}) = \frac{1}{\varepsilon}Z(x,\mathbf{g})\,w_k(x,\mathbf{g}). \tag{38}$$

Also, the target quadratic term always contributes

$$\frac{\partial}{\partial g_k}\sum_{j=1}^n \beta_j\Big(\tfrac{g_j^2}{2\rho_2} - g_j\Big) = \beta_k\Big(\tfrac{g_k}{\rho_2} - 1\Big). \tag{39}$$

Gradient, KL-source. From (30), using (38),

$$\frac{\partial}{\partial g_k}Z(x,\mathbf{g})^\alpha = \alpha Z^{\alpha-1}\frac{\partial Z}{\partial g_k} = \frac{\alpha}{\varepsilon}Z^\alpha w_k = \frac{1}{\rho_1+\varepsilon}Z^\alpha w_k,$$

hence

$$\nabla_k\mathcal{J}_{\mathrm{KL}}(\mathbf{g}) = \int_{\mathcal{X}} Z(x,\mathbf{g})^\alpha\, w_k(x,\mathbf{g})\mathrm{d}\mu(x) + \beta_k\Big(\tfrac{g_k}{\rho_2} - 1\Big). \tag{40}$$

Gradient, χ^2-source. Let $w(x,\mathbf{g}) := W(U(x,\mathbf{g}))$ with U defined in (33). Since $U(x,\mathbf{g}) = \frac{\rho_1}{\varepsilon} e^{\rho_1/\varepsilon} Z(x,\mathbf{g})$, we have

$$\frac{\partial w}{\partial g_k} = W'(U)\frac{\partial U}{\partial g_k} = \frac{w}{U(1+w)} \cdot \frac{U}{Z} \cdot \frac{\partial Z}{\partial g_k} = \frac{w}{Z(1+w)} \cdot \frac{1}{\varepsilon} B_k.$$

Differentiate the integrand in (37):

$$\frac{\partial}{\partial g_k}\Big(w + \tfrac{1}{2}w^2\Big) = (1+w)\frac{\partial w}{\partial g_k} = \frac{w}{Z} \cdot \frac{1}{\varepsilon} B_k = \frac{w}{\varepsilon}\, w_k.$$

Therefore

$$\nabla_k \mathcal{J}_{\chi^2}(\mathbf{g}) = \int_{\mathcal{X}} \frac{\varepsilon}{\rho_1} W\big(U(x,\mathbf{g})\big)\, w_k(x,\mathbf{g}) \mathrm{d}\mu(x) + \beta_k\Big(\tfrac{g_k}{\rho_2} - 1\Big). \tag{41}$$

Summary in density form. For KL-source, define $\alpha = \frac{\varepsilon}{\rho_1+\varepsilon}$ and

$$\sigma_{\mathrm{KL}}(x,\mathbf{g}) := Z(x,\mathbf{g})^{\alpha}.$$

For χ^2-source, define U as in (33) and

$$\sigma_{\chi^2}(x,\mathbf{g}) := \frac{\varepsilon}{\rho_1} W(U(x,\mathbf{g})).$$

Then (40) and (41) can be written uniformly as

$$\nabla_k \mathcal{J}(\mathbf{g}) = \int_{\mathcal{X}} \sigma(x,\mathbf{g})\, w_k(x,\mathbf{g}) \mathrm{d}\mu(x) + \frac{\beta_k}{\rho_2} g_k - \beta_k\,.$$

Appendix C. Proof of Theorem 3: Smoothness Bound via Gradient Transport

Theorem 3 Smoothness Bound via Gradient Transport. For all $\mathbf{g} \in \mathbb{R}^n$, the operator norm of the Hessian satisfies:

$$\|\nabla^2 \mathcal{J}(\mathbf{g})\|_{\mathrm{op}} \le \frac{1}{\varepsilon} \|\nabla \mathcal{J}_{\mathrm{trans}}(\mathbf{g})\|_{\infty} + \frac{\beta_{\max}}{\rho_2}. \tag{42}$$

Proof We derive the Hessian in both cases. A common ingredient is the derivative of the softmax weights $w_k(x,\mathbf{g}) = B_k(x,\mathbf{g})/Z(x,\mathbf{g})$. Using $\partial_{g_l} B_k = \frac{1}{\varepsilon}\delta_{kl} B_k$ and the quotient rule,

$$\frac{\partial w_k}{\partial g_l} = \frac{1}{\varepsilon} w_k(\delta_{kl} - w_l). \tag{43}$$

Hessian Derivation: KL-Source Case. In the KL-source case, the transport density is $\sigma(x,\mathbf{g}) = Z(x,\mathbf{g})^{\alpha}$ with $\alpha = \frac{\varepsilon}{\rho_1+\varepsilon}$. First,

$$\frac{\partial \sigma}{\partial g_l} = \alpha Z^{\alpha-1}\frac{\partial Z}{\partial g_l} = \alpha Z^{\alpha-1} \cdot \frac{1}{\varepsilon} Z w_l = \frac{\alpha}{\varepsilon}\sigma w_l.$$

Then, applying the product rule and (43),

$$\begin{aligned}
\frac{\partial}{\partial g_l}(\sigma w_k) &= \Big(\frac{\partial \sigma}{\partial g_l}\Big) w_k + \sigma\Big(\frac{\partial w_k}{\partial g_l}\Big) \\
&= \frac{\alpha}{\varepsilon}\sigma w_l w_k + \frac{\sigma}{\varepsilon} w_k(\delta_{kl} - w_l) \\
&= \frac{\sigma}{\varepsilon}\Big[w_k\delta_{kl} - (1-\alpha) w_k w_l\Big],
\end{aligned}$$

where $1-\alpha > 0$.

Hessian Derivation: χ^2-Source Case. For the semi-dual with χ^2-source obtained by eliminating f, the gradient of the transport term takes the form

$$\nabla_k \mathcal{J}_{\text{trans}}(\mathbf{g}) = \int_{\mathcal{X}} \sigma(x,\mathbf{g})\, w_k(x,\mathbf{g}) \mathrm{d}\mu(x), \qquad \sigma(x,\mathbf{g}) = \frac{\varepsilon}{\rho_1} W(U(x,\mathbf{g})),$$

with

$$U(x,\mathbf{g}) = \frac{\rho_1}{\varepsilon} Z(x,\mathbf{g}) e^{\rho_1/\varepsilon}.$$

We compute $\partial_{g_l}\sigma$. Using $W'(z) = \frac{W(z)}{z(1+W(z))}$ and $\partial_{g_l} U = \frac{1}{\varepsilon} U w_l$ (since $U \propto Z$ and $\partial_{g_l} Z = \frac{1}{\varepsilon} Z w_l$),

$$\begin{aligned}\frac{\partial \sigma}{\partial g_l} &= \frac{\varepsilon}{\rho_1} W'(U) \frac{\partial U}{\partial g_l} = \frac{\varepsilon}{\rho_1} \cdot \frac{W(U)}{U(1+W(U))} \cdot \frac{1}{\varepsilon} U w_l \\ &= \frac{1}{\rho_1} \frac{W(U)}{1+W(U)}\, w_l = \frac{\sigma}{\varepsilon(1+W(U))}\, w_l,\end{aligned}$$

where in the last equality we used $\sigma = \frac{\varepsilon}{\rho_1} W(U)$.

Applying the product rule and (43),

$$\begin{aligned}\frac{\partial}{\partial g_l}(\sigma w_k) &= \Big(\frac{\partial \sigma}{\partial g_l}\Big) w_k + \sigma \Big(\frac{\partial w_k}{\partial g_l}\Big) \\ &= \frac{\sigma}{\varepsilon(1+W)} w_l w_k + \frac{\sigma}{\varepsilon} w_k(\delta_{kl} - w_l) \\ &= \frac{\sigma}{\varepsilon}\Big[w_k \delta_{kl} - \Big(1 - \frac{1}{1+W(U)}\Big) w_k w_l\Big] \\ &= \frac{\sigma}{\varepsilon}\Big[w_k \delta_{kl} - c(x)\, w_k w_l\Big],\end{aligned}$$

with the nonnegative coefficient

$$c(x) := \frac{W(U(x,\mathbf{g}))}{1+W(U(x,\mathbf{g}))} \in [0,1).$$

Unified Spectral Bound. In both cases, the Hessian of the full objective $\mathcal{J}(\mathbf{g})$ (transport term plus the target quadratic penalty $\sum_j \beta_j(\frac{g_j^2}{2\rho_2} - g_j)$) can be written as

$$\nabla^2 \mathcal{J}(\mathbf{g}) = \int_{\mathcal{X}} \frac{\sigma(x,\mathbf{g})}{\varepsilon} \Big(\mathrm{diag}(\mathbf{w}) - c(x)\, \mathbf{w}\mathbf{w}^\top\Big) \mathrm{d}\mu(x) + \frac{1}{\rho_2} \mathrm{diag}(\boldsymbol{\beta}),$$

where $c(x) = 1 - \alpha$ in the KL-source case and $c(x) = \frac{W(U)}{1+W(U)}$ in the χ^2-source case, hence always $c(x) \geq 0$.

Since $\mathbf{w}\mathbf{w}^\top \succeq 0$, subtracting $c(x)\mathbf{w}\mathbf{w}^\top$ decreases eigenvalues:

$$\mathrm{diag}(\mathbf{w}) - c(x)\mathbf{w}\mathbf{w}^\top \preceq \mathrm{diag}(\mathbf{w}).$$

Therefore

$$\begin{aligned}\nabla^2 \mathcal{J}(\mathbf{g}) &\preceq \int_{\mathcal{X}} \frac{\sigma(x,\mathbf{g})}{\varepsilon} \mathrm{diag}(\mathbf{w}) \mathrm{d}\mu(x) + \frac{1}{\rho_2} \mathrm{diag}(\boldsymbol{\beta}) \\ &= \mathrm{diag}\left(\frac{1}{\varepsilon} \int_{\mathcal{X}} \sigma(x,\mathbf{g})\, \mathbf{w}(x,\mathbf{g}) \mathrm{d}\mu(x) + \frac{\boldsymbol{\beta}}{\rho_2}\right).\end{aligned}$$

Recognizing $\int \sigma w_k \mathrm{d}\mu = [\nabla \mathcal{J}_{\mathrm{trans}}(\mathbf{g})]_k$, we obtain

$$\|\nabla^2 \mathcal{J}(\mathbf{g})\|_{\mathrm{op}} \leq \max_k \left(\frac{1}{\varepsilon} [\nabla \mathcal{J}_{\mathrm{trans}}(\mathbf{g})]_k + \frac{\beta_k}{\rho_2} \right) \leq \frac{1}{\varepsilon} \|\nabla \mathcal{J}_{\mathrm{trans}}(\mathbf{g})\|_\infty + \frac{\beta_{\max}}{\rho_2},$$

which concludes the proof. ■

Appendix D. Proof of Lemma 5: Uniform Gradient Bound and Smoothness

Lemma 5: Uniform Gradient Bound and Smoothness. On $\mathcal{K}$, the L_1-norm of the transport gradient is uniformly bounded: $\|\nabla \mathcal{J}_{\mathrm{trans}}(\mathbf{g})\|_1 \leq C_{\mathrm{bound}}$, where:

$$C^{\mathrm{KL}}_{\mathrm{bound}} := \mu(\mathcal{X}) \|\nu\|_1^\alpha \exp\left(\frac{\rho_2 + \delta}{\rho_1 + \varepsilon} \right) \tag{44}$$

$$C^{\chi^2}_{\mathrm{bound}} := \mu(\mathcal{X}) \frac{\varepsilon}{\rho_1} W\left[\frac{\rho_1}{\varepsilon} e^{\rho_1/\varepsilon} \|\nu\|_1 \exp\Big(\frac{\rho_2 + \delta}{\varepsilon} \Big) \right], \tag{45}$$

Consequently, the Hessian is bounded on $\mathcal{K}$, and $\mathcal{J}$ is L-smooth with $L = \mathcal{O}(1/\varepsilon)$.

Proof In both cases of source divergence, our gradient writes

$$[\nabla \mathcal{J}_{\mathrm{trans}}(\mathbf{g})]_k = \int_{\mathcal{X}} \sigma(x, \mathbf{g})\, w_k(x, \mathbf{g})\, d\mu(x)\ ,$$

with $\sigma_{\mathrm{KL}}(x, \mathbf{g}) := Z(x, \mathbf{g})^\alpha$ or $\sigma_{\chi^2}(x, \mathbf{g}) := \frac{\varepsilon}{\rho_1} W(U(x, \mathbf{g}))$.

Crucially, we have $\sum_{k=1}^n w_k(x, \mathbf{g}) = 1$ for any x. Because of this property, determining the L_1 norm, which is simply the sum of these non-negative components, simplifies easily:

$$\|\nabla \mathcal{J}_{\mathrm{trans}}(\mathbf{g})\|_1 = \sum_{k=1}^n \int_{\mathcal{X}} \sigma(x, \mathbf{g})\, w_k(x, \mathbf{g})\, d\mu(x) = \int_{\mathcal{X}} \sigma(x, \mathbf{g}) \underbrace{\left(\sum_{k=1}^n w_k(x, \mathbf{g}) \right)}_{=1} d\mu(x).$$

Thus, the problem reduces to finding a uniform bound for the integral of the scalar density $\sigma(x, \mathbf{g})$.

Bounding Z. The behavior of $\sigma(x, \mathbf{g})$ in both cases is driven by the potential function $Z(x, \mathbf{g}) = \sum_j B_j(x, \mathbf{g})$. We recall that:

$$B_j(x, \mathbf{g}) = \beta_j \exp\left(\frac{g_j - c(x, y_j)}{\varepsilon} \right).$$

We can bound this term uniformly by utilizing the problem constraints. Since the cost is non-negative ($c \geq 0$) and the algorithm enforces $g_j \leq \rho_2 + \delta$, we have:

$$Z(x, \mathbf{g}) \leq \left(\sum_{j=1}^n \beta_j \right) \exp\left(\frac{\rho_2}{\varepsilon} \right) = \|\nu\|_1 \exp\left(\frac{\rho_2 + \delta}{\varepsilon} \right).$$

Let's denote this upper bound constant as $Z_{\max}$.

With Z bounded, we can now bound the total mass $\int \sigma\, d\mu$ for each geometry to conclude:

Case KL: Here, the density is defined as $\sigma(x, \mathbf{g}) = Z(x, \mathbf{g})^\alpha$. Since $Z(x, \mathbf{g}) \leq Z_{\max}$, the gradient norm is directly bounded by:

$$\|\nabla \mathcal{J}_{\text{trans}}\|_1 \leq \int_{\mathcal{X}} Z_{\max}^\alpha \, d\mu(x) = \mu(\mathcal{X}) \|\nu\|_1^\alpha \exp\left(\frac{\alpha(\rho_2 + \delta)}{\varepsilon}\right).$$

Substituting $\alpha = \frac{\varepsilon}{\rho_1 + \varepsilon}$ yields the final bound $\mu(\mathcal{X}) \|\nu\|_1^\alpha \exp\left(\frac{\rho_2 + \delta}{\rho_1 + \varepsilon}\right)$.

Case χ^2: For the χ^2-source semi-dual obtained by eliminating f, the transport-gradient density is

$$\sigma(x, \mathbf{g}) = \frac{\varepsilon}{\rho_1} W(U(x, \mathbf{g})), \qquad U(x, \mathbf{g}) = \frac{\rho_1}{\varepsilon} Z(x, \mathbf{g}) e^{\rho_1/\varepsilon}.$$

Since W is increasing on $\mathbb{R}_+$ and $Z \leq Z_{\max}$, we have $U(x, \mathbf{g}) \leq U_{\max}$ with

$$U_{\max} := \frac{\rho_1}{\varepsilon} Z_{\max} e^{\rho_1/\varepsilon} = \frac{\rho_1}{\varepsilon} \|\nu\|_1 \exp\left(\frac{\rho_2 + \delta}{\varepsilon}\right).$$

Therefore,

$$\|\nabla \mathcal{J}_{\text{trans}}(\mathbf{g})\|_1 = \int_{\mathcal{X}} \frac{\varepsilon}{\rho_1} W(U(x, \mathbf{g})) \mathrm{d}\mu(x) \leq \mu(\mathcal{X}) \, \frac{\varepsilon}{\rho_1} W(U_{\max}),$$

which is a uniform bound under the constraint $g_j \leq \rho_2 + \delta$. Moreover, one can check numerically that this constant is close to 1. ■

D.1. Corollary - Proof of Lemma 8 : Bounded Variance

Proposition 8 :Variance Bound of Mini-Batch Gradient. Let $\widehat{\nabla} \mathcal{J}(\mathbf{g})$ be the mini-batch gradient estimator computed with batch size $b \geq 1$, as defined in Eq. (9). For any $\mathbf{g} \in \mathcal{K}$, using the uniform bound C_{bound} from Lemma 5, the variance is bounded by:

$$\mathbb{E}\left[\|\widehat{\nabla} \mathcal{J}(\mathbf{g}) - \nabla \mathcal{J}(\mathbf{g})\|_2^2\right] \leq \frac{4 C_{\text{bound}}^2}{b}. \tag{46}$$

Proof From Lemma 5, we have the uniform L_1-norm bound on the estimation error for any sample realization:

$$\|\widehat{\nabla} \mathcal{J}(\mathbf{g}) - \nabla \mathcal{J}(\mathbf{g})\|_1 \leq 2 C_{\text{bound}}.$$

Using the norm inequality $\|\cdot\|_2 \leq \|\cdot\|_1$, the squared Euclidean error for a single sample ($b = 1$) is bounded almost surely by $(2C_{\text{bound}})^2$. Since $\widehat{\nabla} \mathcal{J}(\mathbf{g})$ is the average of b i.i.d. estimators, the variance of the mean scales by $1/b$:

$$\mathbb{E}\left[\|\widehat{\nabla} \mathcal{J}(\mathbf{g}) - \nabla \mathcal{J}(\mathbf{g})\|_2^2\right] = \frac{1}{b} \mathrm{Var}(\widehat{\nabla}_{\text{single}}) \leq \frac{(2C_{\text{bound}})^2}{b}.$$

■

Appendix E. Proof of Proposition 7: Generalized self-concordance of the semi-dual

Proposition 7 : Generalized self-concordance. The semi-dual $\mathcal{J}$ is generalized self-concordant. That is, for $M = \frac{2+3\alpha}{\varepsilon}$ for KL source, and $M = \frac{6}{\varepsilon}$ for χ^2 , we have for any $\mathbf{g} \in \mathbb{R}^n$ and any direction $\mathbf{h} \in \mathbb{R}^n$:

$$\left|\nabla^3 \mathcal{J}(\mathbf{g})[\mathbf{h}, \mathbf{h}, \mathbf{h}]\right| \leq M \|\mathbf{h}\|_\infty \langle \mathbf{h}, \nabla^2 \mathcal{J}(\mathbf{g}) \mathbf{h} \rangle.$$

E.1. Case 1: $D_1 = \mathrm{KL},\ D_2 = \chi^2$

Proof For clarity, we recall the notations

$$B_j(\mathbf{g}) := \beta_j \exp\left(\frac{g_j - c(x, y_j)}{\varepsilon}\right), \qquad Z(\mathbf{g}) := \sum_{j=1}^{n} B_j(\mathbf{g}), \qquad \tau(\mathbf{g}) := Z(\mathbf{g})^\alpha,$$

with $\beta_j > 0$. The softmax weights are

$$w_j(\mathbf{g}) := \frac{B_j(\mathbf{g})}{Z(\mathbf{g})} \in \Delta^n.$$

We now introduce some notions regarding directional derivatives:
For a direction $\mathbf{h} \in \mathbb{R}^n$, denote directional derivatives by $\partial_{\mathbf{h}}$ and define

$$w_{\mathbf{h}} := \langle w, \mathbf{h} \rangle, \qquad w_{\mathbf{h}^2} := \langle w, \mathbf{h}^2 \rangle, \qquad w_{\mathbf{h}^3} := \langle w, \mathbf{h}^3 \rangle,$$

where $\mathbf{h}^2 = (h_1^2, \ldots, h_n^2)$ and $\mathbf{h}^3 = (h_1^3, \ldots, h_n^3)$. Let $L := \|\mathbf{h}\|_\infty$.

We start by giving the directional derivatives of τ.

Derivatives of Z: Consider $\mathbf{g}(t) = \mathbf{g} + t\mathbf{h}$. Then

$$Z(t) = \sum_{j=1}^{n} B_j(\mathbf{g}) \exp\left(\frac{t h_j}{\varepsilon}\right).$$

Differentiating at $t = 0$ yields

$$\partial_{\mathbf{h}} Z = Z'(0) = \frac{1}{\varepsilon} \sum_j B_j(\mathbf{g}) h_j = \frac{1}{\varepsilon} Z(\mathbf{g})\, w_{\mathbf{h}},$$

$$\partial_{\mathbf{h}}^2 Z = Z''(0) = \frac{1}{\varepsilon^2} \sum_j B_j(\mathbf{g}) h_j^2 = \frac{1}{\varepsilon^2} Z(\mathbf{g})\, w_{\mathbf{h}^2},$$

$$\partial_{\mathbf{h}}^3 Z = Z'''(0) = \frac{1}{\varepsilon^3} \sum_j B_j(\mathbf{g}) h_j^3 = \frac{1}{\varepsilon^3} Z(\mathbf{g})\, w_{\mathbf{h}^3}.$$

Derivatives of $\tau = Z^\alpha$: Using the chain rule for $\tau(t) = Z(t)^\alpha$,

$$\tau' = \alpha Z^{\alpha-1} Z', \qquad \tau'' = \alpha(\alpha-1) Z^{\alpha-2} (Z')^2 + \alpha Z^{\alpha-1} Z'',$$

$$\tau''' = \alpha(\alpha-1)(\alpha-2) Z^{\alpha-3} (Z')^3 + 3\alpha(\alpha-1) Z^{\alpha-2} Z' Z'' + \alpha Z^{\alpha-1} Z'''.$$

Substituting the expressions from the derivatives of Z, and writing $\tau = Z^\alpha$, we get:

$$\partial_{\mathbf{h}} \tau = \frac{\alpha}{\varepsilon} \tau\, w_{\mathbf{h}},$$

$$\partial_{\mathbf{h}}^2 \tau = \frac{\alpha}{\varepsilon^2} \tau \Big(w_{\mathbf{h}^2} - (1-\alpha) w_{\mathbf{h}}^2 \Big).$$

Define

$$D := w_{\mathbf{h}^2} - (1-\alpha) w_{\mathbf{h}}^2.$$

Then

$$\partial_{\mathbf{h}}^2 \tau = \frac{\alpha}{\varepsilon^2} \tau D.$$

Moreover, $D \geq 0$ since

$$D = (w_{\mathbf{h}^2} - w_{\mathbf{h}}^2) + \alpha w_{\mathbf{h}}^2 = \mathrm{Var}_w(H) + \alpha(\mathbb{E}_w H)^2 \geq 0,$$

where H is the random variable taking values h_j with probabilities w_j.

A direct substitution into the third-derivative formula also gives

$$\partial_{\mathbf{h}}^3 \tau = \frac{\alpha}{\varepsilon^3} \tau \Big(w_{\mathbf{h}^3} + 3(\alpha - 1) w_{\mathbf{h}} w_{\mathbf{h}^2} + (\alpha - 1)(\alpha - 2) w_{\mathbf{h}}^3 \Big).$$

Central-moment rewrite: As in Bercu and Bigot (2021), we will substitute moments into the expansion of the third-derivative.

Let H be the random variable with $\mathbb{P}(H = h_j) = w_j$. Define

$$m := \mathbb{E}[H] = w_{\mathbf{h}}, \qquad \sigma^2 := \mathrm{Var}(H) = w_{\mathbf{h}^2} - w_{\mathbf{h}}^2, \qquad \kappa_3 := \mathbb{E}[(H - m)^3].$$

Then $\mathbb{E}[H^2] = \sigma^2 + \mu^2$ and $\mathbb{E}[H^3] = \kappa_3 + 3\mu\sigma^2 + m^3$. With these identities, one checks that the bracket in $\partial_{\mathbf{h}}^3 \tau$ equals

$$N := \kappa_3 + 3\alpha m \sigma^2 + \alpha^2 m^3.$$

Hence

$$\partial_{\mathbf{h}}^3 \tau = \frac{\alpha}{\varepsilon^3} \tau \, N, \qquad D = \sigma^2 + \alpha m^2.$$

Bounding $|N|$ by D: Let $L_h = \|\mathbf{h}\|_\infty$. Since $|H| \leq L_h$ almost surely, we have $|m| \leq L_h$ and also $|H - m| \leq |H| + |m| \leq 2L_h$ almost surely. Therefore,

$$|\kappa_3| = \big|\mathbb{E}[(H - m)^3]\big| \leq \mathbb{E}[|H - m|^3] \leq 2L \, \mathbb{E}[(H - m)^2] = 2L_h \, \sigma^2.$$

Using $|m| \leq L_h$ and $\alpha \in (0, 1]$:

$$\begin{aligned} |N| &\leq |\kappa_3| + 3\alpha |m| \sigma^2 + \alpha^2 |m|^3 \\ &\leq 2L_h \sigma^2 + 3\alpha L_h \sigma^2 + \alpha^2 L_h \mu^2 \\ &\leq L_h \Big((2 + 3\alpha)\sigma^2 + \alpha m^2 \Big) \\ &\leq (2 + 3\alpha) L_h (\sigma^2 + \alpha m^2) \\ &= (2 + 3\alpha) \|\mathbf{h}\|_\infty D. \end{aligned}$$

Combining $\partial_{\mathbf{h}}^2 \tau = \frac{\alpha}{\varepsilon^2} \tau D$ and $\partial_{\mathbf{h}}^3 \tau = \frac{\alpha}{\varepsilon^3} \tau N$ with the bound $|N| \leq (2 + 3\alpha)\|\mathbf{h}\|_\infty D$, we obtain

$$|\partial_{\mathbf{h}}^3 \tau| = \frac{\alpha}{\varepsilon^3} \tau |N| \leq \frac{\alpha}{\varepsilon^3} \tau \, (2 + 3\alpha) \|\mathbf{h}\|_\infty D = \frac{2 + 3\alpha}{\varepsilon} \, \|\mathbf{h}\|_\infty \, \partial_{\mathbf{h}}^2 \tau.$$

This proves that τ is quasi-self-concordant with parameter

$$M = \frac{2 + 3\alpha}{\varepsilon}.$$

By differentiation under the integral, from the boundedness of the measure μ, integrating the pointwise inequality yields

$$|\partial_{\mathbf{h}}^3 \mathcal{J}(\mathbf{g})| \leq \frac{2 + 3\alpha}{\varepsilon} \, \|\mathbf{h}\|_\infty \, \partial_{\mathbf{h}}^2 \mathcal{J}(\mathbf{g}).$$

■

E.2. Case 2: $D_1 = D_2 = \chi^2$

Proof We keep the same notations as in Case 1:

$$B_j(\mathbf{g}) = \beta_j \exp\left(\frac{g_j - c(x, y_j)}{\varepsilon}\right), \qquad Z(\mathbf{g}) = \sum_{j=1}^{n} B_j(\mathbf{g}), \qquad w_j(\mathbf{g}) = \frac{B_j(\mathbf{g})}{Z(\mathbf{g})} \in \Delta^n,$$

and for a direction $\mathbf{h} \in \mathbb{R}^n$ we denote

$$m := w_{\mathbf{h}} = \langle w, \mathbf{h}\rangle, \qquad w_{\mathbf{h}^2} = \langle w, \mathbf{h}^2\rangle, \qquad \sigma^2 := w_{\mathbf{h}^2} - m^2, \qquad \kappa_3 := \mathcal{E}[(H-m)^3], \qquad L := \|\mathbf{h}\|_\infty,$$

where H takes values h_j with probabilities w_j. Along $\mathbf{g}(t) = \mathbf{g} + t\mathbf{h}$, the directional derivatives of Z are unchanged:

$$\partial_{\mathbf{h}} Z = \frac{1}{\varepsilon} Z\, m, \qquad \partial^2_{\mathbf{h}} Z = \frac{1}{\varepsilon^2} Z\, w_{\mathbf{h}^2}, \qquad \partial^3_{\mathbf{h}} Z = \frac{1}{\varepsilon^3} Z\, w_{\mathbf{h}^3}.$$

We also reuse the same probabilistic bound as in Case 1: since $|H| \le L$ a.s., one has

$$|\kappa_3| \le \mathcal{E}[|H - m|^3] \le 2L\, \mathcal{E}[(H-m)^2] = 2L\, \sigma^2. \tag{47}$$

Specific changes. For the χ^2-source semi-dual obtained by eliminating f, the *transport* integrand is

$$\tau_{\chi^2}(\mathbf{g}) := \frac{\varepsilon^2}{\rho_1}\Big(W(U(\mathbf{g})) + \tfrac{1}{2} W(U(\mathbf{g}))^2\Big), \qquad U(\mathbf{g}) := \frac{\rho_1}{\varepsilon} e^{\rho_1/\varepsilon} Z(\mathbf{g}).$$

Equivalently, $\tau_{\chi^2} = \psi(Z)$ with

$$\psi(z) := \frac{\varepsilon^2}{\rho_1}\Big(W(az) + \tfrac{1}{2} W(az)^2\Big), \qquad a := \frac{\rho_1}{\varepsilon} e^{\rho_1/\varepsilon}.$$

Write $\omega := W(az) \ge 0$ (so $\omega e^\omega = az$). Using $W'(u) = \frac{W(u)}{u(1+W(u))}$ and the identity $u = az$, one obtains the explicit derivatives

$$\psi'(z) = \frac{\varepsilon^2}{\rho_1}\frac{\omega}{z}, \qquad \psi''(z) = -\frac{\varepsilon^2}{\rho_1}\frac{\omega^2}{z^2(1+\omega)}, \qquad \psi'''(z) = \frac{\varepsilon^2}{\rho_1}\frac{\omega^3(3+2\omega)}{z^3(1+\omega)^3}. \tag{48}$$

Second derivative along $\mathbf{h}$. By the one-dimensional chain rule,

$$\partial^2_{\mathbf{h}} \tau_{\chi^2} = \psi''(Z)(\partial_{\mathbf{h}} Z)^2 + \psi'(Z)\partial^2_{\mathbf{h}} Z.$$

Substituting (48) and the derivatives of Z yields

$$\begin{aligned}\partial^2_{\mathbf{h}} \tau_{\chi^2} &= \left(-\frac{\varepsilon^2}{\rho_1}\frac{\omega^2}{Z^2(1+\omega)}\right)\left(\frac{Z^2 m^2}{\varepsilon^2}\right) + \left(\frac{\varepsilon^2}{\rho_1}\frac{\omega}{Z}\right)\left(\frac{Z w_{\mathbf{h}^2}}{\varepsilon^2}\right) \\ &= \frac{1}{\rho_1}\left(\omega w_{\mathbf{h}^2} - \frac{\omega^2}{1+\omega} m^2\right) = \frac{\omega}{\rho_1}\left(\sigma^2 + \frac{m^2}{1+\omega}\right) \ge 0. \end{aligned} \tag{49}$$

Third derivative along $\mathbf{h}$ and central-moment simplification. Similarly,

$$\partial^3_{\mathbf{h}} \tau_{\chi^2} = \psi'''(Z)(\partial_{\mathbf{h}} Z)^3 + 3\psi''(Z)\partial_{\mathbf{h}} Z\, \partial^2_{\mathbf{h}} Z + \psi'(Z)\partial^3_{\mathbf{h}} Z.$$

Substituting (48) and the derivatives of Z gives

$$\partial_{\mathbf{h}}^3 \tau_{\chi^2} = \frac{1}{\rho_1 \varepsilon} \left[\omega w_{\mathbf{h}^3} - \frac{3\omega^2}{1+\omega} m w_{\mathbf{h}^2} + \frac{\omega^3(3+2\omega)}{(1+\omega)^3} m^3 \right].$$

Using $w_{\mathbf{h}^2} = \sigma^2 + m^2$ and $w_{\mathbf{h}^3} = \kappa_3 + 3m\sigma^2 + m^3$, the cubic terms in m^3 yield the identity

$$\partial_{\mathbf{h}}^3 \tau_{\chi^2} = \frac{\omega}{\rho_1 \varepsilon} \left[\kappa_3 + \frac{3m}{1+\omega} \sigma^2 + \frac{m^3}{(1+\omega)^3} \right]. \tag{50}$$

Generalized self-concordance bound. Using (47), $|m| \leq L$, and $(1+\omega)^{-1} \leq 1$, we get

$$\left| \kappa_3 + \frac{3m}{1+\omega} \sigma^2 \right| \leq |\kappa_3| + 3|m|\sigma^2 \leq 2L\sigma^2 + 3L\sigma^2 = 5L\sigma^2.$$

Moreover, since $(1+\omega)^{-3} \leq (1+\omega)^{-1}$ and $|m| \leq L$, we also have

$$\left| \frac{m^3}{(1+\omega)^3} \right| \leq \frac{|m|}{1+\omega} m^2 \leq L \frac{m^2}{1+\omega}.$$

Combining these two bounds with (50) yields

$$|\partial_{\mathbf{h}}^3 \tau_{\chi^2}| \leq \frac{\omega}{\rho_1 \varepsilon} L \left(5\sigma^2 + \frac{m^2}{1+\omega} \right).$$

Since $5\sigma^2 + \frac{m^2}{1+\omega} \leq 6 \left(\sigma^2 + \frac{m^2}{1+\omega} \right)$ and using (49),

$$\partial_{\mathbf{h}}^2 \tau_{\chi^2} = \frac{\omega}{\rho_1} \left(\sigma^2 + \frac{m^2}{1+\omega} \right),$$

we obtain the pointwise inequality

$$|\partial_{\mathbf{h}}^3 \tau_{\chi^2}| \leq \frac{6}{\varepsilon} L \, \partial_{\mathbf{h}}^2 \tau_{\chi^2}.$$

Therefore, τ_{χ^2} is quasi-self-concordant with parameter $M = 6/\varepsilon$. By differentiation under the integral, the same bound transfers to the transport functional $\mathcal{J}_{\text{trans}}(\mathbf{g}) = \int_{\mathcal{X}} \tau_{\chi^2}(x, \mathbf{g}) \mathrm{d}\mu(x)$:

$$|\partial_{\mathbf{h}}^3 \mathcal{J}_{\text{trans}}(\mathbf{g})| \leq \frac{6}{\varepsilon} \|\mathbf{h}\|_\infty \, \partial_{\mathbf{h}}^2 \mathcal{J}_{\text{trans}}(\mathbf{g}).$$

■

As a corollary of generalized self-concordance, we have enhanced control over the Hessian; see, for instance, Proposition 8 in Sun and Tran-Dinh (2019). However, here, this is with respect to the infinity norm instead of the Euclidean norm.

Corollary 13 *Noting* $M = \frac{2+3\alpha}{\varepsilon}$ *for the* KL *source case,* $M = \frac{6}{\varepsilon}$ *else, we have for any* $\mathbf{g}, \mathbf{g}_2 \in \mathbb{R}^n$:

$$e^{-M\|\mathbf{g}_2 - \mathbf{g}_1\|_\infty} \nabla^2 \mathcal{J}(\mathbf{g}_1) \preceq \nabla^2 \mathcal{J}(\mathbf{g}_2) \preceq e^{M\|\mathbf{g}_2 - \mathbf{g}_1\|_\infty} \nabla^2 \mathcal{J}(\mathbf{g}_1)$$

Appendix F. Proof of Proposition 11: (L_0, L_1)-smoothness

Proof Fix $g \in \mathbb{R}^n$ and consider the segment

$$g_s := g - s\lambda \nabla J(g), \qquad s \in [0,1],$$

with $\lambda > 0$. Then $\|g_s - g\|_\infty = s\lambda \|\nabla J(g)\|_\infty$.
By generalized self-concordance (Corollary 13 stated with $\|\cdot\|_\infty$), we have

$$\nabla^2 J(g_s) \preceq \exp\big(M\|g_s - g\|_\infty\big)\, \nabla^2 J(g) = \exp\big(Ms\lambda \|\nabla J(g)\|_\infty\big)\, \nabla^2 J(g).$$

Choose the step size as $\lambda := 1/\widetilde{L}(g)$ where

$$\widetilde{L}(g) := e\left(\frac{1}{\varepsilon}\|\nabla J_{\mathrm{trans}}(g)\|_\infty + \frac{\beta_{\max}}{\rho_2}\right) \; + \; M\|\nabla J(g)\|_\infty.$$

Then $M\lambda\|\nabla J(g)\|_\infty \le 1$, hence for all $s \in [0,1]$,

$$\nabla^2 J(g_s) \preceq e\, \nabla^2 J(g), \quad \text{and thus} \quad \sup_{s\in[0,1]} \|\nabla^2 J(g_s)\|_{\mathrm{op}} \le e\, \|\nabla^2 J(g)\|_{\mathrm{op}}.$$

Next, we upper bound $\|\nabla^2 J(g)\|_{\mathrm{op}}$ by the local smoothness proxy computed from $\nabla J_{\mathrm{trans}}(g)$. By Theorem 3,

$$\|\nabla^2 J(g)\|_{\mathrm{op}} \le \frac{1}{\varepsilon}\|\nabla J_{\mathrm{trans}}(g)\|_\infty + \frac{\beta_{\max}}{\rho_2}.$$

Therefore,

$$\sup_{s\in[0,1]} \|\nabla^2 J(g_s)\|_{\mathrm{op}} \le e\left(\frac{1}{\varepsilon}\|\nabla J_{\mathrm{trans}}(g)\|_\infty + \frac{\beta_{\max}}{\rho_2}\right) \le \widetilde{L}(g).$$

Finally, for any two points g_1, g_2 on the segment $\{g_s : s \in [0,1]\}$, the mean value theorem yields

$$\|\nabla J(g_1) - \nabla J(g_2)\| \le \left(\sup_{s\in[0,1]} \|\nabla^2 J(g_s)\|_{\mathrm{op}}\right) \|g_1 - g_2\| \le \widetilde{L}(g)\, \|g_1 - g_2\|.$$

This concludes the proof. ■

Appendix G. Proof of Theorem 9 : Convergence of PASGD

Theorem 9: Convergence of PASGD. Let the step sizes be chosen as $\eta_t = Ct^{-\gamma}$ with $\gamma \in (1/2, 1)$. Under Setting 1 and the projection onto $\mathcal{K}$, the averaged iterate $\bar{\mathbf{g}}_T$ converges to the optimum $\mathbf{g}^\star$ in objective value with an expected error of:

$$\mathbb{E}\left[\mathcal{J}(\bar{\mathbf{g}}_T) - \mathcal{J}(\mathbf{g}^\star)\right] = \mathcal{O}\left(\frac{n\rho_2^2}{\varepsilon T}\right).$$

Proof We study the projected SGD recursion

$$g_{t+1} = \Pi_{\mathcal{K}}(g_t - \gamma_t \widehat{\nabla}\mathcal{J}(g_t)), \qquad \bar{g}_T = \frac{1}{T}\sum_{t=1}^{T} g_t, \qquad \gamma_t = Ct^{-\beta},\ \beta \in (1/2, 1).$$

For the non-averaged iterates, the projection brings no additional difficulty: $\Pi_{\mathcal{K}}$ is 1-Lipschitz, so all standard Robbins–Monro estimates based on a one-step expansion remain valid. We therefore rely on the results of Gadat and Panloup (2017), and we start by verifying their martingale increment assumption.

Verification of $(\mathrm{HSC}_{\Sigma_p})$ for all p: Let

$$\xi_{t+1} := \widehat{\nabla}\mathcal{J}(g_t) - \nabla\mathcal{J}(g_t), \qquad \mathcal{F}_t := \sigma(g_0, X_1, \dots, X_t)$$

be the noise and the natural filtration. We assume (by construction of the algorithm) that $g_t \in \mathcal{K}$ for all t.

For every integer $p \geq 1$, the condition $(\mathrm{HSC}_{\Sigma_p})$ of Gadat and Panloup (2017, Sec. 1.3.3) holds with

$$\Sigma_p := (2C_{\mathrm{bound}})^{2p}.$$

Indeed, by Lemma 5, on $\mathcal{K}$ the transport gradient has a uniform ℓ_1 bound: for any $g \in \mathcal{K}$ and any realization of the mini-batch,

$$\|\widehat{\nabla}\mathcal{J}_{\mathrm{trans}}(g)\|_1 \leq C_{\mathrm{bound}}, \qquad \|\nabla\mathcal{J}_{\mathrm{trans}}(g)\|_1 \leq C_{\mathrm{bound}}.$$

Since the quadratic term in $\nabla\mathcal{J}$ is deterministic, the noise satisfies

$$\|\widehat{\nabla}\mathcal{J}(g) - \nabla\mathcal{J}(g)\|_1 = \|\widehat{\nabla}\mathcal{J}_{\mathrm{trans}}(g) - \nabla\mathcal{J}_{\mathrm{trans}}(g)\|_1 \leq \|\widehat{\nabla}\mathcal{J}_{\mathrm{trans}}(g)\|_1 + \|\nabla\mathcal{J}_{\mathrm{trans}}(g)\|_1 \leq 2C_{\mathrm{bound}}.$$

Using $\|\cdot\|_2 \leq \|\cdot\|_1$, we obtain $\|\xi_{t+1}\|_2 \leq 2C_{\mathrm{bound}}$ almost surely, hence for all $p \geq 1$,

$$\mathbb{E}\Big[\|\xi_{t+1}\|_2^{2p} \mid \mathcal{F}_t\Big] \leq (2C_{\mathrm{bound}})^{2p} = \Sigma_p.$$

Finally, since $1 + \mathcal{J}(g_t)^p \geq 1$, we also have

$$\mathbb{E}\Big[\|\xi_{t+1}\|_2^{2p} \mid \mathcal{F}_t\Big] \leq \Sigma_p\big(1 + \mathcal{J}(g_t)^p\big),$$

which is exactly $(\mathrm{HSC}_{\Sigma_p})$ in the sense of Gadat and Panloup (2017).

As a consequence, Proposition 1.1 in Gadat and Panloup (2017) applies and yields, for any $p \geq 1$,

$$\mathbb{E}\big[\|g_t - g^\star\|^{2p}\big] \;\lesssim\; \gamma_t^p. \tag{51}$$

The projection is asymptotically negligible: While the projection is harmless for the analysis of the non-averaged recursion, it is convenient to show that it becomes asymptotically inactive. Fix $\delta = 1$ and denote $\mathcal{K} = \mathcal{K}_1 = \{g : \; g_k \leq \rho_2 + 1\}$. Consider the event that projection is active at time t:

$$\mathcal{P}_t := \Big\{g_t - \gamma_t\widehat{\nabla}\mathcal{J}(g_t) \notin \mathcal{K}\Big\}.$$

Since $g^\star \in \mathcal{K}_0$ (Proposition 4) and $\mathcal{K}_0 \subset \mathcal{K}$, this event can only happen when g_t is sufficiently far from $g^\star$. In particular, there exists a constant $r > 0$ (depending on ρ_2 and δ only) such that $\mathcal{P}_t \subset \{\|g_t - g^\star\|_\infty \geq r\}$ for all t large enough. Therefore, for any integer $p \geq 1$, Markov's inequality and (51) give

$$\mathbb{P}(\mathcal{P}_t) \leq \mathbb{P}(\|g_t - g^\star\|_\infty \geq r) \leq \frac{\mathbb{E}\|g_t - g^\star\|^{2p}}{r^{2p}} \;\lesssim\; \frac{\gamma_t^p}{r^{2p}}.$$

Since p is arbitrary and $\gamma_t = t^{-\beta}$ with $\beta \in (1/2, 1)$, we can make $\mathbb{P}(\mathcal{P}_t) = o(t^{-a})$ for any prescribed $a > 0$ by choosing p large enough. Therefore, the projection affects a vanishing fraction of iterates.

Averaged iterates and objective error: We can now invoke Corollary 1.1 in Gadat and Panloup (2017), which gives the standard Polyak–Ruppert behavior for $\bar{g}_T$. Noting $\mathbf{H} = \nabla^2 \mathcal{J}(\mathbf{g}^\star)$, we have :

$$\mathbb{E}[\|\bar{\mathbf{g}}_T - \mathbf{g}^\star\|^2] \leq \frac{\mathrm{Tr}(\mathbf{H}^{-1}\mathbf{\Sigma}\mathbf{H}^{-1})}{T} + o(1/T). \tag{52}$$

Given the bounded noise variance $\mathbf{\Sigma}$ (coming from Lemma 5), this term scales as $\mathcal{O}(\rho_2^2 n^2/T)$.

We now convert (52) into an objective bound using the generalized self-concordance of the semi-dual (Proposition 7). Locally, the smoothness at $g^\star$ satisfies $L(g^\star) \asymp 1/(n\varepsilon)$ (Corollary 6), and Corollary 13 ensures that along a neighborhood of $g^\star$,

$$L(\bar{\mathbf{g}}_T) \;\leq\; \exp\Big(M\|\bar{g}_T - g^\star\|_\infty\Big)\, L(\mathbf{g}^\star), \qquad M = \frac{2+3\alpha}{\varepsilon} \;\text{(KL-source)}.$$

Splitting on the event $\{\|\bar{g}_T - g^\star\| \leq \varepsilon\}$ and using the above local control yields

$$\mathbb{E}\big[\mathcal{J}(\bar{\mathbf{g}}_T) - \mathcal{J}^\star\big] \;\leq\; \frac{e^{M\varepsilon} L(g^\star)}{2}\, \mathbb{E}\big[\|\bar{\mathbf{g}}_T - \mathbf{g}^\star\|^2\big] \;+\; \frac{C_1}{\varepsilon}\, \mathbb{P}\big(\|\bar{\mathbf{g}}_T - \mathbf{g}^\star\| \geq \varepsilon\big),$$

for a finite constant C_1 (depending only on $\mathcal{K}$ through the crude curvature bound on $\mathcal{K}$).

Finally, using Theorem 4.4 in Godichon-Baggioni (2019), we have high-order moment of averaged iterates

$$\mathbb{E}\big[\|\bar{\mathbf{g}}_t - \mathbf{g}^\star\|^{2p}\big] \;\lesssim\; \frac{1}{t^p}. \tag{53}$$

Again, using Markov's inequality with high order moment gives

$$\mathbb{P}\big(\|\bar{g}_T - g^\star\| \geq \varepsilon\big) = o(t^{-a})$$

for all a, which concludes.

■

Remark. This adaptivity argument uses a step-size schedule $\gamma_t \propto t^{-\beta}$ with $\beta < 1$ to ensure good control of higher moments and tail probabilities, which would have not been possible with the non-averaged iterates, using $\gamma_t \propto \frac{1}{t}$.

Appendix H. Proof of Theorem 12: Adaptive NAG Convergence Rate

Theorem 12 : Adaptive NAG Convergence Rate. Let R be the number of restart. Then, the iterates generated by Algorithm 1 satisfy

$$\mathcal{J}(g_{T+1}) - \mathcal{J}^\star \leq 2^R \left(\mathcal{J}(g_0) - \mathcal{J}^\star + \frac{\beta_{\min}}{2\rho_2}\|g_0 - g^\star\|^2\right) \prod_{t=0}^{T} \left(1 - \sqrt{\frac{\beta_{\min}}{\rho_2 L_t}}\right),$$

Furthermore, the algorithm ensures $y_t \in K_1$ for all t, so using C_{bound} from Lemma 5, we have $L_t \leq \bar{L} = \mathcal{O}(C_{\text{bound}}/\varepsilon)$ for all t. This implies the following rates:

1. **Global Rate:** We have at least the contraction rate $1 - \mathcal{O}(\sqrt{\varepsilon/\beta_{\min}\rho_2})$ for both the objective gap and gradient norm.

2. **Local Rate:** Since $L_t \leq 2\beta_{\max}\left(\frac{1}{\varepsilon} + \frac{1}{\rho_2}\right) + 2\|\nabla \mathcal{J}(g_{T+1})\|$, assuming $\beta_{\min}/\beta_{\max} \simeq 1$ and substituting the contraction rate of the gradient norm into (12) shows that, locally, we have a contraction rate of $1 - \mathcal{O}(\sqrt{\varepsilon/\rho_2})$.

Proof

Contraction of the potential function when we do not restart. The restart schemes permit us to stay in the region K_δ, where we fix $\delta = 1$. We analyze here the contraction, when restart is does not happen.

Following the Lyapunov analysis for accelerated methods (Nesterov, 2015; Bansal and Gupta, 2019), we define the potential function at iteration t as:

$$\Psi_t \triangleq \frac{1}{\sqrt{L_{t-1}}}\left(\mathcal{J}(g_t) - \mathcal{J}^\star + \frac{\alpha}{2}\|z_t - g^\star\|^2\right), \tag{54}$$

where $z_t := \mathbf{g}_t + \left(\sqrt{L_t/\alpha} - 1\right)(\mathbf{g}_t - \mathbf{g}_{t-1})$ is the auxiliary sequence, α is the strong convexity parameter of our function (here = $\frac{\beta_{\min}}{\rho_2}$), L_t is the smoothness bound on the segment $[y_t, \mathbf{g}_{t+1}]$ and by convention $L_{-1} = L_0$.

In Bansal and Gupta (2019), the proof there bounds $\Delta\Psi_t$ by combining (i) one smoothness-based decrease inequality for the gradient step and (ii) an algebraic expansion of the quadratic term in the potential ((Bansal and Gupta, 2019, eqs. (5.25)–(5.27))). In our constrained case, the algebraic part is unchanged; only (i) changes.

Indeed, our update is the projected step $g_{t+1} = \Pi_{\mathcal{K}}(y_t - \frac{1}{L_t}\nabla\mathcal{J}(y_t))$. Define the gradient-mapping residual $\Delta_t := L_t(y_t - g_{t+1})$. By Proposition 11, $\mathcal{J}$ is L_t-smooth on the segment between y_t and g_{t+1}, and by optimality of the projection we obtain the projected descent inequality

$$\mathcal{J}(g_{t+1}) \leq \mathcal{J}(y_t) - \frac{1}{2L_t}\|\Delta_t\|^2,$$

which replaces the unconstrained inequality $f(y_{t+1}) \leq f(x_t) - \frac{1}{2\beta}\|\nabla_t\|^2$ used in (Bansal and Gupta, 2019, p. 28), where $\nabla_t = \nabla f(x_t)$ with their notations.

With this substitution (replace ∇_t by Δ_t and β by L_t), the remainder of the potential-change calculation is identical to (Bansal and Gupta, 2019, eqs. (5.25)–(5.27)), accounting here for the difference between L_{t-1} and L_t, yielding the one-step contraction,

$$\Psi_{t+1} \leq \sqrt{\frac{L_{t-1}}{L_t}}\left(1 - \sqrt{\frac{\alpha}{L_t}}\right)\Psi_t,$$

and unrolling gives

$$\mathcal{J}(g_{T+1}) - \mathcal{J}^\star \leq \Phi_0 \prod_{t=0}^{T}\left(1 - \sqrt{\frac{\alpha}{L_t}}\right).$$

Taking into account the restarts. We implement a safeguard restart to ensure that all smoothness arguments are made inside the bounded region $\mathcal{K}_1$. Concretely, at any iteration t such that the

extrapolated point leaves the safe set, $y_t \notin \mathcal{K}_1$, we restart by resetting the acceleration state (zeroing the momentum):

$$y_t \leftarrow g_t, \qquad g_{t-1} \leftarrow g_t,$$

so that the auxiliary variable definition $z_t = g_t + (\sqrt{L_t/\alpha} - 1)(g_t - g_{t-1})$ is consistent and implies $z_t = g_t$. We also re-initialize the scalar parameters of the Lyapunov construction for the new epoch: we start a new epoch with local index $s = t$ and enforce $L_{s-1} = L_s$ by convention. The projected update $g_{t+1} = \Pi_{\mathcal{K}}(y_t - \frac{1}{L_t}\nabla\mathcal{J}(y_t))$ is then performed from $y_t = g_t \in \mathcal{K}_{0.1} \subset \mathcal{K}_1$, so that Proposition 11 applies on the segment $[y_t, g_{t+1}]$.

Define the unnormalized potential

$$E_t := \sqrt{L_{t-1}}\Psi_t = \mathcal{J}(g_t) - \mathcal{J}^\star + \frac{\alpha}{2}\|z_t - g^\star\|^2.$$

Multiplying the one-step inequality

$$\Psi_{t+1} \leq \sqrt{\frac{L_{t-1}}{L_t}}\left(1 - \sqrt{\frac{\alpha}{L_t}}\right)\Psi_t$$

by $\sqrt{L_t}$ yields the clean contraction

$$E_{t+1} \leq \left(1 - \sqrt{\frac{\alpha}{L_t}}\right) E_t,$$

valid at any iteration where we do not restart (i.e., when $y_t \in \mathcal{K}_1$ and the Lyapunov update is applied). Let $0 = \tau_0 < \tau_1 < \cdots < \tau_R \leq T$ denote the restart times (epoch starts), and set $\tau_{R+1} := T + 1$. On each epoch $[\tau_j, \tau_{j+1})$, unrolling the above inequality gives

$$E_{\tau_{j+1}} \leq \left(\prod_{t=\tau_j}^{\tau_{j+1}-1}\left(1 - \sqrt{\frac{\alpha}{L_t}}\right)\right) E^+_{\tau_j},$$

where $E^+_{\tau_j}$ denotes the value of E *after* the restart initialization at time τ_j (and $E^+_{\tau_0} = E_0$). At each restart time τ_j with $j \geq 1$, we have $z_{\tau_j} = g_{\tau_j}$ and by α-strong convexity $\mathcal{J}(g_{\tau_j}) - \mathcal{J}^\star \geq \frac{\alpha}{2}\|g_{\tau_j} - g^\star\|^2$, hence

$$E^+_{\tau_j} = \mathcal{J}(g_{\tau_j}) - \mathcal{J}^\star + \frac{\alpha}{2}\|g_{\tau_j} - g^\star\|^2 \leq 2\big(\mathcal{J}(g_{\tau_j}) - \mathcal{J}^\star\big) \leq 2E^-_{\tau_j},$$

where $E^-_{\tau_j}$ denotes the value of E just before restarting (same g_{τ_j}, previous z_{τ_j}). Iterating over epochs and using $E_t \geq \mathcal{J}(g_t) - \mathcal{J}^\star$ yields the global bound

$$\mathcal{J}(g_{T+1}) - \mathcal{J}^\star \leq 2^R\left(\mathcal{J}(g_0) - \mathcal{J}^\star + \frac{\alpha}{2}\|z_0 - g^\star\|^2\right)\prod_{t=0}^{T}\left(1 - \sqrt{\frac{\alpha}{L_t}}\right).$$

In particular, since $z_0 = g_0$ at initialization, we have $\|z_0 - g^\star\| = \|g_0 - g^\star\|$. Moreover, $L_t \leq \bar{L}$, where $\bar{L} = O(C_{\text{bound}}/\varepsilon)$ using that all iterations are in $\mathcal{K}_1$. We thus conclude

$$\mathcal{J}(g_{T+1}) - \mathcal{J}^\star \leq 2^R\left(\mathcal{J}(g_0) - \mathcal{J}^\star + \frac{\alpha}{2}\|g_0 - g^\star\|^2\right)\left(1 - \sqrt{\frac{\alpha}{\bar{L}}}\right)^{T+1}.$$

■

Appendix I. Towards Optimal Adaptive Step Sizes

Optimal rates for ASGD are obtained with a learning rate of $\mathcal{O}\left(\frac{1}{L+\beta k}\right)$ for L-smooth and β-strongly convex functions Stich (2019). In this case, the rate is:

$$\mathbb{E}f(\bar{\mathbf{x}}_T) - f^\star + \beta\mathbb{E}\|\mathbf{x}_{T+1} - \mathbf{x}^\star\|^2 = \tilde{\mathcal{O}}\left(LR^2 \exp\left[-\frac{\beta T}{2L}\right] + \frac{\sigma^2}{\beta T}\right) .$$

In our context, if we were near the minimum, this would lead to a complexity of $\mathcal{O}(\rho_2/\varepsilon)$ to eliminate the transient exponential term. This results in a total complexity of $\mathcal{O}(n\rho_2/T)$ for any ε, whereas a global bound would require $\mathcal{O}(\rho_2 n/\varepsilon)$ to handle this exponential term.

In a broader context, the motivation for employing a learning rate of the form $(1/L_t+\beta k)^{-1}$ and leveraging adaptive smoothness was previously investigated by Malitsky and Mishchenko (2019). There, the authors proposed a local estimator of smoothness as a heuristic; however, this approach did not yield theoretical acceleration and resulted in worse constants.

While we will not be able to strictly use or prove this optimal rate here, we remark that for $\mathcal{J}$, another natural choice leads to a similar schedule. A classical schedule for $\gamma_t \propto 1/t$ sets the constant as the inverse of the strong convexity, i.e., $\gamma_t^{sc} = \frac{n}{C\rho_2}$. Concurrently, assuming $C/n = \beta_{\min} = \beta_{\max}$ for clarity, the optimal learning rate from Stich (2019) near the optimum would be:

$$\gamma_t^{\text{opt}} = \frac{1}{1/L(\mathbf{g}^\star) + \beta t} \simeq \frac{n}{C(1/\varepsilon + \rho_2 t)} .$$

We observe that, in this case, the two learning rates differ simply by an offset of $1/\varepsilon$.

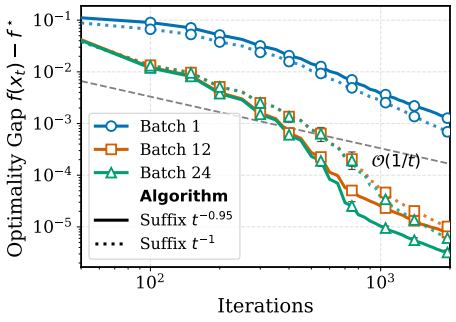

Figure 5: Comparison of ASGD convergence rates using the optimal learning rate schedule $\gamma_t^{\text{opt}} = (1/L(\mathbf{g}^\star) + \beta t)^{-1}$ versus the polynomial decay $\gamma_t = (1/L(\mathbf{g}^\star) + \beta t^b)^{-1}$ with $b = 0.95$. Here $\beta_{\min} = \beta_{\max} = 1/n$ with $n = 2000$. Both methods employ suffix averaging (averaging the last half of iterates), as recommended in Stich (2019). We display results for minibatch sizes of 1, 12, and 24, averaged over 20 independent runs. We observe that both learning rate schedules perform equally well. Variance is omitted from the plot as it is negligible.

Although we do not provide a proof of the true adaptivity of our SGD to local strong convexity for the $\propto 1/t$ learning rate, we provide next a proof of ASGD for any learning rate $\gamma_t \propto 1/t^b$ where $b \in (0, 1)$. We select $b < 1$ to ensure the convergence of all moments of our SGD scheme. This

allows us to demonstrate a mild adaptivity phenomenon by taking $\gamma_t \simeq \frac{n}{C(1/\varepsilon+\rho_2 t^b)}$ with b very close to 1. In contrast, we would not be able to prove this adaptivity with $\gamma_t \propto 1/t$ and it would lead to worse constants in n and $1/\varepsilon$.

Appendix J. Additional Experimental Results

Comparison with Sinkhorn. We benchmark our Adaptive NAG (ANAG) method against the Translation Invariant (TI) Sinkhorn algorithm (Séjourné et al., 2022) on a standard color transfer task. We select 20 random image pairs from the CIFAR-10 dataset, treating pixels as point clouds in RGB space ($n = 4096$). Both solvers are run with regularization $\varepsilon = 0.01$ and marginal penalty $\rho = 10$.

It is important to note that ANAG minimizes the $\mathrm{KL} - \chi^2$ semi-dual, while TI-Sinkhorn solves the standard $\mathrm{KL} - \mathrm{KL}$ formulation. However, in this regime, the resulting optimal couplings and transport costs are nearly identical, justifying a direct comparison of their convergence profiles. We establish a ground truth value $f^\star$ by running each solver for $20,000$ iterations and report the median relative objective gap $(f(x_t) - f^\star)/|f^\star|$ in Figure 6.

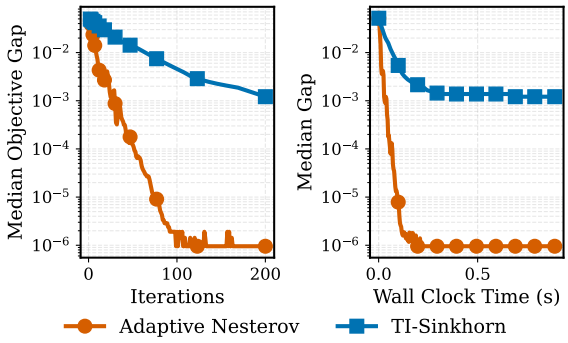

Figure 6: **ANAG vs. TI-Sinkhorn.** Median objective gap convergence on CIFAR-10 color transfer tasks (20 pairs of size $n = 4096$). ANAG demonstrates competitive convergence rates compared to the TI-Sinkhorn algorithm (solving $\mathrm{KL} - \mathrm{KL}$), validating the efficiency of the adaptive scheme on standard semi-discrete tasks.