W-LALM: A LINEARIZED AUGMENTED LAGRANGIAN METHOD FOR CONSTRAINED OPTIMIZATION IN WASSERSTEIN SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Many problems in inference, sampling, and scientific calibration require learning a probability law subject to linear expectation constraints. The natural extension of Wasserstein proximal gradient (WPG) to this problem enforces the constraints inside its Jordan–Kinderlehrer–Otto (JKO) subproblem, which can turn an otherwise tractable backward step into a constrained transport problem. Primal–dual sampling methods avoid this difficulty by moving feasibility into a multiplier update, but their guarantees cover averaged or distributional quantities rather than the full primal–dual sequence, and they do not exploit composite objectives whose nonsmooth part admits a tractable JKO resolvent. We propose the Wasserstein linearized augmented Lagrangian method (W-LALM), which keeps the original unconstrained JKO step and handles feasibility through an explicit step on the smooth part of the augmented Lagrangian together with a multiplier update. Our analysis is built on transport plans rather than transport maps, since optimal Monge maps need not exist for the discrete measures used in particle implementations. This representation yields a plan-based Lyapunov descent estimate, convergence of the full primal–dual sequence to a saddle point under generalized-geodesic convexity, and an $O(1/\varepsilon)$ iteration complexity for the squared residual of a plan-valued Wasserstein Karush–Kuhn–Tucker (KKT) condition that we introduce. W-LALM matches or improves upon the terminal objective of the strongest feasible baseline, while being 8–38$\times$ faster than the constrained-JKO variant of WPG and up to two orders of magnitude faster than the Euclidean augmented-Lagrangian baseline.

1 INTRODUCTION

Learning a probability law rather than a single parameter is central to variational inference, sampling, and scientific calibration. In many such problems, the learned law must also reproduce prescribed statistics: a calibrated posterior should match known moments of the quantity of interest, and an ensemble of physical fields should satisfy conservation laws or reproduce prescribed diagnostics (Bissiri et al., 2016; Evensen, 2003). Motivated by these settings, we study

$$\min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \Phi(\mu) := \int_{\mathbb{R}^d} V(x)\ \mathrm{d}\mu(x) + \mathcal{H}(\mu) \quad \text{subject to} \quad \mathcal{C}(\mu) := \int_{\mathbb{R}^d} Ax\ \mathrm{d}\mu(x) - b = 0, \quad (1)$$

where $\mathcal{P}_2(\mathbb{R}^d)$ denotes the space of Borel probability measures on $\mathbb{R}^d$ with finite second moment, $A \in \mathbb{R}^{m\times d}$, $b \in \mathbb{R}^m$, V is smooth and convex, and $\mathcal{H}$ is possibly nonsmooth and convex along generalized geodesics.

To see the computational difficulty introduced by the constraint, it is useful to first recall the unconstrained composite problem. The Jordan–Kinderlehrer– Otto (JKO) scheme is the Wasserstein analogue of a proximal step (Jordan et al., 1998; Ambrosio et al., 2008), while Wasserstein proximal gradient (WPG) exploits the composite structure by treating the two components differently (Salim et al., 2020; Di Marino et al., 2025):

$$\eta_{k+1} = (\mathrm{Id} - \gamma\nabla V)_{\#}\mu_k, \quad \mu_{k+1} \in \operatorname*{arg\,min}_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \left\{ \mathcal{H}(\mu) + \frac{1}{2\gamma} W_2^2(\mu, \eta_{k+1}) \right\}.$$

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

The smooth potential V is handled by an explicit transport step, while the possibly nonsmooth functional $\mathcal{H}$ is handled by its JKO resolvent. This separation is computationally useful. The forward step only requires evaluations of ∇V, while the backward step retains the original JKO structure of $\mathcal{H}$, so closed-form updates or specialized solvers can be reused. For example, integral functionals induced by proximable functions reduce to pointwise proximal maps, while Gaussian-structured objectives admit specialized Wasserstein updates (Diao et al., 2023; Luu et al., 2025).

The constrained problem raises a new question of how to incorporate the feasibility condition $\mathcal{C}(\mu) = 0$ without losing this forward–backward separation. A straightforward extension of WPG is to incorporate the feasibility condition into the backward step by replacing $\mathcal{H}$ with $\mathcal{H} + I_{\{\mathcal{C}=0\}}$. Keeping the explicit forward step unchanged, the resulting backward update becomes

$$\mu_{k+1} \in \underset{\substack{\mu \in \mathcal{P}_2(\mathbb{R}^d) \\ \mathcal{C}(\mu)=0}}{\arg\min} \left\{ \mathcal{H}(\mu) + \frac{1}{2\gamma} W_2^2(\mu, \eta_{k+1}) \right\}. \tag{2}$$

This is a constrained JKO subproblem, which enforces feasibility at every backward update. Constrained JKO schemes have been studied in the context of moment-constrained Wasserstein flows (Carlen & Gangbo, 2003; Halkiewicz & Rădulescu, 2026), and dynamic optimal-transport and primal–dual techniques provide numerical tools for such subproblems (Benamou et al., 2016; de la Plata et al., 2022). However, imposing the constraints inside the JKO step requires transport, regularization, and feasibility to be handled jointly at every iteration. Even when the original unconstrained JKO resolvent of $\mathcal{H}$ is tractable, the additional constraints can substantially increase the computational cost of the backward update.

This motivates handling feasibility outside the JKO subproblem. A classical approach to constrained optimization is primal–dual gradient descent–ascent, which performs a descent step in the primal variable and an ascent step in the Lagrange multiplier (Polyak, 1970). This idea has also been adapted to constrained sampling in probability space. Liu et al. (2021) implement the primal descent step through Langevin or Stein iterations while updating the multiplier by dual ascent. Chamon et al. (2024) further develop a discrete-time primal–dual Langevin method that can be viewed as a sampling-based counterpart of the classical gradient descent–ascent scheme, with Langevin iterations replacing an exact primal descent step. This primal–dual construction handles the constraint through multiplier updates without introducing a constrained Wasserstein proximal subproblem. The available convergence results for these sampling methods, however, are mainly formulated in terms of averaged, best-iterate, or distributional guarantees, rather than convergence of the full primal–dual iterate sequence. Moreover, these methods are tailored to sampling objectives and do not exploit the composite forward–backward structure in which a possibly nonsmooth functional is handled through a tractable JKO resolvent. For the composite problem considered here, we therefore pursue a different primal–dual construction. Inspired by linearized augmented Lagrangian methods in Euclidean optimization (Yang & Yuan, 2013), we incorporate the gradient of the quadratic penalty into the explicit Wasserstein forward step, retain the original unconstrained JKO resolvent of $\mathcal{H}$, and update the multiplier separately.

This leads to a natural Wasserstein iteration, but its convergence analysis requires new arguments. In Euclidean primal–dual methods, the analysis relies on vector differences and inner-product identities in a fixed linear space. The Wasserstein space $\mathcal{P}_2(\mathbb{R}^d)$ is not a vector space, so these arguments cannot be transferred directly. Existing Wasserstein first-order analyses, including WPG (Salim et al., 2020; Di Marino et al., 2025) and recent mirror and preconditioned gradient methods (Bonet et al., 2024), use transport-map representations under suitable regularity assumptions. However, optimal Monge maps need not exist for atomic or singular measures, including particle approximations, whereas optimal Kantorovich plans always exist. This motivates working directly with transport plans. The main difficulty is then to couple the explicit forward displacement, the JKO transport plan, and the multiplier correction within a single primal–dual estimate.

This leaves two concrete questions, one algorithmic and one theoretical: *can linear expectation constraints be enforced without replacing the original JKO resolvent of $\mathcal{H}$ by a constrained one, and can the resulting primal–dual splitting be analyzed entirely through transport plans?* In this paper, we address these questions by developing the Wasserstein linearized augmented Lagrangian method (W-LALM), which preserves the original unconstrained JKO step, together with a transport-plan Lyapunov framework that establishes full-sequence convergence of the resulting primal–dual iteration.

To position W-LALM relative to representative Wasserstein first-order methods, Table 1 compares their treatment of linear expectation constraints, preservation of the original unconstrained JKO step, transport representations used in convergence analysis, and full-sequence convergence guarantees in the convex setting. The convergence comparison excludes results requiring strong or strict convexity.

Table 1: Structural comparison of Wasserstein optimization methods.

Method	Linear expectation constraints	Unconstrained JKO of $\mathcal{H}$	Transport in analysis	Full-sequence convergence
Wasserstein PG (WPG) (Salim et al., 2020; Di Marino et al., 2025)	–	✓	Map	✓*
Wasserstein MD/PGD (Bonet et al., 2024)	–	–	Map	–
WPG with constrained JKO (CWPG)	✓	–	Map*	✓*
Primal–dual LMC (Chamon et al., 2024)	✓	N/A	Map + diffusion	–
W-LALM	✓	✓	**Plan**	✓

For CWPG, the indicated guarantee follows from applicable WPG theory when $\mathcal{H} + I_{\{\mathcal{C}=0\}}$ satisfies its regularity assumptions. A dash in the convergence column does not exclude other convergence guarantees.

Contributions. Motivated by the algorithmic and analytical gaps above, our main contributions are summarized as follows.

- **A JKO-preserving Wasserstein primal–dual splitting framework.** We propose the Wasserstein linearized augmented Lagrangian method (W-LALM) for linearly constrained composite optimization over probability laws. The method separates constraint handling from the Wasserstein proximal structure, preserving the original unconstrained JKO resolvent of $\mathcal{H}$ while enforcing feasibility through augmented-Lagrangian corrections and multiplier updates. In the absence of constraints, W-LALM reduces to Wasserstein proximal gradient.
- **A transport-plan Lyapunov framework and full-sequence convergence theory.** We develop a transport-plan analysis that does not rely on the existence of optimal Monge maps. The key construction glues the coupling induced by the explicit forward displacement to an optimal transport plan arising from the JKO step through their common intermediate marginal, leading to a unified primal–dual Lyapunov descent mechanism. Under generalized-geodesic convexity, a saddle-point assumption, and suitable step-size conditions, this framework yields asymptotic feasibility, objective convergence, and convergence of the full primal–dual sequence to a saddle point. These results establish full-sequence convergence for the W-LALM iteration, which combines an explicit forward step, an implicit JKO step, and a multiplier update.
- **A plan-valued Wasserstein KKT theory and quantitative first-order complexity.** We formulate a first-order KKT system for constrained Wasserstein optimization through the extended plan-valued subdifferential and establish its equivalence to the saddle-point characterization. This leads to a law-level optimality certificate that jointly quantifies feasibility and stationarity and is directly compatible with the algorithmic iterates. Building on the same Lyapunov structure, we further prove an $O(1/N)$ bound for the averaged squared KKT residual and an $O(1/\varepsilon)$ best-iterate complexity for achieving squared-residual accuracy ε.

Organization. Section 2 recalls the Wasserstein geometry used in the analysis and states the standing assumptions. Section 3 introduces W-LALM, and Section 4 develops the plan-based Lyapunov

framework, the full-sequence convergence theorem, and the KKT-residual complexity bound. Section 5 reports the numerical experiments, and Section 6 concludes. Proofs, technical lemmas, and implementation details are deferred to the appendix.

2 WASSERSTEIN GEOMETRY AND STANDING ASSUMPTIONS

Let $\mathcal{P}(\mathbb{R}^d)$ denote the set of Borel probability measures on $\mathbb{R}^d$, and define $\mathcal{P}_2(\mathbb{R}^d) := \{\mu \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} \|x\|^2 \, \mathrm{d}\mu(x) < \infty\}$. For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, let $\Gamma(\mu, \nu)$ denote the set of couplings between μ and ν, and let $\Gamma_o(\mu, \nu)$ denote the subset of optimal couplings. The quadratic Wasserstein distance is

$$W_2^2(\mu, \nu) = \inf_{\pi \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, \mathrm{d}\pi(x, y).$$

Beyond the metric itself, our analysis relies on three additional ingredients: generalized-geodesic convexity for the JKO step, plan-valued subgradients for first-order optimality, and the weak topology $\mathcal{P}_2^w$ for full-sequence convergence. We recall only the notions needed for the main results and defer further technical details to the appendix.

Generalized geodesics and plan-valued subgradients. Let $\mu_0, \mu_1, \nu \in \mathcal{P}_2(\mathbb{R}^d)$. Choose optimal couplings $\pi_0 \in \Gamma_o(\nu, \mu_0)$ and $\pi_1 \in \Gamma_o(\nu, \mu_1)$, and glue them through their common marginal ν to obtain $\alpha \in \mathcal{P}_2((\mathbb{R}^d)^3)$ on coordinates (x, x_0, x_1), with (x, x_0)- and (x, x_1)-marginals π_0 and π_1, respectively. The curve

$$\mu_t := \big((1-t)x_0 + t x_1\big)_{\#}\alpha, \qquad t \in [0, 1],$$

is a generalized geodesic from μ_0 to μ_1 with base ν. We say that $\mathcal{H}$ is convex along generalized geodesics if, for every $\mu_0, \mu_1, \nu \in \mathrm{dom}(\mathcal{H})$, there exists a generalized geodesic $(\mu_t)_{t\in[0,1]}$ from μ_0 to μ_1 with base ν such that

$$\mathcal{H}(\mu_t) \le (1-t)\mathcal{H}(\mu_0) + t\mathcal{H}(\mu_1), \qquad t \in [0, 1].$$

The same coupling-based geometry also provides the first-order object used in our optimality conditions. The extended Wasserstein subdifferential of Ambrosio et al. (2008) is naturally plan valued: an element

$$\boldsymbol{\xi} \in \boldsymbol{\partial}\mathcal{H}(\mu)$$

is a probability law on position–velocity pairs $(x, v) \in \mathbb{R}^d \times \mathbb{R}^d$ whose first marginal is μ. The precise variational definition is recalled in the appendix. This formulation remains meaningful even when no single optimal transport map represents the relevant displacement. In particular, by the Euler condition for Wasserstein Moreau–Yosida minimizers (Ambrosio et al., 2008, Lemma 10.3.4), if

$$\mu^+ \in \operatorname*{arg\,min}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \mathcal{H}(\mu) + \frac{1}{2\gamma} W_2^2(\mu, \eta) \right\},$$

then, for any $\pi \in \Gamma_o(\eta, \mu^+)$,

$$\left((z, y) \mapsto \left(y, \frac{z - y}{\gamma} \right) \right)_{\#} \pi \in \boldsymbol{\partial}\mathcal{H}(\mu^+).$$

This plan-valued subgradient is the one used later in the Wasserstein KKT system and the associated residual in Section 4.

Weak convergence. Our full-sequence convergence result is formulated in a weaker topology. Following Naldi & Savaré (2021), we denote by $\mathcal{P}_2^w(\mathbb{R}^d)$ the space $\mathcal{P}_2(\mathbb{R}^d)$ endowed with the weak topology characterized below. We write

$$\mu_k \rightharpoonup_{\mathcal{P}_2^w} \mu$$

if

$$\int_{\mathbb{R}^d} \zeta \, \mathrm{d}\mu_k \longrightarrow \int_{\mathbb{R}^d} \zeta \, \mathrm{d}\mu$$

for every $\zeta \in C(\mathbb{R}^d)$ satisfying

$$\lim_{\|x\|\to\infty} \frac{|\zeta(x)|}{1+\|x\|^2} = 0.$$

With these geometric notions in place, we impose the standing assumptions used throughout the algorithmic and convergence analysis.

Assumption 2.1 (Standing assumptions)**.**

1. *$V : \mathbb{R}^d \to \mathbb{R}$ is convex and continuously differentiable, and ∇V is L-Lipschitz continuous.*
2. *$\mathcal{H} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ is proper, W_2-lower semicontinuous, convex along generalized geodesics, and satisfies*
$$\overline{\mathrm{dom}(\mathcal{H})} = \mathcal{P}_2(\mathbb{R}^d).$$
3. *There exists a saddle point $(\mu^\star, \lambda^\star) \in \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^m$ satisfying $\mathcal{C}(\mu^\star) = 0$ and*
$$\Phi(\mu^\star) \le \Phi(\nu) + \langle \lambda^\star, \mathcal{C}(\nu)\rangle \qquad \textit{for every } \nu \in \mathcal{P}_2(\mathbb{R}^d).$$

Remark 2.2. If $(\mu^\star, \lambda^\star)$ satisfies Assumption 2.1(3), then $\mu^\star$ is a global solution of (1). Indeed, for any feasible $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, we have $\mathcal{C}(\nu) = 0$, and hence

$$\Phi(\mu^\star) \le \Phi(\nu) + \langle \lambda^\star, \mathcal{C}(\nu)\rangle = \Phi(\nu).$$

3 Wasserstein linearized augmented Lagrangian method

For $\rho > 0$, define the augmented Lagrangian

$$\mathcal{L}_\rho(\mu, \lambda) = \int_{\mathbb{R}^d} V(x)\, \mathrm{d}\mu(x) + \mathcal{H}(\mu) + \langle \lambda, \mathcal{C}(\mu)\rangle + \frac{\rho}{2}\|\mathcal{C}(\mu)\|^2.$$

Directly minimizing $\mathcal{L}_\rho(\cdot, \lambda)$ would couple $\mathcal{H}$ with the multiplier and quadratic penalty terms, potentially sacrificing the tractability of its original JKO resolvent. W-LALM applies a Wasserstein forward step to the smooth part of $\mathcal{L}_\rho$, retains the original JKO step for $\mathcal{H}$, and updates the multiplier separately. Algorithm 1 presents the resulting scheme for problem (1).

Algorithm 1 Wasserstein linearized augmented Lagrangian method (W-LALM)

1: Choose $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, $\lambda_0 \in \mathbb{R}^m$, $\rho > 0$, and $\gamma > 0$ such that $\delta := 1 - \gamma L - \frac{3}{4}\gamma\rho\|A\|^2 > 0$.
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: $\quad c_k \leftarrow \mathcal{C}(\mu_k)$ and $\widehat{\lambda}_k \leftarrow \lambda_k + \rho c_k$.
4: $\quad P_k(x) \leftarrow x - \gamma(\nabla V(x) + A^\top \widehat{\lambda}_k)$ and $\eta_{k+1} \leftarrow (P_k)_{\#}\mu_k$.
5: $\quad \mu_{k+1} \leftarrow \operatorname*{arg\,min}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \mathcal{H}(\mu) + \frac{1}{2\gamma} W_2^2(\mu, \eta_{k+1}) \right\}$.
6: $\quad \lambda_{k+1} \leftarrow \lambda_k + \rho\mathcal{C}(\mu_{k+1})$.
7: **end for**

Remark 3.1. 1. For $\mu_k \in \mathcal{P}_2(\mathbb{R}^d)$, the moment $\mathcal{C}(\mu_k)$ is finite, and the global Lipschitz continuity of ∇V ensures that P_k has at most linear growth. Thus, $\eta_{k+1} = (P_k)_{\#}\mu_k \in \mathcal{P}_2(\mathbb{R}^d)$. By Items (2)–(3) of Assumption 2.1, each JKO subproblem has a unique minimizer, so Algorithm 1 is well defined for every initial pair $(\mu_0, \lambda_0) \in \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^m$.

2. If $A = 0$ and $b = 0$, the constraint disappears, and Algorithm 1 reduces to the standard Wasserstein proximal gradient method (Salim et al., 2020; Di Marino et al., 2025).
3. Suppose that $\mathcal{H}$ is induced by a proper nonsmooth convex function and that the iteration is initialized with a Dirac measure. Then all iterates remain Dirac measures, and Algorithm 1 recovers the Euclidean composite linearized augmented-Lagrangian method.

4 PLAN-BASED CONVERGENCE AND COMPLEXITY ANALYSIS

The convergence analysis is built around a descent inequality, whose derivation requires addressing several features specific to the Wasserstein setting. First, the relevant displacement fields at successive iterates are associated with different spaces $L^2(\mu_k)$, so the usual Hilbert-space subtraction of primal iterates is not directly available. Second, optimal transport maps need not exist, which prevents a map-based representation of the two-stage update in general. Third, the augmented-Lagrangian structure couples the feasibility residuals c_k and c_{k+1} arising from the forward and multiplier steps.

In the analysis below, we overcome these difficulties through a carefully constructed transport-plan framework, leading to the main primal–dual Lyapunov inequality. This estimate underpins the full-sequence convergence analysis and the iteration-complexity bound for the KKT residual.

4.1 PLAN-BASED LYAPUNOV DESCENT AND FULL-SEQUENCE CONVERGENCE

To establish the main Lyapunov estimate, we first represent the forward and JKO steps on a common transport plan and then combine the resulting plan-based comparison with the multiplier update. Fix $\pi_k \in \Gamma_o(\eta_{k+1}, \mu_{k+1})$. Since the forward coupling $(\mathrm{Id}, P_k)_{\#}\mu_k$ and π_k have the common marginal η_{k+1}, the gluing lemma (Ambrosio et al., 2008, Lemma 5.3.2) yields a plan $\alpha_k \in \mathcal{P}((\mathbb{R}^d)^3)$ whose (x, z)- and (z, y)-marginals are $(\mathrm{Id}, P_k)_{\#}\mu_k$ and π_k, respectively. Define

$$S_k := \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, \mathrm{d}\alpha_k(x, z, y).$$

The (x, y)-marginal of α_k couples μ_k and μ_{k+1}. Hence S_k measures the squared displacement induced by the glued two-stage update, and $W_2^2(\mu_k, \mu_{k+1}) \le S_k$, even though the (x, y)-marginal of α_k need not be optimal.

Having encoded the two-stage primal displacement through S_k, we next introduce the primal–dual energy quantities used to measure progress relative to a saddle point. For a saddle point $(\mu^\star, \lambda^\star)$, let

$$\begin{cases} E_k^\star = W_2^2(\mu_k, \mu^\star) + \frac{\gamma}{\rho} \|\lambda_k - \lambda^\star\|^2, \\ \widetilde{E}_k^\star = E_k^\star - \frac{\gamma\rho}{2} \|c_k\|^2, \\ h_{k+1}^\star = \Phi(\mu_{k+1}) - \Phi(\mu^\star) + \langle \lambda^\star, c_{k+1} \rangle. \end{cases}$$

Here $E_k^\star$ is the natural product-space distance, $\widetilde{E}_k^\star$ is its penalty-corrected form, and $h_{k+1}^\star$ is the saddle-Lagrangian gap. The saddle inequality makes the last quantity nonnegative. With these quantities in place, the following theorem establishes the main plan-based Lyapunov descent estimate.

Theorem 4.1 (Plan-based Lyapunov descent)**.** *Suppose that Assumption 2.1 holds and* $\delta := 1 - \gamma L - \frac{3}{4}\gamma\rho \|A\|^2 > 0$*. Then* $h_{k+1}^\star \ge 0$ *and*

$$\widetilde{E}_{k+1}^\star + \delta S_k + 2\gamma h_{k+1}^\star + \frac{\gamma\rho}{4} \|c_k + c_{k+1}\|^2 \le \widetilde{E}_k^\star. \tag{3}$$

Moreover, $\widetilde{E}_k^\star \ge \left(1 - \frac{\gamma\rho}{2} \|A\|^2\right) W_2^2(\mu_k, \mu^\star) + \frac{\gamma}{\rho} \|\lambda_k - \lambda^\star\|^2$.

Inequality (3) is the master estimate of the analysis. Its dissipative terms control the two-stage primal displacement, the saddle-Lagrangian gap, and consecutive feasibility residuals, while the lower bound ensures coercivity of the corrected energy.

Summing the Lyapunov inequality yields asymptotic regularity and feasibility; in particular, $\sum_k S_k < \infty$, $\sum_k \|c_k\|^2 < \infty$, $W_2(\mu_{k+1}, \mu_k) \to 0$, and $c_k \to 0$. The next theorem strengthens these properties to convergence of the full primal–dual sequence.

Theorem 4.2 (Full-sequence convergence)**.** *Suppose that Assumption 2.1 holds and* $\delta := 1 - \gamma L - \frac{3}{4}\gamma\rho \|A\|^2 > 0$*. Then there exists a saddle point* $(\mu^\infty, \lambda^\infty)$ *such that*

$$\mu_k \rightharpoonup_{\mathcal{P}_2^w} \mu^\infty, \quad \lambda_k \to \lambda^\infty, \quad \Phi(\mu_k) \to \Phi(\mu^\infty).$$

In particular, μ^∞ *is a global solution of* (1)*.*

4.2 PLAN-VALUED KKT CHARACTERIZATION AND RESIDUAL COMPLEXITY

We next formulate a Wasserstein analogue of the classical KKT system for linearly constrained composite optimization. In Euclidean space, the first-order stationarity condition is

$$0 \in \nabla V(x) + A^\top \lambda + \partial h(x).$$

In the Wasserstein setting, the extended subdifferential $\boldsymbol{\partial}\mathcal{H}(\mu)$ is plan valued, so the vector balance above is replaced by a balance on position–velocity plans.

Wasserstein KKT point. A pair $(\bar{\mu}, \bar{\lambda}) \in \mathrm{dom}(\mathcal{H}) \times \mathbb{R}^m$ is called a Wasserstein KKT point if

$$\mathcal{C}(\bar{\mu}) = 0, \quad (\mathrm{Id}, -\nabla V - A^\top \bar{\lambda})_{\#}\bar{\mu} \in \boldsymbol{\partial}\mathcal{H}(\bar{\mu}). \tag{4}$$

Relation to existing formulations. The condition (4) differs from existing Wasserstein optimality formulations in two respects. The map-valued KKT condition of Lanzetti et al. (2025) considers a single differentiable scalar constraint, whereas our formulation uses an extended plan-valued subdifferential and a vector multiplier. The general plan-valued normal-cone condition of Lanzetti et al. (2024) is more abstract; for linear expectation constraints, (4) makes the stationarity balance explicit through the term $A^\top \lambda$.

The following theorem establishes the equivalence between this explicit plan-valued KKT system and the saddle-point characterization.

Theorem 4.3 (KKT characterization)**.** *Under Assumption 2.1 (1)–(2), a pair $(\bar{\mu}, \bar{\lambda}) \in \mathrm{dom}(\mathcal{H}) \times \mathbb{R}^m$ is a Wasserstein KKT point if and only if*

$$\mathcal{C}(\bar{\mu}) = 0, \quad \Phi(\bar{\mu}) \le \Phi(\nu) + \left\langle \bar{\lambda}, \mathcal{C}(\nu) \right\rangle \quad \textit{for every } \nu \in \mathcal{P}_2(\mathbb{R}^d).$$

Consequently, the primal component of every Wasserstein KKT point solves (1)*.*

Hence the limit in Theorem 4.2 is also a Wasserstein KKT point. To quantify first-order optimality for finite iterates, we next introduce an ε-KKT criterion based on a plan-valued residual.

ε-KKT points. For $\boldsymbol{\xi} \in \boldsymbol{\partial}\mathcal{H}(\mu)$, define the squared plan-valued KKT residual

$$\mathcal{R}^2(\mu, \lambda; \boldsymbol{\xi}) := \|\mathcal{C}(\mu)\|^2 + \int_{\mathbb{R}^d \times \mathbb{R}^d} \left\| v + \nabla V(u) + A^\top \lambda \right\|^2 \mathrm{d}\boldsymbol{\xi}(u, v). \tag{5}$$

We call (μ, λ) an ε-KKT point if there exists $\boldsymbol{\xi} \in \boldsymbol{\partial}\mathcal{H}(\mu)$ such that $\mathcal{R}^2(\mu, \lambda; \boldsymbol{\xi}) \le \varepsilon$. Thus, the residual jointly measures feasibility and stationarity, with $\varepsilon = 0$ recovering the Wasserstein KKT system (4).

For each W-LALM iterate, the chosen JKO coupling π_k induces the plan-valued subgradient

$$\boldsymbol{\xi}_{k+1} := \left((z, y) \mapsto \left(y, \frac{z - y}{\gamma} \right) \right)_{\#} \pi_k \in \boldsymbol{\partial}\mathcal{H}(\mu_{k+1}),$$

and we set $\mathcal{R}^2_{k+1} := \mathcal{R}^2(\mu_{k+1}, \lambda_{k+1}; \boldsymbol{\xi}_{k+1})$.

We now quantify how rapidly W-LALM attains approximate first-order optimality in terms of this residual.

Theorem 4.4 (KKT residual complexity)**.** *Under the conditions of Theorem 4.1, there exists a constant $M > 0$, depending on the problem and algorithmic parameters, such that*

$$\frac{1}{N} \sum_{k=0}^{N-1} \mathcal{R}^2_{k+1} \le \frac{M}{N}, \qquad N \ge 1,$$

and, for every $\varepsilon > 0$, there is an index $0 \le k < \lceil M/\varepsilon \rceil$ with $\mathcal{R}^2_{k+1} \le \varepsilon$. Consequently, W-LALM produces an ε-KKT iterate among its first $\lceil M/\varepsilon \rceil$ iterates and therefore has $O(1/\varepsilon)$ iteration complexity.

5 EXPERIMENTS

We compare W-LALM with Wasserstein proximal gradient with constrained JKO (CWPG), primal–dual Langevin Monte Carlo (PD-LMC), and Euclidean Proximal-ALM on generalized-Bayesian calibration. The comparison examines the cost of enforcing linear expectation constraints inside or outside the JKO weight subproblem, together with terminal feasibility and the discrete objective. All methods were implemented in Python 3.14.6 and run on a Windows 10 desktop with an Intel i5-10210U CPU and 8 GB RAM. Code will be released after anonymous review.

Particle implementation. For this experiment, $\mathcal{H}(\mu) = \beta \operatorname{Ent}(\mu)$. We approximate the probability law by $\mu^N = \sum_{i=1}^N w_i \delta_{u_i}$, with $w \in \Delta_N =: \{w \geq 0 : \sum_{i=1}^N w_i = 1\}$, and replace the continuous entropy by the discrete weight-entropy surrogate $\operatorname{Ent}_N(w) = \sum_{i=1}^N w_i \log(N w_i)$. After transporting the particle support to $\widetilde{U}^{k+1}$, the finite-support implementation of W-LALM solves the unconstrained weight subproblem

$$w^{k+1} \in \operatorname*{arg\,min}_{w \in \Delta_N} \left\{ \frac{\mathsf{W}_N^2(w, w^k; \widetilde{U}^{k+1})}{2\gamma} + \beta \sum_i w_i \log(N w_i) \right\}. \tag{6}$$

By contrast, CWPG enforces the linear expectation constraint within its corresponding weight subproblem, requiring $A \sum_i w_i \widetilde{u}_i^{k+1} = b$. As a Euclidean alternative, Proximal-ALM directly optimizes particle locations and weights through an augmented-Lagrangian formulation:

$$\min_{\substack{u_1,\dots,u_N \\ w \in \Delta_N}} \sum_i w_i V(u_i) + \beta \sum_i w_i \log(N w_i) \quad \text{s.t. } A \sum_i w_i u_i = b. \tag{7}$$

Thus, W-LALM retains an unconstrained convex weight subproblem, whereas Proximal-ALM alternates between particle locations and weights in a generally nonconvex joint parametrization. Implementation details for all baselines are provided in Appendix D.

To compare computational performance under a common criterion, we use the movement–feasibility stopping residual

$$r_k = \left[D_k + \|\mathcal{C}(\mu_{k+1})\|^2 \right]^{1/2}, \quad D_k = \sum_{i,j} \Pi_{k,ij} \|u_{k,i} - u_{k+1,j}\|^2.$$

Here, Π_k transports common mass along persistent particle labels and greedily matches the remaining mass, so $W_2^2(\mu_k^N, \mu_{k+1}^N) \leq D_k$. A run stops when $r_k < 10^{-5}$ or reaches the wall-clock limit of 300 seconds. Reported speedup factors are rounded down to integers.

Generalized Bayesian calibration. We consider polynomial regression with a generalized-Bayes objective and linear calibration of the mean parameter (Bissiri et al., 2016). We use the Wine Quality Red (Cortez et al., 2009) and Concrete Compressive Strength (Yeh, 1998) datasets. The Wine Red instance has dimension $d = 110$, 140 particles, and 50 constraints, while the Concrete instance has dimension $d = 110$, 140 particles, and 100 constraints.

Table 2 reports the runtime, terminal discrete objective, stopping residual, and feasibility status. A checkmark indicates that the terminal moment residual satisfies $\|A\bar{u} - b\|_2 < 10^{-8}$, where $\bar{u}$ denotes the particle mean.

Table 2: Generalized-Bayes calibration: discrete objective, stopping residual, feasibility, and runtime.

Dataset	Method	Time (s)	Objective	Final r_k	Feasible
Wine Red	W-LALM	**1.10**	2.82×10^{-2}	9.97×10^{-6}	✓
	CWPG	9.65	2.82×10^{-2}	8.54×10^{-6}	✓
	PD-LMC	300.00*	3.36×10^{-2}	1.67×10^{-2}	–
	Proximal-ALM	186.65	2.83×10^{-2}	9.05×10^{-6}	✓
Concrete	W-LALM	**0.88**	4.33×10^{-2}	9.39×10^{-6}	✓
	CWPG	34.39	4.33×10^{-2}	9.32×10^{-6}	✓
	PD-LMC	300.00*	4.87×10^{-2}	2.61×10^{-2}	–
	Proximal-ALM	72.31	4.34×10^{-2}	9.85×10^{-6}	✓

∗ Reached the 300-second cap before satisfying the stopping criterion.

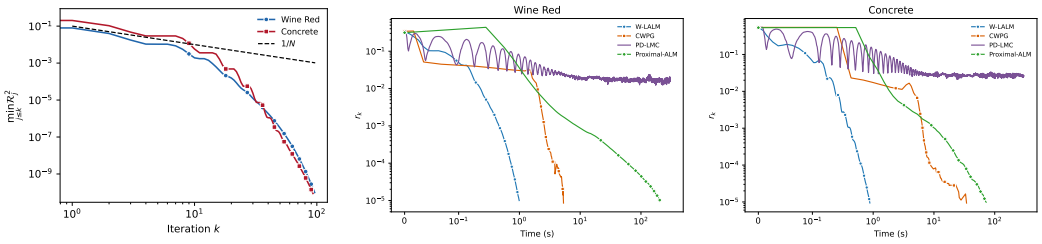

Figure 1: Left: cumulative minimum squared stopping residual for W-LALM on the two calibration data sets, with a $1/N$ reference slope. Middle and right: r_k versus time for all four methods on Wine Red and Concrete, respectively.

On the finite-particle calibration problem, W-LALM and CWPG reach the stopping and mean-feasibility thresholds with nearly identical discrete objectives. W-LALM avoids the inner iterations needed to enforce the moment constraint in CWPG's weight subproblem. Proximal-ALM also reaches both thresholds but spends more time on its coupled location–weight updates. PD-LMC has a higher terminal discrete objective and does not reach the mean-feasibility threshold within the time limit. Each W-LALM or CWPG outer step solves one entropically regularized transport-weight problem by Sinkhorn iteration. CWPG additionally enforces the moment constraint inside each Sinkhorn row update with a damped Newton solve; its Newton systems use conjugate gradients (CG). Table 3 separates these nested iteration counts. The average time per Sinkhorn iteration is the total runtime in Table 2 divided by the cumulative Sinkhorn count; it is an amortized cost that also includes the outer-step work.

Table 3: Generalized-Bayes calibration: outer steps and cumulative inner work for W-LALM and CWPG.

Dataset	Method	Outer steps	Sinkhorn steps	Cumulative CG iterations	Time/Sinkhorn (ms)
Wine Red	W-LALM	96	294	–	3.74
	CWPG	89	538	2.56×10^5	17.94
Concrete	W-LALM	93	318	–	2.78
	CWPG	87	761	1.89×10^6	45.19

The CG count is cumulative over all constrained Sinkhorn updates and all Newton systems solved within those updates. CWPG uses slightly fewer outer steps, but the 100-constraint Concrete run incurs 1.89×10^6 cumulative CG iterations and is substantially slower than W-LALM. The unconstrained W-LALM weight solve avoids this constraint-projection overhead.

6 Conclusion

W-LALM handles linear expectation constraints while keeping the JKO subproblem unconstrained. The method separates the explicit treatment of the smooth potential, the Wasserstein proximal step for the composite term, and the multiplier correction for feasibility. A plan-based Lyapunov analysis establishes full-sequence convergence to a saddle point, and a plan-valued KKT characterization leads to an $O(1/\varepsilon)$ KKT-residual complexity bound. The particle experiments support the computational motivation: across the two generalized-Bayesian calibration tasks, W-LALM reduces runtime relative to CWPG while retaining comparable task metrics. This suggests that moving feasibility outside the JKO solve can be an effective design principle for constrained Wasserstein optimization. Our theory assumes generalized-geodesic convexity, the existence of a saddle point, and a step-size condition determined by global smoothness constants, and the particle discretization contributes an error that the analysis does not control. Relaxing these requirements, together with adaptive penalty and step-size rules and extensions beyond linear expectation constraints, is an interesting direction for future work.

AI USE STATEMENT

Generative AI tools were used to refine and check the language of the manuscript, including a language-editing pass on the present revision, and to assist with LaTeX consistency checks. These tools were not used to alter the mathematical formulation or reasoning of the main results, nor to generate experimental data. The authors reviewed the AI-assisted text and take responsibility for the submitted content.

ETHICS STATEMENT

The experiments use established public tabular benchmark data. They involve no human-subject intervention and make no deployment decision.

REPRODUCIBILITY STATEMENT

The appendix contains the complete Wasserstein definitions, theorem proofs, particle discretization, and subproblem solvers. All datasets, solver parameters, stopping rules, and computational budgets are specified in Section 5 and Appendix D, and the code will be released after anonymous review.

REFERENCES

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Birkhäuser, 2 edition, 2008.

Jean-David Benamou, Guillaume Carlier, and Maxime Laborde. An augmented Lagrangian approach to Wasserstein gradient flows and applications. *ESAIM: Proceedings and Surveys*, 54: 1–17, 2016. doi: 10.1051/proc/201654001.

Ernesto Julián Goldberg Birgin and José Mario Martínez. *Practical Augmented Lagrangian Methods for Constrained Optimization*, volume 10 of *Fundamentals of Algorithms*. Society for Industrial and Applied Mathematics, 2014. doi: 10.1137/1.9781611973365.

Pier Giovanni Bissiri, Chris C. Holmes, and Stephen G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B*, 78:1103–1130, 2016. doi: 10.1111/rssb.12158.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146:459–494, 2014. doi: 10.1007/s10107-013-0701-9.

Clément Bonet, Théo Uscidda, Adam David, Pierre-Cyril Aubin-Frankowski, and Anna Korba. Mirror and preconditioned gradient descent in Wasserstein space. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0798.

Eric Carlen and Wilfrid Gangbo. Constrained steepest descent in the 2-Wasserstein metric. *Annals of Mathematics*, 157:807–846, 2003. doi: 10.4007/annals.2003.157.807.

Luiz Fernando De Oliveira Chamon, Mohammad Reza Karimi, and Anna Korba. Constrained sampling with primal-dual Langevin Monte Carlo. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Laurent Condat. A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 158: 460–479, 2013. doi: 10.1007/s10957-012-0245-9.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Wine quality [dataset]. UCI Machine Learning Repository, 2009.

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

José Antonio Carrillo de la Plata, Katy Craig, Li Wang, and Chaozhen Wei. Primal dual methods for Wasserstein gradient flows. *Foundations of Computational Mathematics*, 22:389–443, 2022. doi: 10.1007/s10208-021-09503-1.

Simone Di Marino, Emanuele Naldi, and Silvia Villa. Inexact JKO and proximal-gradient algorithms in the Wasserstein space. *arXiv:2505.23517*, 2025. URL https://arxiv.org/abs/2505.23517.

Michael Ziyang Diao, Krishna Balasubramanian, Sinho Chewi, and Adil Salim. Forward-backward Gaussian variational inference via JKO in the Bures–Wasserstein space. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 7960–7991, 2023.

Geir Evensen. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, 53:343–367, 2003. doi: 10.1007/s10236-003-0036-9.

Stanisław Marek Sergiusz Halkiewicz and Vicenţiu D. Rădulescu. A constrained Wasserstein gradient flow for aggregation–diffusion dynamics with application to economic adjustment. *Royal Society Open Science*, 13(5):260620, 2026. doi: 10.1098/rsos.260620.

Deren Han. A survey on some recent developments of alternating direction method of multipliers. *Journal of the Operations Research Society of China*, 10:1–52, 2022. doi: 10.1007/s40305-021-00368-3.

Bingsheng He, Feng Ma, and Xiaoming Yuan. Optimal proximal augmented Lagrangian method and its application to full Jacobian splitting for multi-block separable convex minimization problems. *IMA Journal of Numerical Analysis*, 40:1188–1216, 2020. doi: 10.1093/imanum/dry092.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29:1–17, 1998. doi: 10.1137/S0036141096303359.

Nicolas Lanzetti, Antonio Terpin, and Florian Dörfler. Variational analysis in the Wasserstein space. *arXiv:2406.10676*, 2024. URL https://arxiv.org/abs/2406.10676.

Nicolas Lanzetti, Saverio Bolognani, and Florian Dörfler. First-order conditions for optimization in the Wasserstein space. *SIAM Journal on Mathematics of Data Science*, 7:274–300, 2025. doi: 10.1137/23M156687X.

Xingchao Liu, Xin Thomson Tong, and Qiang Liu. Sampling with trustworthy constraints: A variational gradient framework. In *Advances in Neural Information Processing Systems*, volume 34, pp. 23557–23568, 2021.

Hoang Phuc Hau Luu, Hanlin Yu, Bernardo Williams, Marcelo Hartmann, and Arto Klami. Stochastic variance-reduced Gaussian variational inference on the Bures–Wasserstein manifold. In *International Conference on Learning Representations*, pp. 29544–29568, 2025.

Emanuele Naldi and Giuseppe Savaré. Weak topology and Opial property in Wasserstein spaces, with applications to gradient flows and proximal point algorithms of geodesically convex functionals. *Rendiconti Lincei. Matematica e Applicazioni*, 32:725–750, 2021. doi: 10.4171/RLM/955.

Boris Teodorovich Polyak. Iterative methods using lagrange multipliers for solving extremal problems with constraints of the equation type. *USSR Computational Mathematics and Mathematical Physics*, 10:42–52, 1970.

Adil Salim, Anna Korba, and Giulia Luise. The Wasserstein proximal gradient algorithm. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12356–12366, 2020.

Ming Yan. A new primal–dual algorithm for minimizing the sum of three functions with a linear operator. *Journal of Scientific Computing*, 76:1698–1717, 2018. doi: 10.1007/s10915-018-0680-3.

Junfeng Yang and Xiaoming Yuan. Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization. *Mathematics of Computation*, 82:301–329, 2013. doi: 10.1090/S0025-5718-2012-02598-1.

I-Cheng Yeh. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*, 28:1797–1808, 1998. doi: 10.1016/S0008-8846(98)00165-3.

A GUIDE TO THE APPENDIX

The appendix preserves the complete mathematical development of the manuscript: Wasserstein preliminaries, the exact algorithm, assumptions, plan EVI, Lyapunov proof, product-space Opial argument, plan-valued KKT equivalence, and residual-complexity proof.

B WASSERSTEIN PRELIMINARIES AND FULL PROOFS

Throughout this appendix, the primal space is $\mathbb{R}^d$ endowed with a fixed inner product and the multiplier space is $\mathbb{R}^m$ with its Euclidean inner product. For $x, y, z, w \in \mathbb{R}^d$, we use the corresponding inner-product identity

$$2\langle x-y, z-w\rangle = \|x-w\|^2 + \|y-z\|^2 - \|x-z\|^2 - \|y-w\|^2. \tag{8}$$

For $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, the vector integral in $\mathcal{C}(\mu)$ is well defined because $\int \|Ax\|\,\mathrm{d}\mu(x) \le \|A\|\,(\int \|x\|^2\,\mathrm{d}\mu)^{1/2} < \infty$. For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, let $\Gamma(\mu,\nu)$ denote the set of couplings between μ and ν. The quadratic Wasserstein distance between μ and ν is defined by

$$W_2^2(\mu,\nu) := \inf_{\pi\in\Gamma(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2\,\mathrm{d}\pi(x,y).$$

Let $\Gamma_o(\mu,\nu)$ denote the set of quadratic optimal couplings. Since $\mathbb{R}^d$ is Polish, quadratic optimal couplings exist.

For an extended-real functional $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$, its effective domain is $\mathrm{dom}(\mathcal{F}) := \{\mu \in \mathcal{P}_2(\mathbb{R}^d) : \mathcal{F}(\mu) < +\infty\}$. We say that $\mathcal{F}$ is proper if $\mathrm{dom}(\mathcal{F}) \neq \emptyset$. It is W_2-lower semicontinuous if $W_2(\mu_n,\mu) \to 0$ implies $\mathcal{F}(\mu) \le \liminf_{n\to\infty} \mathcal{F}(\mu_n)$. For a finite product space with m components, we denote by ${}^m\pi^i$ the projection onto the i-th component and, more generally, by ${}^m\pi^{i_1,\dots,i_r}$ the projection onto the listed components $(i_1,\dots,i_r)$. For a Borel map T, $T_\#\mu$ denotes the pushforward of μ by T. Equivalently, for every $T_\#\mu$-integrable Borel function f,

$$\int_Y f(y)\,\mathrm{d}(T_\#\mu)(y) = \int_X f(T(x))\,\mathrm{d}\mu(x). \tag{9}$$

Lemma B.1 (Gluing lemma)**.** (Ambrosio et al., 2008, Lemma 5.3.2) *The following gluing statements hold.*

1. *Let $\mu_1,\mu_2,\mu_3 \in \mathcal{P}_2(\mathbb{R}^d)$. If $\gamma^{12} \in \Gamma(\mu_1,\mu_2)$ and $\gamma^{23} \in \Gamma(\mu_2,\mu_3)$, then there exists $\alpha \in \mathcal{P}_2(\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d)$ such that*
$$({}^3\pi^{1,2})_\#\alpha = \gamma^{12}, \quad ({}^3\pi^{2,3})_\#\alpha = \gamma^{23}.$$
2. *Let $\alpha^{12},\alpha^{13} \in \mathcal{P}_2(\mathbb{R}^d\times\mathbb{R}^d)$ have the same first marginal. Then there exists $\beta \in \mathcal{P}_2(\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d)$ such that*
$$({}^3\pi^{1,2})_\#\beta = \alpha^{12}, \quad ({}^3\pi^{1,3})_\#\beta = \alpha^{13}.$$

We first recall the basic Wasserstein notions and variational tools used throughout the paper.

Definition B.2 (Generalized geodesic with a base measure)**.** (Naldi & Savaré, 2021, Definition 6.4) Let $\nu,\mu_0,\mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$, and let $\boldsymbol{\gamma} \in \mathcal{P}_2(\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d)$ satisfy

$$({}^3\pi^{1,2})_\#\boldsymbol{\gamma} \in \Gamma_o(\nu,\mu_0), \quad ({}^3\pi^{1,3})_\#\boldsymbol{\gamma} \in \Gamma_o(\nu,\mu_1).$$

For $t \in [0,1]$, define

$$\mu_t := \big((1-t)^3\pi^2 + t^3\pi^3\big)_\#\boldsymbol{\gamma}. \tag{10}$$

The curve $(\mu_t)_{t\in[0,1]}$ is called a generalized geodesic connecting μ_0 and μ_1 with base measure ν.

The endpoint marginal $({}^3\pi^{2,3})_\#\boldsymbol{\gamma}$ need not be optimal between μ_0 and μ_1. Nevertheless, using the Euclidean identity $\|(1-t)a+tb\|^2 = (1-t)\|a\|^2 + t\|b\|^2 - t(1-t)\|a-b\|^2$ together with the optimality of the $(1,2)$ and $(1,3)$ marginals yields

$$W_2^2(\nu,\mu_t) \le (1-t)W_2^2(\nu,\mu_0) + tW_2^2(\nu,\mu_1) - t(1-t)W_2^2(\mu_0,\mu_1). \tag{11}$$

Definition B.3 (Convexity along generalized geodesics)**.** (Naldi & Savaré, 2021, Definition 6.4) A proper functional $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ is convex along generalized geodesics if, for every $\nu, \mu_0, \mu_1 \in \mathrm{dom}(\mathcal{F})$, there exists a plan $\boldsymbol{\gamma}$ as in Lemma B.2 such that the associated curve (10) satisfies

$$\mathcal{F}(\mu_t) \le (1-t)\mathcal{F}(\mu_0) + t\mathcal{F}(\mu_1) \qquad \text{for every } t \in [0,1].$$

Definition B.4 (Extended Fréchet subdifferential)**.** (Ambrosio et al., 2008, Definition 10.3.1) Let $\boldsymbol{\xi} \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ satisfy $({}^2\pi^1)_\#\boldsymbol{\xi} = \mu$. For $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, set

$$\Gamma_o^{1,3}(\boldsymbol{\xi}, \nu) := \left\{ \boldsymbol{\alpha} \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d) : ({}^3\pi^{1,2})_\#\boldsymbol{\alpha} = \boldsymbol{\xi},\ ({}^3\pi^{1,3})_\#\boldsymbol{\alpha} \in \Gamma_o(\mu, \nu) \right\}.$$

The plan $\boldsymbol{\xi}$ belongs to the extended Fréchet subdifferential $\boldsymbol{\partial}\mathcal{F}(\mu)$ if

$$\mathcal{F}(\nu) - \mathcal{F}(\mu) \ge \inf_{\boldsymbol{\alpha} \in \Gamma_o^{1,3}(\boldsymbol{\xi}, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \langle v, y - x \rangle \, \mathrm{d}\boldsymbol{\alpha}(x, v, y) + o(W_2(\mu, \nu)) \tag{12}$$

as $\nu \to \mu$ in W_2. In addition, by Theorem 10.3.6 in Ambrosio et al. (2008), if $\mathcal{F}$ is lower semicontinuous and convex along generalized geodesics, then the following global variational characterization holds: for every $\nu \in \mathrm{dom}(\mathcal{F})$, there exists $\boldsymbol{\alpha} \in \Gamma_o^{1,3}(\boldsymbol{\xi}, \nu)$ such that

$$\mathcal{F}(\nu) - \mathcal{F}(\mu) \ge \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \langle v, y - x \rangle \, \mathrm{d}\boldsymbol{\alpha}(x, v, y). \tag{13}$$

Lemma B.5 (JKO resolvent properties)**.** *Let $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ be proper, W_2-lower semicontinuous, and convex along generalized geodesics. For every $\eta \in \overline{\mathrm{dom}(\mathcal{F})}$ and $\gamma > 0$, the JKO problem admits a unique minimizer μ^+ (Naldi & Savaré, 2021, Theorem 6.6), namely*

$$\mu^+ = \operatorname*{arg\,min}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \mathcal{F}(\mu) + \frac{1}{2\gamma} W_2^2(\mu, \eta) \right\}. \tag{14}$$

Moreover, the following two properties hold.

1. *For every $\widetilde{\pi} \in \Gamma_o(\mu^+, \eta)$, the $p = 2$ JKO optimality condition (Ambrosio et al., 2008, Lemma 10.3.4) gives*

$$\boldsymbol{\xi}^+ := \left((y, z) \mapsto \left(y, \frac{z - y}{\gamma} \right) \right)_\# \widetilde{\pi} \in \boldsymbol{\partial}\mathcal{F}(\mu^+). \tag{15}$$

   Equivalently, if $\pi \in \Gamma_o(\eta, \mu^+)$, then (15) *holds with $\widetilde{\pi} = ((z, y) \mapsto (y, z))_\#\pi$.*

2. *For every $\nu \in \mathrm{dom}(\mathcal{F})$, the standard one-step JKO inequality (Naldi & Savaré, 2021, Theorem 6.6) gives*

$$W_2^2(\mu^+, \nu) \le W_2^2(\eta, \nu) - W_2^2(\eta, \mu^+) - 2\gamma\big[\mathcal{F}(\mu^+) - \mathcal{F}(\nu)\big]. \tag{16}$$

We next recall the Wasserstein weak topology introduced by Naldi and Savaré Naldi & Savaré (2021), which will be used in the compactness and convergence analysis.

Define the function space

$$\mathcal{C}_2^w(\mathbb{R}^d) := \left\{ \zeta : \mathbb{R}^d \to \mathbb{R} : \zeta \text{ is sequentially weakly continuous, and } \lim_{\|x\| \to \infty} \frac{|\zeta(x)|}{1 + \|x\|^2} = 0 \right\}.$$

The topology of $\mathcal{P}_2^w(\mathbb{R}^d)$ is the initial topology induced by the maps $\mu \mapsto \int \zeta \, \mathrm{d}\mu$ with $\zeta \in \mathcal{C}_2^w(\mathbb{R}^d)$. We write $\mu_n \rightharpoonup_{\mathcal{P}_2^w} \mu$ for convergence in this topology. With respect to this topology, sequential weak lower semicontinuity means that whenever $\mu_n \rightharpoonup_{\mathcal{P}_2^w} \mu$, one has $\mathcal{F}(\mu) \le \liminf_{n\to\infty} \mathcal{F}(\mu_n)$.

Lemma B.6 (Weak topology on $\mathcal{P}_2$)**.** (Naldi & Savaré, 2021, Corollary 3.6, Theorems 4.2, 5.1) *The following facts hold on $\mathcal{P}_2(\mathbb{R}^d)$.*

1. *Every W_2-bounded sequence is relatively sequentially compact in $\mathcal{P}_2^w(\mathbb{R}^d)$.*

2. *For every fixed $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, the map $\mu \mapsto W_2^2(\mu, \nu)$ is sequentially lower semicontinuous with respect to $\mathcal{P}_2^w(\mathbb{R}^d)$. Moreover, every proper W_2-lower-semicontinuous functional that is convex along generalized geodesics is sequentially lower semicontinuous with respect to $\mathcal{P}_2^w(\mathbb{R}^d)$.*

3. *If $\mu_n \rightharpoonup_{\mathcal{P}_2^w} \mu$, then, for every $\nu \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$W_2^2(\mu, \nu) + \liminf_{n\to\infty} W_2^2(\mu_n, \mu) \le \liminf_{n\to\infty} W_2^2(\mu_n, \nu). \tag{17}$$

C Complete algorithm and convergence analysis

Algorithm 2 presents the Wasserstein linearized augmented Lagrangian method. The smooth part of the augmented Lagrangian is treated by an explicit forward map, while $\mathcal{H}$ is treated by the JKO step (18).

Algorithm 2 Wasserstein linearized augmented Lagrangian method (W-LALM)

1: Choose $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, $\lambda_0 \in \mathbb{R}^m$, $\gamma > 0$, and $\rho > 0$ such that $\delta := 1 - \gamma L - \frac{3}{4}\gamma\rho \|A\|^2 > 0$.
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Set $c_k := \mathcal{C}(\mu_k)$ and $\widehat{\lambda}_k := \lambda_k + \rho c_k$.
4: Define $P_k(x) := x - \gamma(\nabla V(x) + A^\top \widehat{\lambda}_k)$ and $\eta_{k+1} := (P_k)_{\#}\mu_k$.
5: Compute the exact JKO step

$$\mu_{k+1} \in \operatorname*{arg\,min}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \mathcal{H}(\mu) + \frac{1}{2\gamma} W_2^2(\mu, \eta_{k+1}) \right\}. \tag{18}$$

6: Update $\lambda_{k+1} := \lambda_k + \rho c_{k+1}$.
7: **end for**

Assumption C.1. *The following assumptions are used throughout the convergence analysis.*

1. *$V : \mathbb{R}^d \to \mathbb{R}$ is convex and continuously differentiable, and ∇V is L-Lipschitz.*
2. *$\mathcal{H} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ is proper, W_2-lower semicontinuous, and convex along generalized geodesics.*
3. *$\overline{\mathrm{dom}(\mathcal{H})} = \mathcal{P}_2(\mathbb{R}^d)$.*
4. *The constrained problem admits a saddle point: there exists $(\mu^\star, \lambda^\star) \in \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^m$ such that*

$$\mathcal{C}(\mu^\star) = 0, \quad \Phi(\mu^\star) \le \Phi(\nu) + \langle \lambda^\star, \mathcal{C}(\nu) \rangle \quad \text{for every } \nu \in \mathcal{P}_2(\mathbb{R}^d). \tag{19}$$

Remark C.2 (Well-posedness of the JKO step). Suppose that $(\mu_k, \lambda_k) \in \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^m$. Since $\int \|x\| \, \mathrm{d}\mu_k(x) < \infty$, the moment $\mathcal{C}(\mu_k)$ and $\widehat{\lambda}_k$ are finite. Global L-Lipschitz continuity gives $\|\nabla V(x)\| \le \|\nabla V(0)\| + L\|x\|$, so

$$\|P_k(x)\| \le (1 + \gamma L)\|x\| + \gamma\big(\|\nabla V(0)\| + \|A\| \, \|\widehat{\lambda}_k\|\big).$$

Thus $\eta_{k+1} = (P_k)_{\#}\mu_k$ belongs to $\mathcal{P}_2(\mathbb{R}^d)$. Assumptions C.1(2)–(3) and Lemma B.5 then give a unique solution $\mu_{k+1} \in \mathcal{P}_2(\mathbb{R}^d)$ to (18). The multiplier update is finite, so $(\mu_{k+1}, \lambda_{k+1}) \in \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^m$. Starting from (μ_0, λ_0), induction proves that every step is well defined; the same lemma yields (16).

Remark C.3.

1. Since ∇V is globally Lipschitz, V has at most quadratic growth and the potential energy $\mu \mapsto \int V \, \mathrm{d}\mu$ is finite on $\mathcal{P}_2(\mathbb{R}^d)$, W_2-continuous, and convex along generalized geodesics. Hence Φ is W_2-lower semicontinuous and convex along generalized geodesics; in particular, it is sequentially lower semicontinuous with respect to $\mathcal{P}_2^w(\mathbb{R}^d)$.
2. If $(\mu^\star, \lambda^\star)$ satisfies (19), then $\mu^\star$ solves (1). Indeed, for every feasible ν one has $\mathcal{C}(\nu) = 0$, and hence $\Phi(\mu^\star) \le \Phi(\nu)$. The saddle-point condition is therefore a global optimality condition for the constrained problem.

Remark C.4.

1. If $A = 0$, feasibility requires $b = 0$, and the multiplier no longer enters the forward step. Algorithm 2 then reduces to the Wasserstein forward–backward iteration

$$\eta_{k+1} = (\mathrm{Id} - \gamma\nabla V)_{\#}\mu_k, \quad \mu_{k+1} = \operatorname*{arg\,min}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \mathcal{H}(\mu) + \frac{1}{2\gamma} W_2^2(\mu, \eta_{k+1}) \right\}.$$

Thus W-LALM contains the unconstrained Wasserstein proximal-gradient method as the zero-constraint case Salim et al. (2020); Di Marino et al. (2025).

2. A Euclidean special case is obtained by choosing a functional supported on Dirac measures. Let $h : \mathbb{R}^d \to (-\infty, +\infty]$ be proper, lower semicontinuous, and convex, and define

$$\mathcal{H}_{\mathrm{D}}(\mu) := \begin{cases} h(x), & \mu = \delta_x, \\ +\infty, & \text{otherwise.} \end{cases}$$

Then $\mathcal{H}_{\mathrm{D}}$ is proper, W_2-lower semicontinuous, and convex along generalized geodesics. Starting from $\mu_0 = \delta_{x_0}$, all iterates remain Dirac measures, $\mu_k = \delta_{x_k}$. Indeed, the forward step sends δ_{x_k} to a Dirac measure and the JKO step for $\mathcal{H}_{\mathrm{D}}$ is the Euclidean proximal map. Since $W_2(\delta_x, \delta_y) = \|x - y\|$ and $\mathcal{C}(\delta_x) = Ax - b$, Algorithm 2 becomes

$$\begin{aligned} c_k &= Ax_k - b, & \widehat{\lambda}_k &= \lambda_k + \rho c_k, \\ x_{k+1} &= \operatorname{prox}_{\gamma h}\big(x_k - \gamma(\nabla V(x_k) + A^\top \widehat{\lambda}_k)\big), & \lambda_{k+1} &= \lambda_k + \rho(Ax_{k+1} - b). \end{aligned}$$

Thus, on the class of Dirac measures, W-LALM reduces exactly to a Euclidean linearized augmented-Lagrangian method. Related proximal and linearized augmented-Lagrangian developments include He et al. (2020); Han (2022), and closely related Euclidean splitting methods include primal–dual schemes Condat (2013); Yan (2018).

C.1 Convergence analysis

For the analysis, we put the explicit forward step and the subsequent JKO step (18) on a common plan. Fix $\pi_k \in \Gamma_o(\eta_{k+1}, \mu_{k+1})$ and set $\gamma_k^{12} := (\mathrm{Id}, P_k)_\# \mu_k \in \Gamma(\mu_k, \eta_{k+1})$. Then

$$({}^2\pi^1)_\# \gamma_k^{12} = \mu_k, \qquad ({}^2\pi^2)_\# \gamma_k^{12} = \eta_{k+1} = ({}^2\pi^1)_\# \pi_k.$$

By Lemma B.1, there exists $\alpha_k \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d)$, with marginals μ_k, η_{k+1} and μ_{k+1}, such that

$$({}^3\pi^{1,2})_\# \alpha_k = \gamma_k^{12}, \qquad ({}^3\pi^{2,3})_\# \alpha_k = \pi_k.$$

The following estimate combines the JKO three-point inequality with the explicit forward displacement through this common plan.

Lemma C.5 (Plan EVI)**.** *Let $\{(\mu_k, \lambda_k)\}_{k \geq 0}$ be generated by Algorithm 2 under Assumption C.1 and the stepsize condition therein. Then, for every $\nu \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\begin{aligned} W_2^2(\mu_{k+1}, \nu) \leq & W_2^2(\mu_k, \nu) - 2\gamma\Big[\Phi(\mu_{k+1}) - \Phi(\nu) + \Big\langle \widehat{\lambda}_k, \mathcal{C}(\mu_{k+1}) - \mathcal{C}(\nu) \Big\rangle\Big] \\ & - (1 - \gamma L) S_k, \end{aligned} \tag{20}$$

where $S_k := \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, \mathrm{d}\alpha_k(x, z, y)$.

Proof. Fix $\nu \in \mathcal{P}_2(\mathbb{R}^d)$. If $\nu \notin \mathrm{dom}(\mathcal{H})$, then $\Phi(\nu) = +\infty$ and (20) is immediate in the extended-real sense. Hence we assume $\nu \in \mathrm{dom}(\mathcal{H})$. By Lemma B.5(ii), applied with $\mathcal{F} = \mathcal{H}$ and $\eta = \eta_{k+1}$, the JKO step (18) gives

$$W_2^2(\mu_{k+1}, \nu) \leq W_2^2(\eta_{k+1}, \nu) - W_2^2(\eta_{k+1}, \mu_{k+1}) - 2\gamma\big[\mathcal{H}(\mu_{k+1}) - \mathcal{H}(\nu)\big]. \tag{21}$$

We next estimate the difference of the two Wasserstein terms. Choose $\chi \in \Gamma_o(\mu_k, \nu)$. Since the first marginal of α_k and the first marginal of χ are both μ_k, Lemma B.1 yields $\beta_k \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d)$, with coordinates (x, z, y, w), such that

$$({}^4\pi^{1,2,3})_\# \beta_k = \alpha_k, \quad ({}^4\pi^{1,4})_\# \beta_k = \chi. \tag{22}$$

In particular,

$$({}^4\pi^{2,4})_\# \beta_k \in \Gamma(\eta_{k+1}, \nu),$$

whereas

$$({}^4\pi^{2,3})_\# \beta_k = ({}^3\pi^{2,3})_\# \alpha_k = \pi_k \in \Gamma_o(\eta_{k+1}, \mu_{k+1}).$$

Hence, by the definition of W_2 and the pushforward identity,

$$W_2^2(\eta_{k+1}, \nu) \leq \int_{\mathbb{R}^d \times \mathbb{R}^d} \|z - w\|^2 \, \mathrm{d}\big(({}^4\pi^{2,4})_\# \beta_k\big)(z, w) = \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \|z - w\|^2 \, \mathrm{d}\beta_k(x, z, y, w), \tag{23}$$

while the optimality of π_k gives

$$W_2^2(\eta_{k+1},\mu_{k+1}) = \int_{\mathbb{R}^d\times\mathbb{R}^d} \|z-y\|^2 \,\mathrm{d}\pi_k(z,y) = \int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \|z-y\|^2 \,\mathrm{d}\beta_k(x,z,y,w). \tag{24}$$

Subtracting (24) from (23), we obtain

$$W_2^2(\eta_{k+1},\nu) - W_2^2(\eta_{k+1},\mu_{k+1}) \le \int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \Big(\|z-w\|^2 - \|z-y\|^2 \Big) \,\mathrm{d}\beta_k(x,z,y,w). \tag{25}$$

We now use the structure of the forward step. From (22) and the definition of α_k,

$$({}^4\pi^{1,2})_\#\beta_k = (\mathrm{Id}, P_k)_\#\mu_k.$$

The definition of pushforward measure implies

$$\begin{aligned}
&\beta_k\big(\{(x,z,y,w)\in\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d : z = P_k(x)\}\big)\\
=&\beta_k\big(({}^4\pi^{1,2})^{-1}\{(x,z) : z = P_k(x)\}\big) = ({}^4\pi^{1,2})_\#\beta_k\big(\{(x,z) : z = P_k(x)\}\big)\\
=&(\mathrm{Id}, P_k)_\#\mu_k\big(\{(x,z) : z = P_k(x)\}\big) = \mu_k\big((\mathrm{Id}, P_k)^{-1}\{(x,z) : z = P_k(x)\}\big)\\
=&\mu_k(\mathbb{R}^d) = 1.
\end{aligned}$$

Therefore

$$z = P_k(x) = x - \gamma\big(\nabla V(x) + A^\top\widehat{\lambda}_k\big), \quad \beta_k\text{-a.s.} \tag{26}$$

Using (8) together with (26), we therefore have

$$\|z-w\|^2 - \|z-y\|^2 = \|x-w\|^2 - \|x-y\|^2 - 2\gamma\left\langle \nabla V(x) + A^\top\widehat{\lambda}_k, y-w\right\rangle, \quad \beta_k\text{-a.s.} \tag{27}$$

For the potential term, convexity of V gives

$$\langle \nabla V(x), x-w\rangle \ge V(x) - V(w),$$

while the L-Lipschitz continuity of ∇V yields, by the descent lemma,

$$\langle \nabla V(x), y-x\rangle \ge V(y) - V(x) - \frac{L}{2}\|y-x\|^2.$$

Hence

$$\langle \nabla V(x), y-w\rangle \ge V(y) - V(w) - \frac{L}{2}\|y-x\|^2. \tag{28}$$

Integrating with respect to β_k and using $({}^4\pi^3)_\#\beta_k = \mu_{k+1}, \quad ({}^4\pi^4)_\#\beta_k = \nu$, gives

$$\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \langle \nabla V(x), y-w\rangle \,\mathrm{d}\beta_k \ge \int_{\mathbb{R}^d} V \,\mathrm{d}\mu_{k+1} - \int_{\mathbb{R}^d} V \,\mathrm{d}\nu - \frac{L}{2}\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \|y-x\|^2 \,\mathrm{d}\beta_k. \tag{29}$$

For the constraint term, using again the marginal identities and the pushforward formula,

$$\begin{aligned}
\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \left\langle A^\top\widehat{\lambda}_k, y-w\right\rangle \mathrm{d}\beta_k &= \left\langle \widehat{\lambda}_k, \int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} A(y-w) \,\mathrm{d}\beta_k\right\rangle\\
&= \left\langle \widehat{\lambda}_k, \int_{\mathbb{R}}^d Ay \,\mathrm{d}\mu_{k+1}(y) - \int_{\mathbb{R}}^d Aw \,\mathrm{d}\nu(w)\right\rangle\\
&= \left\langle \widehat{\lambda}_k, \mathcal{C}(\mu_{k+1}) - \mathcal{C}(\nu)\right\rangle.
\end{aligned} \tag{30}$$

It remains to identify the two quadratic terms. Since $({}^4\pi^{1,4})_\#\beta_k = \chi \in \Gamma_o(\mu_k,\nu)$, the pushforward identity gives

$$\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \|x-w\|^2 \,\mathrm{d}\beta_k = \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-w\|^2 \,\mathrm{d}\chi(x,w) = W_2^2(\mu_k,\nu). \tag{31}$$

Integrating (27) and applying (29), (30), and (31), we obtain

$$\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \Big(\|z-w\|^2 - \|z-y\|^2 \Big)\, \mathrm{d}\beta_k \tag{32}$$

$$\leq W_2^2(\mu_k,\nu) - (1-\gamma L)S_k - 2\gamma\left[\int_{\mathbb{R}^d} V\, \mathrm{d}\mu_{k+1} - \int_{\mathbb{R}^d} V\, \mathrm{d}\nu + \left\langle \widehat{\lambda}_k, \mathcal{C}(\mu_{k+1}) - \mathcal{C}(\nu)\right\rangle\right]. \tag{33}$$

Together with (25), this yields

$$W_2^2(\eta_{k+1},\nu) - W_2^2(\eta_{k+1},\mu_{k+1})$$

$$\leq W_2^2(\mu_k,\nu) - (1-\gamma L)S_k - 2\gamma\left[\int_{\mathbb{R}^d} V\, \mathrm{d}\mu_{k+1} - \int_{\mathbb{R}^d} V\, \mathrm{d}\nu + \left\langle \widehat{\lambda}_k, \mathcal{C}(\mu_{k+1}) - \mathcal{C}(\nu)\right\rangle\right]. \tag{34}$$

Finally, substituting (34) into (21) completes the proof. □

We next establish a Lyapunov descent estimate relative to an arbitrary saddle point. For any reference point $(\nu,\zeta)\in\mathcal{P}_2(\mathbb{R}^d)\times\mathbb{R}^m$, define

$$E_k(\nu,\zeta) := W_2^2(\mu_k,\nu) + \frac{\gamma}{\rho}\|\lambda_k-\zeta\|^2, \quad \widetilde{E}_k(\nu,\zeta) := E_k(\nu,\zeta) - \frac{\gamma\rho}{2}\|c_k\|^2. \tag{35}$$

Let $(\mu^\star,\lambda^\star)$ be any saddle point satisfying (19), and abbreviate $E_k^\star := E_k(\mu^\star,\lambda^\star)$, $\widetilde{E}_k^\star := \widetilde{E}_k(\mu^\star,\lambda^\star)$. Set $h_{k+1}^\star := \Phi(\mu_{k+1}) - \Phi(\mu^\star) + \langle\lambda^\star, c_{k+1}\rangle$. Then $h_{k+1}^\star \geq 0$ by (19).

Theorem C.6 (Lyapunov inequality)**.** *Let $\{(\mu_k,\lambda_k)\}_{k\geq 0}$ be generated by Algorithm 2 under Assumption C.1 and the stepsize condition therein. Then*

$$\widetilde{E}_{k+1}^\star + \delta S_k + 2\gamma h_{k+1}^\star + \frac{\gamma\rho}{4}\|c_k+c_{k+1}\|^2 \leq \widetilde{E}_k^\star. \tag{36}$$

Moreover, with $\kappa := 1-\frac{\gamma\rho}{2}\|A\|^2 > 0$,

$$\widetilde{E}_k^\star \geq \kappa W_2^2(\mu_k,\mu^\star) + \frac{\gamma}{\rho}\|\lambda_k-\lambda^\star\|^2. \tag{37}$$

Proof. We give the algebra because S_k is a plan displacement, rather than an endpoint distance. Since $\mathcal{C}(\mu^\star)=0$, lemma C.5 with $\nu=\mu^\star$ reads

$$W_2^2(\mu_{k+1},\mu^\star) - W_2^2(\mu_k,\mu^\star) \leq -2\gamma\left[\Phi(\mu_{k+1}) - \Phi(\mu^\star) + \left\langle\widehat{\lambda}_k, c_{k+1}\right\rangle\right] - (1-\gamma L)S_k. \tag{38}$$

The dual update gives $\lambda_{k+1}-\lambda^\star = (\lambda_k-\lambda^\star)+\rho c_{k+1}$. Expanding its squared norm and multiplying by γ/ρ yields the exact identity

$$\frac{\gamma}{\rho}\big(\|\lambda_{k+1}-\lambda^\star\|^2 - \|\lambda_k-\lambda^\star\|^2\big) = 2\gamma\langle\lambda_k-\lambda^\star, c_{k+1}\rangle + \gamma\rho\|c_{k+1}\|^2.$$

Adding this identity to (38), splitting $\widehat{\lambda}_k = \lambda_k + \rho c_k$, and using the definition of $h_{k+1}^\star$ cancels the terms containing λ_k. We obtain

$$E_{k+1}^\star - E_k^\star \leq -2\gamma h_{k+1}^\star - (1-\gamma L)S_k - 2\gamma\rho\langle c_k, c_{k+1}\rangle + \gamma\rho\|c_{k+1}\|^2. \tag{39}$$

We first record the W_2-Lipschitz continuity of the linear moment map. For any $\mu,\nu\in\mathcal{P}_2(\mathbb{R}^d)$ and $\pi\in\Gamma(\mu,\nu)$, vector integration and Jensen's inequality give

$$\|\mathcal{C}(\mu)-\mathcal{C}(\nu)\|^2 = \left\|\int A(x-y)\, \mathrm{d}\pi\right\|^2 \leq \|A\|^2\int\|x-y\|^2\, \mathrm{d}\pi.$$

Taking the infimum over π yields

$$\|\mathcal{C}(\mu)-\mathcal{C}(\nu)\| \leq \|A\|\, W_2(\mu,\nu),\ \forall\mu,\nu\in\mathcal{P}_2(\mathbb{R}^d). \tag{40}$$

Since the (x,y)-marginal of α_k couples μ_k and μ_{k+1}, (40) gives

$$\|c_{k+1}-c_k\|^2 \leq \|A\|^2 W_2^2(\mu_k,\mu_{k+1}) \leq \|A\|^2 S_k.$$

Subtracting $\frac{\gamma\rho}{2}(\|c_{k+1}\|^2 - \|c_k\|^2)$ from both sides of (39) changes the left-hand side into $\widetilde{E}^\star_{k+1} - \widetilde{E}^\star_k$. The remaining constraint terms satisfy the exact identity

$$-2\langle c_k, c_{k+1}\rangle + \|c_{k+1}\|^2 - \frac{1}{2}\big(\|c_{k+1}\|^2 - \|c_k\|^2\big) = -\frac{1}{4}\|c_k + c_{k+1}\|^2 + \frac{3}{4}\|c_{k+1} - c_k\|^2.$$

Using the preceding bound on $c_{k+1} - c_k$ leaves the coefficient $1 - \gamma L - \frac{3}{4}\gamma\rho\|A\|^2 = \delta$ in front of S_k, and proves (36).

Finally, since $\mathcal{C}(\mu^\star) = 0$, (40) gives $\|c_k\| \le \|A\| W_2(\mu_k, \mu^\star)$ and hence (37). The stepsize condition implies $\gamma\rho\|A\|^2 < \frac{4}{3}(1 - \gamma L) \le \frac{4}{3}$, so $\kappa = 1 - \frac{\gamma\rho}{2}\|A\|^2 > 1/3$. The proof is completed. □

The Lyapunov estimate yields the global convergence properties of the generated sequence. For the product-space Opial argument, set $a := \gamma/\rho$ and define the product distance

$$D^2((\mu,\lambda),(\nu,\zeta)) := W_2^2(\mu,\nu) + a\|\lambda - \zeta\|^2. \tag{41}$$

Theorem C.7 (Global convergence)**.** *Let $\{(\mu_k, \lambda_k)\}_{k\ge 0}$ be generated by Algorithm 2 under Assumption C.1 and the stepsize condition therein. Then the following statements hold.*

(i)
$$\sum_{k=0}^{\infty} S_k < \infty, \quad \sum_{k=0}^{\infty}\|c_k\|^2 < \infty, \quad W_2(\mu_{k+1}, \mu_k) \to 0, \quad c_k \to 0.$$

(ii) There exists a saddle point $(\mu^\infty, \lambda^\infty)$ such that
$$\mu_k \rightharpoonup_{\mathcal{P}_2^w} \mu^\infty, \quad \lambda_k \to \lambda^\infty \quad , \Phi(\mu_k) \to \Phi(\mu^\infty).$$

Proof. (i) Sum (36) from $k = 0$ to N. By (37), $\widetilde{E}^\star_{N+1} \ge 0$, and hence

$$\delta\sum_{k=0}^{N} S_k + \frac{\gamma\rho}{4}\sum_{k=0}^{N}\|c_k + c_{k+1}\|^2 + 2\gamma\sum_{k=0}^{N} h^\star_{k+1} \le \widetilde{E}^\star_0. \tag{42}$$

Letting $N \to \infty$ yields $\sum_k S_k < \infty$, $\sum_k \|c_k + c_{k+1}\|^2 < \infty$, and $h^\star_{k+1} \to 0$. Moreover,

$$\|c_{k+1} - c_k\|^2 \le \|A\|^2 S_k, \quad \|c_k\|^2 \le \frac{1}{2}\|c_k + c_{k+1}\|^2 + \frac{1}{2}\|c_{k+1} - c_k\|^2,$$

so $\sum_k \|c_k\|^2 < \infty$ and $c_k \to 0$. Since the (x, y)-marginal of α_k is an admissible coupling of μ_k and μ_{k+1},

$$W_2^2(\mu_{k+1}, \mu_k) \le S_k \longrightarrow 0.$$

Since $h^\star_{k+1} \to 0$ and $c_{k+1} \to 0$, it also follows that $\Phi(\mu_k) \to \Phi(\mu^\star)$.

(ii) The monotonicity of $\widetilde{E}^\star_k$ and (37) give

$$W_2^2(\mu_k, \mu^\star) \le \frac{\widetilde{E}^\star_0}{\kappa}, \quad \|\lambda_k - \lambda^\star\|^2 \le \frac{\rho}{\gamma}\widetilde{E}^\star_0,$$

so the primal and multiplier sequences are bounded. To verify relative sequential compactness of $\{(\mu_k, \lambda_k)\}_{k\ge 0}$, consider an arbitrary subsequence $(\mu_{t_j}, \lambda_{t_j})$. By the weak compactness theorem for $\mathcal{P}_2^w(\mathbb{R}^d)$ Naldi & Savaré (2021), its primal component has a further subsequence converging in $\mathcal{P}_2^w(\mathbb{R}^d)$. The corresponding multiplier subsequence is bounded in the finite-dimensional space $\mathbb{R}^m = \mathbb{R}^m$ and therefore has a further convergent subsequence. The primal convergence is preserved under this second extraction. Hence $\{(\mu_k, \lambda_k)\}_{k\ge 0}$ is relatively sequentially compact in $\mathcal{P}_2^w(\mathbb{R}^d) \times \mathbb{R}^m$. Let $(\bar\mu, \bar\lambda)$ be an arbitrary cluster point of this sequence. By definition, there is a subsequence such that

$$\mu_{k_j} \rightharpoonup_{\mathcal{P}_2^w} \bar\mu, \quad \lambda_{k_j} \to \bar\lambda. \tag{43}$$

To identify the primal component, fix $y \in \mathbb{R}^m$. The scalar function $x \mapsto \langle y, Ax\rangle = \langle A^\top y, x\rangle$ belongs to $\mathcal{C}_2^w(\mathbb{R}^d)$, and hence

$$\langle y, \mathcal{C}(\mu_{k_j})\rangle = \int_{\mathbb{R}^d}\langle A^\top y, x\rangle \, \mathrm{d}\mu_{k_j}(x) - \langle y, b\rangle \to \int_{\mathbb{R}^d}\langle A^\top y, x\rangle \, \mathrm{d}\bar\mu(x) - \langle y, b\rangle = \langle y, \mathcal{C}(\bar\mu)\rangle.$$

Thus $\mathcal{C}(\mu_{k_j}) \to \mathcal{C}(\bar{\mu})$. Since $c_{k_j} \to 0$ strongly, $\mathcal{C}(\bar{\mu}) = 0$.

Identification of cluster. Fix an arbitrary $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ with $\Phi(\nu) < \infty$. Rearranging the plan EVI (20) gives

$$\Phi(\mu_{k+1}) - \Phi(\nu) + \left\langle \widehat{\lambda}_k, \mathcal{C}(\mu_{k+1}) - \mathcal{C}(\nu) \right\rangle \le \frac{W_2^2(\mu_k, \nu) - W_2^2(\mu_{k+1}, \nu)}{2\gamma} - \frac{1 - \gamma L}{2\gamma} S_k. \quad (44)$$

Since $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ is fixed, the boundedness of $\{W_2(\mu_k, \mu^\star)\}_{k=1}^\infty$ established above and the triangle inequality yield

$$W_2(\mu_k, \nu) \le W_2(\mu_k, \mu^\star) + W_2(\mu^\star, \nu) \le \left(\frac{\widetilde{E}_0^\star}{\kappa} \right)^{1/2} + W_2(\mu^\star, \nu).$$

Hence $\{W_2(\mu_k, \nu)\}_{k=1}^\infty$ is bounded. Therefore,

$$\left| W_2^2(\mu_k, \nu) - W_2^2(\mu_{k+1}, \nu) \right| \le \left[W_2(\mu_k, \nu) + W_2(\mu_{k+1}, \nu) \right] W_2(\mu_k, \mu_{k+1}) \to 0.$$

Together with $S_k \to 0$, this shows that the right-hand side of (44) converges to zero along $k = k_j$. Moreover,

$$\widehat{\lambda}_{k_j} = \lambda_{k_j} + \rho c_{k_j} \to \bar{\lambda}, \quad c_{k_j+1} \to 0, \quad j \to +\infty.$$

Hence

$$\left\langle \widehat{\lambda}_{k_j}, c_{k_j+1} \right\rangle \to 0, \quad \left\langle \widehat{\lambda}_{k_j}, \mathcal{C}(\nu) \right\rangle \to \left\langle \bar{\lambda}, \mathcal{C}(\nu) \right\rangle, \quad j \to +\infty.$$

We next show that the shifted subsequence has the same weak limit. Since $W_2(\mu_{k_j+1}, \mu_{k_j}) \to 0$ and $\{\mu_{k_j}\}$ is W_2-bounded, the sequence $\{\mu_{k_j+1}\}$ is also W_2-bounded and hence relatively sequentially compact in $\mathcal{P}_2^w(\mathbb{R}^d)$. Let $\mu_{k_{j_\ell}+1} \rightharpoonup_{\mathcal{P}_2^w} \widetilde{\mu}$ be an arbitrary weakly convergent subsubsequence. For every fixed $\sigma \in \mathcal{P}_2(\mathbb{R}^d)$, the triangle inequality gives $|W_2(\mu_{k_{j_\ell}+1}, \sigma) - W_2(\mu_{k_{j_\ell}}, \sigma)| \le W_2(\mu_{k_{j_\ell}+1}, \mu_{k_{j_\ell}}) \to 0$. Since the corresponding distance sequences are bounded, the same holds for their squares. Applying the Opial inequality (17) to $\mu_{k_{j_\ell}} \rightharpoonup_{\mathcal{P}_2^w} \bar{\mu}$ with comparison measure $\widetilde{\mu}$, and to $\mu_{k_{j_\ell}+1} \rightharpoonup_{\mathcal{P}_2^w} \widetilde{\mu}$ with comparison measure $\bar{\mu}$, respectively, and using the preceding vanishing distance difference, we obtain

$$2W_2^2(\bar{\mu}, \widetilde{\mu}) \le 0.$$

Thus $\widetilde{\mu} = \bar{\mu}$. Since every weak cluster point of $\{\mu_{k_j+1}\}$ equals $\bar{\mu}$ and this sequence is relatively sequentially compact, it follows that $\mu_{k_j+1} \rightharpoonup_{\mathcal{P}_2^w} \bar{\mu}$.

Using the weak lower semicontinuity of Φ and then (44), we obtain

$$\Phi(\bar{\mu}) - \Phi(\nu) - \left\langle \bar{\lambda}, \mathcal{C}(\nu) \right\rangle \le \liminf_{j \to \infty} \left[\Phi(\mu_{k_j+1}) - \Phi(\nu) + \left\langle \widehat{\lambda}_{k_j}, c_{k_j+1} - \mathcal{C}(\nu) \right\rangle \right] \le 0.$$

Thus

$$\Phi(\bar{\mu}) \le \Phi(\nu) + \left\langle \bar{\lambda}, \mathcal{C}(\nu) \right\rangle.$$

The inequality is automatic if $\Phi(\nu) = +\infty$. Together with $\mathcal{C}(\bar{\mu}) = 0$, it shows that $(\bar{\mu}, \bar{\lambda})$ is a saddle point; in particular, $\bar{\mu}$ is optimal.

Product Opial argument. For every saddle point $q = (\nu, \zeta)$, the proof of Theorem C.6 applies with q as reference and shows that $\widetilde{E}_k(\nu, \zeta)$ is nonincreasing and therefore convergent. Since

$$D^2((\mu_k, \lambda_k), q) = E_k(\nu, \zeta) = \widetilde{E}_k(\nu, \zeta) + \frac{\gamma\rho}{2} \|c_k\|^2$$

and $c_k \to 0$, the product-distance sequence also converges. Hence

$$\ell(q) := \lim_{k \to \infty} D^2((\mu_k, \lambda_k), q)$$

exists for every saddle point, and hence for every cluster point of $\{(\mu_k, \lambda_k)\}_{k \ge 0}$.

Let $q' = (\mu', \lambda')$ and $q'' = (\mu'', \lambda'')$ be two cluster points. By the preceding argument, both are saddle points, so $\ell(q')$ and $\ell(q'')$ are well defined. By the definition of a cluster point, there exist two, not necessarily related, subsequences (n_j) and (m_j) such that

$$\mu_{n_j} \rightharpoonup_{\mathcal{P}_2^w} \mu', \quad \lambda_{n_j} \to \lambda', \quad \mu_{m_j} \rightharpoonup_{\mathcal{P}_2^w} \mu'', \quad \lambda_{m_j} \to \lambda''.$$

Set

$$u_j := W_2^2(\mu_{n_j}, \mu''), \quad v_j := W_2^2(\mu_{n_j}, \mu').$$

The Opial inequality (17) gives

$$W_2^2(\mu', \mu'') + \liminf_j v_j \le \liminf_j u_j. \tag{45}$$

On the other hand,

$$\begin{aligned} &D^2((\mu_{n_j}, \lambda_{n_j}), q'') - D^2((\mu_{n_j}, \lambda_{n_j}), q') \\ &\quad = u_j - v_j + a\big(\|\lambda_{n_j} - \lambda''\|^2 - \|\lambda_{n_j} - \lambda'\|^2 \big). \end{aligned}$$

Weak convergence of λ_{n_j} and the Euclidean identity gives

$$\|\lambda_{n_j} - \lambda''\|^2 - \|\lambda_{n_j} - \lambda'\|^2 \longrightarrow \|\lambda' - \lambda''\|^2 .$$

Since the left-hand side of the preceding product-distance identity converges to $\ell(q'') - \ell(q')$, the difference $u_j - v_j$ also converges and

$$\lim_{j\to\infty} (u_j - v_j) = \ell(q'') - \ell(q') - a\,\|\lambda' - \lambda''\|^2 .$$

Combining this limit with (45) yields

$$\ell(q'') - \ell(q') - a\,\|\lambda' - \lambda''\|^2 \ge W_2^2(\mu', \mu''),$$

and therefore

$$\ell(q'') - \ell(q') \ge D^2(q', q''). \tag{46}$$

Applying the same calculation along (m_j), with q' and q'' interchanged, gives

$$\ell(q') - \ell(q'') \ge D^2(q', q'').$$

Adding the two inequalities yields $D(q', q'') = 0$, so $q' = q''$.

The sequence $\{(\mu_k, \lambda_k)\}_{k\ge 0}$ is relatively sequentially compact and has exactly one weak cluster point. Denote this point by $q^\infty = (\mu^\infty, \lambda^\infty)$. Hence the entire sequence converges to q^∞ in $\mathcal{P}_2^w(\mathbb{R}^d) \times \mathbb{R}^m$. Since q^∞ is a saddle point, $\Phi(\mu^\infty) = \Phi(\mu^\star)$; the objective convergence established in part (i) therefore gives $\Phi(\mu_k) \to \Phi(\mu^\infty)$. This completes the proof. □

Remark C.8.

1. If $A = 0$ and $b = 0$, Theorem C.7 gives convergence of the unconstrained Wasserstein forward–backward method under the present assumptions. The proof above is written entirely with transport plans and therefore does not require the existence of Monge maps.

2. On the Dirac class described in the preceding remark, Theorem C.7 reduces to the global convergence of a Euclidean linearized augmented-Lagrangian method, with the same type of proximal or linearization restriction as in augmented-Lagrangian analyses He et al. (2020).

C.2 Wasserstein KKT condition and convergence rates

We next characterize saddle points through a plan-valued Wasserstein KKT condition.

The form used here is the Wasserstein counterpart of the Euclidean KKT system. In the Wasserstein setting, an element of $\boldsymbol{\partial}\mathcal{H}(\mu)$ is a plan on the position–velocity variables (x, v). The balance $v + \nabla V(x) + A^\top \lambda = 0$ is therefore represented by the graph plan generated by $-\nabla V(x) - A^\top \lambda$. Accordingly, we call $(\bar\mu, \bar\lambda) \in \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^m$ a Wasserstein KKT point if

$$\mathcal{C}(\bar\mu) = 0, \quad \big(\mathrm{Id}, -\nabla V - A^\top \bar\lambda\big)_{\#}\bar\mu \in \boldsymbol{\partial}\mathcal{H}(\bar\mu). \tag{47}$$

The following theorem shows that this plan-valued KKT system is equivalent to the saddle-point condition used in the convergence analysis.

Theorem C.9 (Plan-valued KKT characterization)**.** *Under Assumption C.1(i)–(ii), let $\bar{\mu} \in \mathrm{dom}(\mathcal{H})$ and $\bar{\lambda} \in \mathbb{R}^m$. Then the pair $(\bar{\mu}, \bar{\lambda})$ satisfies* (47) *if and only if*

$$\mathcal{C}(\bar{\mu}) = 0, \quad \Phi(\bar{\mu}) \le \Phi(\nu) + \langle \bar{\lambda}, \mathcal{C}(\nu) \rangle \quad \textit{for every } \nu \in \mathcal{P}_2(\mathbb{R}^d). \tag{48}$$

Consequently, the primal component of every Wasserstein KKT point solves (1)*.*

Proof. Set

$$T_{\bar{\lambda}}(x) := -\nabla V(x) - A^\top \bar{\lambda}, \qquad \boldsymbol{\xi} := (\mathrm{Id}, T_{\bar{\lambda}})_{\#}\bar{\mu}.$$

Since ∇V is Lipschitz continuous, $T_{\bar{\lambda}}$ has at most linear growth. Hence $T_{\bar{\lambda}} \in L^2(\bar{\mu}; \mathbb{R}^d)$ and therefore $\boldsymbol{\xi} \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$.

Suppose first that (47) holds. Then

$$\boldsymbol{\xi} \in \boldsymbol{\partial}\mathcal{H}(\bar{\mu}).$$

Fix $\nu \in \mathrm{dom}(\mathcal{H})$. By the global characterization (13), there exists $\boldsymbol{\alpha} \in \Gamma_o^{1,3}(\boldsymbol{\xi}, \nu)$ such that

$$\mathcal{H}(\nu) - \mathcal{H}(\bar{\mu}) \ge \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \langle v, y - x \rangle \, \mathrm{d}\boldsymbol{\alpha}(x, v, y). \tag{49}$$

Since $\boldsymbol{\xi} = (\mathrm{Id}, T_{\bar{\lambda}})_{\#}\bar{\mu}$ and $({}^3\pi^{1,2})_{\#}\boldsymbol{\alpha} = \boldsymbol{\xi}$, the definition of the pushforward measure gives

$$\begin{aligned}
&\boldsymbol{\alpha}\left(\{(x, v, y) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d : v = T_{\bar{\lambda}}(x)\}\right) \\
=&\boldsymbol{\alpha}\left(({}^3\pi^{1,2})^{-1}\{(x, v) \in \mathbb{R}^d \times \mathbb{R}^d : v = T_{\bar{\lambda}}(x)\}\right) \\
=&({}^3\pi^{1,2})_{\#}\boldsymbol{\alpha}\left(\{(x, v) \in \mathbb{R}^d \times \mathbb{R}^d : v = T_{\bar{\lambda}}(x)\}\right) \\
=&\boldsymbol{\xi}\left(\{(x, v) \in \mathbb{R}^d \times \mathbb{R}^d : v = T_{\bar{\lambda}}(x)\}\right) \\
=&\bar{\mu}\left((\mathrm{Id}, T_{\bar{\lambda}})^{-1}\{(x, v) \in \mathbb{R}^d \times \mathbb{R}^d : v = T_{\bar{\lambda}}(x)\}\right) \\
=&\bar{\mu}(\mathbb{R}^d) = 1.
\end{aligned}$$

It follows that

$$v = T_{\bar{\lambda}}(x) = -\nabla V(x) - A^\top \bar{\lambda}, \quad \boldsymbol{\alpha}\text{-a.s. } (x, v, y). \tag{50}$$

Because $({}^3\pi^1)_{\#}\boldsymbol{\alpha} = \bar{\mu}$ and $({}^3\pi^3)_{\#}\boldsymbol{\alpha} = \nu$, the pushforward formula and the convexity of V yield

$$\begin{aligned}
\int_{\mathbb{R}^d} V \, \mathrm{d}\nu - \int_{\mathbb{R}^d} V \, \mathrm{d}\bar{\mu} &= \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \big(V(y) - V(x)\big) \, \mathrm{d}\boldsymbol{\alpha}(x, v, y) \\
&\ge \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \langle \nabla V(x), y - x \rangle \, \mathrm{d}\boldsymbol{\alpha}(x, v, y).
\end{aligned} \tag{51}$$

The same marginal identities and the definition of $\mathcal{C}$ give

$$\begin{aligned}
\langle \bar{\lambda}, \mathcal{C}(\nu) - \mathcal{C}(\bar{\mu}) \rangle &= \left\langle \bar{\lambda}, \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} A(y - x) \, \mathrm{d}\boldsymbol{\alpha}(x, v, y) \right\rangle \\
&= \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \langle A^\top \bar{\lambda}, y - x \rangle \, \mathrm{d}\boldsymbol{\alpha}(x, v, y).
\end{aligned} \tag{52}$$

Adding (49), (51), and (52), and then using (50), yields

$$\Phi(\nu) - \Phi(\bar{\mu}) + \langle \bar{\lambda}, \mathcal{C}(\nu) - \mathcal{C}(\bar{\mu}) \rangle \ge \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \langle v + \nabla V(x) + A^\top \bar{\lambda}, y - x \rangle \, \mathrm{d}\boldsymbol{\alpha}(x, v, y) = 0.$$

Since $\mathcal{C}(\bar{\mu}) = 0$, this proves (48); the case $\nu \notin \mathrm{dom}(\mathcal{H})$ is immediate.

Conversely, suppose that (48) holds. Fix $\nu \in \mathrm{dom}(\mathcal{H})$ and $\pi \in \Gamma_o(\bar{\mu}, \nu)$. Since $({}^2\pi^1)_{\#}\pi = \bar{\mu}$ and $({}^2\pi^2)_{\#}\pi = \nu$, the pushforward formula and the L-smoothness of V give

$$\begin{aligned}
\int_{\mathbb{R}^d} V \, \mathrm{d}\nu - \int_{\mathbb{R}^d} V \, \mathrm{d}\bar{\mu} &= \int_{\mathbb{R}^d \times \mathbb{R}^d} \big(V(y) - V(x)\big) \, \mathrm{d}\pi(x, y) \\
&\le \int_{\mathbb{R}^d \times \mathbb{R}^d} \left[\langle \nabla V(x), y - x \rangle + \frac{L}{2} \|y - x\|^2\right] \mathrm{d}\pi(x, y) \\
&= \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle \nabla V(x), y - x \rangle \, \mathrm{d}\pi(x, y) + \frac{L}{2} W_2^2(\bar{\mu}, \nu).
\end{aligned} \tag{53}$$

The same marginal identities, the definition of $\mathcal{C}$, and the adjoint identity yield

$$\begin{aligned}\left\langle \bar{\lambda}, \mathcal{C}(\nu) - \mathcal{C}(\bar{\mu}) \right\rangle &= \left\langle \bar{\lambda}, \int_{\mathbb{R}^d} Ay \ \mathrm{d}\nu(y) - \int_{\mathbb{R}^d} Ax \ \mathrm{d}\bar{\mu}(x) \right\rangle \\ &= \int_{\mathbb{R}^d \times \mathbb{R}^d} \left\langle \bar{\lambda}, A(y-x) \right\rangle \mathrm{d}\pi(x,y) \\ &= \int_{\mathbb{R}^d \times \mathbb{R}^d} \left\langle A^\top \bar{\lambda}, y - x \right\rangle \mathrm{d}\pi(x,y). \end{aligned} \tag{54}$$

Let $G : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$ be defined by

$$G(x,y) := (x, T_{\bar{\lambda}}(x), y), \quad \boldsymbol{\alpha} := G_{\#}\pi.$$

Since $T_{\bar{\lambda}} \in L^2(\bar{\mu}; \mathbb{R}^d)$, we have $\boldsymbol{\alpha} \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d)$. Moreover,

$$({}^3\pi^{1,2})_{\#}\boldsymbol{\alpha} = ({}^3\pi^{1,2} \circ G)_{\#}\pi = (\mathrm{Id}, T_{\bar{\lambda}})_{\#}({}^2\pi^1)_{\#}\pi = (\mathrm{Id}, T_{\bar{\lambda}})_{\#}\bar{\mu} = \boldsymbol{\xi}, \tag{55}$$

$$({}^3\pi^{1,3})_{\#}\boldsymbol{\alpha} = ({}^3\pi^{1,3} \circ G)_{\#}\pi = \pi \in \Gamma_o(\bar{\mu}, \nu). \tag{56}$$

Therefore $\boldsymbol{\alpha} \in \Gamma_o^{1,3}(\boldsymbol{\xi}, \nu)$. Applying the pushforward formula once more gives

$$\int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \langle v, y - x \rangle \ \mathrm{d}\boldsymbol{\alpha}(x,v,y) = \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \langle v, y - x \rangle \ \mathrm{d}G_{\#}\pi = \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle T_{\bar{\lambda}}(x), y - x \rangle \ \mathrm{d}\pi(x,y). \tag{57}$$

Finally, (48) gives

$$\mathcal{H}(\nu) - \mathcal{H}(\bar{\mu}) \geq -\left(\int_{\mathbb{R}^d} V \ \mathrm{d}\nu - \int_{\mathbb{R}^d} V \ \mathrm{d}\bar{\mu} \right) - \left\langle \bar{\lambda}, \mathcal{C}(\nu) - \mathcal{C}(\bar{\mu}) \right\rangle.$$

Combining this inequality with (53), (54), and (57), we obtain

$$\begin{aligned}\mathcal{H}(\nu) - \mathcal{H}(\bar{\mu}) &\geq \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle T_{\bar{\lambda}}(x), y - x \rangle \ \mathrm{d}\pi(x,y) - \frac{L}{2} W_2^2(\bar{\mu}, \nu) \\ &= \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \langle v, y - x \rangle \ \mathrm{d}\boldsymbol{\alpha}(x,v,y) - \frac{L}{2} W_2^2(\bar{\mu}, \nu) \\ &\geq \inf_{\widetilde{\boldsymbol{\alpha}} \in \Gamma_o^{1,3}(\boldsymbol{\xi}, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \langle v, y - x \rangle \ \mathrm{d}\widetilde{\boldsymbol{\alpha}}(x,v,y) - \frac{L}{2} W_2^2(\bar{\mu}, \nu). \end{aligned}$$

Since $W_2^2(\bar{\mu}, \nu) = o(W_2(\bar{\mu}, \nu))$ as $\nu \to \bar{\mu}$, Definition B.4 gives

$$\boldsymbol{\xi} \in \boldsymbol{\partial}\mathcal{H}(\bar{\mu}).$$

Together with $\mathcal{C}(\bar{\mu}) = 0$, this is (47). This completes the proof. □

Theorem C.7 shows that the W-LALM iterates converge to a saddle point, while Theorem C.9 identifies saddle points with Wasserstein KKT points. We now complement these qualitative convergence results with a quantitative estimate of the violation of the KKT system.

For the optimal coupling $\pi_k \in \Gamma_o(\eta_{k+1}, \mu_{k+1})$ fixed in the convergence analysis, the JKO optimality condition in Lemma B.5(i) provides the selected subgradient

$$\boldsymbol{\xi}_{k+1} := \left((z,y) \mapsto \left(y, \frac{z-y}{\gamma} \right) \right)_{\#} \pi_k \in \boldsymbol{\partial}\mathcal{H}(\mu_{k+1}). \tag{58}$$

This naturally leads to the following residual for the KKT system:

$$\mathcal{R}_{k+1}^2 := \|c_{k+1}\|^2 + \int_{\mathbb{R}^d \times \mathbb{R}^d} \left\| v + \nabla V(u) + A^\top \lambda_{k+1} \right\|^2 \mathrm{d}\boldsymbol{\xi}_{k+1}(u,v). \tag{59}$$

The first term measures primal feasibility, while the second measures stationarity with respect to the JKO-selected subgradient. In particular, $\mathcal{R}_{k+1} = 0$ certifies that $(\mu_{k+1}, \lambda_{k+1})$ satisfies the Wasserstein KKT system and hence, by Theorem C.9, is a saddle point.

The next result shows that this KKT residual is controlled by the Lyapunov decrease.

Theorem C.10 (KKT residual complexity)**.** *Let $\{(\mu_k, \lambda_k)\}_{k\geq 0}$ be generated by Algorithm 2 under Assumption C.1 and the stepsize condition therein, and let $\widetilde{\mathcal{R}}_{k+1}$ be defined by* (59)*. Then there exists a constant $C > 0$, depending only on the problem and algorithmic parameters, such that*

$$\mathcal{R}_{k+1}^2 \leq C\big(\widetilde{E}_k^\star - \widetilde{E}_{k+1}^\star\big), \quad k \geq 0. \tag{60}$$

Moreover,

$$\frac{1}{N}\sum_{k=0}^{N-1} \mathcal{R}_{k+1}^2 \leq \frac{C\widetilde{E}_0^\star}{N}, \quad N \geq 1, \tag{61}$$

and

$$\min_{0\leq k<N} \mathcal{R}_{k+1} = o\left(N^{-1/2}\right). \tag{62}$$

Proof. Recall the three-marginal plan $\alpha_k \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d)$ used in the proof of Lemma C.5, satisfying

$$({}^3\pi^{1,2})_\# \alpha_k = (\mathrm{Id}, P_k)_\# \mu_k, \quad ({}^3\pi^{2,3})_\# \alpha_k = \pi_k.$$

By the definition of $\boldsymbol{\xi}_{k+1}$ and the pushforward formula,

$$\begin{aligned}
&\int_{\mathbb{R}^d\times\mathbb{R}^d} \left\|v + \nabla V(u) + A^\top \lambda_{k+1}\right\|^2 \mathrm{d}\boldsymbol{\xi}_{k+1}(u,v) \\
&= \int_{\mathbb{R}^d\times\mathbb{R}^d} \left\|\frac{z-y}{\gamma} + \nabla V(y) + A^\top \lambda_{k+1}\right\|^2 \mathrm{d}\pi_k(z,y) \\
&= \int_{\mathbb{R}^d\times\mathbb{R}^d} \left\|\frac{z-y}{\gamma} + \nabla V(y) + A^\top \lambda_{k+1}\right\|^2 \mathrm{d}\big(({}^3\pi^{2,3})_\#\alpha_k\big)(z,y) \\
&= \int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \left\|\frac{z-y}{\gamma} + \nabla V(y) + A^\top \lambda_{k+1}\right\|^2 \mathrm{d}\alpha_k(x,z,y).
\end{aligned}$$

By the same argument based on the definition of the pushforward measure as in the proof of Lemma C.5, we obtain

$$z = P_k(x) = x - \gamma\big(\nabla V(x) + A^\top \widehat{\lambda}_k\big), \quad \alpha_k\text{-a.s.}$$

Moreover, the multiplier update gives

$$\lambda_{k+1} - \widehat{\lambda}_k = \rho(c_{k+1} - c_k).$$

Hence

$$\frac{z-y}{\gamma} + \nabla V(y) + A^\top \lambda_{k+1} = \frac{x-y}{\gamma} + \nabla V(y) - \nabla V(x) + A^\top(\lambda_{k+1} - \widehat{\lambda}_k) \quad \alpha_k, \text{-a.s.}$$

Using the L-Lipschitz continuity of ∇V and the estimate

$$\|c_{k+1} - c_k\|^2 \leq \|A\|^2 S_k$$

obtained in the proof of Theorem C.6, we obtain

$$\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \left\|\frac{z-y}{\gamma} + \nabla V(y) + A^\top \lambda_{k+1}\right\|^2 \mathrm{d}\alpha_k \leq \left(\gamma^{-1} + L + \rho\|A\|^2\right)^2 S_k. \tag{63}$$

For the feasibility part, we have

$$\begin{aligned}
\|c_{k+1}\|^2 &\leq \frac{1}{2}\|c_k + c_{k+1}\|^2 + \frac{1}{2}\|c_{k+1} - c_k\|^2 \\
&\leq \frac{1}{2}\|c_k + c_{k+1}\|^2 + \frac{1}{2}\|A\|^2 S_k.
\end{aligned}$$

Combining this estimate with (63) yields

$$\mathcal{R}_{k+1}^2 \leq \left[\left(\gamma^{-1} + L + \rho\|A\|^2\right)^2 + \frac{1}{2}\|A\|^2\right] S_k + \frac{1}{2}\|c_k + c_{k+1}\|^2. \tag{64}$$

On the other hand, Theorem C.6 gives

$$\widetilde{E}_k^\star - \widetilde{E}_{k+1}^\star \geq \delta S_k + \frac{\gamma\rho}{4}\left\|c_k + c_{k+1}\right\|^2,$$

where we have dropped the nonnegative term $2\gamma h_{k+1}^\star$. Hence there exists $C > 0$ such that

$$\mathcal{R}_{k+1}^2 \leq C\big(\widetilde{E}_k^\star - \widetilde{E}_{k+1}^\star\big),$$

which proves (60). Summing over $k = 0, \ldots, N-1$ and using $\widetilde{E}_N^\star \geq 0$ gives

$$\sum_{k=0}^{N-1} \mathcal{R}_{k+1}^2 \leq C(\widetilde{E}_0^\star - \widetilde{E}_N^\star) \leq C\widetilde{E}_0^\star.$$

Dividing by N proves (61). Letting $N \to \infty$ also yields

$$\sum_{k=0}^{\infty} \mathcal{R}_{k+1}^2 \leq C\widetilde{E}_0^\star < +\infty.$$

Furthermore,

$$\begin{aligned} N \min_{0\leq k<N} \mathcal{R}_{k+1}^2 &\leq N \min_{\lfloor N/2\rfloor \leq k<N} \mathcal{R}_{k+1}^2 \\ &\leq 2 \sum_{k=\lfloor N/2\rfloor}^{N-1} \mathcal{R}_{k+1}^2 \longrightarrow 0. \end{aligned}$$

Therefore

$$\min_{0\leq k<N} \mathcal{R}_{k+1} = o\left(N^{-1/2}\right).$$

This completes the proof. □

D EXPERIMENTAL DETAILS

Discretization of the W-LALM steps. Given $\mu_k^N = \sum_i w_i^k \delta_{u_i^k}$, the forward pushforward preserves the weights and moves the locations: $\widetilde{u}_i^{k+1} = P_k(u_i^k)$ and $\eta_{k+1}^N = \sum_i w_i^k \delta_{\widetilde{u}_i^{k+1}}$. We approximate the subsequent JKO step by restricting candidate measures to $\mu_w = \sum_i w_i \delta_{\widetilde{u}_i^{k+1}}$, $w \in \Delta_N$, on this transported support. The discrete transport term is

$$\mathsf{W}_N^2(w, w^k; \widetilde{U}^{k+1}) = \min_{\substack{\pi\geq 0 \\ \pi\mathbf{1}=w,\ \pi^\top\mathbf{1}=w^k}} \sum_{i,j} \pi_{ij} \|\widetilde{u}_i^{k+1} - \widetilde{u}_j^{k+1}\|_2^2.$$

Replacing continuous entropy by the weight-entropy surrogate $\sum_i w_i \log(N w_i)$ gives (6). With the support fixed, the transport value and weight entropy are convex in w. Thus the location–weight separation comes from the forward pushforward followed by a fixed-support approximation of the JKO step; the unrestricted JKO minimization also allows the support to change. Proximal-ALM instead discretizes the original constrained problem as (7) and optimizes the coupled variables (U, w).

Generalized Bayesian calibration. Let $\Psi \in \mathbb{R}^{n\times d}$ and $y \in \mathbb{R}^n$ denote the polynomial design matrix and responses on the training split. We use

$$V(u) = \frac{0.05}{2n}\|\Psi u - y\|_2^2 + \frac{0.08}{2}\|u\|_2^2, \qquad h(u) = 0, \qquad \beta = 10^{-4}.$$

The rows of A are the leading m orthonormal eigenvectors of $\Psi^\top\Psi/n$. On a separate calibration split $(\Psi_{\mathrm{cal}}, y_{\mathrm{cal}})$, we compute

$$\hat{\theta}_{\mathrm{cal}} = \arg\min_\theta \left\{ \frac{\|\Psi_{\mathrm{cal}}\theta - y_{\mathrm{cal}}\|_2^2}{2n_{\mathrm{cal}}} + \frac{0.08}{2}\|\theta\|_2^2 \right\}, \qquad b = A\hat{\theta}_{\mathrm{cal}}.$$

Thus $A\mathbb{E}_\mu[u] = b$ calibrates the mean regression coefficients along these data-dependent directions.

Solvers and numerical settings. We solve the discrete transport problems with entropically regularized Sinkhorn iterations (Cuturi, 2013). This transport regularization is distinct from the weight-entropy term in the objective. For CWPG, a support translation makes the current weights feasible before the constrained weight solve, whose equality multiplier is computed by damped Newton iterations with conjugate-gradient linear solves. Proximal-ALM uses alternating proximal-gradient updates for particle locations and entropy-mirror updates for weights (Birgin & Martínez, 2014; Bolte et al., 2014). PD-LMC (Chamon et al., 2024) takes one Langevin step per multiplier update. The implementation details and algorithmic parameters are provided in the released code.

For the reported runs, we use $N = 140$ particles, seed 20260823, and the training/calibration sizes $(1000, 400)$ for Wine Red and $(700, 250)$ for Concrete. The two instances use $m = 50$ and $m = 100$ constraints, respectively. All methods use stopping tolerance 10^{-5} and a 300-second wall-clock budget. The inner numerical tolerance is 10^{-10}; Sinkhorn uses at most 35 iterations per outer step. We set $\beta = 10^{-4}$, $\kappa = 0.08$, loss scale 0.05, and W-LALM penalty $\rho = 0.2$. PD-LMC uses $\eta_k = \eta_0/\sqrt{k+1}$, while Proximal-ALM uses $\rho_{\mathrm{ALM}} = 2$, at most 150 inner sweeps, and checks the inner stationarity condition every five sweeps.