Three-Operator Splitting with Stale Gradients for Faster Non-Linear Optimal Transport

Jacob Lindbäck
EECS, KTH

David Alvarez-Melis
Harvard SEAS

Mikael Johansson
EECS, KTH

Abstract

Scalable optimization for non-linear optimal transport (OT) poses unique challenges; it requires efficient memory management of large matrices, effective parallelization strategies suited for modern accelerators like GPUs, and theoretical guarantees that support practical implementation patterns. To address these challenges, we introduce a new algorithm based on three-operator splitting that reduces gradient computation costs by allowing gradient evaluations to run asynchronously and in parallel with other computations. Using monotone operator theory, we establish new convergence guarantees for this asynchronous adaptation and extend existing results to an important non-convex problem class including Gromov–Wasserstein as a notable example. We validate our method through a series of experiments demonstrating improved accuracy and faster convergence for a broad range of problems

1 INTRODUCTION

Optimal transport (OT) (Villani et al., 2008; Peyré et al., 2019) has become a foundational tool in machine learning and data-intensive scientific fields. It provides a principled and geometrically meaningful way to compare probability distributions by computing the most cost-efficient plan to morph one distribution into another. Unlike simple divergence measures, OT respects the geometry of the underlying space through ground metrics (Peyré et al., 2019; Santambrogio, 2015), making it especially powerful for structured data such as images, point clouds, and graphs. This theoretical elegance has led to widespread use in applications including domain adaptation (Courty et al., 2016), generative modeling (Arjovsky et al., 2017; Genevay et al., 2018), flow matching (Tong et al., 2024), and computational biology (Demetci et al., 2022).

The classical formulation of OT is, however, too restrictive in settings that require non-linear transportation costs. A prominent example is the Gromov-Wasserstein (GW) distance, which compares metric-measure spaces by matching intra-domain distances through a non-convex quadratic cost (Mémoli, 2011). Non-linear costs generally incur significant computational costs, as they eliminate structural properties of the original OT problem. Moreover, while discrete OT reduces to a linear program solvable in polynomial time, computing, e.g., the GW distance requires solving a nonconvex quadratic assignment problem, which is NP-hard in general. The computational burden of non-linear OT has motivated various approximation techniques, using regularization (Peyré et al., 2016), low-rank approximations (Scetbon et al., 2022), and relaxation (Li et al., 2023). Rather than relying on further approximations, in this work, we explore the use of operator splitting methods to solve non-linear OT problems, including Gromov–Wasserstein, more efficiently by exploiting their underlying structure.

We focus on a general class of optimization problems over the transportation polytope, encompassing several important OT problems as special cases. Given two discrete distributions $p \in \Delta_m$, and $q \in \Delta_n$, along with the transportation polytope:

$$\mathcal{T}(p,q) \coloneqq \{\gamma \in \mathbb{R}_+^{m\times n} : \gamma 1_n = p,\, \gamma^\top 1_m = q\},$$

we focus on the problem

$$\min_{\gamma\in\mathcal{T}(p,q)} \quad \ell(\gamma), \tag{1}$$

where $\ell : \mathbb{R}^{m\times n} \to \mathbb{R}$ is a differentiable loss function that encodes some notion of quality of the transportation plan. Several other OT variants can be expressed within the framework of equation 1, including:

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

Gromov–Wasserstein (GW):

$$\ell(\gamma) = \sum_{i,j,i',j'} \left| d_X(x_i, x_{i'}) - d_Y(y_j, y'_j) \right|^p \gamma_{i,j}\gamma_{i',j'},$$

where (X, d_X), (Y, d_Y) are two metric spaces, and $\{x_i\}_{i=1}^m \subset X$, $\{y_j\}_{j=1}^n \subset Y$, are two samples corresponding to the discrete measures $\mu_X = \sum_{i=1}^m p_i \delta_{x_i}$, $\mu_Y = \sum_{j=1}^n q_j \delta_{y_j}$. GW identifies optimal correspondences between the two samples based solely on within-sample distances (Mémoli, 2011).

Information-Theoretic OT (InfoOT):

$$\ell(\gamma) = -\hat{I}_\gamma(X, Y),$$

where $\hat{I}_\gamma(X, Y)$ represents a kernelized version of the mutual information between samples X and Y, with joint distribution determined by the transport plan γ, as proposed by Chuang et al. (2023).

OT with smooth regularization:

$$\ell(\gamma) = \operatorname{tr}\left(D^\top \gamma\right) + r(\gamma),$$

where r is an L-smooth regularizer. Such approaches have been developed for domain adaptation tasks using Laplacian regularization (Courty et al., 2016), ensuring the preservation of cluster structures when aligning training and test domains.

Convex point-cloud registration:

$$\ell(\gamma) = \|K_X\gamma - \gamma K_Y\|^2,$$

where $K_X \in \mathbb{R}^{m\times m}$ and $K_Y \in \mathbb{R}^{n\times n}$ are Gram matrices. This convex problem was proposed by Grave et al. (2019) as a relaxation to the Wasserstein–Procrustes problem.

In practice, combinations of objectives are often used—e.g., Fused GW (Titouan et al., 2019) and Fused InfoOT (Chuang et al., 2023) which incorporate cross-sample information by adding a term corresponding to the classic OT problem, which often yields improved empirical performance.

Algorithmic challenges

The effectiveness of different variations non-linear OT problems hinges on the availability of efficient and robust optimization algorithms. Even for standard OT, scalability is limited by several key factors:

Memory constraints. Optimal transport is fundamentally memory-bound (Mai et al., 2022). Its $O(mn)$ memory footprint leads to substantial accesses to bandwidth-limited memory, and together with operations on large dense matrices, often dominates runtime—even when the underlying computations are simple. Scalable algorithms must minimize global memory access and optimize access patterns to fully exploit memory bandwidth. These issues are further exacerbated in nonlinear OT variants, where large matrix gradients are updated iteratively.

Handling the transportation polytope. The transportation polytope lacks a closed-form Bregman projection. Classical methods that circumvent the use of projections, like network simplex or interior-point algorithms, are hard to parallelize and poorly suited to GPUs, with cubic iteration complexity that makes them impractical at scale. Entropic regularization offers a widely-used alternative: by adding a strictly convex entropy term, the feasible set is smoothed, enabling efficient Sinkhorn iterations (Peyré et al., 2019). These iterations reduce to parallelizable matrix-scaling steps, but the regularization term introduces bias and may cause numerical instability.

Costly gradient computations. Solving nonlinear OT problems typically requires iterative linearization, where gradient evaluation is often the dominant cost. For Gromov–Wasserstein, each gradient evaluation requires $O(m^2n^2)$ operations. Even in the optimized quadratic case, this is reduced to $O(mn^2+m^2n)$ (Peyré et al., 2016), which still remains a significant computational burden.

Nonconvexity. Several key OT variants—including Gromov–Wasserstein—lead to nonconvex optimization problems. This introduces additional challenges in both theoretical analysis and practical algorithm design, as many results that hold for OT or other convex formulations do not readily extend to the nonconvex setting.

To address these challenges, our work makes several key contributions that enable scalable optimal transport. First, we develop a novel optimization algorithm based on three-operator splitting that addresses the computational and theoretical challenges in a unified framework. Our algorithm supports asynchronous gradient computations, significantly reducing computational burden—especially for Gromov–Wasserstein problems where gradient evaluations are costly. Second, we establish new convergence guarantees for three-operator splitting with delayed gradients using monotone operator theory, complemented by specific non-convex guarantees. Third, we describe an implementation designed to exploit modern accelerator hardware through low-overhead memory operations. Our algorithm maintains a single-loop struc-

ture and avoids entropic regularization, naturally producing sparse transport plans and enabling efficient handling of polytope constraints on-the-fly. Together, these contributions lead to a method that is both theoretically sound and empirically superior across a variety of large-scale OT tasks.

Notation. We will reserve $\mathbb{R}^n_+$ for the non-negative orthant, and $\mathbb{R}^{m\times n}_+$ be its matrix counterpart. The n-dimensional simplex is denoted Δ_n, and set of signed transportation plans is defined as $\bar{\mathcal{T}}(p,q) \coloneqq \{\gamma \in \mathbb{R}^{m\times n} : \gamma 1_n = p,\, \gamma^\top 1_m = q\}$. The indicator function associated with a closed and convex set S will be denoted ι_S, that is $\iota_S(x) = 0$, if $x \in S$ and $\iota_S(x) = +\infty$, if $x \notin S$. Further, ∂f denotes the subdifferential of f, i.e. $\partial f(x)$ is the set of subgradients at x. The identity mapping is denoted Id, and the proximal operator associated with a stepsize $\rho > 0$ is denoted and defined, for any $x \in \mathbb{R}^n$: $\operatorname{prox}_{\rho f}(x) \coloneqq \operatorname{argmin}_{y\in\mathbb{R}^n} f(y) + \frac{1}{2\rho}\|x-y\|^2$. We also let $\mathrm{P}_S(\cdot)$ denote the projection onto a closed and convex set S. In particular, we let $\mathrm{P}_{[0,1]^n}(\cdot) = \operatorname{clip}_{[0,\,1]}(\cdot)$, which clamps the input entrywise into the range $[0,1]$.

2 RELATED WORKS

Scalable OT solvers. The introduction of entropy-regularized OT and Sinkhorn iterations by Cuturi (2013) spurred extensive research into memory- and time-efficient algorithms for optimal transport. Altschuler et al. (2017) established near-linear theoretical guarantees, later extended by Dvurechensky et al. (2018) in terms of improved ε-complexity bounds. Orthogonally, operator-splitting methods tailored to GPU memory hierarchies (Mai et al., 2022; Lindbäck et al., 2023; Lu and Yang, 2024) significantly reduce communication overhead and form the computational basis of our approach. Finally, for non-linear convex transportation costs, Ballu and Berthet (2023) proposed the Mirror-Sinkhorn algorithm along with online and stochastic variants.

Gromov–Wasserstein methods. When source and target distributions lie in different metric spaces, the Gromov–Wasserstein (GW) distance (Mémoli, 2011) provides a natural dissimilarity measure. However, its inherent numerical complexity motivated significant research on approximate formulations. Entropic regularization combined with projected gradient descent (Peyré et al., 2016) and proximal-point methods (Xu et al., 2019) have enabled larger-scale applications. Recent work exploited low-rank structure to achieve significant speed-ups (Scetbon et al., 2022), and (Li et al., 2023) considered a relaxed formulation to design a Bregman Alternating Projected Gradient (BAPG) algorithm, a single-loop solver with convergence guarantees and high GPU efficiency.

Operator-splitting techniques. Splitting methods exploit problem structure by decomposing composite objectives into simpler subproblems; see (Ryu and Yin, 2022; Bauschke and Combettes, 2011). In the context of OT, PDHG-based solvers have been developed for transport and barycenter problems (Chambolle and Contreras, 2022), while the closely related Douglas–Rachford splitting method has been adapted for GPU implementations (Mai et al., 2022; Lindbäck et al., 2023). Three-operator splitting (TOS) provides a unifying framework for many of these methods. The foundational Davis–Yin scheme (Davis and Yin, 2017) has been extended with adaptive step sizes (Pedregosa and Gidel, 2018), inexact updates (Zong et al., 2018), and for non-convex/stochastic variants (Yurtsever et al., 2021), which rely on weaker boundedness assumptions than earlier smoothness-based analyses (Bian and Zhang, 2021).

Asynchronous optimization. Relaxing synchronization to increase hardware utilization is well-studied for several optimization methods. For distributed SGD, *Hogwild!* (Recht et al., 2011) enabled lock-free asynchronous updates, and several extensions and generalized theoretical convergence guarantees have been proposed since (e.g. (Mishchenko et al., 2022; Feyzmahdavian and Johansson, 2023)). Other frameworks allow for asynchronous updates of variable blocks, notably for fixed point updates (Peng et al., 2016). Although applicable to operator-splitting techniques, our work focuses on a different setting, where a single global variable is updated using stale gradients, as opposed to variable blocks being updated asynchronously.

3 THREE-OPERATOR SPLITTING

The *three-operator splitting method (TOS)*, also referred to as *Davis & Yin-splitting*, was first proposed as an iterative scheme to solve inclusion problems involving the sums of three maximally monotone operators on Hilbert spaces, where one of the operators is cocoercive (Davis and Yin, 2017). Monotone operator theory (see, *e.g*, (Ryu and Yin, 2022)) provides a powerful theoretical framework for analyzing many classes of algorithms. In the convex optimization setting, TOS can be used to solve problems of the form

$$\min_{x\in\mathbb{R}^n} f(x) + g(x) + h(x), \tag{2}$$

where f and g are closed, convex, and proper functions, and h is differentiable with an L-Lipschitz con-

tinuous gradient. This composite objective corresponds to a monotone inclusion involving the subdifferentials ∂f, ∂g, and the gradient operator ∇h. Under standard assumptions, TOS generates a sequence of iterates that converge to a solution satisfying the optimality condition:

$$0 \in \partial f(x^\star) + \partial g(x^\star) + \nabla h(x^\star).$$

The TOS algorithm proceeds by generating iterates according to the following update rules:

$$\begin{aligned} x_{k+1} &= \mathrm{prox}_{\rho f}\left(y_k\right), \\ z_{k+1} &= \mathrm{prox}_{\rho g}\left(2x_{k+1} - y_k - \rho\nabla h(x_{k+1})\right), \\ y_{k+1} &= y_k + z_{k+1} - x_{k+1}. \end{aligned} \tag{3}$$

Here ρ is a stepsize fulfilling $0 < \rho < 2/L$. This iteration can also be reformulated as a fixed-point update $y_{k+1} = T_{\mathrm{DY}} y_k$ for the operator

$$\begin{aligned} T_{\mathrm{DY}} \coloneqq \mathrm{Id} - \mathrm{prox}_{\rho f}\left(\cdot\right) \\ + \mathrm{prox}_{\rho g}\left(2\mathrm{prox}_{\rho f}\left(\cdot\right) - \mathrm{Id} - \rho\nabla h(\mathrm{prox}_{\rho f}\left(\cdot\right))\right) \end{aligned} \tag{4}$$

Each fixed point of T_{DY} corresponds to a solution of the optimality condition in equation 2. Specifically, as shown in Lemma 3.2 of (Davis and Yin, 2017),

$$\begin{gathered} y^\star = T_{\mathrm{DY}} y^\star, \quad x^\star = \mathrm{prox}_{\rho f}\left(y^\star\right) \\ \Longleftrightarrow \\ -\nabla h(x^\star) \in \partial f(x^\star) + \partial g(x^\star). \end{gathered}$$

Therefore, if the fixed-point iteration converges, a solution can be recovered via $x_{k+1} = \mathrm{prox}_{\rho f}\left(y_k\right)$.

Several well-known splitting methods arise as special cases of TOS. For instance, setting $g = 0$ recovers the proximal gradient method, while setting $h = 0$ yields the Douglas–Rachford splitting method that has been used to derive fast OT solvers in (Mai et al., 2022; Lindbäck et al., 2023).

3.1 Three-operator splitting with delays

The main idea behind our algorithm design is to adapt the TOS scheme to support asynchronous updates where gradients can be reused across iterations, to reduce the computational costs the gradient updates incur. To this end, we introduce a sequence of positive delays $\tau_k \in [k]$, and consider the following asynchronous analogue of equation 3:

$$\begin{aligned} x_{k+1} &= \mathrm{prox}_{\rho f}\left(y_k\right), \\ z_{k+1} &= \mathrm{prox}_{\rho g}\left(2x_{k+1} - y_k - \rho\nabla h(x_{k+1-\tau_k})\right), \\ y_{k+1} &= y_k + z_{k+1} - x_{k+1}. \end{aligned} \tag{5}$$

To cast the optimal transport formulation in equation 1 in the TOS framework, we set $f = \iota_{[0,1]^{m\times n}}$, $g = \iota_{\bar{\mathcal{T}}(p,q)}$, and $h = \ell$, which results in the following delayed iterations

$$\begin{aligned} \gamma_{k+1} &= \mathrm{clip}_{[0,\,1]}\left(y_k\right), \\ \gamma'_{k+1} &= \mathrm{P}_{\bar{\mathcal{T}}(p,q)}\left(2\gamma_{k+1} - y_k - \rho\nabla\ell(\gamma_{k+1-\tau_k})\right), \\ y_{k+1} &= y_k + \gamma'_{k+1} - \gamma_{k+1}. \end{aligned} \tag{6}$$

By introducing auxiliary vectors and scalars, equation 6 can be considerably simplified. Specifically, through additional rank-1 operations, the y-iterate can be expressed

$$y_{k+1} = \gamma_{k+1} + \phi_{k+1}1_n^\top + 1_m\varphi_{k+1}^\top - \rho\texttt{corr}(\nabla\ell(\gamma_{k+1-\tau_k}))$$

where $\phi_{k+1} \in \mathbb{R}^m$, $\varphi_{k+1} \in \mathbb{R}^n$, and

$$\begin{aligned} \texttt{corr}(x) = x - n^{-1}x1_n1_n^\top - m^{-1}1_m1_m^\top x \\ + (mn)^{-1}(1_m^\top x1_n)1_m1_n^\top. \end{aligned} \tag{7}$$

A complete derivation of these steps is provided in the supplementary material, and the resulting algorithm is detailed in Algorithm 1.

Algorithm 1 ATOS with stale gradients

Require: Initialization $\gamma_0 \in \mathcal{T}(p,q)$, ϕ_0, φ_0, $\rho > 0$
1: `stale_grad` $= \texttt{corr}(\nabla\ell(\gamma_0))$, {see equation 7}
2: $a_0 = n\phi_0 + (1_n^\top\varphi_0)1_m$, $b_0 = m\varphi_0 + (1_m^\top\phi_0)1_n$, $\theta_0 = 1_m^\top a_0/(m+n)$
3: $k = 0$
4: **for** $t = 0, 1, 2, \ldots$ **do**
5: $\quad\tau = 0$
6: $\quad$**while** $\tau \le \tau_{\max}$ **or** stopping criteria fulfilled **do**
7: $\quad\quad y_k = \gamma_k + \phi_k 1_n^\top + 1_m\varphi_k^\top - \rho$`stale_grad`
8: $\quad\quad\gamma_{k+1} = \mathrm{clip}_{[0,\,1]}\left(y_k\right)$
9: $\quad\quad r_{k+1} = \gamma_{k+1}1_n - p$, $s_{k+1} = \gamma_{k+1}^\top 1_n - q$, $\eta_{k+1} = 1_m^\top r_{k+1}/(m+n)$
10: $\quad\quad\phi_{k+1} = \left(a_k - 2r_{k+1} + (2\theta_k - \eta_{k+1})1_m\right)/n$
11: $\quad\quad\varphi_{k+1} = \left(b_k - 2s_{k+1} + (2\theta_k - \eta_{k+1})1_n\right)/m$
12: $\quad\quad a_{k+1} = a_k - r_{k+1}$, $b_{k+1} = b_k - s_{k+1}$, $\theta_{k+1} = \theta_k - \eta_{k+1}$
13: $\quad\quad k \leftarrow k+1$, $\tau \leftarrow \tau + 1$
14: $\quad$**end while**
15: $\quad$`stale_grad` $\leftarrow \texttt{corr}(\nabla\ell(\gamma_k))$, {see equation 7}
16: **end for**
17: **return** γ_k

Implementation. The most computationally expensive steps in Algorithm 1, aside from the gradient computations, are the updates

$$\begin{gathered} \gamma_{k+1} = \mathrm{clip}_{[0,\,1]}\left(\gamma_k + \phi_k 1_n^\top + 1_m\varphi_k^\top - \rho\nabla\ell(\gamma_{k-\tau_k})\right), \\ r_k = \gamma_{k+1}1_n, \quad s_k = \gamma_{k+1}^\top 1_m. \end{gathered}$$

This is because they require reading and writing large matrices from global off-chip memory. To mitigate

this, we build on the GPU kernel developed in Mai et al. (2022), which performs all these operations, i.e. updating the transportation plan, and incrementing the row and column sums, while the data is held in low-latency, on-ship memory. In addition, by reusing the gradient an even number of times, the kernel only needs to read the cost matrix from memory every other iteration, reducing the number of memory operations substantially.

Delayed gradients also allow us to overlap the gradient computations completely. This is achieved through double-buffering by maintaining two tuples, each containing a transportation plan and its associated gradient: $(\gamma_{k,1}, g_{k,1})$ and $(\gamma_{k,2}, g_{k,2})$. One tuple is used for gradient computation, while the other is used to update the transport plan. Once the update is complete, the roles of the tuples are swapped and the process is repeated.We implement this in PyTorch using CUDA streams, which allow gradient and OT computations to run concurrently.

4 CONVERGENCE GUARANTEES

While the convergence of three-operator splitting (TOS) methods has been studied extensively, much less is known about their behavior in the presence of delayed gradient information. In this section, we establish convergence results for the delayed variant of TOS introduced in Section 3.1. We first consider the convex setting and characterize the convergence using monotone operator theory. Subsequently, we adapt the guarantees by Yurtsever et al. (2021) for the nonconvex case, leveraging that the feasible set is compact. Proofs are provided in the supplementary material.

The convex case. Analyzing delayed updates using tools from monotone operator theory presents additional challenges. In standard three-operator splitting, each update depends only on the previous iterate y_k. With delayed gradients, however, the update y_{k+1} also depends on a past iterate $y_{k-\tau_k}$ through the term $\nabla h(\mathrm{prox}_{\rho f}(y_{k-\tau_k}))$. This temporal coupling complicates the analysis, as the iteration is no longer governed by a single nonexpansive operator.

To better understand and control the behavior of such delayed iterations, we begin by analyzing the scheme where gradients are either refreshed or reused, resulting in a delay sequence where $\tau_{k+1} \in \{\tau_k + 1, 0\}$. We address other modes of asynchrony in the final part of this section.

First we introduce the following auxiliary operator $\bar{T} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$, which allows us to isolate the effect of the gradient input:

$$\begin{aligned}&\bar{T}(y,u)\\ &:= \mathrm{Id} + \mathrm{prox}_{\rho g}\left(2\mathrm{prox}_{\rho f}(y) - y - \rho u\right) - \mathrm{prox}_{\rho f}(y).\end{aligned}$$

This generalizes the Davis–Yin operator, since

$$\bar{T}(y, \nabla h(\mathrm{prox}_{\rho f}(y))) = T_{\mathrm{DY}} y,$$

and, in the delayed case, $\bar{T}(y_k, \nabla h(x_{k+1-\tau_k}))$ corresponds to the update rule in equation 5. To formalize the iteration with reused gradients, we define the composite operators $T^{(\tau)}$ for $\tau \geq 0$, which perform τ iterations of three-operator splitting while reusing the gradient computed at the first step:

$$\begin{aligned}T^{(0)} &:= \bar{T}(\cdot, \nabla h(\mathrm{prox}_{\rho f}(\cdot))),\\ T^{(\tau)} &:= \bar{T}(T^{(\tau-1)}(\cdot), \nabla h(\mathrm{prox}_{\rho f}(\cdot))), \quad \tau \geq 1. \end{aligned} \tag{8}$$

Importantly, $T^{(\tau)}$ share the same set of fixed points as T_{DY}, regardless of the number of internal iterations $\tau \geq 0$; we formalize this in Lemma 1. Throughout the convex analyses, we assume that f, g, h are closed, convex, and proper, h has L-Lipschitz continuous gradients, and $\mathrm{Fix}\, T_{\mathrm{DY}} \neq \emptyset$.

Lemma 1. *Let $T^{(\tau)}$ be defined according to equation 8, then $\mathrm{Fix}\, T^{(\tau)} = \mathrm{Fix}\, T_{\mathrm{DY}}$.*

Under the assumption that the stepsize lies in the admissible range $\rho < 2/(L(\tau+1)^2)$, we first establish the convergence behavior of $T^{(\tau)}$.

Proposition 1. *Let $\tau \geq 0$ and $\rho < 2/(L(\tau+1)^2)$ be fixed. Then $T^{(\tau)}$ is ν-quasi-nonexpansive, i.e., for any $y^\star \in \mathrm{Fix}\, T_{\mathrm{DY}}$*

$$\|T^{(\tau)} y - y^\star\|^2 \leq \|y - y^\star\|^2 - \nu \|y - T^{(\tau)} y\|^2$$

with

$$\nu = \frac{2 - L(\tau+1)^2 \rho}{2(\tau+1)}.$$

Assuming that the delay at each iteration k is bounded by $\tau_k \leq \tau$, the quasi-nonexpansiveness of $T^{(\tau)}$ allows us to prove summability of the fixed point residuals.

Theorem 1. *Let $y_{k+1} = T^{(\tau_k)} y_k$, where $0 \leq \tau_k \leq \tau$, and $\rho < 2/(L(\tau+1)^2)$. Then, for any $y^\star \in \mathrm{Fix}\, T_{\mathrm{DY}}$*

$$\mathrm{dist}\left(y_{k+1}, \mathrm{Fix}\, T_{\mathrm{DY}}\right) \leq \mathrm{dist}\left(y_k, \mathrm{Fix}\, T_{\mathrm{DY}}\right), \tag{9a}$$

$$\begin{aligned}&\sum_{k=0}^{t} \|T^{(\tau_k)} y_k - y_k\|^2\\ &\leq \frac{2(\tau+1)}{2 - L(\tau+1)^2\rho} \|y_0 - y^\star\|^2. \end{aligned} \tag{9b}$$

We also obtain sublinear convergence rates in both feasibility and objective suboptimality. We subsequently use this to establish the iteration complexity.

Theorem 2. *Let $y_{k+1} = T^{(\tau_k)} y_k$, where $0 \le \tau_k \le \tau$. If $\rho < 2/(L(\tau+1)^2)$, so that Theorem 1 holds, then, for $f^\star + g^\star + h^\star = \inf_{x\in\mathbb{R}^n} f(x) + g(x) + h(x)$, we have*

$$\min_{t\le k} \|x_t - z_t\| \le \frac{c_1}{\sqrt{k}} \|y_0 - y^\star\|$$

and

$$\min_{t\le k} f(x_t) + g(z_t) + h(x_t) - f^\star - g^\star - h^\star \le \frac{c_2}{\sqrt{k}} \|y_0 - y^\star\| + \frac{c_3}{k} \|y_0 - y^\star\|^2$$

where c_1, c_2, c_3 are positive constants dependent on the max delay, the stepsize and the Lipschitz constant. Both bounds are attained simultaneously at some iteration $k' \le k$.

Moreover, there exists a fourth constant $c_4 > 0$, depending on c_1, c_2, c_3 and the initialization, such that for any $\epsilon > 0$, if $k > c_4 \epsilon^{-2}$, there is an iteration $k' \le k$ such that

$$\|x'_k - z'_k\| < \epsilon,$$
$$f(x_{k'}) + g(z_{k'}) + h(x_{k'}) < f^\star + g^\star + h^\star + \epsilon.$$

We also provide an adaptation of Theorem 2 to our setting when applied to non-linear OT problems.

Corollary 1. *Consider equation 6 with bounded delays, i.e. $\tau_k \le \tau$, and $\tau_{k+1} \in \{0, \tau_k + 1\}$. Assume the stepsize satisfies $\rho < 2/(L(\tau+1)^2)$.*

Then, there exists a problem-dependent constant $C_2 > 0$ such that, for any $\epsilon > 0$, if $k > C_2 \epsilon^{-2}$, then there is an iterate $k' \le k$ such that

$$(\|\gamma_{k'} 1_n - p\|^2 + \|\gamma_{k'}^\top 1_m - q\|^2)^{1/2} < \epsilon,$$
$$\ell(\gamma_{k'}) < \ell^\star + \epsilon,$$

where $\ell^\star = \inf_{\gamma\in\mathcal{T}(p,q)} \ell(\gamma)$.

The non-convex case. Since many important applications of non-linear OT are non-convex, including Gromov–Wasserstein, we complement the operator theoretical contributions by adapting the non-convex analysis by (Yurtsever et al., 2021) to the asynchronous setting. By only assuming smoothness of the objective, bounded delays, and compact domains, we establish the following convergence result.

Theorem 3. *Let (x_k, z_k, y_k) be generated by equation 5. Assume that $f = \iota_A$ and $g = \iota_B$ where A and B are closed and convex, A is bounded, and delay is bounded by τ. Further assume h is L-smooth and define $D_1 = \operatorname{diam}(A \cap B)$, $D_0 = \operatorname{dist}(y_0, A \cap B)$, $D = D_0 + D_1$, and $G = \sup_{x\in A} \|\nabla h(x)\| < +\infty$. We use the stepsize*

$$\rho = \frac{D_1}{2(G + LD_1\tau)} K^{-2/3}, \tag{10}$$

where K is the number of iterations. Then, for any tolerance $\epsilon > 0$, if $K > \frac{10D^6}{D_1^3}\epsilon^{-3}$, the following holds for all $x \in A \cap B$:

$$\frac{1}{K}\sum_{k=1}^K \|x_k - z_k\| < \epsilon$$
$$\frac{1}{K}\sum_{k=1}^K \langle \nabla h(x_k),\, x_k - x\rangle < \frac{3(G + LD\tau)}{2}\epsilon.$$

We next specialize this result to our general optimal transport formulation equation 1, using the partition specified in equation 6.

Corollary 2. *Consider equation 6 with bounded delays, i.e., $\tau_k \le \tau$. Then, there exists a problem-dependent constant C_1, such that for any $\epsilon > 0$, if $K > C_1\epsilon^{-3}$, then for all $\gamma \in \mathcal{T}(p,q)$*

$$\frac{1}{K}\sum_{k=1}^K (\|\gamma_k 1_n - p\|^2 + \|\gamma_k^\top 1_m - q\|^2)^{1/2} < \epsilon$$
$$\frac{1}{K}\sum_{k=1}^K \langle \nabla \ell(\gamma_k),\, \gamma_k - \gamma\rangle < \epsilon.$$

Parallel gradient computations

Note that the non-convex analysis only requires the delay sequence to be bounded, meaning that the theory generalizes directly to other modes of asynchrony. One notable example is when the gradient computations run completely in parallel, a setting discussed in the Implementation paragraph. However, adapting the convex analysis to the parallel setting is more challenging since it requires the gradients to be decoupled from the iterates. Specifically, this setting is formalized using

$$\begin{bmatrix} g_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} \nabla h(\operatorname{prox}_{\rho f}(y_k)) \\ T^{(\tau)}_{\text{par.}}(y_k, g_k) \end{bmatrix}$$

where

$$T^{(0)}_{\text{par.}}(\cdot, g) \coloneqq \bar{T}(\cdot, g),$$
$$T^{(\tau)}_{\text{par.}}(\cdot, g) \coloneqq \bar{T}(T^{(\tau-1)}_{\text{par.}}(\cdot, g), g), \quad \tau \ge 1. \tag{11}$$

We show in the Supplementary material that Lemma 1 generalizes to this setting, i.e., the operator of equation 11 shares fixed points with T_{DY}. For the convergence, we note that the update can be interpreted as

an approximate version of $T^{(2\tau)}$, where an old gradient is used during the first τ iterations. Intuitively, it suffices to establish summability of the gradient differences and apply results from Zong et al. (2018); we leave a formal analysis to future work.

5 NUMERICAL RESULTS

In this section, we evaluate the performance of our method by running it on a series of problems. We first demonstrate the numerical advantages of reusing gradients in an ablation study on simulated data, followed by experiments on large-scale graph alignment problems and single-cell multi-omics translation tasks. We have chosen these experiments to demonstrate that our asynchronous adaptation of TOS outperforms relevant baselines in both speed and accuracy of the computed transportation plans. We focus specifically on large-scale problems that require GPU acceleration for practical runtimes, and thus restrict our comparisons to methods with readily available GPU implementations. All experiments were carried out on an NVIDIA RTX 5000 ADA generation, with 32GB of global memory. For entropic Gromov–Wasserstein, we compare with the POT implementation. For all methods, we use PyTorch with GPU acceleration enabled.

Throughout our experiments, we refer to our method as ATOS, short for asynchronous three-operator splitting. For ATOS, we will treat the number of iterations for which a gradient is reused as a fixed hyperparameter, denoted by τ. Moreover, the maximum number of outer iterations, which corresponds to the number of gradient calls for ATOS, will be denoted by N.

For a problem of size $m \times n$, we use the following constant stepsize $\rho = \rho_0 \frac{1}{(m+n)(1+\|\nabla h(\gamma_0)\|_\infty)}$, where ρ_0 is a sufficiently small scale factor, which we treat as a hyperparameter.

Ablation study. To compare ATOS with its synchronous counterpart, we test it on synthetic Gromov–Wasserstein problems. We generate datasets of size 4000 consisting of 5 isotropic Gaussian clusters randomly placed in $[-10, 10]^2$, each with noise level 1. Each experiment is repeated with 5 different random seeds. The algorithm terminates if the residual falls below 10^{-5} and the ℓ_1-norm of the difference between two consecutive iterates is less than $5\sqrt{\tau} \times 10^{-4}$. If these stopping criteria are not satisfied, the algorithm stops after 2000 iterations. Figure 1 summarizes the results, showing that ATOS with delayed gradients converges substantially faster than its synchronous variant for relevant step sizes. Additional experiments using different problem sizes and noise levels are added to the supplementary material.

Graph alignment. Gromov–Wasserstein is well-suited for graph comparison tasks, as node relationships are typically defined within each individual graph rather than across graphs. To demonstrate the applicability of our method, we apply it to a graph alignment problem using Gromov-Wasserstein. Inspired by the recent work of Li et al. (2023), we consider a dataset of real-world graphs representing social media interactions on Reddit (Yanardag and Vishwanathan, 2015).

The task involves aligning each graph with a noisy version of itself, using only the adjacency matrices. We introduce noise by randomly inserting an additional 10% edges and nodes, following the same experimental setup as in (Li et al., 2023). Since our primary goal is to design a scalable algorithm, we focus exclusively on graphs with at least 1,000 nodes, resulting in 202 graphs in total. As (Li et al., 2023) conduct extensive comparisons with other solvers, including (Peyré et al., 2016; Xu et al., 2019), and demonstrate that their BAPG method is both faster and more reliable, we omit those methods from our evaluation and compare only against BAPG.

Experimental setup. We consider 202 graphs from the Reddit database. First, we align each graph with itself. For the noisy counterpart, we repeat the process with five different random seeds and report the average results along with the maximum deviations from the means. For BAPG, we set $\rho = 0.1$, consistent with the choice in (Li et al., 2023). We also test several configurations for the number of inner and outer iterations, as shown in Figure 2. Finally, all methods use a relative error tolerance of 10^{-5} as the stopping criterion. The stepsize scale is set to $\rho_0 = 2/\tau$

Our results, presented in Figure 2, show that our method yields alignments that are more accurate than those of BAPG—while also being faster, especially on larger graphs.

Single-cell multi-omics translation Recent advances in single-cell technologies enable vast amounts of multi-modal data collection at single-cell resolution. Even when derived from the same cell population, cross-modality correspondences are typically unknown, and the data often differ in nature and dimensionality—limiting the applicability of classical alignment methods. Gromov–Wasserstein offers a compelling solution as it relies only on intra-modality (within-sample) distances to infer cross-modality correspondences. We adopt this approach, introduced by Demetci et al. (2022), to align ATAC and RNA embeddings from the bone marrow dataset (Luecken et al., 2021), accessed via the `moscot` package (Klein

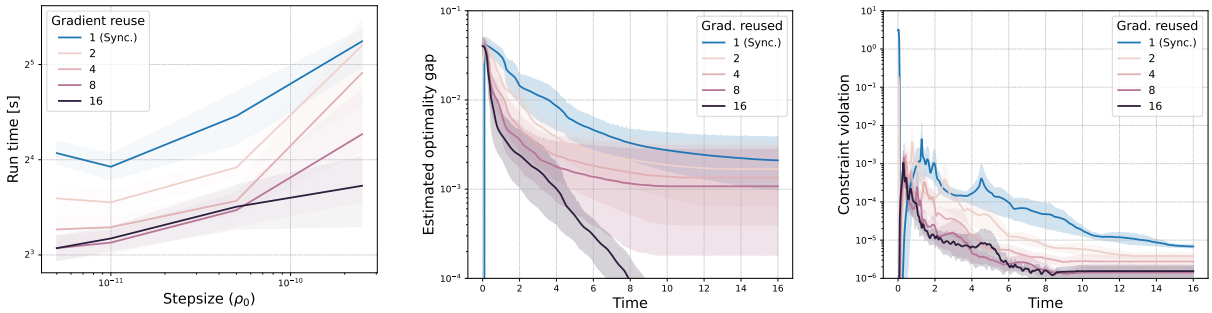

Figure 1: Ablation study evaluating the impact of gradient reuse when applying ATOS to the Gromov–Wasserstein problem. Experiments were conducted on simulated datasets of size 4000, repeated over 5 random seeds.

Method	Type	Iter. / grad.	N. iter	Raw data Time[s]	Raw data Acc.	Noisy data Time[s]	Noisy data Acc.
BAPG			1000	2.24	0.241	2.64 ± 0.001	0.219 ± 0.0001
			1500	3.43	0.283	4.03 ± 0.002	0.266 ± 0.0003
			2000	**4.62**	**0.299**	**5.44** ± 0.002	**0.285** ± 0.0003
ATOS	Stale	1 (Sync.)	2000	2.98	0.182	3.40 ± 0.006	0.126 ± 0.0010
		2	1500	**2.62**	**0.298**	**2.96** ± 0.005	**0.293** ± 0.0003
		4	1200	2.72	0.290	3.03 ± 0.004	0.286 ± 0.0003
		8	800	2.65	0.267	2.87 ± 0.003	0.263 ± 0.0005
		16	600	3.26	0.245	3.45 ± 0.003	0.242 ± 0.0005
	Par.	1	2000	3.17	0.181	3.59 ± 0.010	0.146 ± 0.0005
		2	1500	**2.77**	**0.297**	**3.12** ± 0.007	**0.293** ± 0.0007
		4	1200	2.84	0.290	3.15 ± 0.006	0.286 ± 0.0002
		8	800	2.74	0.266	2.96 ± 0.004	0.263 ± 0.0004
		16	600	3.32	0.244	3.51 ± 0.004	0.241 ± 0.0003

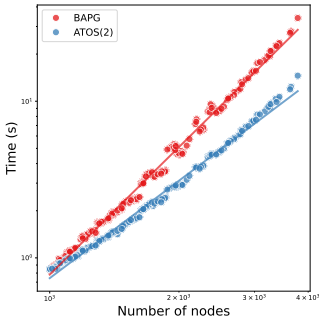

Figure 2: **Table (left):** Summary statistics of the graph alignment accuracy and wall-clock times for the Gromov-Wasserstein (GW) computations, executed on GPU. ATOS refers to our proposed method using both stale and parallel (par.) gradients. We benchmark both against BAPG, and highlight the best performing version for each category. **Figure (right):** Scalability plot showing how the runtime of the two best-performing methods increases with problem size.

et al., 2025). We compare our method—using $\tau = 2$, and $\rho_0 = 0.00025$—with the entropic PGD solver, a widely used baseline, evaluating performance using the FOSCTTM score (Demetci et al., 2022) (lower is better) to measure correspondence accuracy. Additionally, we assess alignment quality via the Sinkhorn Divergence (Genevay et al., 2018) between the translated and target embeddings. Since there is a trade-off between runtime and these accuracy metrics, we present our results using Pareto frontiers. The frontiers for both metrics can be found in Figure 3, where it is evident that the entropic PGD solver fails to achieve comparable performance within the same runtime budget. We generate the frontiers by tuning the regularization parameter for EGW, and the maximum number of iterations for ATOS.

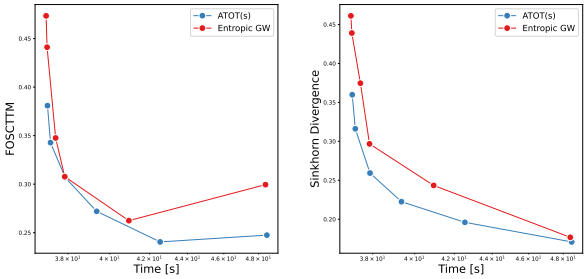

Figure 3: Multi-omics translation on the full bone marrow dataset. We illustrate both the runtime and the quality of the resulting correspondences using: **(Left):** the FOSCTTM score, and **(Right):** the Sinkhorn divergence, which measures the similarity between the aligned dataset and its target (lower is better for both scores).

6 CONCLUDING REMARKS

In this work, we introduce new algorithms based on three-operator splitting (TOS) for efficiently solving various OT problems. By extending TOS to handle stale gradients, we reduce gradient calls and enable fully parallel computation. Our GPU-adapted implementation is backed by two complementary convergence guarantees, offering both reliability and practical efficiency. Given their scalability and flexibility, we expect these methods to advance the state of the art in OT and related areas.

Our work opens several promising directions for future research. In particular, better leveraging parallel gradient computations requires a deeper understanding of how to optimally allocate resources between gradient and OT steps. A potentially more effective solution is to design a unified kernel that computes both the OT plan and gradients jointly, rather than using separate streams. Additionally, adaptive step sizes could further speed up convergence, though incorporating them remains challenging in our setting and is left for future work.

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

Altschuler, J., Niles-Weed, J., and Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. *Advances in neural information processing systems*, 30.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR.

Ballu, M. and Berthet, Q. (2023). Mirror Sinkhorn: Fast online optimization on transport polytopes. In *International Conference on Machine Learning*, pages 1595–1613. PMLR.

Bauschke, H. H. and Combettes, P. L. (2011). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer-Verlag.

Beck, A. (2017). *First-order methods in optimization*. SIAM.

Bian, F. and Zhang, X. (2021). A three-operator splitting algorithm for nonconvex sparsity regularization. *SIAM Journal on Scientific Computing*, 43(4):A2809–A2839.

Chambolle, A. and Contreras, J. P. (2022). Accelerated Bregman primal-dual methods applied to optimal transport and Wasserstein barycenter problems. *SIAM Journal on Mathematics of Data Science*, 4(4):1369–1395.

Chuang, C.-Y., Jegelka, S., and Alvarez-Melis, D. (2023). InfoOT: Information maximizing optimal transport. In *International Conference on Machine Learning*, pages 6228–6242. PMLR.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Davis, D. and Yin, W. (2017). A three-operator splitting scheme and its optimization applications. *Set-valued and variational analysis*, 25:829–858.

Demetci, P., Santorella, R., Sandstede, B., Noble, W. S., and Singh, R. (2022). SCOT: Single-cell multi-omics alignment with optimal transport. *Journal of computational biology*, 29(1):3–18.

Dvurechensky, P., Gasnikov, A., and Kroshnin, A. (2018). Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In *International conference on machine learning*, pages 1367–1376. PMLR.

Feyzmahdavian, H. R. and Johansson, M. (2023). Asynchronous iterations in optimization: New sequence results and sharper algorithmic guarantees. *Journal of Machine Learning Research*, 24(158):1–75.

Genevay, A., Peyré, G., and Cuturi, M. (2018). Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR.

Grave, E., Joulin, A., and Berthet, Q. (2019). Unsupervised alignment of embeddings with Wasserstein Procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890. PMLR.

Klein, D., Palla, G., Lange, M., Klein, M., Piran, Z., Gander, M., Meng-Papaxanthos, L., Sterr, M., Saber, L., Jing, C., et al. (2025). Mapping cells through time and space with moscot. *Nature*, pages 1–11.

Li, J., Tang, J., Kong, L., Liu, H., Li, J., So, A. M.-C., and Blanchet, J. (2023). A convergent single-loop algorithm for relaxation of Gromov–Wasserstein in graph data. *The 11th International Conference on Learning Representations*.

Lindbäck, J., Wang, Z., and Johansson, M. (2023). Bringing regularized optimal transport to lightspeed: a splitting method adapted for GPUs. *Advances in Neural Information Processing Systems*, 36:26845–26871.

Lu, H. and Yang, J. (2024). PDOT: A practical primal-dual algorithm and a gpu-based solver for optimal transport. *arXiv preprint arXiv:2407.19689*.

Luecken, M. D., Burkhardt, D. B., Cannoodt, R., Lance, C., Agrawal, A., Aliee, H., Chen, A. T., Deconinck, L., Detweiler, A. M., Granados, A. A., et al. (2021). A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (Round 2)*.

Mai, V. V., Lindbäck, J., and Johansson, M. (2022). A fast and accurate splitting method for optimal transport: analysis and implementation. *The 10th International Conferencee of Learning Representation*.

Mémoli, F. (2011). Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11:417–487.

Mishchenko, K., Bach, F., Even, M., and Woodworth, B. E. (2022). Asynchronous sgd beats minibatch sgd under arbitrary delays. *Advances in Neural Information Processing Systems*, 35:420–433.

Pedregosa, F. and Gidel, G. (2018). Adaptive three operator splitting. In *International Conference on Machine Learning*, pages 4085–4094. PMLR.

Peng, Z., Xu, Y., Yan, M., and Yin, W. (2016). Arock: an algorithmic framework for asynchronous parallel coordinate updates. *SIAM Journal on Scientific Computing*, 38(5):A2851–A2879.

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

Peyré, G., Cuturi, M., and Solomon, J. (2016). Gromov–Wasserstein averaging of kernel and distance matrices. In *International conference on machine learning*, pages 2664–2672. PMLR.

Recht, B., Re, C., Wright, S., and Niu, F. (2011). Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. *Advances in neural information processing systems*, 24.

Ryu, E. K. and Yin, W. (2022). *Large-scale convex optimization: algorithms & analyses via monotone operators*. Cambridge University Press.

Santambrogio, F. (2015). *Optimal transport for applied mathematicians*, volume 87. Springer.

Scetbon, M., Peyré, G., and Cuturi, M. (2022). Linear-time Gromov Wasserstein distances using low rank couplings and costs. In *International Conference on Machine Learning*, pages 19347–19365. PMLR.

Titouan, V., Courty, N., Tavenard, R., and Flamary, R. (2019). Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*, pages 6275–6284. PMLR.

Tong, A., Fatras, K., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Wolf, G., and Bengio, Y. (2024). Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*.

Villani, C. et al. (2008). *Optimal transport: old and new*, volume 338. Springer.

Xu, H., Luo, D., Zha, H., and Duke, L. C. (2019). Gromov–Wasserstein learning for graph matching and node embedding. In *International conference on machine learning*, pages 6932–6941. PMLR.

Yanardag, P. and Vishwanathan, S. (2015). Deep graph kernels. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1365–1374.

Yurtsever, A., Mangalick, V., and Sra, S. (2021). Three operator splitting with a nonconvex loss function. In *International Conference on Machine Learning*, pages 12267–12277. PMLR.

Zong, C., Tang, Y., and Cho, Y. J. (2018). Convergence analysis of an inexact three-operator splitting algorithm. *Symmetry*, 10(11):563.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes**
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **No** (link to repository will be included in camera-ready version).

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes**
   (b) Complete proofs of all theoretical results. **Yes**
   (c) Clear explanations of any assumptions. **Yes**

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **Yes**—all instructions are available, and and URL to a repository will be included in the camera-ready version.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes** (only applicable to hyperparameter choices)
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Yes**
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Yes**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. **Yes**
   (b) The license information of the assets, if applicable. **Yes**
   (c) New assets either in the supplemental material or as a URL, if applicable. **Not Applicable**
   (d) Information about consent from data providers/curators. **Yes**
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not Applicable**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. **Not Applicable**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not Applicable**
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not Applicable**

Three-Operator Splitting with Stale Gradients for Faster Non-Linear Optimal Transport

A ALGORITHM DERIVATION

Inspired by Mai et al. (2022), we leverage the following projection formula to simplify the algorithm

$$\mathrm{P}_{\bar{\mathcal{T}}(p,q)}(\gamma) = \gamma - n^{-1}\left((\gamma 1_n - p)\,1_n^\top - \delta 1_m 1_n^\top\right) - m^{-1}\left(1_m(\gamma^\top 1_m - q)^\top - \delta 1_m 1_n^\top\right), \tag{12}$$

where $\delta = 1_m^\top(\gamma 1_n - p)/(m+n) = 1_n^\top\left(\gamma^\top 1_m - q\right)/(m+n)$.

Recall that

$$\begin{aligned}
\gamma_{k+1} &= \mathrm{clip}_{[0,\,1]}(y_k),\\
\gamma'_{k+1} &= \mathrm{P}_{\bar{\mathcal{T}}(p,q)}(2\gamma_{k+1} - y_k - \rho\nabla h(\gamma_{k+1-\tau_k})),\\
y_{k+1} &= y_k + \gamma'_{k+1} - \gamma_{k+1}.
\end{aligned}$$

Since $\mathrm{P}_{\bar{\mathcal{T}}(p,q)}(\gamma)$ can be partitioned into $\mathrm{P}_{\bar{\mathcal{T}}(p,q)}(\gamma) = \gamma - \Delta(\gamma)$, where Δ is the rank-1 updates detailed in equation 12, i.e.,

$$\Delta(\gamma) = n^{-1}\left((\gamma 1_n - p)\,1_n^\top - \delta 1_m 1_n^\top\right) + m^{-1}\left(1_m(\gamma^\top 1_m - q)^\top - \delta 1_m 1_n^\top\right),$$

we can eliminate the γ'-update, resulting in

$$\begin{aligned}
\gamma_{k+1} &= \mathrm{clip}_{[0,\,1]}(y_k),\\
y_{k+1} &= \gamma_{k+1} - \rho\nabla h(\gamma_{k+1-\tau_k}) - \Delta(2\gamma_{k+1} - y_k - \rho\nabla h(\gamma_{k+1-\tau_k})).
\end{aligned}$$

Next, we define the following quantities associated with y_k:

$$\begin{aligned}
a_k &= y_k 1_n - p,\\
b_k &= y_k^\top 1_n - q,\\
\theta_k &= 1_m^\top a_k/(m+n),
\end{aligned}$$

Similarly, for γ_k, we let

$$r_k = \gamma_k 1_n - p,\ s_k = \gamma_k^\top 1_m - q,\ \text{and } \eta_k = 1_m^\top r_k/(m+n) = 1_n^\top s_k/(m+n).$$

These auxillary variables allow us to express

$$\begin{aligned}
\Delta(2\gamma_{k+1} - y_k - \rho\nabla h(\gamma_{k+1-\tau_k})) &= n^{-1}(2r_{k+1} - a_k - (2\eta_{k+1} - \theta_k)1_m)1_n^\top\\
&\quad + m^{-1}1_m(2s_{k+1} - b_k - (2\eta_{k+1} - \theta_k)1_n)^\top\\
-\rho(n^{-1}\nabla h(\gamma_{k+1-\tau_k})1_n 1_n^\top + m^{-1}1_m 1_m^\top\nabla h(\gamma_{k+1-\tau_k}) &- (mn)^{-1}(1_m^\top\nabla h(\gamma_{k+1-\tau_k})1_n)1_m 1_n^\top)
\end{aligned}$$

By denoting the terms:

$$\begin{aligned}
\phi_{k+1} &= n^{-1}(a_k - 2r_{k+1} - (\theta_k - 2\eta_{k+1})1_m)1_n^\top,\\
\varphi_{k+1} &= m^{-1}1_m(b_k - 2s_{k+1} - (\theta_k - 2\eta_{k+1})1_n)^\top,\\
\texttt{grad_corr}_{k+1} &= n^{-1}\nabla h(\gamma_{k+1-\tau_k})1_n 1_n^\top + m^{-1}1_m 1_m^\top\nabla h(\gamma_{k+1-\tau_k}) - (mn)^{-1}(1_m^\top\nabla h(\gamma_{k+1-\tau_k})1_n)1_m 1_n^\top,
\end{aligned}$$

then the preceding update can be written compactly as

$$y_{k+1} = \gamma_{k+1} + \phi_{k+1}1_n^\top + 1_m\varphi_{k+1}^\top - \rho(\nabla h(\gamma_{k+1-\tau_k}) - \texttt{grad\_corr}_{k+1}).$$

Hence, we can eliminate y iterate as well, yielding

$$\gamma_{k+1} = \text{clip}_{[0,\,1]}\left(\gamma_k + \phi_k 1_n^\top + 1_m\varphi_k^\top - \rho(\nabla h(\gamma_{k-\tau_k}) - \texttt{grad\_corr}_{k+1})\right) \tag{13}$$

Since $\texttt{grad\_corr}_{k+1}1_n = \nabla h(\gamma_{k-\tau_k})1_n$, and $\texttt{grad\_corr}_{k+1}^\top 1_m = \nabla h(\gamma_{k-\tau_k})^\top 1_m$, it follows that the ancillary variables can be updated according to $a_{k+1} = a_k - r_k$, $b_{k+1} = b_k - s_k$, and $\theta_{k+1} = \theta_k - \eta_k$. Using that $\texttt{corr}(\nabla h(\gamma_{k-\tau_k})) = \nabla h(\gamma_{k-\tau_k}) - \texttt{grad\_corr}_{k+1}$ gives the updates of Algorithm 1.

B CONVERGENCE GUARANTEES

We begin this appendix by deriving results used in proofs for both the convex and the non-convex case.

Proximal inequality. The following inequality (see e.g. (Beck, 2017, Theorem 6.9) for more details) will be used extensively in establishing the convergence results. For any $x \in \mathbb{R}^n$, and with $x^+ = \text{prox}_{\rho f}(x)$, we have

$$\rho^{-1}\left\langle x - x^+,\, u - x^+\right\rangle \le f(u) - f(x^+), \quad \text{for all } u \in \mathbb{R}^n. \tag{14}$$

In particular, suppose x^+, z^+, and y^+ are generated by

$$\begin{aligned} x^+ &= \text{prox}_{\rho f}(y) \\ z^+ &= \text{prox}_{\rho g}\left(2x^+ - y - \rho\nabla h(x^-)\right) \\ y^+ &= y + z^+ - x^+ \end{aligned} \tag{15}$$

then, by applying equation 14, we obtain,

$$\begin{aligned} \rho^{-1}\left\langle y - x^+,\, u - x^+\right\rangle &\le f(u) - f(x^+) \\ \rho^{-1}\left\langle 2x^+ - y - \rho\nabla h(x^-) - z^+,\, u - z^+\right\rangle &\le g(u) - g(z^+). \end{aligned}$$

Using the update $y^+ = y + z^+ - x^+$, we rewrite the inequalities as

$$\begin{aligned} \rho^{-1}\left\langle y - x^+,\, u - x^+\right\rangle &\le f(u) - f(x^+) \\ \rho^{-1}\left\langle x^+ - y^+ - \rho\nabla h(x^-),\, u - z^+\right\rangle &\le g(u) - g(z^+). \end{aligned} \tag{16}$$

These inequalities form the basis for the following useful result.

Lemma 2. *Let x^+, z^+, y^+ be given by the update rules in equation 15. Then for all $u \in \mathbb{R}^n$*

$$\begin{aligned} \frac{1}{2\rho}\left\|y^+ - u\right\|^2 - \frac{1}{2\rho}\left\|y - u\right\|^2 + \frac{1}{2\rho}\left\|y^+ - y\right\|^2 + \left\langle\nabla h(x^-),\, z^+ - u\right\rangle \\ \le f(u) + g(u) - f(x^+) - g(z^+). \end{aligned} \tag{17}$$

Proof. As

$$\rho^{-1}\left\langle y - x^+,\, u - x^+\right\rangle = \rho^{-1}\left\langle y^+ - x^+,\, u - x^+\right\rangle + \rho^{-1}\left\langle y - y^+,\, u - x^+\right\rangle,$$

we add the inequalities in equation 16 and rewrite as follows

$$\rho^{-1}\left\langle y^+ - x^+,\, y^+ - y\right\rangle + \rho^{-1}\left\langle y - y^+,\, u - x^+\right\rangle + \left\langle\nabla h(x^-),\, z^+ - u\right\rangle \tag{18}$$

$$\le f(u) + g(u) - f(x^+) - g(z^+). \tag{19}$$

This simplifies to:

$$\rho^{-1}\left\langle y^+ - y,\, y^+ - u\right\rangle + \left\langle\nabla h(x^-),\, z^+ - u\right\rangle \le f(u) + g(u) - f(x^+) - g(z^+).$$

By using the cosine identity: $2\langle a-b, c-b\rangle = \|a-b\|^2 + \|c-b\|^2 - \|a-c\|^2$ to the first term, we obtain the desired results:

$$\frac{1}{2\rho}\left\|y^+ - u\right\|^2 - \frac{1}{2\rho}\|y-u\|^2 + \frac{1}{2\rho}\left\|y^+ - y\right\|^2 + \left\langle \nabla h(x^-),\, z^+ - u\right\rangle \\ \leq f(u) + g(u) - f(x^+) - g(z^+),$$

□

Bridging residuals. When solving OT problems of the form specified in equation 1, marginal constraint violations are often more insightful—and therefore preferred— over conditions on the form $\|\gamma_k - \gamma_k'\|$. The following Lemma provides a simple way to bridge these two types of conditions.

Lemma 3. *Let $(\gamma_k, \gamma_k', y_k)$ be generated by equation 6. Then it holds that*

$$\left(\|\gamma_k 1_n - p\|^2 + \left\|\gamma_k^\top 1_m - q\right\|^2\right)^{1/2} \leq (m+n)^{1/2}\|\gamma_k - \gamma_k'\|.$$

Proof. Consider the linear operator $\mathcal{A}: \mathbb{R}^{m\times n} \to \mathbb{R}^{m+n}$ defined $\mathcal{A}\gamma \coloneqq (\gamma 1_m; \gamma^\top 1_n)$. Then we can express the marginal constraint condition as

$$\begin{aligned}\left(\|\gamma_k 1_n - p\|^2 + \left\|\gamma_k^\top 1_m - q\right\|^2\right)^{1/2} &= \|\mathcal{A}\gamma_k - (p;q)\| \\ &= \|\mathcal{A}(\gamma_k - \gamma_k')\| \\ &\leq \|\mathcal{A}\|_{\text{op.}}\|\gamma_k - \gamma_k'\|\end{aligned}$$

where $\|\mathcal{A}\|_{\text{op.}}$ denotes the operator norm of $\mathcal{A}$. Since $\|\mathcal{A}\|_{\text{op.}} = (m+n)^{1/2}$, we thus have

$$\left(\|\gamma_k 1_n - p\|^2 + \left\|\gamma_k^\top 1_m - q\right\|^2\right)^{1/2} \leq (m+n)^{1/2}\|\gamma_k - \gamma_k'\|,$$

which completes the proof. □

We will now start with the convex case.

B.1 The convex case

Assuming h is convex allows us to use additional tools from monotone operator theory. The following result concerns the averagedness of the three-operator splitting operator—defined according to equation 4—and will be key when establishing the convergence results. For a detailed proof, see (Davis and Yin, 2017, Proposition 3.1).

Lemma 4. *Let T_{DY} be defined as in equation 4. If $\rho \in (0,\ 2/L)$, then T_{DY} is α-averaged with $\alpha = 2/(4-\rho L)$, That is, for any $x, y \in \mathbb{R}^n$, we have*

$$\|T_{\text{DY}}x - T_{\text{DY}}y\|^2 \leq \|x-y\|^2 - \frac{1-\alpha}{\alpha}\|(\text{Id} - T_{\text{DY}})x - (\text{Id} - T_{\text{DY}})y\|^2.$$

Before establishing Lemma 1, we will need the following two results:

Lemma 5. *For any $u \in \mathbb{R}^n$, $\bar{T}(\cdot, u)$ is firmly non-expansive.*

Proof. Letting $\bar{g} = g + \langle u, \cdot\rangle$, we have

$$\begin{aligned}\bar{T}(\cdot, u) &= \text{Id} + \text{prox}_{\rho g}\left(2\text{prox}_{\rho f}(\cdot) - \text{Id} - \rho u\right) - \text{prox}_{\rho f}(\cdot) \\ &= \text{Id} + \text{prox}_{\rho \bar{g}}\left(2\text{prox}_{\rho f}(\cdot) - \text{Id}\right) - \text{prox}_{\rho f}(\cdot).\end{aligned}$$

This is the Douglas-Rachford operator associated with the functions f and $\bar{g}$, which is firmly non-expansive (see e.g., (Bauschke and Combettes, 2011, Proposition 26.1). □

We will also need the following Lemma that characterize fixed point sets of certain operator compositions. For a complete proof, see e.g. (Bauschke and Combettes, 2011, Corollary 4.51).

Lemma 6. *Let $\{T_k\}_{k=1}^s$ be a finite collection of α-averaged operators from $\mathbb{R}^n$ to $\mathbb{R}^n$. If $\bigcap_k \operatorname{Fix} T_k \neq \emptyset$, the set of fixed point of the finite concatenation $T = T_s \circ T_{s-1} \circ \cdots \circ T_1$, fulfills*

$$\operatorname{Fix} T = \bigcap_{i=1}^{s} \operatorname{Fix} T_i.$$

Now we will use these two auxiliary Lemmas to prove Lemma 1.

Proof of Lemma 1.

Proof. We start by fixing $y^\star \in \operatorname{Fix} T_{\mathrm{DY}}$, and let $x^\star = \operatorname{prox}_{\rho f}(y^\star)$, then

$$\begin{aligned} y^\star &= y^\star + \operatorname{prox}_{\rho g}\left(2\operatorname{prox}_{\rho f}(y^\star) - y^\star - \rho \nabla h(\operatorname{prox}_{\rho f}(y^\star))\right) - \operatorname{prox}_{\rho f}(y^\star) \\ &= \bar{T}(y^\star, \nabla h(\operatorname{prox}_{\rho f}(y^\star)) \end{aligned} \tag{20}$$

Therefore, $T^{(\tau)} y^\star = y^\star$, establishing that $y^\star \in \operatorname{Fix} T^{(\tau)}$.

Conversely, if we assume $y^\star \in \operatorname{Fix} T^{(\tau)}$ and letting $x^\star = \operatorname{prox}_{\rho}(y^\star)$, then

$$y^\star = \bar{T}(\cdot, \nabla h(x^\star)) \circ \bar{T}(\cdot, \nabla h(x^\star)) \circ \cdots \circ \bar{T}(\cdot, \nabla h(x^\star)) y^\star. \tag{21}$$

Since $\bar{T}(\cdot, \nabla h(x^\star))$ firmly non-expansive due to Lemma 5, and hence 1/2-averaged, we can use Lemma 6 to deduce that $y^\star = \bar{T}(\cdot, \nabla h(x^\star)) y^\star$. This, in turn, implies that $y^\star = T_{\mathrm{DY}} y^\star$, which completes the proof.

□

Establishing that the parallel gradient counterpart $T^{(\tau)}_{\mathrm{par.}}$, given by equation 11, shares fixed points with T_{DY} is analogous. Specifically, let $y^\star \in \operatorname{Fix} T_{\mathrm{DY}}$ and $g^\star = \nabla h(\operatorname{prox}_{\rho f}(y^\star))$. From equation 11, it follows that $y^\star = \bar{T}(y^\star, \nabla h(\operatorname{prox}_{\rho f}(y^\star))) = \bar{T}(y^\star, g^\star)$.. Consequently, $y^\star = T^{(\tau)}_{\mathrm{par.}}(y^\star, g^\star)$, meaning that indeed $(y^\star, g^\star)$ is a fixed point.

Conversely, suppose $(y^\star, g^\star)$ associated with $T^{(\tau)}_{\mathrm{par.}}$, i.e.,

$$\begin{bmatrix} g^\star \\ y^\star \end{bmatrix} = \begin{bmatrix} \nabla h(\operatorname{prox}_{\rho f}(y^\star)) \\ T^{(\tau)}_{\mathrm{par.}}(y^\star, g^\star) \end{bmatrix}.$$

This implies that $y^\star = T^{(\tau)}_{\mathrm{par.}}(y^\star, \nabla h(\operatorname{prox}_{\rho f}(y^\star)))$, which yields the exact same condition as equation 21. Therefore, analogous to the second part of the proof of Lemma 1, we thus have that $y^\star = \operatorname{Fix} T_{\mathrm{DY}}$.

Establishing that $T^{(\tau)}$ is ν-quasi-nonexpansive. Our proof results in conditions that depend on a recursive relationship that is non-trivial to expand. To address this, we adopt an approximate approach. The following result allows us to significantly simplify the computations while still capturing the correct dependence on the delay.

Lemma 7. *Let $s \geq 0$, and define $\eta_t(s) = s + (2 - \eta_{t-1}(s))^{-1}$ for $t \geq 1$ and $\eta_0(s) = s$. Then for any $t \geq 0$, we have that:*

$$\eta_t(0) = 1 - \frac{1}{t+1}, \quad \eta_t'(0) = \frac{(t+2)(2t+3)}{6(t+1)}. \tag{22}$$

Moreover, for any $t = 0, 1, 2, \ldots$, assumed that $0 < s < (t+1)^{-2}$, then $\eta_t(s) < 1$. In particular,

$$\eta_t(s) \leq 1 - (t+1)\left(\frac{1}{(t+1)^2} - s\right).$$

In addition, $\eta_t(s) > 1$ if $s > \kappa(t+1)^{-2}$, for some $\kappa \in (1, 3]$, meaning that decay complexity is tight.

Proof. Note that $\eta_{t+1}(0) = 1/(2 - \eta_t(0))$, and $\eta_0(0) = 0$. By unrolling the recursion, we get that $\eta_{t+1}(0) = \frac{t+1}{t+2}$, and hence $\eta_t(0) = 1 - \frac{1}{t+1}$.

For the derivative, we have the recursive relationship

$$\eta'_{t+1}(s) = 1 + \frac{\eta'_t(s)}{(2 - \eta_t(s))^2},$$

By substituting $\eta_t(0) = 1 - \frac{t}{t+1}$ we get

$$\eta'_{t+1}(0) = 1 + \frac{\eta'_t(0)}{(2 - \eta_t(0))^2} = 1 + \eta'_t(0)\frac{(t+1)^2}{(t+2)^2}.$$

Using that $\eta'_0(0) = 1$, we can unroll the recursion to get that

$$\eta'_{t+1}(0) = 1 + \frac{(t+1)^2}{(t+2)^2} + \frac{t^2}{(t+2)^2} + \frac{(t-1)^2}{(t+2)^2} + \cdots + \frac{1}{(t+2)^2} = \frac{1}{(t+2)^2}\sum_{n=1}^{t+2} n^2$$

If we use that $\sum_{n=1}^{N} n^2 = N(N+1)(2N+1)/6$, we can reexpress the derivative as

$$\eta'_{t+1}(0) = \frac{(t+3)(2t+5)}{6(t+2)}$$

We obtain the desired result through reindexing.

The second part is proven via induction. Clearly, as $\eta_0(s) = s$, the assertion holds for $t = 0$,. Therefore, we assume that it holds for all $t = 0, 1, \ldots, p$ for some $p \geq 0$. That is

$$\eta_t(s) < 1 \text{ if } s < \frac{1}{(t+1)^2} \text{ for } t = 1, 2, \ldots p,$$

and in particular,

$$\eta_t(s) < 1 \text{ if } s < \frac{1}{(p+2)^2} \text{ for } t = 1, 2, \ldots p.$$

A direct consequence of the induction assumption is thus that η_{p+1} is continuous for $s < \frac{1}{(p+2)^2}$. Moreover, for any $q \leq p+1$, we have

$$\eta'_q(s) = 1 + \frac{1}{(2 - \eta_{q-1}(s))^2}\eta'_{q-1}(s).$$

When $s < (p+2)^{-2}$, we thus have

$$\begin{aligned}\eta'_q(s) &< 1 + \underbrace{\frac{1}{(2 - \eta_{q-1}(s))^2}}_{<1}\eta'_{q-1}(s) \\ &\leq 1 + \eta'_{q-1}(s).\end{aligned}$$

Therefore, since $\eta'_0(s) = 1$, iterating the inequality yields

$$\eta'_{p+1}(s) \leq p + 2.$$

Furthermore

$$\eta''_{p+1}(s) = \underbrace{\frac{1}{(2 - \eta_p(s))^2}}_{>0}\eta''_p(s) + \underbrace{\frac{2}{(2 - \eta_p(s))^3}(\eta'_p(s))^2}_{>0}. \tag{23}$$

Since $\eta_0''(s) = 0$, we have that $\eta_{p+1}''(s) \geq 0$ when $s < (p+2)^{-2}$, and is thus convex in this interval. In particular

$$\eta_{p+1}(0) \geq \eta_{p+1}(s) + \eta_{p+1}'(s)(0 - s)$$

or

$$\begin{aligned}\eta_{p+1}(s) &\leq \eta_{p+1}(0) + \eta_{p+1}'(s)s \\ &\leq 1 - \frac{1}{p+2} + (p+2)s\end{aligned}$$

Therefore, $\eta_{p+1}(s) < 1$ if $s < (p+2)^{-2}$, which completes the first part of induction proof.

We will consider two cases to establish the decay complexity for $\eta_t(s)$. First, we assume that $\eta_{t-1}(s) < 2$ in some covering interval $s < m(t+1)^{-2}$, where $m \geq 3$. Then $\eta_{t-1}(s)$ is smooth and convex when $s < m(t+1)^{-2}$. This allows us to adopt the first-order convexity characterization. Specifically, for any s such that $3(p+1)^{-2} \leq s < m(p+1)^{-2}$ we have

$$\begin{aligned}\eta_t(s) &\geq \eta_t(0) + \eta_t'(0)s \\ &= 1 - \frac{1}{t+1} + \frac{(t+1)(2t+3)}{6(t+1)}s \\ &\geq 1 - \frac{1}{t+1} + \frac{(t+2)(2t+3)}{6(t+1)}\frac{3}{(t+1)^2} \\ &= 1 - \frac{1}{t+1} + \frac{1}{t+1}\frac{(2t+3)(t+2)}{(2t+2)(t+1)}. \\ &> 1,\end{aligned}$$

Therefore, the claim holds for the first case.

For the other case, there is an $s^\star \in [t^{-2}, 3(t+1)^{-2}$, such that $\eta_{t-1}(s^\star) = 2$, and $\eta_{t-1}(s) < 2$ when $s < s^\star$. Since $\eta_t(s) \to +\infty$, as $s \to s^\star$, and $\eta_t(s)$ is continuous for $s < s^\star$, there exists an $s_1 \in ((t+1)^{-2}, s^\star)$ such that $\eta_t(s_1) > 1$. As $s^\star < 3(t+1)^{-2}$, when have that $s_1 = \kappa(t+1)^{-2}$, for some $\kappa \in (1, 3]$, which completes the proof. □

To simplify the notation, we define $T_1 = \operatorname{prox}_{\rho g}(\cdot)$, $T_2 = \operatorname{prox}_{\rho f}(\cdot)$, $U = \mathrm{Id} - T_2$, $W = \rho\nabla h \circ T_2$, and $V = 2T_2 - \mathrm{Id} - W$. Using these definitions, we can compactly express the three-operator splitting operator, given by equation 4, as

$$T_{\mathrm{DY}} = U + T_1 \circ V.$$

In addition, we consider the operators $S, \bar{S} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ defined $\bar{S}(y', y) = (2T_2 - I)y' - Wy$, and

$$\begin{aligned}S(y', y) &\coloneqq Uy' + T_1 \underbrace{((2T_2 - I)y' - Wy)}_{\coloneqq \bar{S}(y', y)} &(24)\\ &= Uy' + T_1\bar{S}(y', y). &(25)\end{aligned}$$

In particular, note that $S(y, y) = T_{\mathrm{DY}}y$, and more generally, $T^{(\tau+1)}y = S(T^{(\tau)}y, y)$. The following bound on the normed difference between evaluations of S will thus be instrumental when proving that $T^{(\tau)}$ is ν-quasi-nonexpansive.

Lemma 8. *Let $y_1, y_2, z_1, z_2 \in \mathbb{R}^n$, and $\eta > 0$, then*

$$\begin{aligned}\|S(z_1, y_1) - S(z_2, y_2)\|^2 \leq \|y_1 - y_2\|^2 &- (1 - \eta)\|(y_1 - S(z_1, y_1)) - (y_2 - S(z_2, y_2))\|^2 \\ &+ \eta^{-1}\|(\mathrm{Id} + W)y_1 - (\mathrm{Id} + W)y_2 - (z_1 - z_2)\|^2 \\ &- 2\langle y_1 - y_2, (\mathrm{Id} + W)y_1 - (\mathrm{Id} + W)y_2 - (z_1 - z_2)\rangle \\ &+ 2\langle Uz_1 - Uz_2, Wy_1 - Wy_2\rangle.\end{aligned}$$

Proof. By using that U and T_1 are firmly non-expansive (see e.g. (Bauschke and Combettes, 2011, Proposition 23.8), we obtain

$$\begin{aligned}\|S(z_1,y_1)-S(z_2,y_2)\|^2 &= \|Uz_1-Uz_2\|^2 + \left\|T_1\bar{S}(z_1,y_1)-T_1\bar{S}(z_2,y_2)\right\|^2 \\ &\quad + 2\left\langle Uz_1-Uz_2,\, T_1\bar{S}(z_1,y_1)-T_1\bar{S}(z_2,y_2)\right\rangle \\ &\le \langle Uz_1-Uz_2,\, z_1-z_2\rangle + \left\langle T_1\bar{S}(z_1,y_1)-T_1\bar{S}(z_2,y_2),\, \bar{S}(z_1,y_1)-\bar{S}(z_2,y_2)\right\rangle \\ &\quad + 2\left\langle Uz_1-Uz_2,\, T_1\bar{S}(z_1,y_1)-T_1\bar{S}(z_2,y_2)\right\rangle \\ &= \langle Uz_1-Uz_2,\, z_1-z_2\rangle \\ &\quad + \langle T_1\bar{S}(z_1,y_1)-T_1\bar{S}(z_2,y_2), (2Uz_1+\bar{S}(z_1,y_1))-(2Uz_2+\bar{S}(z_2,y_2))\rangle.\end{aligned}$$

Since $2Uy'+\bar{S}(y',y)=y'-Wy$, we can simplify the bound as follows

$$\begin{aligned}\|S(z_1,y_1)-S(z_2,y_2)\|^2 &\le \langle Uz_1-Uz_2,\, z_1-z_2\rangle + \left\langle T_1\bar{S}(z_1,y_1)-T_1\bar{S}(z_2,y_2),\, (z_1-z_2)-(Wy_1-Wy_2)\right\rangle \\ &= \langle S(z_1,y_1)-S(z_2,y_2),\, z_1-z_2-(Wy_1-Wy_2)\rangle + \langle Uz_1-Uz_2,\, Wy_1-Wy_2\rangle \\ &= \langle S(z_1,y_1)-S(z_2,y_2),\, y_1-y_2\rangle \\ &\quad + \langle S(z_1,y_1)-S(z_2,y_2),\, (z_1-z_2)-(y_1-y_2)-(Wy_1-Wy_2)\rangle \\ &\quad + \langle Uz_1-Uz_2,\, Wy_1-Wy_2\rangle\end{aligned}$$

Since

$$\begin{aligned}\langle S(z_1,y_1)-S(z_2,y_2),\, y_1-y_2\rangle &= \frac{1}{2}\|S(z_1,y_1)-S(z_2,y_2)\|^2 + \frac{1}{2}\|y_1-y_2\|^2 \\ &\quad -\frac{1}{2}\|(y_1-S(z_1,y_1))-(y_2-S(z_2,y_2))\|^2,\end{aligned}$$

after rearranging terms, we obtain

$$\begin{aligned}\|S(z_1,y_1)-S(z_2,y_2)\|^2 &\le \|y_1-y_2\|^2 - \|(y_1-S(z_1,y_1))-(y_2-S(z_2,y_2))\|^2 \\ &\quad + 2\langle S(z_1,y_1)-S(z_2,y_2),\, (z_1-z_2)-(y_1-y_2)-(Wy_1-Wy_2)\rangle \\ &\quad + 2\langle Uz_1-Uz_2,\, Wy_1-Wy_2\rangle \\ &\le \|y_1-y_2\|^2 - \|(y_1-S(z_1,y_1))-(y_2-S(z_2,y_2))\|^2 \\ &\quad + 2\langle (y_1-S(z_1,y_1))-(y_2-S(z_2,y_2)), (\mathrm{Id}+W)y_1-(\mathrm{Id}+W)y_2-(z_1-z_2)\rangle \\ &\quad - 2\langle y_1-y_2,\, (\mathrm{Id}+W)y_1-(\mathrm{Id}+W)y_2-(z_1-z_2)\rangle + 2\langle Uz_1-Uz_2,\, Wy_1-Wy_2\rangle\end{aligned}$$

Applying Young's inequality—that is, for $\eta>0$

$$2\langle a,\, b\rangle \le \eta\|a\|^2 + \eta^{-1}\|b\|^2,$$

to the first inner product, using the parameter $\eta\in(0,1)$, yields

$$\begin{aligned}\|S(z_1,y_1)-S(z_2,y_2)\|^2 &\le \|y_1-y_2\|^2 - (1-\eta)\|(y_1-S(z_1,y_1))-(y_2-S(z_2,y_2))\|^2 \\ &\quad + \eta^{-1}\|(\mathrm{Id}+W)y_1-(\mathrm{Id}+W)y_2-(z_1-z_2)\|^2 \\ &\quad - 2\langle y_1-y_2,\, (\mathrm{Id}+W)y_1-(\mathrm{Id}+W)y_2-(z_1-z_2)\rangle \\ &\quad + 2\langle Uz_1-Uz_2,\, Wy_1-Wy_2\rangle.\end{aligned}$$

□

Finally, the following equivalent characterization of ν-quasi-nonexpansive operators will be used in our proof.

Lemma 9. *Let T be a ν-quasi-nonexpansive operator, meaning that there exists $\nu>0$, such that for any fixed point $y^\star$ of T, we have*

$$\|Ty-y^\star\|^2 \le \|y-y^\star\|^2 - \nu\|Ty-y\|^2,\ \textit{for all } y\in\mathbb{R}^n.$$

Then we have the following equivalent characterization

$$(1+\nu)\|Ty-y\|^2 + 2\langle y-y^\star,\, Ty-y\rangle \le 0,\ \textit{for all } y\in\mathbb{R}^n.$$

Proof. The assertion is quickly proven by expanding the quadratic of the left-hand side and rearranging the remaining terms. □

With these auxiliary results established, we are now ready to prove Proposition 1.

Proof of Proposition 1. Let $y^\star \in \operatorname{Fix} T_{\mathrm{DY}}$, which is, due to Lemma 1, also a fixed point to $T^{(q)}$ for $q = 0, 1, \dots \tau{+}1$. Let $y \in \mathbb{R}^n$. Since $T^{(\tau+1)}y = S(T^{(\tau)}y, y)$, we invoke Lemma 8 with $y_1 = y$, $y_2 = y^\star$, $z_1 = T^{(\tau)}y$, and $z_2 = T^{(\tau)}y^\star = y^\star$, yielding

$$\begin{aligned}
\left\|T^{(\tau+1)}y - y^\star\right\|^2 &= \left\|S(T^{(\tau)}y, y) - S(y^\star, y^\star)\right\|^2 \\
&\le \|y - y^\star\|^2 - (1-\eta)\left\|y - T^{(\tau+1)}y\right\|^2 \\
&\quad + \eta^{-1}\left\|(\mathrm{Id}+W)y - (\mathrm{Id}+W)y^\star - (T^{(\tau)}y - y^\star)\right\|^2 \\
&\quad - 2\left\langle y - y^\star,\, (\mathrm{Id}+W)y - (\mathrm{Id}+W)y^\star - (T^{(\tau)}y - y^\star)\right\rangle \\
&\quad + 2\left\langle UT^{(\tau)}y - Uy^\star,\, Wy - Wy^\star\right\rangle \\
&= \|y - y^\star\|^2 - (1-\eta)\left\|y - T^{(\tau+1)}y\right\|^2 \\
&\quad + \eta^{-1}\Big(\left\|y - T^{(\tau)}y + Wy - Wy^\star\right\|^2 + 2\eta\left\langle UT^{(\tau)}y - Uy^\star,\, Wy - Wy^\star\right\rangle \\
&\qquad - 2\eta\left\langle y - y^\star,\, y - T^{(\tau)}y + Wy - Wy^\star\right\rangle\Big) \\
&= \|y - y^\star\|^2 - (1-\eta)\left\|y - T^{(\tau+1)}y\right\|^2 + \eta^{-1}M_\eta^{(\tau)}(y, y^\star),
\end{aligned} \tag{26}$$

where

$$\begin{aligned}
M_\eta^{(\tau)}(y, y^\star) &\coloneqq \left\|y - T^{(\tau)}y + Wy - Wy^\star\right\|^2 \\
&\quad + 2\eta\left\langle UT^{(\tau)}y - Uy^\star,\, Wy - Wy^\star\right\rangle - 2\eta\left\langle y - y^\star,\, y - T^{(\tau)}y + Wy - Wy^\star\right\rangle.
\end{aligned} \tag{27}$$

Therefore, to establish that that $T^{(\tau+1)}$ is ν-quasi-nonexpansive, we characterize a set of stepsizes and η-parameters such that $M_\eta^{(t)}$ is non-positive. To this end, we rewrite $M_\eta^{(t)}$ as

$$\begin{aligned}
M_\eta^{(\tau)} &= \left\|y - T^{(\tau)}y + Wy - Wy^\star\right\|^2 + 2\eta\left\langle UT^{(\tau)}y - Uy^\star,\, Wy - Wy^\star\right\rangle \\
&\quad - 2\eta\left\langle y - y^\star,\, y - T^{(\tau)}y + Wy - Wy^\star\right\rangle \\
&= \left\|y - T^{(\tau)}y\right\|^2 + 2\left\langle y - T^{(\tau)}y,\, Wy - Wy^\star\right\rangle + \|Wy - Wy^\star\|^2 \\
&\quad + 2\eta\left\langle UT^{(\tau)}y - Uy^\star,\, Wy - Wy^\star\right\rangle - 2\eta\left\langle y - y^\star,\, y - T^{(\tau)}y + Wy - Wy^\star\right\rangle
\end{aligned} \tag{28}$$

Since $U = \mathrm{Id} - T_2$, we have

$$\begin{aligned}
&\left\langle UT^{(\tau)}y - Uy^\star,\, Wy - Wy^\star\right\rangle \\
&= \left\langle y - y^\star - (y - T^{(\tau)}y) - (T_2y - T_2y^\star) + T_2y - T_2T^{(\tau)}y,\, Wy - Wy^\star\right\rangle.
\end{aligned}$$

Substituting this back into equation 28 and collecting terms yields

$$\begin{aligned}
M_\eta^{(\tau)} &= \left\|y - T^{(\tau)}y\right\|^2 - 2\eta\left\langle y - y^\star,\, y - T^{(\tau)}y\right\rangle \\
&\quad + \|Wy - Wy^\star\|^2 - 2\eta\left\langle T_2y - T_2y^\star,\, Wy - Wy^\star\right\rangle \\
&\quad + 2(1-\eta)\left\langle y - T^{(\tau)}y,\, Wy - Wy^\star\right\rangle + 2\eta\left\langle T_2y - T_2T^{(\tau)}y,\, Wy - Wy^\star\right\rangle \\
&= \left\|y - T^{(\tau)}y\right\|^2 - 2\eta\left\langle y - y^\star,\, y - T^{(\tau)}y\right\rangle \\
&\quad + \|Wy - Wy^\star\|^2 - 2\eta\left\langle T_2y - T_2y^\star,\, Wy - Wy^\star\right\rangle \\
&\quad + 2\left\langle (1-\eta)(y - T^{(\tau)}y) + \eta(T_2y - T_2T^{(\tau)}y),\, Wy - Wy^\star\right\rangle.
\end{aligned}$$

Invoking Young's inequality again, with parameter $\theta >$, to the last inner product yields

$$\begin{aligned}
M_\eta^{(\tau)} \le \left\|y - T^{(\tau)}y\right\|^2 - 2\eta\left\langle y - y^\star,\, y - T^{(\tau)}y\right\rangle + (1+\theta)\|Wy - Wy^\star\|^2 - 2\eta\left\langle T_2y - T_2y^\star,\, Wy - Wy^\star\right\rangle \\
+ \theta^{-1}\left\|(1-\eta)(y - T^{(\tau)}y) + \eta(T_2y - T_2T^{(\tau)}y)\right\|^2.
\end{aligned}$$

We subsequently apply Jensen's inequality to the last term, and leverage that T_2 is non-expansive, which yields

$$\begin{aligned}
M_\eta^{(\tau)} &\le \left\|y - T^{(\tau)}y\right\|^2 - 2\eta\left\langle y - y^\star,\, y - T^{(\tau)}y\right\rangle + (1+\theta)\|Wy - Wy^\star\|^2 - 2\eta\left\langle T_2y - T_2y^\star,\, Wy - Wy^\star\right\rangle \\
&\quad + \theta^{-1}(1-\eta)\left\|y - T^{(\tau)}y\right\|^2 + \theta^{-1}\eta\left\|T_2y - T_2T^{(\tau)}y\right\|^2 \\
&\le \left\|y - T^{(\tau)}y\right\|^2 - 2\eta\left\langle y - y^\star,\, y - T^{(\tau)}y\right\rangle + (1+\theta)\|Wy - Wy^\star\|^2 - 2\eta\left\langle T_2y - T_2y^\star,\, Wy - Wy^\star\right\rangle \\
&\quad + \theta^{-1}\left\|y - T^{(\tau)}y\right\|^2 \\
&= (1+\theta^{-1})\left\|y - T^{(\tau)}y\right\|^2 - 2\eta\left\langle y - y^\star,\, y - T^{(\tau)}y\right\rangle + (1+\theta)\|Wy - Wy^\star\|^2 - 2\eta\left\langle T_2y - T_2y^\star,\, Wy - Wy^\star\right\rangle.
\end{aligned} \tag{29}$$

Recall that $W = \rho\nabla h(T_2\cdot)$. Since ∇h L-smooth and convex it is also is cocoersive. That is, for any $x_1, x_2 \in \mathbb{R}^n$, we have

$$\frac{1}{L}\|\nabla h(x_1) - \nabla h(x_2)\|^2 \le \langle \nabla h(x_1) - \nabla h(x_2),\, x_1 - x_2\rangle.$$

We will use this to bound all the terms in equation 29 associated with gradients.

$$\begin{aligned}
&(1+\theta)\|Wy - Wy^\star\|^2 - 2\eta\left\langle T_2y - T_2y^\star,\, Wy - Wy^\star\right\rangle \\
&= (1+\theta)\rho^2\|\nabla h(T_2y) - \nabla h(T_2y^\star)\|^2 - 2\eta\rho\left\langle T_2y - T_2y^\star,\, \nabla h(T_2y) - \nabla h(T_2y^\star)\right\rangle \\
&\le (1+\theta)\rho^2\|\nabla h(T_2y) - \nabla h(T_2y^\star)\|^2 - \frac{2\eta\rho}{L}\|\nabla h(T_2y) - \nabla h(T_2y^\star)\|^2 \\
&= \rho\left((1+\theta)\rho - \frac{2\eta}{L}\right)\|\nabla h(T_2y) - \nabla h(T_2y^\star)\|^2.
\end{aligned} \tag{30}$$

For the remaining terms, we apply an induction argument. When $t = 0$, then $T^{(0)} = T_{\mathrm{DY}}$, which is averaged with parameter $\alpha_0 = 2/(4-\rho L)$ due to Lemma 4. This implies that T_{DY} is ν_0-quasi-nonexpansive with $\nu_0 = (2-\rho L)/2$. Let $\eta_0 = 1 - \nu_0 = \rho L/2$, invoking Lemma 4, thus gives

$$-2\left\langle y - y^\star,\, y - T^{(0)}y\right\rangle \le -(1+\nu_0)\left\|y - T^{(0)}y\right\|^2 = -(2-\eta_0)\left\|y - T^{(0)}y\right\|^2 \tag{31}$$

Substituting the bounds equation 30 and equation 31, to equation 29 then yields

$$M_\eta^{(0)} \le \left(1 + \theta^{-1} - \eta(2-\eta_0)\right)\left\|y - T_1^{(0)}y\right\|^2 + \rho\left((1+\theta)\rho - \frac{2\eta}{L}\right)\|\nabla h(T_2y) - \nabla h(T_2y^\star)\|^2. \tag{32}$$

Hence, $M_\eta^{(0)} \leq 0$ if

$$\begin{aligned} 1 + \theta^{-1} - \eta(2 - \eta_0) &\leq 0 \\ (1+\theta)\rho - \frac{2\eta}{L} &\leq 0 \end{aligned}$$

or, by letting $s = \rho L/2$, these inequalities are equivalent to

$$\begin{aligned} (1 + \theta^{-1})(2 - \eta_0)^{-1} &\leq \eta \\ (1+\theta)s &\leq \eta \end{aligned}$$

We set $\theta = 1/(s(2-\eta_0))$ to maximize the allowed range for η, which gives the lower bound

$$s + \frac{1}{2 - \eta_0} \leq \eta.$$

Let $\eta_1 \coloneqq s + 1/(2-\eta_0)$, then we the condition simplifies

$$\eta_1 \leq \eta.$$

Substituting θ, and η_1 into equation 32 then yields

$$\begin{aligned} M_\eta^{(0)} &\leq (2-\eta_0)\left(s + \frac{1}{2-\eta_0} - \eta\right)\left\|y - T^{(0)}y\right\|^2 + \frac{4^2 s}{L^2}\left(s + \frac{1}{2-\eta_0} - \eta\right)\|\nabla h(T_2 y) - \nabla h(T_2 y^\star)\|^2. \\ &= (2-\eta_0)(\eta_1 - \eta)\left\|y - T^{(0)}y\right\|^2 + \frac{4^2 s}{L^2}(\eta_1 - \eta)\|\nabla h(T_2 y) - \nabla h(T_2 y^\star)\|^2 \end{aligned}$$

Therefore, the smallest possible η ensuring that $M_\eta^{(0)} \leq 0$, is when $\eta = \eta_1$, which results in $M_{\eta_1}^{(0)} = 0$. Consequently, when $\eta = \eta_1$, equation 26 reduces to

$$\left\|T^{(1)}y - y^\star\right\|^2 \leq \|y - y^\star\|^2 - (1-\eta_1)\left\|y - T^{(1)}y\right\|^2$$

Let $\nu_1 = 1 - \eta_1$, then $T_1^{(1)}$ is ν_1-quasi-averaged if $\eta_1 < 1$. This gives the following condition on the stepsize:

$$\eta_1 = s + \frac{1}{2-s} < 1 \iff s < \frac{3-\sqrt{5}}{2}$$

or equivalently $\rho < \frac{3-\sqrt{5}}{L}$.

Therefore, indeed, the Proposition holds for $t = 1$, since if when $\rho < \frac{2}{L(1+t)^2} = \frac{1}{2L}$, then clearly $\rho < \frac{3-\sqrt{5}}{L}$, which ensures $\eta_1 < 1$. Hence, $T_1^{(1)}$ ν_1-quasi-nonexpansive under the stepsize condition stated in the Proposition.

If we now assume that the Proposition holds for all $t \leq p-1$. Then all aforementioned results generalize for $t = p$ with

$$\eta_p = s + \frac{1}{2 - \eta_{p-1}}$$

and

$$M_\eta^{(p-1)} \leq (2-\eta_{p-1})(\eta_p - \eta)\left\|y - T^{(p-1)}y\right\|^2 + \frac{4^2 s}{L^2}(\eta_p - \eta)\|\nabla h(T_2 y) - \nabla h(T_2 y^\star)\|^2. \tag{33}$$

In particular, if we set $\eta = \eta_p$, then

$$\left\|T^{(p)}y - y^\star\right\|^2 \leq \|y - y^\star\|^2 - (1-\eta_p)\left\|y - T^{(p)}y\right\|^2.$$

Using the recursion from Lemma 7, we obtain that $\eta_p < 1$ if $s < 1/(p+1)^2$, or equivalently, $\rho < 2/(L(p+1)^2)$ implies $M_\eta^{(p-1)} \leq 0$. Moreover, since $\eta_p < 1 - (p+1)((p+1)^{-2} - s)$, the best averagedness parameter derived using this proving technique is bounded below according to $\nu_p \geq (p+1)((p+1)^{-2} - s) = \frac{2 - L(p+1)^2\rho}{2(p+1)}$, which completes the proof. □

We can strengthen the result further, which will later be necessary when deriving convergence rates.

Corollary 3. *Consider $T^{(\tau)}$ for $t \geq 1$, and let $\epsilon \in (0,1)$. Then if $\rho < \frac{2\epsilon}{L(t+1)^2}$, by letting, $s = \rho L/2$, meaning that $s \in (0, \epsilon/(t+1)^2)$, we have*

$$\left\|T^{(\tau)}y - y^\star\right\|^2 \leq \|y - y^\star\|^2 - c_1 \left\|y - T^{(\tau)}y\right\|^2 - c_2 \left\|y - T^{(\tau-1)}y\right\|^2 - c_3 \left\|\nabla h(T_2 y) - \nabla h(T_2 y^\star)\right\|^2$$

where

$$\begin{aligned} c_1 &= \frac{1-\epsilon}{t+1}, \\ c_2 &= \frac{1}{2}\left(\epsilon - (t+1)^2 s\right), \\ c_3 &= \frac{2s}{L^2}\left(\epsilon - (t+1)^2 s\right), \end{aligned}$$

Proof. Resubstiting equation 33 into equation 26 yields

$$\begin{aligned} \left\|T^{(\tau)}y - y^\star\right\|^2 &\leq \|y - y^\star\|^2 - (1-\eta)\left\|y - T^{(\tau)}y\right\|^2 + \eta^{-1} M_\eta^{(\tau-1)}(y, y^\star) \\ &\leq \|y - y^\star\|^2 - (1-\eta)\left\|y - T^{(\tau)}y\right\|^2 \\ +\eta^{-1}\left((2-\eta_{t-1})(\eta_t - \eta)\left\|y - T^{(\tau-1)}y\right\|^2 + \frac{4^2 s}{L^2}(\eta_t - \eta)\left\|\nabla h(T_2 y) - \nabla h(T_2 y^\star)\right\|^2\right) \end{aligned}$$

Since $\eta_t \leq \eta < 1$ by construction, all but the first term are negative. Set η to the maximum value permitted by the stepsize range in Proposition 1, i.e.

$$\eta = 1 - \left.\frac{1-(t+1)^2 s}{t+1}\right|_{s=\epsilon/(t+1)^2} = 1 - \frac{1-\epsilon}{t+1},$$

and since $(2 - \eta_{t-1}) > 1$, we get

$$\begin{aligned} \left\|T^{(\tau)}y - y^\star\right\|^2 &\leq \|y - y^\star\|^2 - \frac{1-\epsilon}{t+1}\left\|y - T^{(\tau)}y\right\|^2 - \frac{t+1}{1+\epsilon}\left(\frac{\epsilon - (t+1)^2 s}{t+1}\right)\left\|y - T^{(\tau-1)}y\right\|^2 \\ &\quad - \frac{t+1}{1+\epsilon}\frac{4s}{L^2}\left(\frac{\epsilon - (t+1)^2 s}{t+1}\right)\left\|\nabla h(T_2 y) - \nabla h(T_2 y^\star)\right\|^2 \\ &= \|y - y^\star\|^2 - \frac{1-\epsilon}{t+1}\left\|y - T^{(\tau)}y\right\|^2 - \left(\frac{\epsilon - (t+1)^2 s}{1+\epsilon}\right)\left\|y - T^{(\tau-1)}y\right\|^2 \\ &\quad - \frac{4s}{L^2}\left(\frac{\epsilon - (t+1)^2 s}{1+\epsilon}\right)\left\|\nabla h(T_2 y) - \nabla h(T_2 y^\star)\right\|^2. \end{aligned}$$

Substituting $-1/(1+\epsilon) < -1/2$ into the inequality, yields the desired result. □

Proof of Theorem 1. Proofs of variants of Theorem 1 appear in many textbooks on monotone operators, e.g. Ryu and Yin (2022). However, these proofs often assume that the same averaged operator is used in every iteration—a condition that does not necessarily hold in our setting, as gradient delays are allowed to vary. Therefore, for completeness, we provide an adapted proof for our setting here.

Proof. Since $y_{k+1} = T^{(\tau_k)} y_k$, by invoking Proposition 1, we have that, for any fixed point $y^\star \in \operatorname{Fix} T_{\mathrm{DY}}$,

$$\|y_{k+1} - y^\star\|^2 \leq \|y_k - y^\star\|^2 - \nu_{\tau_k}\|y_{k+1} - y_k\|^2$$

where $\nu_{\tau_k} = \frac{2 - L(\tau_k+1)^2\rho}{2(\tau_k+1)} > 0$. Consequently, we also have

$$\|y_{k+1} - y^\star\|^2 \leq \|y_k - y^\star\|^2,$$

and by minimizing both sides with respect to $y^\star \in \operatorname{Fix} T_{\mathrm{DY}}$, we establish *1.*

As $\tau_k \leq \tau$, letting $\nu_\tau = \frac{2-L(\tau+1)^2\rho}{2(\tau+1)} > 0$ yields the lower bound $\nu_{\tau_k} \geq \nu_\tau$, which subsequently yields the following uniform bound:

$$\|y_{k+1} - y^\star\|^2 \leq \|y_k - y^\star\|^2 - \nu_\tau \|y_{k+1} - y_k\|^2,$$

or equivalently,

$$\|y_{k+1} - y_k\|^2 \leq \frac{1}{\nu_\tau}\left(\|y_k - y^\star\|^2 - \|y_{k+1} - y^\star\|^2\right).$$

As the right-hand side of the inequality is telescoping, we average of the $k+1$ first inequalities, resulting in

$$\begin{aligned}\frac{1}{k+1}\sum_{t=0}^{k}\|y_{t+1} - y_t\|^2 &\leq \frac{1}{\nu_\tau(k+1)}\left(\|y_0 - y^\star\|^2 - \|y_{k+1} - y^\star\|^2\right)\\ &\leq \frac{1}{\nu_\tau(k+1)}\|y_0 - y^\star\|^2.\end{aligned}$$

Since $y^\star \in \operatorname{Fix} T_{\mathrm{DY}}$ is arbitrary, we minimize the right-hand side of the inequality with respect to $y^\star$. By substituting ν_τ we obtain

$$\frac{1}{k+1}\sum_{t=0}^{k}\|y_{t+1} - y_t\|^2 \leq \frac{2(\tau+1)}{(2-L(\tau+1)^2\rho)(k+1)}\operatorname{dist}^2(y_0, \operatorname{Fix} T_{\mathrm{DY}}),$$

which establishes *2.* □

Proof of Theorem 2 Let $y_{k+1} = T^{(\tau_k)}y_k$, $x_{k+1} = T_2 y_k$, $y^\star \in T_{\mathrm{DY}}$ be a fixed point, and $x^\star = T_2 y^\star$. For the given stepsize, let $\epsilon \in (\frac{\rho L(\tau+1)^2}{2}, 1)$. Then we invoke Corollary 3 to obtain

$$\begin{aligned}\|y_{k+1} - y^\star\|^2 \leq \|y_k - y^\star\|^2 - c_1\|y_{k+1} - y_k\|^2 - c_2\left\|y_k - T_1^{(\tau_k)}y_k\right\|^2\\ - c_3\|\nabla h(x_{k+1}) - \nabla h(x^\star)\|^2.\end{aligned}$$

where $c_1, c_2, c_3 > 0$ are given in Corollary 3. Rearranging the terms, and summing the inequalities gives

$$\sum_{t=0}^{k}\left(c_1\|y_{t+1} - y_t\|^2 + c_2\left\|y_t - T^{(\tau_t - 1)}y_t\right\|^2 + c_3\|\nabla h(x_{t+1}) - \nabla h(x^\star)\|^2\right) \leq \|y_0 - y^\star\|^2 \tag{34}$$

Therefore, all three quantities are summable, which will be crucial in establishing the rates.

Consider the last inner update associated with $T_1^{(\tau_k)}$:

$$\begin{aligned}y'_k &= T^{(\tau_k - 1)}y_k,\\ x'_{k+1} &= \operatorname{prox}_{\rho f}(y'_k),\\ z'_{k+1} &= \operatorname{prox}_{\rho g}\left(2x'_{k+1} - y'_k - \rho\nabla h(\operatorname{prox}_{\rho f}(y_k))\right),\\ y_{k+1} &= y'_k + z'_{k+1} - x'_{k+1}.\end{aligned}$$

The iterates x'_{k+1} and z'_{k+1}, are of main interest, since they correspond to last primal solution processed by the algorithm. If we now apply equation 18 in Lemma 2 using $(x^+, z^+, y^+) = (x'_{k+1}, z'_{k+1}, y_{k+1})$, $y = y'_k$, $x^- = \nabla h(\operatorname{prox}_{\rho f}(y_k))$ and $u = x^\star$, which yield

$$\begin{aligned}\rho^{-1}\langle y_{k+1} - y'_k, y_{k+1} - x^\star\rangle + \langle\nabla h(\operatorname{prox}_{\rho f}(y_k)), z'_{k+1} - x^\star\rangle\\ \leq f^\star + g^\star - f(x'_{k+1}) - g(z'_{k+1}).\end{aligned}$$

Moreover, as h is smooth and convex, we have

$$\left\langle \nabla h(x'_{k+1}),\, x^\star - x'_{k+1} \right\rangle \leq h(x^\star) - h(x'_{k+1}).$$

By adding these two inequalities and rearranging terms we obtain

$$\begin{aligned} &f(x'_{k+1}) + g(z'_{k+1}) + h(x'_{k+1}) - (f^\star + g^\star + h^\star) \\ \leq\, &-\rho^{-1} \left\langle y_{k+1} - y'_k,\, y_{k+1} - x^\star \right\rangle - \left\langle \nabla h(\mathrm{prox}_{\rho f}(y_k)),\, z'_{k+1} - x^\star \right\rangle - \left\langle \nabla h(x'_{k+1}),\, x^\star - x'_{k+1} \right\rangle \end{aligned}$$

Notice that the terms involving gradients can be expressed

$$\begin{aligned} &- \left\langle \nabla h(\mathrm{prox}_{\rho f}(y_k)),\, z'_{k+1} - x^\star \right\rangle - \left\langle \nabla h(x'_{k+1}),\, x^\star - x'_{k+1} \right\rangle \\ =\, &- \left\langle \nabla h(\mathrm{prox}_{\rho f}(y_k)),\, y_{k+1} - y'_k \right\rangle + \left\langle \nabla h(x'_{k+1}) - \nabla h(\mathrm{prox}_{\rho f}(y_k)),\, x'_{k+1} - x^\star \right\rangle. \end{aligned}$$

Therefore, we can bound the suboptimality by

$$\begin{aligned} &f(x'_{k+1}) + g(z'_{k+1}) + h(x'_{k+1}) - (f^\star + g^\star + h^\star) \\ &\leq -\rho^{-1} \left\langle y_{k+1} - y'_k,\, y_{k+1} + \rho \nabla h(\mathrm{prox}_{\rho f}(y_k)) - x^\star \right\rangle \\ &\quad + \left\langle \nabla h(x'_{k+1}) - \nabla h(\mathrm{prox}_{\rho f}(y_k)),\, x'_{k+1} - x^\star \right\rangle \\ &\leq \rho^{-1} \left\| y_{k+1} - y'_k \right\| \left\| y_{k+1} + \rho \nabla h(\mathrm{prox}_{\rho f}(y_k)) - x^\star \right\| + \left\| \nabla h(x'_{k+1}) - \nabla h(\mathrm{prox}_{\rho f}(y_k)) \right\| \left\| x'_{k+1} - x^\star \right\| \\ &\leq \rho^{-1} \left\| y_{k+1} - y'_k \right\| \left\| y_{k+1} + \rho \nabla h(\mathrm{prox}_{\rho f}(y_k)) - x^\star \right\| + L \left\| y'_k - y_k \right\| \left\| y_k - y^\star \right\|. \end{aligned}$$

Since y_{k+1} is obtained through applying a DR operator τ_k times. Assuming that this operator has a fixed point, say $y^\star_k$, then the first term decays as $\|y_{k+1} - y'_k\| \leq \|y_k - y^\star_k\| / \sqrt{1+\tau_k}$. However, we have no results quantifying the rate of $\|y_k - y^\star_k\|$, so we leave this direction for future work. Moreover, for our algorithm specifically, we can use equation 13 to write $y_{k+1} + \rho \nabla h(\mathrm{prox}_{\rho f}(y_k) = \phi_k 1_n^\top + 1_m \varphi_k^\top$, which is helpful in deriving tighter bounds. Also this is left for future work.

To derive rates without adding any additional assumptions, we use the conservative bound $\|y_{k+1} - y'_k\| \leq \|y_{k+1} - y_k\| + \|y'_k - y_k\|$. Moreover, we use that

$$\begin{aligned} &\left\| y_{k+1} + \rho \nabla h(\mathrm{prox}_{\rho f}(y_k)) - x^\star \right\| \\ &\leq \|y_{k+1} - y^\star\| + \rho \left\| \nabla h(\mathrm{prox}_{\rho f}(y_k)) - \nabla h(x^\star) \right\| + \|x^\star - y^\star\| + \rho \|\nabla h(x^\star)\|. \end{aligned}$$

Since $\|y_k - y^\star\| \leq \|y_0 - y^\star\|$, we thus have the bound

$$\begin{aligned} &f(x'_{k+1}) + g(z'_{k+1}) + h(x'_{k+1}) - (f^\star + g^\star + h^\star) \\ &\leq \left(\rho^{-1} \|y_0 - y^\star\| + \rho^{-1} \|x^\star - y^\star\| + \|\nabla h(x^\star)\| \right) \|y_{k+1} - y_k\| \\ &\quad + \left((L + \rho^{-1}) \|y_0 - y^\star\| + \rho^{-1} \|x^\star - y^\star\| + \|\nabla h(x^\star)\| \right) \|y'_k - y_k\| \\ &\quad + \left(\|y_{k+1} - y_k\| + \|y'_k - y_k\| \right) \left\| \nabla h(\mathrm{prox}_{\rho f}(y_k)) - \|\nabla h(x^\star)\| \right\| \\ &\leq C_k \sqrt{c_1 \|y_{k+1} - y_k\|^2 + c_2 \|y'_k - y_k\|^2 + c_3 \left\| \nabla h(\mathrm{prox}_{\rho f}(y_k)) - \|\nabla h(x^\star)\| \right\|^2}, \end{aligned}$$

where C_k is defined in terms of the coefficients:

$$\begin{aligned} a_1 &= \left(\rho^{-1} \|y_0 - y^\star\| + \rho^{-1} \|x^\star - y^\star\| + \|\nabla h(x^\star)\| \right) \\ a_2 &= \left((L + \rho^{-1}) \|y_0 - y^\star\| + \rho^{-1} \|x^\star - y^\star\| + \|\nabla h(x^\star)\| \right) \\ a_{3,k} &= \left(\|y_{k+1} - y_k\| + \|y'_k - y_k\| \right) \end{aligned}$$

via

$$C_k = \left(\frac{a_1^2}{c_1} + \frac{a_2^2}{c_2} + \frac{a_{3,k}^2}{c_3} \right)^{1/2}.$$

By equation 34, there exists a $k' \leq k$, such that

$$c_1 \left\| y_{k'+1} - y'_k \right\|^2 + c_2 \left\| y'_{k'} - y_{k'} \right\|^2 + c_3 \left\| \nabla h(\text{prox}_{\rho f}(y_{k'})) - \|\nabla h(x^\star)\| \right\|^2 \leq \frac{1}{k+1} \left\| y_0 - y^\star \right\|^2.$$

Moreover

$$C_k \leq \frac{a_1}{\sqrt{c_1}} + \frac{a_2}{\sqrt{c_2}} + \frac{a_{3,k}}{\sqrt{c_3}}$$

Which thus gives that

$$\begin{aligned} &f(x'_{k'+1}) + g(z'_{k'+1}) + h(x'_{k'+1}) - (f^\star + g^\star + h^\star) \\ &\leq \left(\frac{a_1}{\sqrt{c_1}} + \frac{a_2}{\sqrt{c_2}} + \frac{a_{3,k}}{\sqrt{c_3}} \right) \left\| y_0 - y^\star \right\| (k+1)^{-1/2}. \end{aligned}$$

Recall that

$$\begin{aligned} c_1 &= \frac{1-\epsilon}{\tau+1}, \\ c_2 &= \frac{1}{2}\left(\epsilon - (\tau+1)^2 s\right), \\ c_3 &= \frac{2s}{L^2}\left(\epsilon - (\tau+1)^2 s\right), \end{aligned}$$

By choosing e.g. $\epsilon = 1/2$, and resubstituting $s = \rho L/2$, we obtain

$$\begin{aligned} c_1 &= \frac{1}{2(\tau+1)}, \\ c_2 &= \frac{1}{4}\left(1 - (\tau+1)^2 \rho L\right), \\ c_3 &= \frac{\rho}{2L}\left(1 - (\tau+1)^2 \rho L\right) \end{aligned}$$

Then

$$\begin{aligned} \frac{a_1}{\sqrt{c_1}} &= \sqrt{2(\tau+1)} \Big(\rho^{-1} \left\| y_0 - y^\star \right\| + \rho^{-1} \left\| x^\star - y^\star \right\| + \|\nabla h(x^\star)\| \Big), \\ \frac{a_2}{\sqrt{c_2}} &= 2(1 - (\tau+1)^2 \rho L)^{-1/2} \Big((L + \rho^{-1}) \left\| y_0 - y^\star \right\| + \rho^{-1} \left\| x^\star - y^\star \right\| + \|\nabla h(x^\star)\| \Big), \\ \frac{a_{3,k}}{\sqrt{c_3}} &= \sqrt{2L}(\rho(1 - (\tau+1)^2 \rho L))^{-1/2} \Big(\left\| y_{k+1} - y_k \right\| + \left\| y'_k - y_k \right\| \Big) \end{aligned}$$

Since

$$\begin{aligned} &\Big(\left\| y_{k'+1} - y_{k'} \right\| + \left\| y'_{k'} - y_{k'} \right\| \Big) \\ &\leq \left(\frac{1}{c_1} + \frac{1}{c_2} \right)^{1/2} \sqrt{c_1 \left\| y_{k'+1} - y_{k'} \right\|^2 + c_2 \left\| y'_{k'} - y_{k'} \right\|^2 + c_3 \left\| \nabla h(\text{prox}_{\rho f}(y_{k'})) - \nabla h(x^\star) \right\|^2} \\ &\leq \left(\frac{1}{c_1} + \frac{1}{c_2} \right)^{1/2} \left\| y_0 - y^\star \right\| (k+1)^{-1/2} \\ &\leq \left(\frac{1}{\sqrt{c_1}} + \frac{1}{\sqrt{c_2}} \right) \left\| y_0 - y^\star \right\| (k+1)^{-1/2}, \end{aligned}$$

we thus have

$$\frac{a_{3,k'}}{\sqrt{c_3}} \leq \sqrt{\frac{2L}{\rho}}\left(\sqrt{\frac{2(\tau+1)}{1-(\tau+1)^2\rho L}} + \frac{2}{1-(\tau+1)^2\rho L}\right)\|y_0 - y^\star\|\,(k+1)^{-1/2}$$

Putting everything together yields

$$\begin{aligned} f(x'_{k'+1}) + g(z'_{k'+1}) + h(x'_{k'+1}) - (f^\star + g^\star + h^\star) \\ \leq (M_1 + M_2)(k+1)^{-1/2} + M_3 k^{-1} \end{aligned}$$

where

$$\begin{aligned} M_1 &= \sqrt{2(\tau+1)}\left(\rho^{-1}\|y_0 - y^\star\| + \rho^{-1}\|x^\star - y^\star\| + \|\nabla h(x^\star)\|\right)\|y_0 - y^\star\|, \\ M_2 &= \sqrt{\frac{1}{2(1-(\tau+1)^2\rho L)}}\left((L+\rho^{-1})\|y_0 - y^\star\| + \rho^{-1}\|x^\star - y^\star\| + \|\nabla h(x^\star)\|\right)\|y_0 - y^\star\|, \\ M_3 &= \sqrt{\frac{2L}{\rho}}\left(\sqrt{\frac{2(\tau+1)}{1-(\tau+1)^2\rho L}} + \frac{2}{1-(\tau+1)^2\rho L}\right)\|y_0 - y^\star\|^2. \end{aligned}$$

For the quantity $\|x_{k+1} - z_{k+1}\|$, we use that

$$\begin{aligned} \left\|x'_{k+1} - z'_{k+1}\right\| &= \|y_{k+1} - y'_k\| \\ &\leq \|y_{k+1} - y_k\| + \|y'_k - y_k\| \\ &\leq \left(\frac{1}{\sqrt{c_1}} + \frac{1}{\sqrt{c_2}}\right)\|y_0 - y^\star\|\,(k+1)^{-1/2} \\ &= \left(\sqrt{2(\tau+1)} + \frac{2}{\sqrt{1-2(\tau+1)^2 s)}}\right)\|y_0 - y^\star\|\,(k+1)^{-1/2} \\ &= \left(\sqrt{2(\tau+1)} + \frac{2\sqrt{L}}{\sqrt{L-(\tau+1)^2\rho)}}\right)\|y_0 - y^\star\|\,(k+1)^{-1/2}. \end{aligned}$$

Establishing the Corollary is a direct consequence of Lemma 3, and that $f(x'_{k+1}) = g(z'_{k+1}) = 0$, and $h = \ell$

B.2 The non-convex case.

To establish Theorem 3 and Corollary 2, we will need the following Lemmas

Lemma 10. *Let $f = \iota_A$ $g = \iota_B$, and u be an arbitrary point in $A \cap B$. Assuming that the diameter of the intersection is bounded, i.e., $D_1 = \operatorname{diam}(A \cap B) < +\infty$, we let $D_0 = \operatorname{dist}(y_0, A \cap B)$, and $D = D_0 + D_1$. Then, if (x_k, y_k, z_k) is defined by equation 5, we have*

$$\frac{1}{2\rho}\left(\frac{1}{K}\sum_{k=1}^{K}\|x_k - z_k\|\right)^2 + \frac{1}{K}\sum_{k=1}^{K}\langle \nabla h(x_{k-\tau_k}),\, z_k - u\rangle \leq \frac{1}{2\rho K}D^2. \tag{35}$$

Proof. We invoke Lemma 2 by letting

$$(x, z, y) = (x_k, z_k, y_k), \quad (x^+, z^+, y^+) = (x_{k+1}, z_{k+1}, y_{k+1}), \quad x^- = x_{k+1-\tau_{k+1}}.$$

Notice that this results in the right-hand side being constantly zero, as $x_{k+1} \in A$, $z_{k+1} \in B$, and $u \in A \cap B$. Consequently, by summing equation 17 from $k = 0, 1, \ldots, K-1$, and rearranging the remaining terms we obtain

$$\frac{1}{2\rho}\sum_{k=0}^{K-1}\|y_{k+1} - y_k\|^2 + \sum_{k=0}^{K-1}\left\langle \nabla h(x_{k+1-\tau_{k+1}}),\, z_{k+1} - u\right\rangle \leq \frac{1}{2\rho}\|y_0 - u\|^2 - \frac{1}{2\rho}\|y_k - u\|^2. \tag{36}$$

By using the update rule $y_{k+1} - y_k = z_{k+1} - x_{k+1}$, and dividing both sides by K, we can invoke Jensen's inequality on the first term, yielding

$$\frac{1}{2\rho}\left(\frac{1}{K}\sum_{k=1}^{K}\|x_k - z_k\|\right)^2 + \frac{1}{K}\sum_{k=1}^{K}\langle \nabla h(x_{k-\tau_k}),\, z_k - u\rangle \le \frac{1}{2\rho K}\|y_0 - u\|^2 - \frac{1}{2\rho K}\|y_k - u\|^2 \le \frac{1}{2\rho K}\|y_0 - u\|^2 . \tag{37}$$

Let $v \in A \cap B$, then

$$\begin{aligned}\|y_0 - u\|^2 &= \|y_0 - v + v - u\|^2 \\ &\le \|y_0 - v\|^2 + \|v - u\|^2 + 2\|y_0 - v\|\,\|v - u\| \\ &\le \|y_0 - v\|^2 + 2D_1\|y_0 - v\| + D_1^2.\end{aligned}$$

Taking the supremum of both sides with respect to v gives that

$$\begin{aligned}\|y_0 - u\|^2 &\le \sup_{v\in A\cap B}(\|y_0 - v\|^2 + 2D_1\|y_0 - v\| + D_1^2) \\ &\le D_0^2 + 2D_0D_1 + D_1^2 \\ &= D^2.\end{aligned}$$

Resubstituting this back into equation 37 yields the desired result. □

We will also need the following result, which bounds the accumulated gradient errors induced by the delays.

Lemma 11. *Let $K \ge 1$, $\tau_k \le \tau$ for all $k = 1, 2, \dots K$. If h is and L-smooth, and the iterates are generated by equation 5, then*

$$\sum_{k=1}^{K}\|\nabla h(x_k) - \nabla h(x_{k-\tau_k})\| \le L\tau\sum_{k=1}^{K}\|x_k - z_k\| . \tag{38}$$

Proof.

$$\begin{aligned}\|\nabla h(x_k) - \nabla h(x_{k-\tau_k})\| &\le L\|x_k - x_{k-\tau_k}\| \\ &= L\left\|\sum_{t=1}^{\tau_k} x_{k+1-t} - x_{k-t}\right\| \\ &\le L\sum_{t=1}^{\tau_k}\|x_{k+1-t} - x_{k-t}\| .\end{aligned}$$

Since $x_{n+1} = \operatorname{prox}_{\rho f}(y_n)$, and the proximal operator is non-expansive, we have

$$\begin{aligned}\|\nabla h(x_k) - \nabla h(x_{k-\tau_k})\| &\le L\sum_{t=1}^{\tau_k}\|x_{k+1-t} - x_{k-t}\| \\ &\le L\sum_{t=1}^{\tau_k}\|y_{k-t} - y_{k-t-1}\| \\ &= L\sum_{t=1}^{\tau_k}\|x_{k-t} - z_{k-t}\| .\end{aligned}$$

For notational convenience, assume $\|x_k - z_k\| = 0$ for $k \leq 0$. Then

$$\begin{aligned}\sum_{k=1}^{K} \|\nabla h(x_k) - \nabla h(x_{k-\tau_k})\| &\leq L \sum_{k=1}^{K} \sum_{t=1}^{\tau_k} \|x_{k-t+1} - z_{k-t+t}\| \\ &\leq L \sum_{k=1}^{K} \sum_{t=1}^{\tau} \|x_{k-t} - z_{k-t}\| \\ &\leq L\tau \sum_{k=1}^{K} \|x_k - z_k\| .\end{aligned}$$

□

With these auxiliary results in place, we are now ready to derive the main results.

Proof (of Theorem 3). Let $x \in A \cap B$, then we have that

$$\begin{aligned}&\langle \nabla h(x_{k-\tau_k}),\, z_k - x\rangle \\ &= \langle \nabla h(x_k),\, x_k - x\rangle + \langle \nabla h(x_{k-\tau_k}),\, z_k - x_k\rangle + \langle \nabla h(x_{k-\tau_k}) - \nabla h(x_k),\, x_k - x\rangle \\ &\geq -\|\nabla h(x_k)\| \|x_k - x\| - \|\nabla h(x_{k-\tau_k})\| \|x_k - z_k\| - \|\nabla h(x_{k-\tau_k}) - \nabla h(x_k)\| \|x_k - x\| . \end{aligned} \tag{39}$$

Specifically,

$$\langle \nabla h(x_{k-\tau_k}),\, z_k - x\rangle \geq -GD - G\|x_k - z_k\| - \|\nabla h(x_{k-\tau_k}) - \nabla h(x_k)\| D.$$

By substituting this into equation 35 with $u = x$, we obtain

$$\begin{aligned}\frac{1}{2\rho}\left(\frac{1}{K}\sum_{k=1}^{K}\|x_k - z_k\|\right)^2 - GD_1 - \frac{D_1}{K}\sum_{k=1}^{K}\|\nabla h(x_{k-\tau_k}) - \nabla h(x_k)\| - \frac{G}{K}\sum_{k=1}^{K}\|x_k - z_k\| \\ \leq \frac{1}{2\rho K} D^2.\end{aligned}$$

Invoking Lemma 11, and rearranging terms yield

$$\begin{aligned}\frac{1}{2\rho}\left(\frac{1}{K}\sum_{k=1}^{K}\|x_k - z_k\|\right)^2 - (G + LD_1\tau)\frac{1}{K}\sum_{k=1}^{K}\|x_k - z_k\| \\ \leq \frac{1}{2\rho K} D^2 + GD_1.\end{aligned}$$

By solving this inequality for $\frac{1}{K}\sum_{k=1}^{K}\|x_k - z_k\|$ we obtain

$$\begin{aligned}&\frac{1}{K}\sum_{k=1}^{K}\|x_k - z_k\| \\ &\leq \rho\,(G + LD_1\tau) + \sqrt{\rho^2\,(G + LD_1\tau)^2 + \frac{1}{K}D^2 + 2D_1\rho G}.\end{aligned}$$

By substituting the stepsize

$$\rho = \frac{D_1}{2(G + LD_1\tau)} K^{-2/3}$$

we obtain

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\|x_k - z_k\| &\leq \frac{D_1}{2K^{2/3}} + \sqrt{\frac{D_1^2}{4K^{4/3}} + \frac{1}{K}D^2 + \frac{D_1^2}{K^{2/3}}} \\ &\leq \frac{D_1}{2K^{2/3}} + \frac{D_1}{K^{1/3}}\left(1 + \frac{1}{8K^{2/3}} + \frac{D^2}{2D_1^2K^{1/3}}\right) \\ &= \frac{D_1}{K^{1/3}} + \frac{D_1^2 + D^2}{2D_1K^{2/3}} + \frac{D_1}{8K}.\end{aligned} \tag{40}$$

Since $D_1 \leq D$, by bounding $K^{-2/3} \leq K^{-1/3}$, and $K^{-1} \leq K^{-1/3}$, we obtain

$$\frac{D_1}{K^{1/3}} + \frac{D_1^2 + D^2}{2D_1 K^{2/3}} + \frac{D_1}{8K} \leq \frac{17D^2}{8D_1} K^{-1/3}.$$

Deriving the simplified rate is analogous to how it is derived for the stationarity condition, we use equation 39 to express

$$\langle \nabla h(x_k),\, x_k - x \rangle \leq \langle \nabla h(x_{k-\tau_k}),\, z_k - x \rangle + G \left\| x_k - z_k \right\| + \left\| \nabla h(x_{k-\tau_k}) - \nabla h(x_k) \right\| D.$$

Using Lemma 10 and Lemma 11 yields

$$\begin{aligned}
&\frac{1}{K} \sum_{k=1}^{K} \langle \nabla h(x_k),\, x_k - x \rangle \\
&\leq \frac{1}{K} \sum_{k=1}^{K} \langle \nabla h(x_{k-\tau_k}),\, z_k - x \rangle + G \left\| x_k - z_k \right\| + \left\| \nabla h(x_{k-\tau_k}) - \nabla h(x_k) \right\| D_1 \\
&\leq \frac{1}{2\rho K} D^2 + (G + LD_1\tau) \frac{1}{K} \sum_{k=1}^{K} \left\| x_k - z_k \right\|.
\end{aligned}$$

By using equation 40, we obtain the rate

$$\begin{aligned}
&\frac{1}{K} \sum_{k=1}^{K} \langle \nabla h(x_k),\, x_k - x \rangle \\
&\leq (G + LD\tau) \frac{D^2}{D_1 K^{1/3}} + (G + LD\tau) \left(\frac{D_1}{K^{1/3}} + \frac{D_1^2 + D^2}{2D_1 K^{2/3}} + \frac{D_1}{8K} \right) \\
&= (G + LD\tau) \left(\frac{D_1^2 + D^2}{D_1 K^{1/3}} + \frac{D_1^2 + D^2}{2D_1 K^{2/3}} + \frac{D_1}{8K} \right) \\
&\leq (G + LD\tau) \frac{25D^2}{8D_1} K^{-1/3},
\end{aligned}$$

where the last step is analogous to how the simplified rate was derived for the averaged residual: $\frac{1}{K} \sum_{k=1}^{K} \left\| x_k - z_k \right\|$.

If we choose

$$K > \frac{10D^6}{D_1^3} \epsilon^{-3} > \left(\frac{8D_1}{17D^2} \epsilon \right)^{-3}$$

then

$$\frac{1}{K} \sum_{k=1}^{K} \left\| x_k - z_k \right\| < \epsilon.$$

Moreover

$$\begin{aligned}
\frac{1}{K} \sum_{k=1}^{K} \langle \nabla h(x_k),\, x_k - x \rangle &< (G + LD\tau) \frac{25}{17} \epsilon^{-3} \\
&< \frac{3(G + LD\tau)}{2} \epsilon^{-3}
\end{aligned}$$

□

C ADDITIONAL EXPERIMENTS—ABLATION STUDY

This section presents additional results from the ablation study detailed in Section 5, obtained using simulated datasets with varying problem sizes and noise levels.

Different problem sizes

In addition to the experiments in Section 5, which considered datasets of size 4000, we reran the ablation study with problem sizes of 1000 (Figure 4) and 2000 (Figure 5). The results are consistent with those reported in Figure 1.

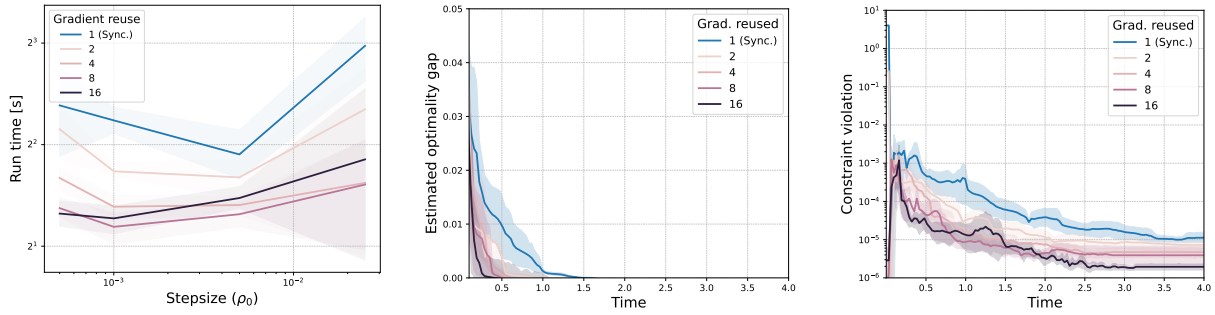

Figure 4: Ablation study evaluating the impact of reusing gradients when applying ATOS to the Gromov–Wasserstein problem. Experiments were conducted on simulated datasets of size 1000, repeated over 5 random seeds.

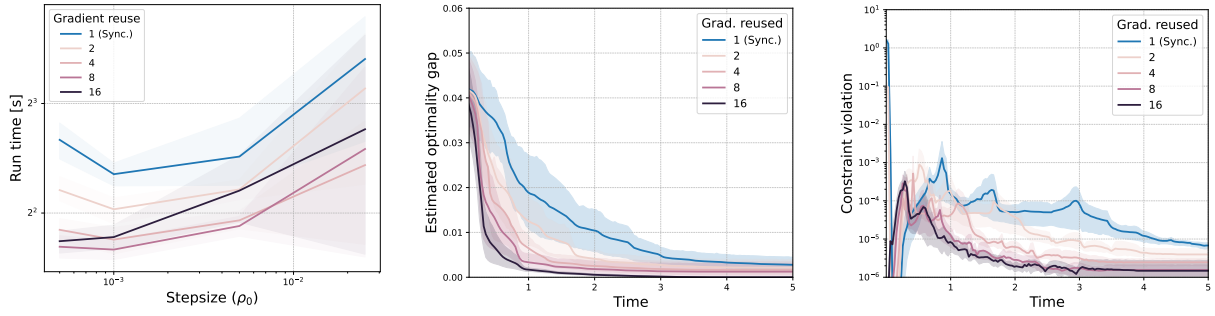

Figure 5: Ablation study evaluating the impact of reusing gradient when applying ATOS to the Gromov–Wasserstein problem. Experiments were conducted on simulated datasets of size 2000, repeated over 5 random seeds.

Different noise levels

We also reran the experiments using different noise levels for the simulated isotropic Gaussians. Varying the noise changes the geometry of the underlying optimization problem, and thus provides a stronger case for the generality of our results. These experiments were conducted on datasets of size 4000, with a lower noise level of 0.5 (Figure 6) and a higher noise level of 2 (Figure 7). The results remain consistent with those presented in the main paper, although the benefits of reusing gradients vary slightly across regimes.

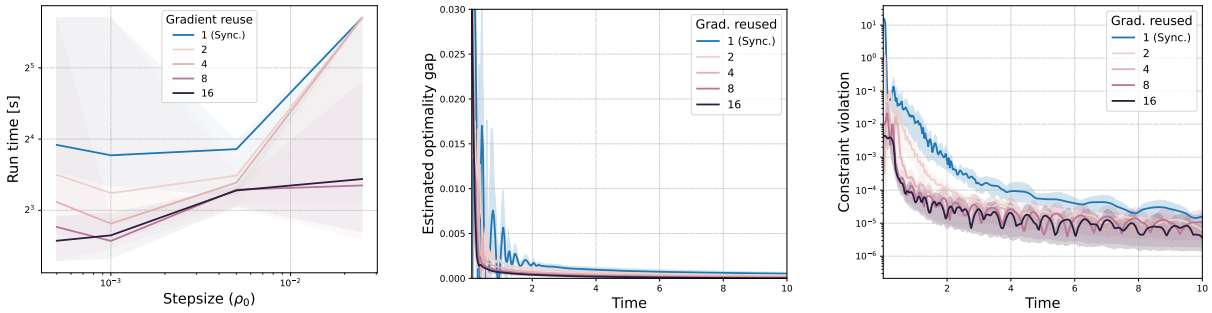

Figure 6: Ablation study evaluating the impact of reusing gradients when applying ATOS to the Gromov–Wasserstein problem. Experiments were conducted on simulated datasets of size 4000, repeated over 5 random seeds. The noise level used in this experiment was 0.5

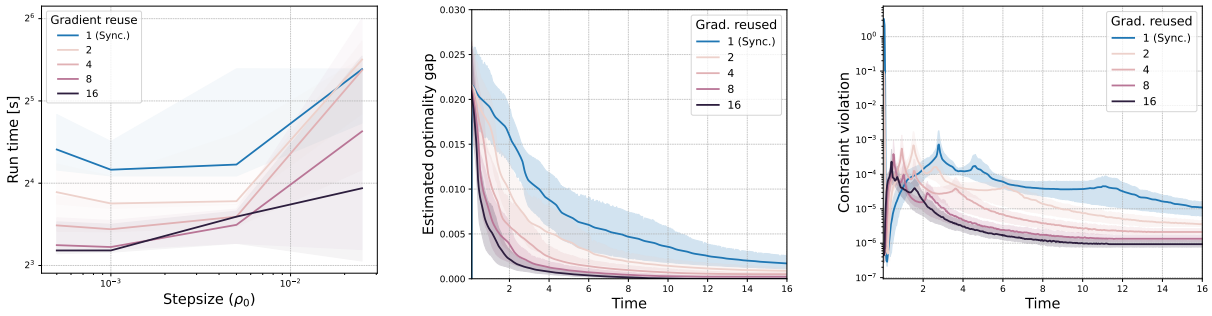

Figure 7: Ablation study evaluating the impact of reusing gradients when applying ATOS to the Gromov–Wasserstein problem. Experiments were conducted on simulated datasets of size 4000, repeated over 5 random seeds. The noise level used in this experiment was 2