POINT SOURCE LOCALISATION WITH UNBALANCED OPTIMAL TRANSPORT

Tuomo Valkonen*

Abstract Replacing the quadratic proximal penalty familiar from Hilbert spaces by an unbalanced optimal transport distance, we develop forward-backward type optimisation methods in spaces of Radon measures. We avoid the actual computation of the optimal transport distances through the use of transport three-plans and the rough concept of transport subdifferentials. The resulting algorithm has a step similar to the sliding heuristics previously introduced for conditional gradient methods, however, now non-heuristically derived from the geometry of the space. We demonstrate the improved numerical performance of the approach.

1 INTRODUCTION

We continue the quest started in [35] to understand the challenges Banach space geometries pose to the realisation of forward-backward type optimisation methods and the proximal step. We do this by focussing on the point source localisation problem [9, 27]

$$\min_{\mu\in\mathscr{M}(\Omega)} \frac{1}{2}\|A\mu - b\|_Y^2 + \alpha\|\mu\|_{\mathscr{M}} + \delta_{\geq 0}(\mu), \tag{1.1}$$

where $A \in \mathbb{L}(\mathscr{M}(\Omega); Y)$ is a bounded, linear forward-operator from the space of Radon measures $\mathscr{M}(\Omega)$ on $\Omega \subset \mathbb{R}^n$ to a Hilbert space Y of measurements b. The parameter $\alpha > 0$ controls the sparsity of the solution via the Radon norm regularisation term, while the final positivity constraint slightly simplifies the technical details by concentrating on sources only, avoiding sinks.

Our hunch when starting the work on [35] was that the weak-$*$ topology would be the natural topology for working with measures and, therefore, any forward-backward method should try to replace the Hilbert-space quadratic penalisation in the forward-backward method

$$x^{k+1} := \operatorname*{arg\,min}_x \; F(x^k) + \langle \nabla F(x^k), x - x^k\rangle + G(x) + \frac{1}{2\tau}\|x - x^k\|^2 \tag{1.2}$$

for the minimisation of $F + G$, by a metrisation of weak-$*$ convergence. Wasserstein distances provide such metrisations [29], however, we were in the work leading up to [35] unable to yet make them work. Instead, we introduced "particle-to-wave" operators $\mathscr{D} \in \mathbb{L}(\mathscr{M}(\Omega); C_0(\Omega))$ from Radon measures to the predual space of continuous functions that vanish at infinity, and defined (semi-)norms and (semi-)inner products with the help of these operators. That way, we could make optimisation in arbitrary normed spaces appear like optimisation in Hilbert spaces. Practically, we took $\mathscr{D}$ as a convolution operator, similarly to the construction of Maximum Mean Discrepancy (MMD) norms for optimal transport [33].

Now, in this work we finally introduce a forward-backward method based on (conventional, non-MMD) unbalanced optimal transport of measures [3, 15, 16, 24, 26, 30, 31, 33]. Practically, such an optimal transport based algorithm implements the sliding heuristics introduced in [18] for conditional gradient

*Research Center in Mathematical Modeling and Optimization (MODEMAT), Quito, Ecuador; Department of Mathematics and Statistics, University of Helsinki, Finland; *and* Department of Mathematics, Escuela Politécnica Nacional, Quito, Ecuador. tuomo.valkonen@iki.fi, ORCID: 0000-0001-6683-3572

methods [5, 7, 8, 18, 20, 28] in a less heuristic way, tied to the geometry of the space. Our primary variant of unbalanced optimal transport, which we introduce in Section 2, still includes $\mathscr{D}$ as the marginal cost. We will, however, also consider the Radon norm $\|\cdot\|_{\mathscr{M}}$ as a marginal cost, with weaker convergence guarantees. Indeed, we will see in Section 7 that numerically the $\mathscr{D}$-norm is more effective. However, it is much easier to prove that the differential of the data term F is Lipschitz with respect to the Radon norm. This is the term $\frac{1}{2}\|A\mu - b\|^2$ in (1.1).

We never actually compute the optimal transport plans $\gamma \in \mathscr{M}(\Omega^2)$ that realise the optimal transport distances. Instead, the forward-backward algorithm that we develop in Sections 4 and 5, is based on the "weaving" of suboptimal *three-plans* $\lambda \in \mathscr{M}(\Omega^3)$ between the current iterate μ^k, the next iterate μ^{k+1}, and a comparison point $\bar{\mu}$, typically a solution to (1.1). We also employ *transport subdifferential* ideas adapted from [1] and updated to the unbalanced setting. We briefly review these ideas and then develop relevant subdifferential-type and transport smoothness estimates in Section 3. Before illustrating numerical performance in Section 7, we also present an extension of our method to product spaces in Section 6. This allows treating problems with auxiliary variables as well as dual variables in a primal-dual splitting method. For the weaker subdifferential convergence results of Section 4, we do not assume the convexity of the data term F. For the $O(1/N)$ function value convergence results of Section 5 we make that assumption.

Besides the aforementioned conditional gradient methods and [35], other algorithms for (1.1) include [10, 11, 21]. In [13, 14] Wasserstein gradient flows are considered for probability measures modelling a discrete set of possible source locations. A gradient flow of probability measures based on a balanced transport distance very similar to our unbalanced transport distance, has been recently introduced in [23]. Although their theoretical time-discretised numerical method is a cyclic proximal point method, their practically employed numerical method, based on a fixed number of interacting particles, bears parallels to ours, but lacks its finer details, and is presented without convergence proof. Optimal transport regularised dynamic point source location problems—not algorithms based on optimal transport—are considered in [6]. As optimal transport plans can be related to geodesics with respect to a Wasserstein distance, our overall approach can also be related to optimisation methods on manifolds [4, 34] and general spaces [25]. Our work could also be combined with [19] for single-loop PDE-constrained and bilevel optimisation with measures.

NOTATION AND ELEMENTARY RESULTS

We denote the extended reals by $\overline{\mathbb{R}} := [-\infty, \infty]$, and the space of finite Radon measures on a locally compact Borel measurable set $\Omega \subset \mathbb{R}^n$ by $\mathscr{M}(\Omega)$. The non-negative radon measures we denote by $\mathscr{M}_+(\Omega)$. We write $\|\cdot\|_{\mathscr{M}(\Omega)}$ or, for short, $\|\cdot\|_{\mathscr{M}}$ for the Radon norm. The subset $\mathscr{P}(\Omega) \subset \mathscr{M}(\Omega)$ consists of probability measures, and the subspace $\mathscr{F}(\Omega) \subset \mathscr{M}(\Omega)$ of *discrete measures* $\mu = \sum_{k=1}^n \alpha_k \delta_{x_k}$ for any $n \in \mathbb{N}$, where the weights $\alpha_k \in \mathbb{R}$, and locations $x_k \in \Omega$. Here, δ_x is the Dirac measure with mass one at the point $x \in \Omega$. For a $\mu \in \mathscr{M}(\Omega)$, $\operatorname{sign}\mu := d\mu/d|\mu|$ denotes the Radon–Nikodym derivative of μ respect to its own total variation measure $|\mu|$. For a μ-integrable functions ρ, we write $[\rho * \mu](x) := \int \rho(x-y)d\mu(y)$ for the convolution.

With $F : X \to R$ Fréchet differentiable on a normed space X, we write $F'(x) \in X^*$ for the Fréchet derivative at $x \in X$. Here X^* is the dual space to X. We call F *pre-differentiable* if $F'(x) \in X_*$ for X_* a designated *predual space* of X, satisfying $X = (X_*)^*$. We have $(X_*)^{**} = X^*$, so X_* canonically injects into X^*. For a convex function $F : X \to \overline{\mathbb{R}}$, we write $\partial F : X \rightrightarrows X_*$ for the (set-valued) pre-subdifferential map, defined as $\partial F(x) = \{x_* \in X_* \mid F(\tilde{x}) - F(x) \geq \langle x_* | \tilde{x} - x\rangle \text{ for all } \tilde{x} \in X\}$. We write $\delta_C : X \to \overline{\mathbb{R}}$ for the $\{0, \infty\}$-valued indicator function of a set $C \subset X$.

We write $\langle x, x'\rangle$ for the inner product between two elements x and x' of a Hilbert space X, and $\langle x^* | x\rangle := \langle x^* | x\rangle_{X^*,X} := x^*(x)$ for the dual product between elements of a normed space X and its dual or predual X^*. The space $\mathbb{L}(X; Y)$ stands for bounded linear operators between two vector spaces X and

Y. The identity operator is $\mathrm{Id} \in \mathbb{L}(X;X)$. With Y a Hilbert space for simplicity, we call $A \in \mathbb{L}(X;Y)$ *pre-adjointable* if there exists a *pre-adjoint* $A_* : \mathbb{L}(X_*;Y)$ whose mixed Hilbert–Banach adjoint $(A_*)^* = A$. In other words $\langle x|A_*z\rangle = \langle Ax, z\rangle$ for all $z \in Y$ and $x \in X$.

A predual of $\mathscr{M}(\Omega)$ is the space $C_c(\Omega)$ of continuous functions with compact support. We write $C_0(\Omega)$ for continuous functions on Ω that vanish at infinity, which is also a predual space of $\mathscr{M}(\Omega)$ [22, Theorem 1.200]. We also set

$$C_0^1(\Omega) := \{f \in C_0(\Omega) \mid f \text{ is Fréchet differentiable with } f' \in C_0(\Omega;\mathbb{R}^n)\}$$

and, for a $\mu \in \mathscr{M}(\Omega)$,

$$C_0'(\Omega,\mu) := \{f \in C_0(\Omega) \mid f \text{ is Fréchet differentiable for } \mu\text{-a.e. } x \in \Omega\}. \tag{1.3}$$

The space $C_0^1(\Omega)$ is a dense subspace of $C_0(\Omega)$, as seen by taking convolutions with a differentiable bump function.

For a map $f : X \to Y$ and measure μ on X, we define the *push-forward* measure $f_\#\mu$ by

$$[f_\#\mu](A) := \mu(\{x \in X \mid f(x) \in A\}) \quad (A \subset Y \text{ measurable}).$$

We will typically take f as a projection $\pi^{i_1,\dots,i_k}(x_1,\dots,x_n) := (x_{i_1},\dots,x_{i_k})$, the diagonal constructor $\mathrm{diag}(x) := (x,x)$, or the reversal $\mathrm{rev}(x,y) = (y,x)$

2 AN APPROACH TO UNBALANCED TRANSPORT

We now introduce our approach to unbalanced optimal transport. We start by recalling basic definitions of balanced optimal transport in Section 2.1. We then define our variant of unbalanced optimal transport in Section 2.2, and prove the existence of minimising transport and the metrisation of weak-$*$ convergence in Sections 2.3 and 2.4.

2.1 BALANCED OPTIMAL TRANSPORT THEORY

Balanced optimal transport theory treats the cost of transporting the total mass of one probability measure to another [29]. Let $c : \Omega^2 \to [0,\infty)$ be some symmetric *transport cost* that models the cost of transporting a unit mass between two points of $\Omega \subset \mathbb{R}^n$. For example, $c = c_p$ for some $p \geq 1$ and

$$c_p(x,y) := \frac{1}{p}|x-y|_p^p.$$

Given $\gamma \in \mathscr{M}_+(\Omega^2)$, we then define the Monge–Kantorovich *transport cost* between $\mu, \nu \in \mathscr{M}_+(\Omega)$ as

$$T_c(\mu,\nu) := \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\Omega^2} c(x,y)\, d\gamma(x,y)$$

over the set of *transport plans* $\Gamma(\mu,\nu) := \{\gamma \in \mathscr{M}_+(\Omega^2) \mid \pi^0_\#\gamma = \mu,\ \pi^1_\#\gamma = \nu\}$. These plans indicate the direct paths from x to y, along which mass is transported, and the corresponding amount. They are restricted to transport all mass of μ exactly to ν. Indeed, $\Gamma(\mu,\nu) = \emptyset$ if $\|\mu\|_{\mathscr{M}} \neq \|\nu\|_{\mathscr{M}}$.

The p-Wasserstein distance is now defined through

$$W_p(\mu,\nu) := T_{c_p}(\mu,\nu)^{1/p}.$$

Restricted to probability measures, this distance metricises the topology of weak-$*$ convergence. For more details, we refer to [29]. If we want to use W_p to measure the distance between two iterates μ^k and μ^{k+1} in an optimisation algorithm, such as replacing the quadratic penalty in (1.2), we are limited to working with probability measures. Therefore, we need a theory of unbalanced optimal transport that does not impose this restriction of equal masses. Such theories have been introduced in, e.g., [3, 15, 16, 24, 26, 30, 31, 33].

2.2 BASIC DEFINITIONS

Let c be as above, and let $E : \mathcal{M}(\Omega)^2 \to [0,\infty)$ be some (possibly nonsymmetric) *marginal cost*. We then define for $\mu_0, \mu_1 \in \mathcal{M}(\Omega)$ the *unbalanced transport cost*[1]

$$T_{c,E}(\mu_0,\mu_1) := \inf_{\gamma \in \mathcal{M}(\Omega^2)} V_{c,E}(\mu_0,\mu_1;\gamma),$$

where

$$V_{c,E}(\mu_0,\mu_1;\gamma) := \int_{\Omega^2} c(x,y)\, d|\gamma|(x,y) + E(\mu_0 - \pi^0_\#\gamma, \mu_1 - \pi^1_\#\gamma).$$

For the marginal cost, we consider the squared Radon metric

$$E_{\mathcal{M}}(\nu_0,\nu_1) := \frac{1}{2}\|\nu_1 - \nu_0\|^2_{\mathcal{M}}$$

as well as Bregman divergences

$$E_J(\nu_0,\nu_1) = B^\omega_J(\nu_0,\nu_1) := \begin{cases} J(\nu_1) - J(\nu_0) - \langle \omega | \nu_1 - \nu_0 \rangle, & \omega \in \partial J(\nu_0), \\ \infty, & \omega \notin \partial J(\nu_0), \end{cases} \tag{2.1}$$

for some convex, proper, and weakly-$*$ lower semicontinuous $J : \mathcal{M}(\Omega) \to \overline{\mathbb{R}}$. If J is differentiable, we simply write $B_J(\nu_0,\nu_1) := B_J^{J'(\nu_0)}(\nu_0,\nu_1)$.

In particular, following [35], let $\mathscr{D} \in \mathbb{L}(\mathcal{M}(\Omega); C_0(\Omega))$ be self-adjoint and positive semi-definite, i.e.,

$$\langle \mathscr{D}x | y \rangle_{C_0(\Omega),\mathcal{M}(\Omega)} = \langle x | \mathscr{D}y \rangle_{C_0(\Omega),\mathcal{M}(\Omega)} \quad \text{and} \quad \langle \mathscr{D}x | x \rangle_{C_0(\Omega),\mathcal{M}(\Omega)} \geq 0 \quad \text{for all} \quad x, y \in \mathcal{M}(\Omega).$$

Then the semi-inner product and semi-norm

$$\langle x, x \rangle_{\mathscr{D}} := \langle \mathscr{D}x | x \rangle_{C_0(\Omega),\mathcal{M}(\Omega)} \quad \text{and} \quad \|x\|_{\mathscr{D}} := \sqrt{\langle x, x \rangle_{\mathscr{D}}}$$

are well-defined [35]. We may, for example, take $\mathscr{D}\mu = \rho * \mu$ for a symmetric and self-adjoint convolution kernel ρ. For $J_{\mathscr{D}} := \frac{1}{2}\|\cdot\|^2_{\mathscr{D}}$, we then have

$$E_{\mathscr{D}}(\nu_0,\nu_1) := B_{J_{\mathscr{D}}}(\nu_0,\nu_1) = \frac{1}{2}\|\nu_1 - \nu_0\|^2_{\mathscr{D}}.$$

For both $\star = \mathcal{M}, \mathscr{D}$, we have

$$V_{c,\star}(\mu_0,\mu_1;\gamma) = \int_{\Omega^2} c(x,y)\, d\gamma(x,y) + \frac{1}{2}\|\mu_1 - \mu_0 - (\pi^1_\# - \pi^0_\#)\gamma\|^2_\star.$$

These are invariant with respect to diagonal additions ($\mathrm{diag}_\# \mu$ for a $\mu \in \mathcal{M}(\Omega)$) to the transport γ.

2.3 EXISTENCE OF MINIMISING TRANSPORTS

We call E *anti-diagonally coercive* if $\|\mu^k - \nu^k\|_{\mathcal{M}} \to \infty$ implies $E(\mu^k,\nu^k) \to \infty$. The squared Radon metric $E_{\mathcal{M}}$ is obviously anti-diagonally coercive, but $E_{\mathscr{D}}$ requires further assumptions.

Lemma 2.1. *Suppose $0 \leq c$ and E is anti-diagonally coercive and weakly-$*$ lower semicontinuous in $\mathcal{M}(\Omega)^2$. Then there exists $\gamma \in \mathcal{M}(\Omega^2)$ such that $T_{c,E}(\mu_0,\mu_1) = V_{c,E}(\mu_0,\mu_1;\gamma)$.*

[1] We do not call this a "distance", since we will concentrate on $c(x,y) = c_2(x,y) = \frac{1}{2}|x-y|^2$ without taking the corresponding square root of the integral.

Proof. If $T_{c,E}(\mu_0,\mu_1)=\infty$, there is nothing to prove, as by definition $V_{c,E}(\mu_0,\mu_1;\gamma)=\infty$ for all $\gamma\in\mathscr{M}(\Omega^2)$. So suppose $T_{c,E}(\mu_0,\mu_1)<\infty$. Then there exists a minimising sequence $\{\gamma^k\}_{k\in\mathbb{N}}$ with

$$\int_{\Omega^2} c\,d|\gamma^k| + E(\mu_0-\pi^0_\#\gamma^k,\mu_1-\pi^1_\#\gamma^k) = V_{c,E}(\mu_0,\mu_1;\gamma^k) \leq T_{c,E}(\mu_0,\mu_1)+1/k \quad \text{for all} \quad k\in\mathbb{N}.$$

This implies the boundedness of $\{E(\mu_0-\pi^0_\#\gamma^k,\mu_1-\pi^1_\#\gamma^k)\}_{k\in\mathbb{N}}$. By the anti-diagonal coercivity of E, also $\{\|\mu_1-\mu_0-(\pi^1_\#-\pi^0_\#)\gamma^k\|_{\mathscr{M}}\}_{k\in\mathbb{N}}$ must then be bounded. Thus, we can assume $\{\gamma^k\}_{k\in\mathbb{N}}$ to be bounded; if not, we could replace the sequence by a bounded and still minimising sequence $\{\gamma^k+\operatorname{diag}_\#\theta^k\}_{k\in\mathbb{N}}$ for some $\theta^k\in\mathscr{M}(\Omega)$. By the Banach–Alaoglu theorem, we may thus extract a subsequence, unrelabelled, convergent weakly-$*$ in $\mathscr{M}(\Omega^2)$ to some γ, and such that $\{|\gamma^k|\}_{k\in\mathbb{N}}$ also converge weakly-$*$ to some $\bar\gamma\in\mathscr{M}_+(\Omega^2)$. Since, for any $\varphi\in C_0(\Omega)2$, we have $|\gamma^k|(\varphi)\geq\gamma^k(\varphi$, passing to the limit, also $\bar\gamma\geq|\gamma|$. Since $\pi^i_\#:\mathscr{M}(\Omega^2)\to\mathscr{M}(\Omega)$ is a bounded linear functional, $\{\pi^i_\#\gamma^k\}_{k\in\mathbb{N}}$ converges weakly-$*$ to $\pi^i_\#\gamma$ for $i=0,1$. Now, as $\gamma\mapsto\int c\,d\gamma$ is linear, $|\gamma^k|\overset{*}{\rightharpoonup}\bar\gamma\geq|\gamma|$, and E is assumed weakly-$*$ lower semicontinuous, it follows that $V_{c,E}(\mu_0,\mu_1;\,\cdot\,)$, is weakly-$*$ lower semicontinuous. Consequently, it reaches its minimum at γ. □

2.4 WEAK-$*$ CONVERGENCE

Lemma 2.2. *Suppose $0\leq c$, $0\leq E$, and that $\mu^k-\nu^k\overset{*}{\rightharpoonup}0$ implies $E(\nu^k,\mu^k)\to 0$. Then $\mu^k-\nu^k\overset{*}{\rightharpoonup}0$ implies $T_{c,E}(\nu^k,\mu^k)\to 0$.*

Proof. Let $\{(\mu^k,\nu^k)\}_{k\in\mathbb{N}}$ be such that $\mu^k-\nu^k\overset{*}{\rightharpoonup}0$. Then $E(\nu^k,\mu^k)\to 0$. Since $0\leq T_{c,E}(\nu^k,\mu^k)\leq V_{c,E}(\nu^k,\mu^k;0)=E(\nu^k,\mu^k)$, it follows, as claimed, that $T_{c,E}(\nu^k,\mu^k)\to 0$. □

Lemma 2.3. *Suppose $\varepsilon\|x-y\|^p\leq c(x,y)$ for some $\varepsilon>0$ and $p>1$ for all $x,y\in\Omega$; $0\leq E$; and that $E(\nu^k,\mu^k)\to 0$ with $\{(\nu^k,\mu^k)\}_{k\in\mathbb{N}}$ bounded implies $\mu^k-\nu^k\overset{*}{\rightharpoonup}0$. Then $T_{c,E}(\nu^k,\mu^k)\to 0$ with $\{(\nu^k,\mu^k)\}_{k\in\mathbb{N}}$ bounded implies $\mu^k-\nu^k\overset{*}{\rightharpoonup}0$.*

Proof. Let $\{(\mu^k,\nu^k)\}_{k\in\mathbb{N}}$ be bounded with $T_{c,E}(\nu^k,\mu^k)\to 0$. Then for some $\gamma^k\in\mathscr{M}(\Omega^2)$, we have

$$E(\nu^k-\pi^0_\#\gamma^k,\mu^k-\pi^1_\#\gamma^k)\to 0 \quad \text{and} \quad \int_{\Omega^2} c\,d|\gamma^k|\to 0. \tag{2.2}$$

Let $\varphi\in C_c^\infty(\Omega)$. Then φ is Lipschitz with some factor L_φ. Using Hölder's inequality, it follows that

$$\begin{aligned}
|\langle(\pi^1_\#-\pi^0_\#)\gamma^k|\varphi\rangle| &= \left|\int_{\Omega^2}\varphi(y)-\varphi(x)\,d\gamma^k(x,y)\right| \leq L_\varphi\int_{\Omega^2}\|x-y\|d|\gamma^k|(x,y)\\
&\leq L_\varphi\|\gamma^k\|^{1-1/p}_{\mathscr{M}(\Omega^2)}\left(\int_{\Omega^2}\|x-y\|^p d|\gamma^k|(x,y)\right)^{1/p}\\
&\leq L_\varphi\varepsilon^{-1/p}\|\gamma^k\|^{1-1/p}_{\mathscr{M}(\Omega^2)}\left(\int_{\Omega^2}c(x,y)d|\gamma^k|(x,y)\right)^{1/p}.
\end{aligned}$$

By the second part of (2.2), it follows that $\langle(\pi^1_\#-\pi^0_\#)\gamma^k|\varphi\rangle\to 0$. Since $C_c^\infty(\Omega)$ is dense in $C_0(\Omega)$, we obtain the weak-$*$ convergence of $(\pi^1_\#-\pi^0_\#)\gamma^k$ to zero. In particular, $\{(\pi^1_\#-\pi^0_\#)\gamma^k\}_{k\in\mathbb{N}}$ must be bounded. Our assumptions and the first part of (2.2) then imply $(\mu^k-\nu^k)-(\pi^1_\#-\pi^0_\#)\gamma^k\overset{*}{\rightharpoonup}0$. Since $(\pi^1_\#-\pi^0_\#)\gamma^k\overset{*}{\rightharpoonup}0$, necessarily $\mu^k-\nu^k\overset{*}{\rightharpoonup}0$. □

We now concentrate on the special case $c=c_2$ for our marginal energies of primary interest.

Corollary 2.4. *$T_{c_2,E_{\mathscr{M}}}(\nu^k,\mu^k)\to 0$ with $\{(\nu^k,\mu^k)\}_{k\in\mathbb{N}}$ bounded implies $\mu^k-\nu^k\overset{*}{\rightharpoonup}0$.*

Proof. The conditions of Lemma 2.3 hold with $p=2$ and $\varepsilon=1/2$. □

Corollary 2.5. *Let $0 \not\equiv \rho \in C_0(\mathbb{R}^n) \cap L^2(\mathbb{R}^n)$ be symmetric and positive semi-definite and presentable as the autoconvolution $\rho = \rho^{1/2} * \rho^{1/2}$ for some $\rho^{1/2} \in L^2(\mathbb{R}^n) \cap C_0(\mathbb{R}^n)$. On a bounded domain $\Omega \subset \mathbb{R}^n$, let $\mathscr{D} \in \mathbb{L}(\mathscr{M}(\Omega); C_0(\Omega))$ be defined by $\mathscr{D}\mu = \rho * \mu$ for $\mu \in \mathscr{M}(\Omega)$. Then*

(i) If $\rho^{1/2} \in C_c(\mathbb{R}^n)$, then $\mu^k \overset{*}{\rightharpoonup} \mu$ weakly-$*$ in $\mathscr{M}(\Omega)$ implies $T_{c_2,E_{\mathscr{D}}}(\mu^k, \mu) \to 0$.

(ii) If $\rho^{1/2}$ is locally Lipschitz, and $\operatorname{span}\{x \mapsto \rho^{1/2}(x-y) \mid y \in \Omega\}$ is dense in $C_0(\Omega)$, then $T_{c_2,E_{\mathscr{D}}}(\mu^k, \mu) \to 0$ with $\{\mu^k\}_{k\in\mathbb{N}} \subset \mathscr{M}(\Omega)$ bounded implies $\mu^k \overset{*}{\rightharpoonup} \mu$ weakly-$*$ in $\mathscr{M}(\Omega)$.

Proof. Let $E(\mu, \nu) := \frac{1}{2}\|\nu - \mu\|_{\mathscr{D}}$. Then [35, Theorem 2.4] verifies the conditions of Lemmas 2.2 and 2.3 regarding E. The rest follow from those lemmas. □

3 ESTIMATES AND EXPANSIONS

We now study subdifferential-type lower estimates that incorporate unbalanced transport, as well as transport-relative smoothness of functions. We first review existing "transport subdifferentials" in Section 3.1, then develop relevant estimates for the unbalanced transport costs $V_{c,E}$ in Section 3.2. Based on this, in Section 3.3 we suggest a definition of an unbalanced transport subdifferential. In the final Section 3.4, we discuss relevant concepts of smoothness, that we will be using.

3.1 BALANCED TRANSPORT SUBDIFFERENTIALS

In [1], a theory is presented for (extended) Fréchet subdifferentials of functions of probability measures with respect to three-transport plans. Specifically, with $\mu_1, \mu_2 \in \mathscr{P}(\Omega)$, denote the set of optimal transports from μ_1 to μ_2 by

$$\Gamma_o(\mu_1, \mu_2) := \left\{ \gamma \in \Gamma(\mu_1, \mu_2) \,\middle|\, T_{c_2}(\mu_1, \mu_2) = \int c_2(x, y)\, d\gamma(x, y) \right\}.$$

Then [1, Definition 10.3.1] defines the (extended) Fréchet subdifferential of $F : \mathscr{P}(\Omega) \to (-\infty, \infty]$ at μ_1 to be the set $\partial F(\mu_1)$ of transport plans $\gamma \in \mathscr{M}_+(\Omega^2)$ satisfying $\pi^0_\# \gamma = \mu_1$ and[2]

$$F(\mu_2) - F(\mu_1) \geq \inf \left\{ \int_{\Omega^3} \langle x, z - y \rangle d\lambda(x, y, z) + o(W_2(\mu_1, \mu_2)) \,\middle|\, \begin{array}{l} \lambda \in \mathscr{P}(\Omega^3),\ \pi^{1,0}_\# \lambda = \gamma, \\ \pi^{1,2}_\# \lambda \in \Gamma_o(\mu_1, \mu_2) \end{array} \right\}. \tag{3.1}$$

According to [1, Theorem 10.3.6], for geodesically convex F the o-term can be omitted.

On the other hand, with $\mu_0, \mu_1, \mu_2 \in \mathscr{P}(\Omega)$, let $\lambda \in \mathscr{P}(\Omega^3)$ be such that $\pi^{0,2}_\# \lambda \in \Gamma_0(\mu_0, \mu_2)$ and $\pi^{0,1}_\# \lambda \in \Gamma_0(\mu_0, \mu_1)$. Then, using Pythagoras' identity, we expand

$$\begin{aligned} W_2^2(\mu_0, \mu_2) - W_2^2(\mu_0, \mu_1) &= \frac{1}{2} \int_{\Omega^2} |z - x|^2 d\pi^{0,2}_\# \lambda(x, z) - \frac{1}{2} \int_{\Omega^2} |y - x|^2 d\pi^{0,1}_\# \lambda(x, y) \\ &= \frac{1}{2} \int_{\Omega^3} |z - x|^2 - |y - x|^2 d\lambda(x, y, z) \\ &= \int_{\Omega^3} \langle y - x, z - y \rangle + \frac{1}{2}|z - y|^2 d\lambda(x, y, z). \end{aligned}$$

Hence,

$$W_2^2(\mu_0, \mu_2) - W_2^2(\mu_0, \mu_1) \geq \int_{\Omega^3} \langle y - x, z - y \rangle d\lambda(x, y, z) + W_2^2(\mu_1, \mu_2).$$

[2] For simplicity we consider here only the exponent $p = 2$, and a bounded set Ω. The definition of Γ_o on [1, page 14] has a typing mistake. The element of $\mathscr{P}(\Omega^3)$ should be μ instead of γ. For consistency, our indexing also differs from [1].

In analogy with standard convex subdifferentials of c_2, this suggests a different definition of a "transport subdifferential": $g(y,x) = y - x$ should be a transport subdifferential of $W_2^2(\nu, \cdot)$. We will base our approach on this latter idea, extending it to the setting of unbalanced transport. However, to avoid introducing high computational costs, we will not work with the optimal squared distances W_2^2, and more generally, $T_{c,E}$, but with $V_{c,E}$.

3.2 TRANSPORT COSTS

With an eye on the Pythagoras' or three-point identity satisfied by Hilbert space norms, and its replacement, we now derive three-point identities and inequalities for the non-marginalised transport cost $V_{c_2,E}(\mu_0, \cdot; \cdot)$. We assume to be given an energy $E : \mathcal{M}(\Omega)^2 \to \mathbb{R}$ that satisfies the bound

$$E(\mu, \bar{\mu}) \geq \langle \omega | \bar{\mu} - \nu \rangle + E(\mu, \nu) + E(\nu, \bar{\mu}) \quad (\omega \in \partial E(\mu, \cdot)(\nu)) \tag{3.2}$$

and is convex in the second parameter. This is satisfied by $E = B_J$ a Bregman divergence, for simplicity with a smooth generator J; see Lemma A.1. In particular, $E_{\mathscr{D}}$ satisfies (3.2), while $E_{\mathscr{M}}$ does not.

We will apply the next theorem with $\mu_0 = \mu^k$ and $\mu_1 = \mu^{k+1}$ the iterates of our proposed algorithm, and μ_2 a comparison measure, e.g., an optimal solution. Observe that $\bar{V}^{i,j}_{c_2,E}(\mu_i, \mu_j; \lambda) \geq V_{c_2,E}(\mu_i, \mu_j; \pi^{i,j}_\# \lambda)$ with equality if sign λ only depends on the spatial variables with indices i and j, in particular, if $\lambda \geq 0$.

Theorem 3.1. *Let E be as above and $J : \mathcal{M}(\Omega) \to \overline{\mathbb{R}}$ be convex, proper, weak-$*$ lower semicontinuous. Pick $\mu_0, \mu_1, \mu_2 \in \mathcal{M}(\Omega)$ as well as $\lambda \in \mathcal{M}(\Omega^3)$. Let*

$$\begin{aligned}\bar{V}^{i,j}_{c_2,E}(\mu_i, \mu_j; \lambda) &:= V_{c_2,E}(\mu_i, \mu_j; \pi^{i,j}_\# \lambda) + \int_{\Omega^2} c_2 d(\pi^{i,j}_\# |\lambda| - |\pi^{i,j}_\# \lambda|). \\ &= \int_{\Omega^2} c_2(x,y)\, d\pi^{i,j}_\# |\lambda|(x,y) + E(\mu_i - \pi^i_\# \lambda, \mu_j - \pi^j_\# \lambda).\end{aligned}$$

Then for any $\omega \in \partial E(\mu_0 - \pi^0_\# \lambda_{01}, \cdot)(\mu_1 - \pi^1_\# \lambda_{01})$, we have

$$\begin{aligned}\bar{V}^{0,2}_{c_2,E}(\mu_0, \mu_2; \lambda) - \bar{V}^{0,1}_{c_2,E}(\mu_0, \mu_1; \lambda) \geq &\int_{\Omega^3} \langle y - x, z - y \rangle\, d|\lambda|(x,y,z) \\ &+ \langle \omega | \mu_2 - \mu_1 - (\pi^2_\# - \pi^1_\#)\lambda \rangle + \bar{V}^{1,2}_{c_2,E}(\mu_1, \mu_2; \lambda).\end{aligned} \tag{3.3}$$

Proof. We expand

$$\bar{V}^{0,2}_{c_2,E}(\mu_0, \mu_2; \lambda) - \bar{V}^{0,1}_{c_2,E}(\mu_0, \mu_1; \lambda) = D + I \tag{3.4}$$

for

$$\begin{aligned}I &:= \frac{1}{2}\int_{\Omega^2} |z-x|^2\, d\pi^{0,2}_\# |\lambda|(x,z) - \frac{1}{2}\int_{\Omega^2} |y-x|^2\, d\pi^{0,1}_\# |\lambda|(x,y) \\ &= \frac{1}{2}\int_{\Omega^3} [|z-x|^2 - |y-x|^2]\, d|\lambda|(x,y,z) \\ &= \int_{\Omega^3} \langle y - x, z - y \rangle + \frac{1}{2}|z-y|^2\, d|\lambda|(x,y,z) \\ &= \int_{\Omega^3} \langle y - x, z - y \rangle\, d|\lambda|(x,y,z) + \frac{1}{2}\int_{\Omega^2} |z-y|^2\, d\pi^{1,2}_\# |\lambda|(y,z)\end{aligned} \tag{3.5}$$

and

$$\begin{aligned}D &:= E(\mu_0 - \pi^0_\# \lambda, \mu_2 - \pi^2_\# \lambda) - E(\mu_0 - \pi^0_\# \lambda, \mu_1 - \pi^1_\# \lambda) \\ &\geq E(\mu_1 - \pi^1_\# \lambda, \mu_2 - \pi^2_\# \lambda) + \langle \omega | (\mu_2 - \pi^2_\# \lambda) - (\mu_1 - \pi^1_\# \lambda) \rangle.\end{aligned} \tag{3.6}$$

Combined, (3.4) to (3.6) yield the claim. □

Remark 3.2. In the previous theorem, we can replace c_2 by any Bregman divergence B_j for a convex and smooth $j : \Omega \to \mathbb{R}$, and $\langle y - x, z - y\rangle$ by $\langle D_2 B_j(x, y), z - y\rangle$.

The following variant only considers two-transport, and omits the comparison measure μ_2.

Theorem 3.3. *Let $E : \mathcal{M}(\Omega)^2 \to \mathbb{R}$ be convex in the second parameter and satisfy $E(\nu, \nu) = 0$ for all ν. Then for any $\mu_0, \mu_1 \in \mathcal{M}(\Omega)$; $\gamma \in \mathcal{M}(\Omega^2)$; and $\omega \in \partial E(\pi^0_\#\gamma - \mu_0, \cdot)(\pi^1_\#\gamma - \mu_1)$, we have*

$$0 \geq \int_{\Omega^3} -|y - x|^2 \, d|\gamma|(x, y) + \langle \omega | \mu_0 - \mu_1 - (\pi^1_\# - \pi^0_\#)\gamma\rangle + V_{2c_2,E}(\mu_0, \mu_1; \gamma) \\ + E(\mu_0 - \pi^0_\#\gamma, \cdot + \mu_0 - \pi^0_\#\gamma)^*(\omega).$$

Proof. By the Fenchel–Young equality

$$E(\mu_0 - \pi^0_\#\gamma, \mu_1 - \pi^1_\#\gamma) + E(\mu_0 - \pi^0_\#\gamma, \cdot)^*(\omega) = \langle \omega | \mu_1 - \pi^1_\#\gamma\rangle.$$

By the properties of Fenchel conjugates (e.g., [17, Lemma 5.7 (ii)]), this rearranges as

$$0 = E(\mu_0 - \pi^0_\#\gamma, \mu_1 - \pi^1_\#\gamma) + E(\mu_0 - \pi^0_\#\gamma, \cdot + \mu_0 - \pi^0_\#\gamma)^*(\omega) + \langle \omega | \mu_0 - \mu_1 - (\pi^0_\# - \pi^1_\#)\gamma\rangle.$$

This and the construction of $V_{2c_2,E}$ give the claim. □

Remark 3.4. If $E(\mu, \nu) = J(\mu - \nu)$, we have $E(\mu_0 - \pi^0_\#\gamma, \cdot + \mu_0 - \pi^0_\#\gamma)^*(\omega) = J^*(-\omega)$.

3.3 TRANSPORT SUBDIFFERENTIALS

Theorem 3.5. *Let $G : \mathcal{M}(\Omega) \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous. Then, for any $\mu_2, \mu_1 \in \mathcal{M}(\Omega)$ as well as $\lambda \in \mathcal{M}(\Omega^3)$, for all differentiable $w \in \partial G(\mu_1)$, we have*

$$G(\mu_2) - G(\mu_1) \geq \int_{\Omega^3} \langle \nabla w(y), z - y\rangle \, d\lambda(x, y, z) + \langle w | \mu_2 - \mu_1 - (\pi^2_\# - \pi^1_\#)\lambda\rangle + r(w, \lambda), \tag{3.7}$$

where $r(w, \lambda) := \int_{\Omega^3} B_w(y, z) \, d\lambda(x, y, z)$.

Proof. By the definition of the convex subdifferential, we have

$$G(\mu_2) - G(\mu_1) \geq \langle w | \mu_2 - \mu_1\rangle = \int_{\Omega^3} w(z) - w(y) \, d\lambda(x, y, z) + \langle w | \mu_2 - \mu_1 - (\pi^2_\# - \pi^1_\#)\lambda\rangle \\ = \int_{\Omega^3} \langle \nabla w(y), z - y\rangle \, d\lambda(x, y, z) + \langle w | \mu_2 - \mu_1 - (\pi^2_\# - \pi^1_\#)\lambda\rangle + r(w, \lambda).$$ □

Remark 3.6 (Intepretation: transport subdifferentials). Taking for simplicity $\lambda \geq 0$, we can interpret (3.3) as (g, ω) for $g(y, x) := y - x$ being a ($V_{c_2,E}$-strong) *unbalanced transport subdifferential* of $V_{c_2,E}(\mu_0, \cdot\,; \cdot)$. Likewise, if w were convex and $\lambda \geq 0$, so that $r(w, \lambda) \geq 0$, we could interpret (g, w) for $g(x, y) := \nabla w(x)$ as an unbalanced transport subdifferential of G. In both cases, g is the transport component of the subdifferential, and w or ω is the marginal component, only evaluated against $\mu_2 - \mu_1 - (\pi^2_\# - \pi^1_\#)\lambda$.

If general, we cannot expect w to be convex. Then, to define a weaker form of an unbalanced transport subdifferential, similarly to (3.1), it would be possible to assume $r(w, \lambda)$ to be small by considering proximal subdifferentials or restricting the set of three-plans λ such that $(\pi^2_\# - \pi^1_\#)\lambda \approx 0$. The algorithm that we present in Section 4 will require such restrictions, however, we refrain at this stage from proposing an explicit definition of a transport subdifferential.

3.4 TRANSPORT SMOOTHNESS

As we intend to derive forward-backward type methods for (1.1), we will need to make more precise the various flavours of smoothness required from the data term. For our main concept smoothness, we say that a convex and pre-differentiable $F : \mathscr{M}(\Omega) \to \mathbb{R}$ is (L, ℓ)-*smooth* with respect to E and c, if for all $\mu, \nu \in \mathscr{M}(\Omega)$ and $\gamma \in \mathscr{M}(\Omega^2)$, we have

$$B_F(\mu + (\pi^1_\# - \pi^0_\#)\gamma, \nu) \leq V_{\ell c, LE}(\mu, \nu; \gamma).$$

Example 3.7. Let $F(\mu) = \frac{1}{2}\|A\mu - b\|^2$ for some $A \in \mathbb{L}(\mathscr{M}(\Omega); Y)$ and a Hilbert space Y. Then

$$B_F(\mu + (\pi^1_\# - \pi^0_\#)\gamma, \nu) = \frac{1}{2}\|A(\nu - \mu - (\pi^1_\# - \pi^0_\#)\gamma)\|^2.$$

On the other hand, for $E(\mu, \nu) = \frac{1}{2}\|\nu - \mu\|^2_{\mathscr{D}}$ and a $\mathscr{D} \in \mathbb{L}(\mathscr{M}(\Omega); C_0(\Omega))$, we have

$$V_{\ell c_2, LE}(\mu, \nu; \gamma) = \ell \int_{\Omega^2} c_2(x, y)\, d|\gamma|(x, y) + \frac{L}{2}\|\nu - \mu - (\pi^1_\# - \pi^0_\#)\gamma\|^2_{\mathscr{D}}$$

Therefore, F is (L, ℓ)-smooth with respect to E and c_2 if

$$A_* A \leq L\mathscr{D} \quad \text{and} \quad \ell \geq 0.$$

We will also employ smoothness with respect to transport plans only, requiring F' to be *firmly transport Lipschitz*, meaning that for some $\Theta > 0$ and a semi-norm $\|\cdot\|_*$ on $\mathscr{M}(\Omega)$, we have

$$\langle F'(\mu + (\pi^1_\# - \pi^0_\#)\gamma) - F'(\mu)|\Delta\rangle \leq \Theta_F \sqrt{|\gamma|(c_2)\|\gamma\|_{\mathscr{M}}}\|\Delta\|_* \quad \text{with} \tag{3.8a}$$

$$\langle F'(\mu + (\pi^1_\# - \pi^0_\#)\gamma) - F'(\mu)|(\pi^1_\# - \pi^0_\#)\gamma\rangle \leq \Theta_F^2 |\gamma|(c_2)\|\gamma\|_{\mathscr{M}} \tag{3.8b}$$

for all $\gamma \in \mathscr{M}(\Omega^2)$ and $\Delta, \mu \in \mathscr{M}(\Omega)$. In the next lemma, which proves this result for specific cases of $F(\mu) = \frac{1}{2}\|A\mu - b\|^2$, the operator A can model, e.g., a sensor grid of m sensors whose spatial sensitivities are modelled by the functions ψ_i.

Lemma 3.8. *Let $F(\mu) = \frac{1}{2}\|A\mu - b\|^2_{\mathbb{R}^m}$, where $A\mu = (\mu(\psi_1), \dots, \mu(\psi_m))$ for some L_i-Lipschitz ψ_i, $(i = 1, \dots, m)$. Then F' is firmly transport Lipschitz with $\Theta_F^2 = 2\sum_{i=1}^m L_i^2$ and $\|\Delta\|_* = \|A\Delta\|_2$.*

If $L_i = L$ are equal, then we can take $\Theta = 4N_\psi L^2$, where the maximum number of overlapping supports

$$N_\psi := \max\{\#P \mid y \in \Omega,\ P \subset \{1, \dots, m\},\ \psi_k(y) \neq 0 \text{ for all } k \in P\}.$$

Proof. Let $\gamma \in \mathscr{M}(\Omega^2)$ and $\Delta \in \mathscr{M}(\Omega)$. By rearrangements, the Lipschitz assumption, and Jensen's inequality, we get

$$\begin{aligned}
\langle F'(\mu + (\pi^1_\# - \pi^0_\#)\gamma) - F'(\mu)|\Delta\rangle &= \langle A(\pi^1_\# - \pi^0_\#)\gamma|A\Delta\rangle = \int_{\Omega^2} \sum_{i=1}^m (\psi_i(y) - \psi_i(x))[A\Delta]_i d\gamma(x, y) \\
&\leq \int_{\Omega^2} \sum_{i=1}^m L_i |[A\Delta]_i| |x - y| d|\gamma|(x, y) \leq \sqrt{\sum_{i=1}^m L_i^2} \cdot \int_{\Omega^2} \|A\Delta\|_2 \cdot |x - y| d|\gamma|(x, y) \\
&\leq \sqrt{\sum_{i=1}^m L_i^2} \cdot \|A\Delta\|_2 \sqrt{\|\gamma\|_{\mathscr{M}} \int_{\Omega^2} |x - y|^2 d|\gamma|(x, y)} = \sqrt{2\sum_{i=1}^m L_i^2} \cdot \|A\Delta\|_2 \sqrt{|\gamma|(c_2)\|\gamma\|_{\mathscr{M}}}.
\end{aligned}$$

This proves (3.8a). The part (3.8b), follows by observing that with $\Delta = (\pi^1_\# - \pi^0_\#)$ we have $\langle A(\pi^1_\# - \pi^0_\#)\gamma|A\Delta\rangle = \|A\Delta\|_2^2$, dividing the above result by $\|A\Delta\|_2$, and squaring.

The more refined constant is achieved by observing that at most $2N_\psi$ components of the sum inside the first integral above are nonzero for each (x, y). □

Similarly, we can form transport estimates for convolution operators $\mathscr{D}$:

Lemma 3.9. *Let $\mathscr{D} = \rho *$ for some $\rho : \mathbb{R}^n \to \mathbb{R}$ with L_ρ-Lipschitz gradient. Then, for all $\gamma \in \mathscr{M}(\Omega^2)$,*

$$\|(\pi^1_\# - \pi^0_\#)\gamma\|^2_{\mathscr{D}} \leq 2L_\rho |\gamma|(\Omega^2) \int_{\Omega^2} c_2(x_1, x_2)\, d|\gamma|(x_1, x_2).$$

Proof. Expanding and using the mean value theorem, we get

$$\begin{aligned}
\|(\pi^1_\# - \pi^0_\#)\gamma\|^2_{\mathscr{D}} &= \int_\Omega \int_\Omega \rho(y - x)\, d[(\pi^1_\# - \pi^0_\#)\gamma](y)\, d[(\pi^1_\# - \pi^0_\#)\gamma](x) \\
&= \int_\Omega \int_{\Omega^2} \rho(x - y_1) - \rho(x - y_2)\, d\gamma(y_1, y_2)\, d[(\pi^1_\# - \pi^0_\#)\gamma](x) \\
&= \int_{\Omega^2} \int_{\Omega^2} \rho(x_1 - y_1) - \rho(x_2 - y_1) + \rho(x_2 - y_2) - \rho(x_1 - y_2)\, d\gamma(y_1, y_2)\, d\gamma(x_1, x_2) \\
&= \int_{\Omega^2} \int_{\Omega^2} \int_0^1 \langle \rho'(tx_1 + (1-t)x_2 - y_1) - \rho'(tx_1 + (1-t)x_2 - y_2), x_1 - x_2 \rangle \\
&\qquad\qquad dt\, d\gamma(y_1, y_2)\, d\gamma(x_1, x_2).
\end{aligned}$$

Thus, the assumed Lipschitz property of ρ yields, as claimed

$$\begin{aligned}
\|(\pi^1_\# - \pi^0_\#)\gamma\|^2_{\mathscr{D}} &\leq \int_{\Omega^2} \int_{\Omega^2} L_\rho \|y_1 - y_2\| \|x_1 - x_2\|\, d|\gamma|(y_1, y_2)\, d|\gamma|(x_1, x_2) \\
&\leq L_\rho \int_{\Omega^2} \int_{\Omega^2} c_2(x_1, x_2) + c_2(y_1, y_2)\, d|\gamma|(y_1, y_2)\, d|\gamma|(x_1, x_2) \\
&= 2L_\rho |\gamma|(\Omega^2) \int_{\Omega^2} c_2(x_1, x_2)\, d|\gamma|(x_1, x_2).
\end{aligned}$$

□

4 SLIDING FORWARD–BACKWARD SPLITTING

With a particular interest in the instance (1.1), we now develop a forward-backward splitting method based on unbalanced optimal transport for the general problem

$$\min_{\mu \in \mathscr{M}(\Omega)} F(\mu) + G(\mu), \tag{4.1}$$

where F satisfies several smoothness properties, while G is a general convex, and possibly nonsmooth function. In this section we do not assume the convexity of F. We require that the differentials $F'(\mu) \in C^1_0(\Omega)$, and, through $C'_0(\Omega, \mu)$ defined in (1.3), that other predual objects are differentiable at relevant points. The assumption on $F'(\mu)$ could similarly be relaxed, and the differentiability could be relaxed to a bounded linear approximation property (encoded by bounds on integrals of Bregman divergences $B_{F'(\mu)}$, etc.).

In the next Section 4.1, we introduce the abstract form of our algorithm, as well as our main assumptions. These are largely shared with the next Section 5.3, where we study the convergence rates of function values under additional restrictions. Here, in Section 4.2, we prove the convergence of subdifferentials.

4.1 GENERIC ALGORITHM AND MAIN ASSUMPTIONS

On each step of our algorithm, we will, roughly speaking, take a forward step with respect to F, and a proximal step with respect to G, in the $V_{c_2,E}$ "metric". The step is determined by the marginal and transport components of to the transport subdifferentials of Section 3.3, although we do not make the

connection explicit here, and, indeed, need to add some technical complications, as the forward step is taken not at the current iterate μ^k, but at the iterate transported by some $\gamma^{k+1} \in \mathcal{M}(\Omega^2)$ as

$$\check{\mu}^k := \mu^k + (\pi^1_\# - \pi^0_\#)\gamma^{k+1}.$$

The transportable mass will be controlled through the curvature of F at $\mu \in \mathcal{M}(\Omega)$ along $\gamma \in \mathcal{M}(\Omega^2)$, defined as

$$\mathcal{K}_F(\mu, \gamma) := \gamma(B_{F'(\mu)}) = \int_{\Omega^2} B_{F'(\mu)}(x, y) d\gamma(x, y).$$

We also write

$$v^k := F'(\mu^k), \quad \text{and} \quad \check{v}^k := F'(\check{\mu}^k).$$

Given an energy $E : \mathcal{M}(\Omega) \times \mathcal{M}(\Omega) \to \mathbb{R}$ that satisfies (3.2), we seek

$$\text{(4.2a)} \qquad \begin{cases} \mu^{k+1} \in \mathcal{M}(\Omega), \quad \gamma^{k+1} \in \mathcal{M}(\Omega^2), \quad w^{k+1} \in \partial G(\mu^{k+1}) \cap C'_0(\Omega, \pi^1_\# \gamma^{k+1}), \quad \text{and} \\ \omega^{k+1} \in \partial E(\mu^k - \pi^0_\# \gamma^{k+1}, \cdot)(\mu^{k+1} - \pi^1_\# \gamma^{k+1}) \cap C'_0(\Omega, \pi^1_\# \gamma^{k+1}) \end{cases}$$

that satisfy for some $\check{C}, \ell_F, \ell_r \geq 0$, step lengths $\theta, \tau > 0$, and tolerances $\varepsilon^{k+1} > 0$ that

$$\text{(4.2b)} \qquad y = x - \theta\tau \operatorname{sign} \gamma^{k+1}(x, y)[\nabla v^k(x) + \nabla w^{k+1}(y)] \quad \text{for} \quad \gamma^{k+1}\text{-a.e. } (x, y),$$

$$\text{(4.2c)} \qquad -\varepsilon^{k+1} \leq \tilde{\varepsilon}^{k+1} \leq \varepsilon^{k+1} \quad \text{for} \quad \tilde{\varepsilon}^{k+1} := \tau[\check{v}^k + w^{k+1}] + \omega^{k+1},$$

$$\text{(4.2d)} \qquad 2\ell_F |\gamma^{k+1}|(c_2) \geq \mathcal{K}_F(\mu^k, \gamma^{k+1}) + \mathcal{K}_F(\mu^{k+1}, \operatorname{rev}_\# \gamma^{k+1}) - B_F(\check{\mu}^k, \mu^k), \quad \text{and}$$

$$\text{(4.2e)} \qquad \begin{aligned} \check{C}\varepsilon^{k+1} \geq \check{r}^{k+1} := &-\langle \tilde{\varepsilon}^{k+1} | \check{\mu}^k - \mu^{k+1} \rangle - \tau \operatorname{rev}_\# \gamma^{k+1}(B_{v^{k+1}+w^{k+1}}) \\ &+ \tau[\langle v^k - \check{v}^k | (\pi^0_\# - \pi^1_\#)\gamma^{k+1} \rangle - \ell_r |\gamma^{k+1}|(c_2)]. \end{aligned}$$

We will make this abstract, implicit algorithm more explicit as we progress.

Similarly to [35], for an exact forward-backward step with respect to the energy E, starting from $\check{\mu}^k$, we would have $\tilde{\varepsilon}^{k+1} = 0$. Indeed,

$$\text{(4.3)} \qquad \tilde{\varepsilon}^{k+1} \in F'(\check{\mu}^k) + \partial G(\mu^{k+1}) + \partial E(\mu^k - \pi^0_\# \gamma^{k+1}, \cdot)(\mu^{k+1} - \pi^1_\# \gamma^{k+1}).$$

Both (4.2c) and the first term of $\check{r}^{k+1}$ control the inexactness of this *marginal step*.

The second *convexity error control* term of $\check{r}^{k+1}$ is non-positive if $v^{k+1} + w^{k+1} \in \partial[F + G](\mu^{k+1})$ is convex—or if the target points y improve its value over the source points x, direction of improvement depending on sign $\lambda(x, y, z)$, Recalling that both conditional gradient methods and the forward-backward method of [35] insert new points at the minima or maxima of v^{k+1}, it seems reasonable that this would be the case, at least locally, if γ^{k+1} correctly identifies the spikes of μ^{k+1}. We will, however, never explicitly construct w^{k+1}, so, in Section 5.3, we will split this term into two more easily controllable parts. The last term of $\check{r}^{k+1}$ is a *curvature error control*. As we will show in Lemma 5.11, this term can be made non-positive if F' is firmly transport Lipschitz.

Part (4.2b) is a *transport step*. For $s = \operatorname{sign} \gamma^{k+1}(x, y)$, it realises the spatial forward-backward update $y = \operatorname{prox}_{s\theta\tau w^{k+1}}(x - s\theta\tau v^k(x))$. Since we cannot compute it without knowing w^{k+1}, and we cannot impose $\tilde{\varepsilon}^{k+1} \approx 0$ without knowing γ^{k+1}, in practise we will make the ansatz $\nabla w^{k+1}(y) = 0$ on $\operatorname{supp} \gamma^{k+1}$ to compute γ^{k+1}. Then, if, after solving for w^{k+1} by imposing (4.2c), this does not hold, we will remove (x, y) from $\operatorname{supp} \gamma^{k+1}$, and repeat. This same procedure can be used to bound the convexity error control term.

The part (4.2d) is a version of the standard step length condition $\tau L \leq 1$ for the transport component of the step: the Lipschitz-like factor ℓ_F will be used to control θ through $\tau\theta\ell_F \leq 1$. However, since typically $\|v^k\|_\infty = O(\|\mu^k\|)$, we also allow reducing transport to satisfy this condition. Alternatively, it would be possible to include a corresponding term in $\check{r}^{k+1}$.

Algorithm 1 Point insertion and weight adjustment for $\mathscr{D}$-marginal term [35]

Require: $\check{\mu} \in \mathscr{Z}(\Omega)$, $\check{v} \in C_0(\Omega)$, $\alpha, \tau, \varepsilon > 0$ on a domain $\Omega \subset \mathbb{R}^n$. A self-adjoint and positive semi-definite $\mathscr{D} \in \mathbb{L}(\mathscr{M}(\Omega); C_0^1(\Omega))$.

function INSERT_AND_ADJUST($\check{\mu}$, $\check{v}$, α, τ, ε, $\mathscr{D}$)
  Decompose $\sum_{x\in S} \alpha_x \delta_x := \check{\mu}$.
  Initialise $\mu := \check{\mu}$.
  repeat
    Form $\vec{\eta} := ([\tau\check{v} - \mathscr{D}\mu](x))_{x\in S} \in \mathbb{R}^{\#S}$ and $D := ([\mathscr{D}\delta_y](x))_{x,y\in S} \in \mathbb{R}^{\#S\times\#S}$.
    Find $\vec{\beta} = (\beta_x)_{x\in S} \in \mathbb{R}^{\#S}$ solving $\min f$ to the accuracy $\inf_{g\in\partial f(\vec{\beta})} \|g\|_\infty \leq \kappa\varepsilon/(1 + \|\vec{\beta}\|_1)$ for

$$f(\vec{\beta}) := \frac{1}{2}\langle \vec{\beta}, D\vec{\beta}\rangle + \langle \vec{\eta}, \vec{\beta}\rangle + \tau\alpha\|\vec{\beta}\|_1 + \delta_{\geq 0}(\vec{\beta}).$$

    Let $\mu := \sum_{x\in S} \beta_x \delta_x$
    Find $\bar{x}$ (approximately) minimising $\tau\check{v} + \mathscr{D}(\mu - \check{\mu})$. ▷ *For example, branch-and-bound.*
    Let $S := S \cup \{\bar{x}\}$ ▷ *$\bar{x}$ will only be inserted into μ if the next bounds check fails.*
  until $\tau\check{v}(\bar{x}) + \theta + [\mathscr{D}(\mu - \check{\mu})](\bar{x}) \geq -\varepsilon$
  return μ
end function

Example 4.1 (Rough algorithm for the basic point source localisation problem). When $G(\mu) = \alpha\|\mu\|_{\mathscr{M}} + \delta_{\geq 0}(\mu)$ as in (1.1), and we take $E = E_{\mathscr{D}}$ for a self-adjoint and positive semi-definite $\mathscr{D} \in \mathbb{L}(\mathscr{M}(\Omega); C_0^1(\Omega))$, we can use Algorithm 1 from [35] for the marginal step (4.2c). It inserts new spikes into $\check{\mu}^k$ to form μ^{k+1}, and, as a finite-dimensional convex subproblem, optimises the weights of the spikes, until the condition is satisfied. Thus, given tolerances ε^{k+1}, each step of the overall algorithm would proceed as follows:

1. Form an initial γ^{k+1} based on $\pi^0_\#\gamma^{k+1} = \mu^k$ and (4.2b) holding with the ansatz $\nabla w^{k+1}(y) = 0$.
2. Choose a fractional error control parameter $\kappa \in (0, 1)$, and form

$$\mu^{k+1} := \text{INSERT\_AND\_ADJUST}(\check{\mu}^k, \check{v}^k, \alpha, \tau, \kappa\varepsilon^{k+1}, \mathscr{D}).$$

3. If it turns out that there exist $(x, y) \in \pi^1_\#\gamma^{k+1} \setminus \operatorname{supp}\mu^{k+1}$, we may not have $\nabla w^{k+1}(y) = 0$, so we remove the violating pair from γ^{k+1} and try the previous step again. Likewise, if (4.2e) or (4.2d) does not hold, we reduce the transport and try again.

We will make this algorithm more precise in Section 5.4. The insertion Algorithm 1, including modifications to satisfy (4.2c) for $G = \alpha\|\cdot\|_{\mathscr{M}}$ without the positivity constraint, are further discussed in [35].

Example 4.2 (Radon marginal term). When $E = E_{\mathscr{M}}$, we have

$$\begin{aligned}\omega^{k+1} &\in \partial E(\mu^k - \pi^0_\#\gamma^{k+1}, \cdot)(\mu^{k+1} - \pi^1_\#\gamma^{k+1}) \cap C_0'(\Omega, \pi^1_\#\gamma^{k+1}) \\ &= \{\|\mu^{k+1} - \check{\mu}^k\|_{\mathscr{M}}\varphi \mid \varphi \in C_0'(\Omega, \pi^1_\#\gamma^{k+1}),\ \|\varphi\|_\infty \leq 1,\ \langle \mu^{k+1} - \check{\mu}^k | \varphi\rangle = \|\mu^{k+1} - \check{\mu}^k\|_{\mathscr{M}}\}.\end{aligned}$$

Through the multiplier $\|\mu^{k+1} - \check{\mu}^k\|_{\mathscr{M}}$ and φ being merely bounded outside the support of $\check{\mu}^k$, inserting just one new spike to $\check{\mu}^k$ to form μ^{k+1}, gives enough degrees of freedom to satisfy (4.2c). Thus, we may modify Algorithm 1 to Algorithm 2 and update

$$\mu^{k+1} := \text{INSERT\_AND\_ADJUST\_RADON}(\check{\mu}^k, \check{v}^k, \alpha, \tau, \kappa\varepsilon^{k+1}).$$

We collect our core assumptions, shared with Section 5, in the following.

Algorithm 2 Point insertion and weight adjustment for Radon marginal term

Require: $\check{\mu} \in \mathcal{Z}(\Omega)$, $\check{v} \in C_0(\Omega)$, $\tau, \alpha, \varepsilon > 0$ on a domain $\Omega \subset \mathbb{R}^n$.

1: **function** INSERT_AND_ADJUST_RADON($\check{\mu}$, $\check{v}$, α, τ, ε)
2: Find $\bar{x}$ (approximately) minimising $\check{v}$. ▷ *For example, branch-and-bound.*
3: Decompose $\sum_{x\in S} \alpha_x \delta_x := \check{\mu} + 0\delta_{\bar{x}}$.
4: From $\vec{\eta} := (\tau\check{v}(x))_{x\in S} \in \mathbb{R}^{\#S}$.
5: Find $\vec{\beta} = (\beta_x)_{x\in S} \in \mathbb{R}^{\#S}$ solving $\min f$ to the accuracy $\inf_{g\in\partial f(\vec{\beta})} \|g\|_\infty \leq \kappa\varepsilon/(1 + \|\vec{\beta}\|_1)$ for

$$f(\vec{\beta}) := \frac{1}{2}\|\vec{\beta} - \vec{\alpha}\|_1^2 + \langle\vec{\eta}, \vec{\beta}\rangle + \tau\alpha\|\vec{\beta}\|_1 + \delta_{\geq 0}(\vec{\beta}).$$

6: **return** $\mu := \sum_{x\in S} \beta_x \delta_x$
7: **end function**

Assumption 4.3. We have:

(i) $E : \mathcal{M}(\Omega)\times\mathcal{M}(\Omega) \to [0,\infty]$ is convex in the second parameter and $E(v, v) = 0$ for all $v \in \mathcal{M}(\Omega)$.
(ii) $G : \mathcal{M}(\Omega) \to \overline{\mathbb{R}}$ is convex, proper, and lower semicontinuous.
(iii) $F : \mathcal{M}(\Omega) \to \mathbb{R}$ is Fréchet differentiable with $F'(\mu) \in C_0^1(\Omega)$ for all μ, and (L, ℓ)-smooth with respect to B_J and c_2, i.e., for all $\mu, v \in \mathcal{M}(\Omega)$ and $\gamma \in \mathcal{M}(\Omega^2)$ we have

$$B_F(\mu + (\pi_\#^1 - \pi_\#^0)\gamma, v) \leq V_{\ell c_2, LE}(\mu, v; \gamma).$$

(iv) The tolerances $\{\varepsilon^{k+1}\}_{k\in\mathbb{N}} \subset [0,\infty)$ satisfy $\lim_{N\to\infty} \frac{1}{N}\sum_{k=0}^{N-1} \varepsilon^{k+1} = 0$.

An advantage of $E_{\mathcal{M}}$ over $E_{\mathcal{D}}$ is that with the former (iii) merely requires F' to be L-Lipschitz with respect to the Radon norm. The latter depends on a more difficult condition:

Example 4.4. Let $F(\mu) = \frac{1}{2}\|A\mu - b\|_Y^2$ with $A \in \mathbb{L}(\mathcal{M}(\Omega); Y)$ for a Hilbert space Y, and $E = E_{\mathcal{D}}$. Then the (L, ℓ)-smoothness of Assumption 4.3 (iii) is treated in Example 3.7 and holds when $A_*A \leq L\mathcal{D}$ and $\ell \geq 0$.

4.2 MONOTONICITY AND SUBDIFFERENTIAL CONVERGENCE

We now prove a very weak but easily obtained form of convergence for the abstract algorithm.

We will in this first phase measure convergence of iterates through

$$\check{V}_k(\mu, v; \gamma) := V_{(2\theta^{-1}-\tau[\ell+\ell_r])c_2,(1-\tau L)E}(\mu, v; \gamma) - \tau[\mathscr{K}_F(\mu, \gamma) + \mathscr{K}_F(v, \mathrm{rev}_\#\gamma) - B_F(\mu + (\pi_\#^1 - \pi_\#^0)\gamma, \mu)]. \tag{4.4}$$

To measure the convergence of subdifferentials, we will require a Fenchel pre-conjugate $\mathscr{E}_k^* : C_0(\Omega) \to \overline{\mathbb{R}}$ of

$$\mathscr{E}_k(\mu) := E(\mu^k - \pi_\#^0\gamma^{k+1}, \mu + \mu^k - \pi_\#^0\gamma^{k+1}).$$

To ensure the usefulness of these distances, we require the following additions to Assumption 4.3.

Assumption 4.5. We have:

(i) The step lengths $\tau, \theta > 0$, the parameters ℓ and L from Assumption 4.3, and ℓ_F and ℓ_r from (4.2), satisfy $\tau L < 1$ and $\theta\tau[\ell + \ell_r + 2\ell_F] < 2$.
(ii) F' is continuous with respect to E in the sense that, for any sequence $\{(\mu^k, \gamma^{k+1})\}_{k\in\mathbb{N}} \subset \mathcal{M}(\Omega)\times\mathcal{M}(\Omega^2)$, the convergence $E(\mu^k - \pi_\#^0\gamma^{k+1}, \mu^{k+1} - \pi_\#^1\gamma^{k+1}) \to 0$ implies $F'(\mu^k + (\pi_\#^1 - \pi_\#^0)\gamma^{k+1}) - F'(\mu^{k+1}) \to 0$.
(iii) For any sequence $\{\omega^k\}_{k\in\mathbb{N}} \subset C_0(\Omega)$, the convergence $\mathscr{E}_k^*(\omega^k) \to 0$ implies $\|\omega^k\|_{C_0(\Omega)} \to 0$.
(iv) $\inf[F + G] > -\infty$.

The following bound is immediate:

Lemma 4.6. *Suppose Assumption 4.3 (i) and 4.5 (i) as well as (4.2d) hold. Then*

$$\check{V}_k(\mu^k, \mu^{k+1}; \gamma^{k+1}) \geq \varepsilon V_{c_2,E}(\mu^k, \mu^{k+1}; \gamma^{k+1}) \geq \varepsilon T_{c_2,E}(\mu^k, \mu^{k+1}; \gamma^{k+1}) \geq 0$$

for $\varepsilon := \{1 - \tau L, 2\theta^{-1} - \tau[\ell + \ell_r + 2\ell_F]\} > 0$.

Example 4.7. For $E(\nu, \mu) = \frac{1}{2}\|\cdot\|^2_{\mathscr{D}}$, where $\mathscr{D} \in \mathbb{L}(\mathscr{M}(\Omega); C_0(\Omega))$ is self-adjoint and positive semi-definite, we have $\mathscr{E}_k(\mu) = \frac{1}{2}\|\mu\|^2_{\mathscr{D}}$. If $F(\mu) = \frac{1}{2}\|A\mu - b\|^2_Y$ in a Hilbert space Y, then Assumption 4.5 (ii) holds under the assumption $A_*A \leq L\mathscr{D}$ already seen in Example 4.4. Let then $c := \|\mathscr{D}\|_{\mathbb{L}(\mathscr{M}(\Omega);C_0(\Omega))}$. By the properties of conjugates (e.g., [17, Lemmas 5.4 and 5.7]), we have

$$\frac{2}{c}\|\omega\|^2_{C_0(\Omega)} = \frac{c}{2}\|\omega/c\|^2_{C_0(\Omega)} = \left(\frac{c}{2}\|\cdot\|^2_{\mathscr{M}(\Omega)}\right)^*(\omega) \leq \left(\frac{1}{2}\|\cdot\|^2_{\mathscr{D}}\right)^*(\omega) = \mathscr{E}_k^*(\omega).$$

Thus, also Assumption 4.5 (iii) holds.

Example 4.8. For $E(\nu, \mu) = \frac{1}{2}\|\cdot\|^2_{\mathscr{M}}$, we have $\mathscr{E}_k(\mu) = \frac{1}{2}\|\mu\|^2_{\mathscr{M}}$ and $\mathscr{E}_k^*(\mu) = \frac{1}{2}\|\mu\|^2_{C_0(\Omega)}$; see [17, Lemma 5.4]. Assumption 4.5 (ii) and (iii) clearly hold if F' is, e.g., Lipschitz in the Radon norm.

The following technical lemma will also be useful in Section 5. It, therefore, uses three-plans instead of just the two-plan γ^{k+1}. We denote the set of compatible three-plans by

$$\Lambda_{01}(\gamma) := \{\lambda_0 + p_\#\tilde{\gamma} \mid \lambda_0 \in \mathscr{M}(\Omega^3),\ \tilde{\gamma} \in \mathscr{M}(\Omega^2),\ \pi^{0,1}_\#\lambda_0 = \gamma \text{ with } p(x,z) = (x,z,z)\}. \tag{4.5}$$

Roughly speaking, the three-plan can either transport mass from x to y as γ^{k+1}, or return the mass to z.

Lemma 4.9. *Suppose Assumption 4.3 (ii) and (iii) as well as (4.2a) and (4.2b) hold. Then, for any* $\mu \in \mathscr{M}(\Omega^2)$ *and* $\lambda \in \Lambda_{01}(\gamma^{k+1})$*, we have*

$$\begin{aligned}\frac{1}{\theta}\int_{\Omega^3}\langle y - x, z - y\rangle\, d|\lambda|(x,y,z) + \langle \omega^{k+1}|\mu - \mu^{k+1} - (\pi^2_\# - \pi^1_\#)\lambda\rangle \\ = \tau[\mathscr{K}_F(\mu^k, \pi^{0,2}_\#\lambda) - \mathscr{K}_F(\mu^k, \pi^{0,1}_\#\lambda) - \mathscr{K}_F(\mu^{k+1}, \pi^{1,2}_\#\lambda)] \\ - \tau\langle \check{v}^k + w^{k+1}|\mu - \mu^{k+1}\rangle - \tau\ell_r\pi^{0,1}_\#|\lambda|(c_2) - r^{k+1}(\mu,\lambda),\end{aligned} \tag{4.6}$$

where, for $\tilde{\varepsilon}^{k+1}$ *defined in (4.2c),*

$$\begin{aligned}r^{k+1}(\lambda,\mu) := -\langle\tilde{\varepsilon}^{k+1}|\mu - \mu^{k+1} - (\pi^2_\# - \pi^1_\#)\lambda\rangle - \tau\pi^{1,2}_\#\lambda(B_{v^{k+1}+w^{k+1}}) \\ + \tau[\langle v^k - \check{v}^k|(\pi^2_\# - \pi^1_\#)\lambda\rangle - \ell_r\pi^{0,1}_\#|\lambda|(c_2)].\end{aligned} \tag{4.7}$$

Proof. By Lemma A.1,

$$\begin{aligned}\langle\nabla\check{v}^k(y) - \nabla v^k(x), z - y\rangle &= \langle\nabla\check{v}^k(y) - \nabla v^k(y), z - y\rangle + B_{v^k}(x,z) - B_{v^k}(x,y) - B_{v^k}(y,z) \\ &= [v^k - \check{v}^k](y) - [v^k - \check{v}^k](z) + B_{v^k}(x,z) - B_{v^k}(x,y) - B_{\check{v}^k}(y,z).\end{aligned}$$

Hence,

$$\begin{aligned}\int\langle\nabla\check{v}^k(y) - \nabla v^k(x), z - y\rangle\, d\lambda(x,y,z) = \pi^{1,2}_\#\lambda(B_{v^{k+1}-\check{v}^k}) - \langle v^k - \check{v}^k|(\pi^2_\# - \pi^1_\#)\lambda\rangle \\ + \mathscr{K}_F(\mu^k, \pi^{0,2}_\#\lambda) - \mathscr{K}_F(\mu^k, \pi^{0,1}_\#\lambda) - \mathscr{K}_F(\mu^{k+1}, \pi^{1,2}_\#\lambda).\end{aligned} \tag{4.8}$$

Using (4.2b), the assumption $\lambda \in \Lambda_{01}(\gamma^{k+1})$, and (4.8), we get

$$
\begin{aligned}
(4.9)\qquad & \frac{1}{\tau\theta}\int_{\Omega^3} \langle y-x, z-y\rangle \, d|\lambda|(x,y,z) = \frac{1}{\tau\theta}\int_{\Omega^3} \langle y-x, z-y\rangle \operatorname{sign}\lambda(x,y,z)\, d\lambda(x,y,z) \\
&= \int_{\Omega^3} \langle \nabla\check{v}^k(y) - \nabla v^k(x) - \nabla[\check{v}^k + w^{k+1}](y), z-y\rangle \, d\lambda(x,y,z) \\
&= \int_{\Omega^3} \langle \nabla\check{v}^k(y) - \nabla v^k(x), z-y\rangle \, d\lambda(x,y,z) + \pi^{1,2}_\#\lambda(B_{\check{v}^k+w^{k+1}}) - \langle \check{v}^k + w^{k+1} | (\pi^2_\# - \pi^1_\#)\lambda\rangle \\
&= \mathscr{K}_F(\mu^k, \pi^{0,2}_\#\lambda) - \mathscr{K}_F(\mu^k, \pi^{0,1}_\#\lambda) - \mathscr{K}_F(\mu^{k+1}, \pi^{1,2}_\#\lambda) - \langle v^k - \check{v}^k | (\pi^2_\# - \pi^1_\#)\lambda\rangle \\
&\quad + \pi^{1,2}_\#\lambda(B_{v^{k+1}+w^{k+1}}) - \langle \check{v}^k + w^{k+1} | (\pi^2_\# - \pi^1_\#)\lambda\rangle \\
&= \mathscr{K}_F(\mu^k, \pi^{0,2}_\#\lambda) - \mathscr{K}_F(\mu^k, \pi^{0,1}_\#\lambda) - \mathscr{K}_F(\mu^{k+1}, \pi^{1,2}_\#\lambda) - \tau^{-1}\ell_r\pi^{0,1}_\#|\lambda|(c_2) \\
&\quad - \tau^{-1}\langle \tilde{\varepsilon}^{k+1} | \mu - \mu^{k+1} - (\pi^2_\# - \pi^1_\#)\lambda\rangle - \langle \check{v}^k + w^{k+1} | (\pi^2_\# - \pi^1_\#)\lambda\rangle - \tau^{-1} r^{k+1}(\mu,\lambda).
\end{aligned}
$$

Multiplying (4.9) by τ and rearranging the two last dual products establishes the claim. □

We can now prove a transport-adapted version of a standard quasi-monotonicity result. In the proof, we construct the three-plan $\lambda^{k+1}_{\mathrm{rev}} := p_\#\gamma^{k+1}$ for $p(x,y) = (x,y,x)$, which satisfies

$$
r^{k+1}(\lambda^{k+1}_{\mathrm{rev}}, \mu^k) = \check{r}^{k+1}. \tag{4.10}
$$

This will be important in Section 5.3, where we seek to control this term.

Lemma 4.10 (Value quasi-monotonicity). *Suppose Assumption 4.3 (i) to (iii) hold, and that $k \in \mathbb{N}$ and $(\mu^k, \gamma^{k+1}) \in \mathscr{M}(\Omega) \times \mathscr{M}(\Omega^2)$ are given. Then $\mathscr{E}^*_k \geq 0$, and, if (4.2a) and (4.2b) hold,*

$$
\tau[F+G](\mu^{k+1}) + \mathscr{E}^*_k(\omega^{k+1}) + \check{V}_k(\mu^k, \mu^{k+1}; \gamma^{k+1}) \leq \tau[F+G](\mu^k) + \check{r}^{k+1}. \tag{4.11}
$$

Proof. As $\mathscr{E}_k(0) = 0$ by Assumption 4.3 (i), we have, from the definition, $\mathscr{E}^*_k \geq 0$. Let then ω^{k+1} be as in (4.2a), and abbreviate $\lambda = \lambda^{k+1}_{\mathrm{rev}}$. Then $\lambda \in \Lambda_{01}(\gamma^{k+1})$, and $(\pi^2_\# - \pi^0_\#)\lambda = 0$. Thus, $\mathscr{K}_F(\mu^k, \pi^{0,2}_\#\lambda) = 0$, and using Theorem 3.3 and Lemma 4.9 with $\mu = \mu^k$, we get

$$
\begin{aligned}
(4.12)\qquad 0 &\geq V_{2\theta^{-1}c_{2,E}}(\mu^k, \mu^{k+1}; \gamma^{k+1}) + \mathscr{E}^*_k(\omega^{k+1}) \\
&\quad - \frac{1}{\theta}\int_{\Omega^3} |x-y|^2 \, d|\gamma^{k+1}|(x,y) + \langle \omega^{k+1} | \mu^k - \mu^{k+1} - (\pi^1_\# - \pi^0_\#)\gamma^{k+1}\rangle \\
&= V_{2\theta^{-1}c_{2,E}}(\mu^k, \mu^{k+1}; \gamma^{k+1}) + \mathscr{E}^*_k(\omega^{k+1}) \\
&\quad + \frac{1}{\theta}\int_{\Omega^3} \langle y-x, z-y\rangle \, d|\lambda|(x,y,z) + \langle \omega^{k+1} | \mu^k - \mu^{k+1} - (\pi^2_\# - \pi^1_\#)\lambda\rangle \\
&\geq V_{2\theta^{-1}c_{2,E}}(\mu^k, \mu^{k+1}; \gamma^{k+1}) + \mathscr{E}^*_k(\omega^{k+1}) - \tau\mathscr{K}_F(\mu^k, \gamma^{k+1}) - \tau\mathscr{K}_F(\mu^{k+1}, \mathrm{rev}_\#\gamma^{k+1}) \\
&\quad - \tau\langle \check{v}^k + w^{k+1} | \mu^k - \mu^{k+1}\rangle - \tau\ell_r|\gamma^{k+1}|(c_2) - r^{k+1}(\mu^k, \lambda).
\end{aligned}
$$

Since F has Lipschitz derivative by Assumption 4.3 (iii), the descent inequality holds. Also using the convexity/subdifferentiability of G (Assumption 4.3 (ii)), we get

$$
[F+G](\mu^k) - [F+G](\mu^{k+1}) \geq \langle \check{v}^k + w^{k+1} | \mu^k - \mu^{k+1}\rangle + B_F(\check{\mu}^k, \mu^k) - B_F(\check{\mu}^k, \mu^{k+1}).
$$

By Assumption 4.3 (iii), we also have

$$
B_F(\check{\mu}^k, \mu^{k+1}) \leq V_{\ell c_2, LE}(\mu^k, \mu^{k+1}; \gamma^{k+1}),
$$

so we get

$$
[F+G](\mu^k) - [F+G](\mu^{k+1}) \geq \langle \check{v}^k + w^{k+1} | \mu^k - \mu^{k+1}\rangle + B_F(\check{\mu}^k, \mu^k) - V_{\ell c_2, LE}(\mu^k, \mu^{k+1}; \gamma^{k+1}). \tag{4.13}
$$

To establish the claim, we add this inequality multiplied by τ to (4.12), and then use the definition of $\check{V}_k$ from (4.4). □

Theorem 4.11 (Subdifferential convergence). *Suppose Assumptions 4.3 and 4.5 hold, and that* $\{(\mu^k, \gamma^{k+1})\}_{k\in\mathbb{N}}$ *are generated through the satisfaction of (4.2). Then* $\inf_{w\in\partial G(\mu^{k+1})} \|F'(\mu^{k+1}) + w\|_{C_0(\Omega)} \to 0$.

Proof. Let $N \in \mathbb{N}$. We apply Lemma 4.10 for all $k = 0, \dots, N-1$. Summing (4.11) over $k = 0, \dots, N-1$ and using $\check{r}^{k+1} \leq \check{C}\varepsilon^{k+1}$, we obtain

$$\tau[F+G](\mu^N) + \sum_{k=0}^{N-1} \left(\mathscr{E}_k^*(\omega^{k+1}) + \check{V}_k(\mu^k, \mu^{k+1}; \gamma^{k+1}) \right) \leq \tau[F+G](\mu^0) + \check{C} \sum_{k=0}^{N-1} \varepsilon^{k+1}.$$

By Assumption 4.5 (iv), $[F+G](\mu^N) \geq \inf[F+G] > -\infty$. Minding Assumption 4.3 (iv), it follows for some constant C, independent of N, that

$$\sum_{k=0}^{N-1} \left(\mathscr{E}_k^*(\omega^{k+1}) + \check{V}_k(\mu^k, \mu^{k+1}; \gamma^{k+1}) \right) \leq C.$$

By Lemmas 5.3 and 4.10, we have $\check{V}_k(\mu^k, \mu^{k+1}; \gamma^{k+1}) \geq 0$ and $\mathscr{E}_k^*(\omega^{k+1}) \geq 0$. Thus, letting $N \to \infty$ above, we see that both converge to zero. By Lemma 4.6, this implies $V_{c_2,E}(\mu^k, \mu^{k+1}; \gamma^{k+1}) \to 0$, hence $E(\mu^k - \pi^0_\# \gamma^{k+1}, \mu^{k+1} - \pi^1_\# \gamma^{k+1}) \to 0$. Assumption 4.5 (ii) then implies $F'(\mu^{k+1}) - F'(\check{\mu}^k) \to 0$. By (4.2c) we now have $\omega^{k+1} + \tau[F'(\mu^{k+1}) + w^{k+1}] = \tau[F'(\mu^{k+1}) - F'(\check{\mu}^k)] + \check{\varepsilon}^{k+1} \to 0$. As Assumption 4.5 (iii) establishes $\omega^{k+1} \to 0$, it follows that $F'(\mu^{k+1}) + w^{k+1} \to 0$. This implies the claim. □

The following corollary can be useful for verifying various assumptions in an inductive manner.

Corollary 4.12. *Suppose Assumptions 4.3 and 4.5 hold. Pick* $N \in \mathbb{N}$, *and for initial* $\mu^0 \in \mathscr{M}(\Omega)$, *generate* $\{(\mu^{k+1}, \gamma^{k+1})\}_{k=0}^{N-1}$ *through the satisfaction of (4.2). If* $\inf F \geq i_F > -\infty$ *and* $G \geq \varphi(\|\mu\|)$ *for a coercive* $\varphi : [0,\infty) \to [0,\infty)$, *then* $\sup_{k=0,\dots,N-1} \|\mu^{k+1}\| \leq m_\mu$ *for a constant* m_μ *independent of* N.

Proof. The proof of Theorem 4.11 establishes $\tau[F+G](\mu^N) \leq \tau[F+G](\mu^0) + \check{C}\sum_{k=0}^{N-1} \varepsilon^{k+1}$. The claim now follows after using Assumption 4.3 (iv) and the bounds on F and G. □

5 SUBLINEAR CONVERGENCE OF FUNCTION VALUES

We continue from Section 4 to show $O(1/N)$ convergence rates under additional technical requirements. Under suitable second-order growth assumptions, the work here could be extended to linear convergence. We will work with three-plans $\lambda^{k+1} \in \mathscr{M}(\Omega^3)$ to transport between the triples $(\mu^k, \mu^{k+1}, \bar{\mu})$ for a reference point $\bar{\mu}$. First, in Section 5.1 we discuss the evolution or "weaving" of these three-plans, and introduce distances for measuring the convergence of iterates, as well as additional assumptions needed to prove the convergence of the generic method in Section 5.2. After this we discuss in Section 5.3 how to bound the remainder term $\check{r}^{k+1}$ from (4.2e), and its more general version that we need in this section. Finally, in Section 5.4, we discuss a specific algorithm for point source localisation with the $\mathscr{D}$-norm as a marginal term.

5.1 DISTANCES AND WEAVING OF THREE-PLANS

We recall the definition (4.5) of three-plans $\lambda^{k+1} \in \Lambda_{01}(\gamma^{k+1})$ compatible with the two-plan γ^{k+1}. The projected transport $\pi^{1,2}_\# \lambda^{k+1}$ from μ^{k+1} to $\bar{\mu}$ will, subject to diagonal modifications, be "weaved" into the transport $\pi^{0,2}_\# \lambda^{k+2}$ of the next step, likewise from μ^{k+1} to $\bar{\mu}$. We express this restriction as $\lambda^{k+2} \in \Lambda^+(\lambda^{k+1})$ for

$$\Lambda^+(\lambda) := \left\{ \tilde{\lambda} \in \mathscr{M}(\Omega^3) \;\middle|\; \begin{array}{l} \pi^{0,2}_\# \tilde{\lambda} = \pi^{1,2}_\# \lambda + \operatorname{diag}_\# v \text{ for some } v \in \mathscr{M}(\Omega) \\ \pi^{0,2}_\# |\tilde{\lambda}| \leq \pi^{1,2}_\# |\lambda| + \operatorname{diag}_\# \bar{v} \text{ for some } \bar{v} \in \mathscr{M}(\Omega) \end{array} \right\}. \tag{5.1}$$

Table 1: The conditions that specific instances of the generic algorithm need to be secure for different types of convergence. For the function value convergences, the satisfaction of the remainder condition through Section 5.3 gives an indirect marginal condition; otherwise, there is none. The remainder condition for subdifferential convergence can be satisfied directly.

Convergence	Variables	Transport	Marginal	Curvature	Remainder	Result
Subdifferential	|	|	(4.2c)	(4.2d)	(4.2e)	Theorem 4.11
Ergodic $O(1/N)$	(4.2a)	(4.2b)	↱ (5.14)	(5.2a) (5.2b)	(5.2c)	Theorem 5.7
Value $O(1/N)$	|	|	↱ (5.14)	(4.2d) (5.2a) (5.2b)	(5.2d)	Corollary 5.8
Observations			indirect	Remark 5.15	Section 5.3	

Lemma 5.1. *There exists a* $\lambda^{k+1} \in \Lambda_{01}(\gamma^{k+1}) \cap \Lambda^+(\lambda^k)$.

Proof. We decompose $\gamma^{k+1} = \gamma_+^{k+1} - \gamma_-^{k+1}$ and $\lambda^k = \lambda_+^k - \lambda_-^k$, where the component measures are non-negative. We will construct non-negative measures $\lambda_\pm^{k+1} \in \Lambda_{01}(\gamma_\pm^{k+1}) \cap \Lambda^+(\lambda_\pm^k)$, and then set $\lambda^{k+1} = \lambda_+^{k+1} - \lambda_-^{k+1}$. By construction, $\lambda^{k+1} \in \Lambda_{01}(\gamma^{k+1})$. Because the supports of $\gamma_\pm^{k+1}$ are disjoint, as are those of $\lambda_\pm^k$, so are the supports of $\lambda_\pm^{k+1}$. It follows that also $\lambda^{k+1} \in \Lambda^+(\lambda^k)$. For the rest of the proof, we may therefore assume, w.log, that $\gamma^{k+1} \geq 0$ and $\lambda^k \geq 0$.

Abbreviate $\bar\gamma := \pi_\#^{1,2}\lambda^k$. By the Lebesgue decomposition and Radon–Nikodym theorems, we can write $\pi_\#^0\gamma^{k+1} = f\pi_\#^0\bar\gamma + \nu_s$, where $\nu_s \perp \pi_\#^0\bar\gamma$ and $f \geq 0$ is Borel measurable. Let $\tilde\gamma := \min\{1, f\}\bar\gamma + \mathrm{diag}_\# \nu$ for $\nu := \pi_\#^0\gamma^{k+1} - \min\{1, f\}\pi_\#^0\bar\gamma$. Then $\pi_\#^0\tilde\gamma = \pi_\#^0\gamma^{k+1}$, so by [1, Lemma 5.3.2], there exists $0 \leq \lambda_0 \in \mathscr{M}(\Omega^3)$ with $\pi_\#^{0,2}\lambda = \tilde\gamma$ and $\pi_\#^{0,1}\lambda = \gamma^{k+1}$. Let then $\lambda := \lambda_0 + p_\# \max\{0, 1-f\}\bar\gamma$ for $p(x,z) = (x,z,z)$. This has the structure of the definition of $\Lambda_{01}(\gamma^{k+1})$ in (4.5). Moreover, we have $\pi_\#^{0,2}\lambda = \bar\gamma + \mathrm{diag}_\# \nu$, which satisfies the definition of $\Lambda^+(\lambda^k)$ in (5.1). □

For r^{k+1} defined in (4.7), and $\check r^{k+1}$ defined in (4.2e) and dependent on the parameter ℓ_r, we need to modify parts of (4.2), as follows:

For some $C_{\mathscr{K}}, \ell_r, C_{\bar\mu}, \ell_F > 0$ and a comparison point $\bar\mu \in \mathscr{M}(\Omega)$ of interest (typically a minimiser), for a $\lambda^{k+1} \in \Lambda_{01}(\gamma^{k+1}) \cap \Lambda^+(\lambda^k)$ we have

$$\ell_F|\gamma^{k+1}|(c_2) \geq \mathscr{K}_F(\mu^k, \gamma^{k+1}), \tag{5.2a}$$

$$-C_{\mathscr{K}} \leq \ell_F\pi_\#^{0,2}|\lambda^{k+1}|(c_2) - \mathscr{K}_F(\mu^k, \pi_\#^{0,2}\lambda^{k+1}) \quad \text{and} \tag{5.2b}$$

$$C_{\bar\mu}\varepsilon^{k+1} \geq r^{k+1}(\lambda^{k+1}, \bar\mu) \quad \text{for ergodic convergence, or} \tag{5.2c}$$

$$C_{\bar\mu}\varepsilon^{k+1} \geq (k+1)\check r^{k+1} + r^{k+1}(\lambda^{k+1}, \bar\mu) \quad \text{for non-ergodic convergence.} \tag{5.2d}$$

We summarise the exact conditions that each form of convergence needs to satisfy in Table 1.

We will, besides Assumption 4.3, require:

Assumption 5.2. We have:

(i) F is convex.

(ii) The step lengths $\tau, \theta > 0$, the parameters ℓ and L from Assumption 4.3, and ℓ_F and ℓ_r from (4.2) and (5.2), satisfy $\tau L \leq 1$ and $\theta\tau[\ell + \ell_r + \ell_F] \leq 1$.

(iii) Besides Assumption 4.3 (i), E satisfies (3.2).

The latter condition rules out the Radon norm marginal term $E_{\mathscr{M}}$.

Recalling the definition of $\bar V^{i,j}_{\theta^{-1}c_2,E}$ from Theorem 3.1, we introduce between μ^{k+i} and $\bar\mu$ for $i = 0, 1$ the distances

$$\begin{aligned} V_k^{i,2}(\mu, \nu; \lambda) &:= \bar V^{i,2}_{\theta^{-1}c_2,E}(\mu, \nu; \lambda) - \tau\mathscr{K}_F(\mu, \pi_\#^{i,2}\lambda) \\ &= \theta^{-1}\pi_\#^{i,2}|\lambda|(c_2) - \tau\pi_\#^{i,2}\lambda(B_{F'(\mu)}) + E(\mu - \pi_\#^i\lambda, \nu - \pi_\#^2\lambda) \end{aligned} \tag{5.3}$$

For distances between μ^k and μ^{k+1}, we recall $\check{V}_k$ defined in (4.4) and also introduce

$$V_k^{0,1}(\mu,\nu;\lambda) := \bar{V}^{0,1}_{\theta^{-1}c_2,E}(\mu,\nu;\lambda) - \tau[V_{\ell c_2,LE}(\mu,\nu;\pi_\#^{0,1}\lambda) + \mathscr{K}_F(\mu,\pi_\#^{0,1}\lambda) + \ell_r\pi_\#^{0,1}|\lambda|(c_2)] \\ = (\theta^{-1} - \tau[\ell+\ell_r])\pi_\#^{0,1}|\lambda|(c_2) - \tau\pi_\#^{0,1}\lambda(B_{F'(\mu)}) + (1-\tau L)E(\mu - \pi_\#^0\lambda, \nu - \pi_\#^1\lambda). \tag{5.4}$$

The following lower bounds are immediate:

Lemma 5.3. *Suppose Assumption 4.3 (i), 5.2 (ii), as well as (5.2a) and (5.2b) hold. Then, for any $k \in \mathbb{N}$ and $\nu \in \mathscr{M}(\Omega)$, we have*

(i) $V_k^{0,2}(\mu^k, \nu; \lambda^{k+1}) \geq -C_{\mathscr{K}}$.

(ii) $V_k^{0,1}(\mu^k, \mu^{k+1}; \lambda^{k+1}) \geq 0$.

If also (4.2d) holds, then

(iii) $\check{V}_k(\mu^k, \mu^{k+1}; \gamma^{k+1}) \geq 0$.

The weaving compatibility condition $\lambda^{k+1} \in \Lambda^+(\lambda^k)$ allows us to rebalance the three-plan before commencing with the next step, as follows:

Assumption 5.4. For all $k \in \mathbb{N}$, for given $\mu^{k+1} \in \mathscr{M}(\Omega)$, $\gamma^{k+1} \in \mathscr{M}(\Omega^2)$, and $\lambda^{k+1} \in \Lambda_{01}(\gamma^{k+1})$, the choice of $\gamma^{k+2} \in \mathscr{M}(\Omega^2)$ and $\lambda^{k+2} \in \Lambda_{01}(\gamma^{k+2}) \cap \Lambda^+(\lambda^{k+1})$ is such that

$$V_k^{1,2}(\mu^{k+1}, \bar{\mu}; \lambda^{k+1}) \geq V_{k+1}^{0,2}(\mu^{k+1}, \bar{\mu}; \lambda^{k+2}). \quad \text{for all} \quad \bar{\mu} \in \mathscr{M}(\Omega). \tag{5.5}$$

The next lemma proves Assumption 5.4 for our typical choice of E.

Lemma 5.5. *Let $E = E_{\mathscr{D}}$ for a self-adjoint $\mathscr{D} \in \mathbb{L}(\mathscr{M}(\Omega); C_0(\Omega))$. Then Assumption 5.4 holds for any choice of $\gamma^{k+2} \in \mathscr{M}(\Omega^2)$ and $\lambda^{k+2} \in \Lambda_{01}(\gamma^{k+2}) \cap \Lambda^+(\lambda^{k+1})$.*

Proof. Inserting (5.3) into (5.5) and expanding, we need to show that

$$0 \leq \int_{\Omega^2} \frac{1}{\theta} c_2(x,z)\, d(\pi_\#^{1,2}|\lambda^{k+1}| - \pi_\#^{0,2}|\lambda^{k+2}|)(x,z) - \tau \int_{\Omega^2} B_{F'(\mu^{k+1})}(x,z)\, d(\pi_\#^{1,2}\lambda^{k+1} - \pi_\#^{0,2}\lambda^{k+2})(x,z) \\ + \frac{1}{2}\|\bar{\mu} - \mu^{k+1} - (\pi_\#^2 - \pi_\#^1)\lambda^{k+1}\|_{\mathscr{D}}^2 - \frac{1}{2}\|\bar{\mu} - \mu^{k+1} - (\pi_\#^2 - \pi_\#^0)\lambda^{k+2}\|_{\mathscr{D}}^2.$$

By the definition of $\Lambda^+(\lambda^{k+1}) \ni \lambda^{k+2}$ in (5.1), we have $\pi_\#^{0,2}\lambda^{k+2} = \pi_\#^{1,2}\lambda^{k+1} + \operatorname{diag} \nu$ and $\pi_\#^{0,2}|\lambda^{k+2}| \leq \pi_\#^{1,2}|\lambda^{k+1}| + \operatorname{diag} \bar{\nu}$ for some $\nu, \bar{\nu} \in \mathscr{M}(\Omega)$, hence also $(\pi_\#^2 - \pi_\#^0)\lambda^{k+2} = (\pi_\#^2 - \pi_\#^1)\lambda^{k+1}$. This shows that the first line is non-negative, and the rest zero. □

5.2 CONVERGENCE OF FUNCTION VALUES

We proceed with a transport-aware version of a standard descent estimate.

Lemma 5.6 (Descent estimate to a reference point). *Suppose Assumption 4.3 (i) to (iii) and 5.2 (iii) hold, and that $k \in \mathbb{N}$ and $(\mu^k, \gamma^{k+1}) \in \mathscr{M}(\Omega) \times \mathscr{M}(\Omega^2)$ are given. If (4.2a) and (4.2b) hold, then for any $\mu \in \mathscr{M}(\Omega)$, and any $\lambda \in \Lambda_{01}(\gamma^{k+1})$,*

$$\tau[F+G](\mu) - \tau[F+G](\mu^{k+1}) \geq V_k^{1,2}(\mu^{k+1}, \mu; \lambda) - V_k^{0,2}(\mu^k, \mu; \lambda) + V_k^{0,1}(\mu^k, \mu^{k+1}; \lambda) - r^{k+1}(\lambda, \mu). \tag{5.6}$$

Proof. Theorem 3.1 establishes

$$\bar{V}^{0,2}_{\theta^{-1}c_2,E}(\mu^k, \mu; \lambda) - \bar{V}^{0,1}_{\theta^{-1}c_2,E}(\mu^k, \mu^{k+1}; \lambda) \\ \geq \frac{1}{\theta}\int_{\Omega^3} \langle y-x, z-y\rangle\, d|\lambda|(x,y,z) + \langle \omega^{k+1} | \mu - \mu^{k+1} - (\pi_\#^2 - \pi_\#^1)\lambda\rangle + \bar{V}^{1,2}_{\theta^{-1}c_2,E}(\mu^{k+1}, \mu; \lambda).$$

Combining this with (4.6) of Lemma 4.9, we get

$$\begin{aligned} 0 \geq \bar{V}^{1,2}_{\theta^{-1}c_2,E}(\mu^{k+1},\mu;\lambda) - \bar{V}^{0,2}_{\theta^{-1}c_2,E}(\mu^k,\mu;\lambda) + \bar{V}^{0,1}_{\theta^{-1}c_2,E}(\mu^k,\mu^{k+1};\lambda) \\ + \tau[\mathscr{K}_F(\mu^k,\pi^{0,2}_\#\lambda) - \mathscr{K}_F(\mu^k,\pi^{0,1}_\#\lambda) - \mathscr{K}_F(\mu^{k+1},\pi^{1,2}_\#\lambda)] \\ - \tau\langle \check{v}^k + w^{k+1} | \mu - \mu^{k+1}\rangle - \tau\ell_r\pi^{0,1}_\#|\lambda|(c_2) - r^{k+1}(\mu,\lambda). \end{aligned} \tag{5.7}$$

Using the definitions (5.3) and (5.4), we obtain

$$\begin{aligned} 0 \geq V^{1,2}_k(\mu^{k+1},\mu;\lambda) - V^{0,2}_k(\mu^k,\mu;\lambda) + V^{0,1}_k(\mu^k,\mu^{k+1};\lambda) - r^{k+1}(\mu,\lambda). \\ - \tau\langle \check{v}^k + w^{k+1} | \mu - \mu^{k+1}\rangle + V_{\ell c_2,LE}(\mu^k,\mu^{k+1};\gamma^{k+1}) \end{aligned} \tag{5.8}$$

Analogously to (4.13), we prove

$$[F+G](\mu) - [F+G](\mu^{k+1}) \geq \langle \check{v}^k + w^{k+1} | \mu - \mu^{k+1}\rangle + B_F(\check{\mu}^k,\mu) - V_{\ell c_2,LE}(\mu^k,\mu^{k+1};\gamma^{k+1}). \tag{5.9}$$

Since $B_F(\check{\mu}^k,\mu) \geq 0$ by the convexity[3] of F, the claim now follows by combining (5.9) multiplied by τ with (5.8). □

We are now ready to state our main convergence theorems for the abstract algorithm (4.2) for (4.1).

Theorem 5.7 (Ergodic function value convergence). *Suppose Assumptions 4.3, 5.2 and 5.4 hold, and let $\bar{\mu} \in \mathscr{M}(\Omega)$. For an initial $\mu^0 \in \mathscr{M}(\Omega)$, generate $\{(\mu^{k+1},\gamma^{k+1})\}_{k\in\mathbb{N}}$ through the satisfaction of the "Ergodic $O(1/N)$" conditions of Table 1. Pick $N \in \mathbb{N}$ and $\lambda^1 \in \Lambda_{01}(\gamma^1)$. Then for $\tilde{\mu}^N := \frac{1}{N}\sum_{k=0}^{N-1}\mu^{k+1}$, we have*

$$[F+G](\tilde{\mu}^N) - [F+G](\bar{\mu}) \leq \frac{1}{\tau N}V^{0,2}_0(\mu^0,\bar{\mu};\lambda^1) + \frac{C_{\mathscr{K}}}{\tau N} + \frac{C_{\bar{\mu}}}{\tau N}\sum_{k=0}^{N-1}\varepsilon^{k+1} \tag{5.10}$$

In particular, $[F+G](\tilde{\mu}^N) \to \inf[F+G]$ at the rate $O(1/N)$.

Proof. We have assumed to be given some $\lambda^1 \in \Lambda_{01}(\gamma^1)$; for example, $\lambda^1 = \lambda^1_{\mathrm{rev}}$, as constructed in Lemma 5.6. For all $k \geq 1$, we let $\lambda^{k+1} \in \Lambda_{01}(\gamma^{k+1}) \cap \Lambda^+(\lambda^k)$ be determined by (5.2) and the rebalancing Assumption 5.4.

Pick any $k \in \{0,\dots,N-1\}$. We apply Lemma 5.6 with $\mu = \bar{\mu}$ and $\lambda = \lambda^{k+1}$ to obtain (5.6). Using Assumption 5.4 and Lemma 5.3 (ii) there, we obtain

$$\tau[F+G](\bar{\mu}) - \tau[F+G](\mu^{k+1}) \geq V^{0,2}_{k+1}(\mu^{k+1},\bar{\mu};\lambda^{k+2}) - V^{0,2}_k(\mu^k,\bar{\mu};\lambda^{k+1}) - r^{k+1}(\lambda^{k+1},\bar{\mu}). \tag{5.11}$$

Using (5.2c) and summing over $k = 0,\dots,N-1$ in (5.11), we thus obtain

$$V^{0,2}_N(\mu^N,\bar{\mu};\lambda^{N+1}) + \sum_{k=0}^{N-1}\tau[F+G](\mu^{k+1}) \leq V^{0,2}_0(\mu^0,\bar{\mu};\lambda^1) + \tau N[F+G](\bar{\mu}) + \sum_{k=0}^{N-1}C_{\bar{\mu}}\varepsilon^{k+1}.$$

An application of Lemma 5.3 (i) and Jensen's inequality then establishes (5.10). Assumption 4.3 (iv) now establishes the convergence rate claim. □

Corollary 5.8 (Function value convergence). *Suppose Assumption 4.3, 4.5 (i), 5.2, and 5.4 hold, and let $\bar{\mu} \in \mathscr{M}(\Omega)$. For an initial $\mu^0 \in \mathscr{M}(\Omega)$, generate $\{(\mu^{k+1},\gamma^{k+1})\}_{k\in\mathbb{N}}$ through the satisfaction of the "Value $O(1/N)$" conditions of Table 1. Then for any $N \in \mathbb{N}$ and $\lambda^1 \in \Lambda_{01}(\gamma^1)$, we have*

$$[F+G](\mu^N) - [F+G](\bar{\mu}) \leq \frac{1}{N\tau}V^{0,2}_0(\mu^0,\bar{\mu};\lambda^1) + \frac{C_{\mathscr{K}}}{\tau N} + \frac{C_{\bar{\mu}}}{\tau N}\sum_{k=0}^{N-1}\varepsilon^{k+1}.$$

In particular, $[F+G](\mu^N) \to \inf[F+G]$ at the rate $O(1/N)$.

[3] This is the only place where we require the convexity. With some effort, the assumption could be made local; compare [19].

Proof. We follow the proof of Theorem 5.7 until (5.11). We then repeatedly apply (4.11) of Lemma 4.10 in (5.11) and sum over $k = 0, \dots, N-1$ to obtain

$$V_N^{0,2}(\mu^N, \bar{\mu}; \lambda^{N+1}) + b_N + N\tau[F+G](\mu^N) \leq V_0^{0,2}(\mu^0, \bar{\mu}; \lambda^1) + N\tau[F+G](\bar{\mu}) + a_N \tag{5.12}$$

for

$$a_N := \sum_{k=0}^{N-1} \left(\sum_{j=k+1}^{N-1} \check{r}^{j+1} + r^{k+1}(\lambda^{k+1}, \bar{\mu}) \right) = \sum_{k=0}^{N-1} \left((k+1)\check{r}^{k+1} + r^{k+1}(\lambda^{k+1}, \bar{\mu}) \right)$$

and

$$b_N := \sum_{k=0}^{N-1} \sum_{j=k+1}^{N-1} \check{V}_j(\mu^j, \mu^{j+1}; \gamma^{j+1}).$$

By Lemma 4.6 and 5.3 (iii), we have $b_N \geq 0$. Dividing (5.12) by $N\tau$ and using (5.2d) and Lemma 5.3 (i), the claim thus follows. □

The following corollary will be useful for inductively verifying various assumptions.

Corollary 5.9. *Suppose Assumptions 4.3, 5.2 and 5.4 hold, and let $\bar{\mu} \in \mathscr{M}(\Omega)$. Pick $N \in \mathbb{N}$, and for initial $\mu^0 \in \mathscr{M}(\Omega)$ and $\lambda^1 \in \Lambda_{01}(\gamma^1)$, generate $\{(\mu^{k+1}, \gamma^{k+1})\}_{k=0}^{N-1}$ through the satisfaction of either the "Ergodic $O(1/N)$" or "Value $O(1/N)$" conditions of Table 1. If $\inf F \geq i_F > -\infty$ and $G \geq \varphi(\|\mu\|)$ for a coercive $\varphi : [0,\infty) \to [0,\infty)$, then $\sup_{k=0,\dots,N-1} \|\mu^k\| \leq m_\mu$ for a constant m_μ independent of N.*

Proof. We first show the "Ergodic $O(1/N)$" case. Using $[F+G](\mu^k) \geq [F+G](\bar{\mu})$ in (5.11) in the proof of Theorem 5.7 establishes

$$\tau[F+G](\mu^{k+1}) + V_{k+1}(\mu^{k+1}, \bar{\mu}; \pi_\#^{0,2}\lambda^{k+2}) \leq \tau[F+G](\mu^k) + V_k(\mu^k, \bar{\mu}; \pi_\#^{0,2}\lambda^{k+1}) + r^{k+1}(\lambda^{k+1}, \bar{\mu}).$$

Summing this over $k = 0, \dots, N-1$, using Assumption 4.3 (iv), Lemma 5.3 (i), and the bounds on F and G, we obtain

$$\tau i_F - C_{\mathscr{K}} + \tau\varphi(\|\mu^N\|_{\mathscr{M}}) \leq \tau[F+G](\mu^0) + V_0(\mu^0, \bar{\mu}; \pi_\#^{0,2}\lambda^1) + \sum_{k=0}^{N-1} C\varepsilon^{k+1}.$$

Minding Assumption 4.3 (iv), we thus establish $\|\mu^N\| \leq m_\gamma$ for a constant m_γ independent of N.

The "Value $O(1/N)$" case follows similarly from (5.12). □

5.3 CONTROLS ON THE CURVATURE AND REMAINDER

We now seek to enforce the curvature and remainder conditions of Table 1. The approaches we present here are just some of many options. We recall from (4.2e), (4.7) and (4.10) the expressions for $r^{k+1}(\lambda, \mu)$ and $\check{r}^{k+1}$, and from (4.2c) that $\tilde{\varepsilon}^{k+1} = \tau[\check{v}^k + w^{k+1}] + \omega^{k+1}$. We can rearrange

$$\tau B_{v^{k+1}+w^{k+1}} = B_{\tilde{\varepsilon}^{k+1}} - B_{\omega^{k+1} - \tau[v^{k+1} - \check{v}^k]}, \tag{5.13}$$

whose summands we find more practical to control individually, due to their structure.

To control the marginal inexactness $\tilde{\varepsilon}^{k+1}$, for some $C' > 0$, we enforce

$$-\varepsilon^{k+1} \leq \tilde{\varepsilon}^{k+1} \leq \varepsilon^{k+1}, \quad \text{i.e., (4.2c)}, \tag{5.14a}$$

$$C'\varepsilon^{k+1} \geq \langle \mu^{k+1} - \pi_\#^1\gamma^{k+1} | \tilde{\varepsilon}^{k+1} \rangle, \quad \text{and} \tag{5.14b}$$

$$C'\varepsilon^{k+1} \geq \int_{\Omega^2} \sup_{z\in\Omega} [-B_{\tilde{\varepsilon}^{k+1}}(y,z) \operatorname{sign} \gamma^{k+1}(x,y)] \, d|\gamma^{k+1}|(x,y). \tag{5.14c}$$

We can now bound $r^{k+1}(\lambda,\mu)$ through simpler components. If we take $|\pi^0_\# \gamma^{k+1}| \le |\mu^k|$, as we will, then the bound $\|\gamma^{k+1}\|_{\mathscr{M}} \le m_\gamma$ in the next lemma follows from Corollary 4.12 or Corollary 5.9.

Lemma 5.10. *Suppose (5.14) holds, and we ensure $\|\gamma^{k+1}\|_{\mathscr{M}} \le m_\gamma$ for some $m_\gamma \ge 0$. If, moreover, for some $C_{\mathrm{cur}}, C_{\mathrm{con}} \ge 0$, we have*

$$C_{\mathrm{cur}}\varepsilon^{k+1} \ge r^{k+1}_{\mathrm{cur}} := \tau[\langle v^k - \check{v}^k | (\pi^2_\# - \pi^1_\#)\lambda^{k+1}\rangle - \ell_r \pi^{0,1}_\# |\lambda^{k+1}|(c_2)] \quad \textit{and} \tag{5.15a}$$

$$C_{\mathrm{con}}\varepsilon^{k+1} \ge r^{k+1}_{\mathrm{con}} := \int_{\Omega^2} \sup_{z\in\Omega} [B_{\omega^{k+1}-\tau[v^{k+1}-\check{v}^k]}(y,z)\,\operatorname{sign}\gamma^{k+1}(x,y)]\, d|\gamma^{k+1}|(x,y), \tag{5.15b}$$

then (5.2c) holds, more precisely, for any $\mu \in \mathscr{M}(\Omega)$,

$$r^{k+1}(\lambda^{k+1},\mu) \le (\|\mu\|_{\mathscr{M}} + m_\gamma + 1 + C' + C_{\mathrm{con}} + C_{\mathrm{cur}})\varepsilon^{k+1}. \tag{5.15c}$$

Likewise, if

$$C_{\mathrm{cur}}\varepsilon^{k+1} \ge \check{r}^{k+1}_{\mathrm{cur}} := \tau[\langle v^k - \check{v}^k | (\pi^0_\# - \pi^1_\#)\gamma^{k+1}\rangle - \ell_r |\gamma^{k+1}|(c_2)] \quad \textit{and} \tag{5.16a}$$

$$C_{\mathrm{con}}\varepsilon^{k+1} \ge \check{r}^{k+1}_{\mathrm{con}} := \int_{\Omega^2} B_{\omega^{k+1}-\tau[v^{k+1}-\check{v}^k]}(y,x)\, d\gamma^{k+1}(x,y), \tag{5.16b}$$

then (4.2e) holds, more precisely, for any $\mu \in \mathscr{M}(\Omega)$,

$$\check{r}^{k+1} \le (\|\mu\|_{\mathscr{M}} + m_\gamma + 1 + C' + C_{\mathrm{con}} + C_{\mathrm{cur}})\varepsilon^{k+1}. \tag{5.16c}$$

Finally, (5.2d) holds if, besides (5.15a) and (5.15b), we have $C_{\mathrm{con}}\varepsilon^{k+1} \ge (k+1)\check{r}^{k+1}_{\mathrm{con}}$ and $C_{\mathrm{cur}}\varepsilon^{k+1} \ge (k+1)\check{r}^{k+1}_{\mathrm{cur}}$.

Proof. Let $\lambda \in \Lambda_{01}(\gamma^{k+1})$. By the definition of $\Lambda_{01}(\gamma^{k+1})$ in (4.5), we have $(\pi^2_\# - \pi^1_\#)\lambda = (\pi^2_\# - \pi^1_\#)\lambda_0$ for some $\lambda_0 \in \mathscr{M}(\Omega^3)$ with $\pi^{0,1}_\#\lambda_0 = \gamma^{k+1}$. Hence, (5.13) and the definition (4.7) of $r^{k+1}(\lambda,\mu)$ yield

$$\begin{aligned} r^{k+1}(\lambda,\mu) = {} & -\langle \tilde{\varepsilon}^{k+1} | \mu - \mu^{k+1} - (\pi^2_\# - \pi^1_\#)\lambda_0\rangle + \pi^{1,2}_\#\lambda_0(B_{\omega^{k+1}-\tau[v^{k+1}-\check{v}^k]} - B_{\tilde{\varepsilon}^{k+1}}) \\ & + \tau[\langle v^k - \check{v}^k | (\pi^2_\# - \pi^1_\#)\lambda\rangle - \ell_r \pi^{0,1}_\# |\lambda|(c_2)]. \end{aligned} \tag{5.17}$$

We have

$$\pi^{1,2}_\#\lambda_0(-B_{\tilde{\varepsilon}^{k+1}}) = \int_{\Omega^3} -B_{\tilde{\varepsilon}^{k+1}}(y,z)\, d\lambda_0(x,y,z) \le \int_{\Omega^3} \sup_{z\in\Omega}(-B_{\tilde{\varepsilon}^{k+1}}(y,z)\operatorname{sign}\gamma^{k+1}(x,y))\, d|\gamma^{k+1}|(x,y),$$

and likewise for $\pi^{1,2}_\#\lambda_0(B_{\omega^{k+1}-\tau[v^{k+1}-\check{v}^k]})$. Using $\|\pi^2_\#\lambda\|_{\mathscr{M}} \le \|\gamma^{k+1}\|_{\mathscr{M}} \le m_\gamma$ and (5.14), we thus estimate

$$\begin{aligned} \langle \tilde{\varepsilon}^{k+1} | \mu^{k+1} - \mu + (\pi^2_\# - \pi^1_\#)\lambda_0\rangle - \pi^{1,2}_\#\lambda_0(B_{\tilde{\varepsilon}^{k+1}}) &\le \varepsilon^{k+1}\|\mu - \pi^2_\#\lambda_0\|_{\mathscr{M}} + \langle \tilde{\varepsilon}^{k+1} | \mu^{k+1} - \pi^1_\#\gamma^{k+1}\rangle + C'\varepsilon^{k+1} \\ &\le (\|\mu\|_{\mathscr{M}} + m_\gamma + 1 + C')\varepsilon^{k+1}. \end{aligned}$$

Using this with $\lambda = \lambda^{k+1}$, (5.15a), and (5.15b) in (5.17) yields (5.15c) To prove (5.16c) we recall (5.2c), fix $\lambda = \lambda^{k+1}_{\mathrm{rev}}$ above, and do not take the suprema. The proof of (5.2d) is immediate from (5.15) and (5.16). □

We now seek to ensure the required *curvature transport error bounds* on $\check{r}^{k+1}_{\mathrm{cur}}$ or r^{k+1}_{cur} and *convexity transport error bounds* on $\check{r}^{k+1}_{\mathrm{con}}$ or r^{k+1}_{con}. Generally, the curvature error bounds can be satisfied *a priori*, before solving for μ^{k+1}, ω^{k+1}, w^{k+1}, while the convexity error bounds can only be checked *a posteriori*, and their satisfaction may require adjusting the transport γ^{k+1}, and repeating the process.

For our typical quadratic-affine data term, the firm transport Lipschitz property required by the next lemma is proved in Lemma 3.8.

Lemma 5.11 (Curvature transport error control for subdifferential convergence only). *Suppose F' is firmly transport Lipschitz in the sense (3.8), and that $\|\gamma^{k+1}\|_{\mathscr{M}} \leq m_\gamma$ as in Lemma 5.10. Then (5.16a) holds with $C_{\mathrm{cur}} = 0$ if we take $\ell_r = \Theta_F^2 m_\gamma$. Moreover,*

$$-B_F(\check{\mu}^k, \mu^k) \leq \begin{cases} 0, & F \text{ is convex,} \\ \frac{1}{2}\Theta_F^2 |\gamma^{k+1}|(c_2)\|\gamma^{k+1}\|_{\mathscr{M}}, & \text{otherwise.} \end{cases}$$

Proof. That (5.16a) holds as stated, as an immediate consequence of (3.8b).

If F is convex, $B_F(\check{\mu}^k, \mu^k) \geq 0$ by the definitions of the Bregman divergence and the convex subdifferential. Otherwise, letting $\Delta := (\pi^1_\# - \pi^0_\#)\gamma^{k+1}$, we estimate with (3.8a) that

$$\begin{aligned} -B_F(\check{\mu}^k, \mu^k) &= F(\check{\mu}^k) - F(\mu) + \langle F'(\check{\mu}^k) | -\Delta\rangle = \int_0^1 \langle F'(\mu^k + (1-t)\Delta) - F'(\mu^k + \Delta) | \Delta\rangle \, dt \\ &\leq \int_0^1 t\Theta_F^2 |\gamma^{k+1}|(c_2)\|\gamma^{k+1}\| \, dt = \frac{1}{2}\Theta_F^2 |\gamma^{k+1}|(c_2)\|\gamma^{k+1}\|. \end{aligned}$$

□

Example 5.12 (Curvature transport error control for $O(1/N)$ function value convergence). Assuming that F' is firmly transport Lipschitz in the sense (3.8) with $\|\cdot\|_* \leq M\|\cdot\|_{\mathscr{M}}$ for some $M \geq 0$, we get

$$\langle v^k - \check{v}^k | (\pi^2_\# - \pi^1_\#)\lambda^{k+1}\rangle \leq \sqrt{\Theta_F^2 |\gamma^{k+1}|(c_2)\|\gamma^{k+1}\|}_{\mathscr{M}} \cdot \|(\pi^2_\# - \pi^1_\#)\gamma^{k+1}\|_*.$$

By the definition of $\Lambda_{01}(\gamma^{k+1})$ in (4.5), we have $(\pi^2_\# - \pi^1_\#)\lambda = (\pi^2_\# - \pi^1_\#)\lambda_0$ for some $\lambda_0 \in \mathscr{M}(\Omega^3)$ with $\pi^{0,1}_\#\lambda_0 = \gamma^{k+1}$. Thus, $\|(\pi^2_\# - \pi^1_\#)\lambda^{k+1}\|_* \leq 2M\|\gamma^{k+1}\|_{\mathscr{M}}$. Taking $\ell_F = 0$, (5.15a) therefore holds if we ensure $\|\gamma^{k+1}\|\sqrt{|\gamma^{k+1}|(c_2)} \leq C_\gamma \varepsilon^{k+1}$ for some $C_\gamma \geq 0$. Practically, we expect $|\gamma^{k+1}|(c_2)$ to become small as the algorithm advances.

The next example motivates the split (5.13).

Example 5.13 (Convexity transport error control). Let F be twice Fréchet pre-differentiable in the sense that $F''(\zeta) \in \mathbb{L}(\mathscr{M}(\Omega); C_0(\Omega))$ for all $\zeta \in \mathscr{M}(\Omega)$. Then $\omega^{k+1} - \tau[v^{k+1} - \check{v}^k] = M\Delta^k$. for

$$\Delta^k := \mu^{k+1} - \mu^k - (\pi^1_\# - \pi^0_\#)\gamma^{k+1}, \quad M := \mathscr{D} - \tau F''(\zeta^k), \quad \text{and some} \quad \zeta^k \in [\mu^k, \check{\mu}^k].$$

Define $\nabla M(y) \in \mathbb{L}(\mathscr{M}(\Omega); C_0(\Omega; \mathbb{R}^n))$ by $\nabla M(y)\mu := \nabla[M\mu](y)$. If ∇M is L-Lipschitz, then $\nabla[M\Delta^k]$ is $L\|\Delta^k\|_{\mathscr{M}}$-Lipschitz, and we get

$$r^{k+1}_{\mathrm{con}} \leq \int_\Omega \sup_{z\in\Omega} \frac{L\|\Delta^k\|_{\mathscr{M}}}{2}\|z - y\|^2 \, d|\pi^1_\#\gamma^{k+1}|(y) \leq \frac{L}{2}(\operatorname{diam}\Omega)^2\|\gamma^{k+1}\|_{\mathscr{M}}\|\Delta^k\|_{\mathscr{M}}.$$

Hence, to ensure (5.15a) or (5.16a), it suffices for some $c_{\mathrm{con}} > 0$ to bound

$$\|\gamma^{k+1}\|_{\mathscr{M}}\|\Delta^k\|_{\mathscr{M}} \leq c_{\mathrm{con}}\varepsilon^{k+1}.$$

This can be done a posteriori, after computing μ^{k+1}, by reducing $\gamma^{k+1} = \gamma^{k+1,j}$ as necessary, and recomputing μ^{k+1} with the new $\gamma^{k+1} = \gamma^{k+1,j+1}$. If μ^{k+1} can be shown to be bounded, and we ensure $\|\gamma^{k+1,j+1}\| \leq \rho\|\gamma^{k+1,j}\|$ for some $\rho \in (0,1)$, this adaptation procedure will eventually stop.

Example 5.14 (Lipschitz gradient of particle-to-wave operator based on convolution). Let $\mathscr{D}\mu = \rho * \mu$ for $\rho \in C_0^1(\Omega)$ with L-Lipschitz gradient. Then $\nabla[\mathscr{D}(y)]\mu = \nabla[\rho * \mu](y) = [\nabla\rho * \mu](y)$, so $\nabla\mathscr{D}$ is L-Lipschitz, as required by the previous example.

We return to the Lipschitz properties of $\nabla F''(\zeta^k)$ for the quadratic-affine F after the next remark.

Remark 5.15 (Curvature bounds). We can always satisfy the curvature bounds (4.2d) and (5.2a) by taking $\gamma^{k+1} = 0$. To satisfy these conditions with a non-zero transport, and to satisfy (5.2b), it is sufficient that, for some $\ell_F^{(1)}, \ell_F^{(2)} \geq 0$, we have

$$|B_{v^k}(x,y)| \leq \ell_F^{(1)} c_2(x,y) \quad \text{for} \quad \gamma\text{-a.e.}(x,y),$$

where $\gamma = \gamma^{k+1}$ or $\gamma = \pi_\#^{0,2}\lambda$. To show (4.2d), we also need

$$-B_F(\mu^k + \gamma^{k+1}, \mu^k) \leq \ell_F^{(2)} \int_{\Omega^2} c_2(x,y)\, d|\gamma^{k+1}|(x,y)$$

Lemma 5.11 shows the latter to hold with $\ell_F^{(2)} = 0$ when F is convex, and otherwise with $\ell_F^{(2)} = \frac{1}{2}\Theta_F^2 m_\gamma$. By the descent lemma, the former holds if ∇v^k is $\ell_F^{(2)}$-Lipschitz. The next example shows how this can be ensured for our typical data term.

Example 5.16 (Lipschitz properties of a simple convolution sensor grid). Let $F(\mu) = \frac{1}{2}\|A\mu - b\|^2$ with $A \in \mathbb{L}(\mathscr{M}(\Omega); Y)$ for a Hilbert space Y have a pre-adjoint $A_* \in \mathbb{L}(Y; C_0^1(\Omega))$. Then $\nabla[A_*A](y)\mu = \nabla[A_*A\mu](y) = [\nabla A_*](y)A\mu$, where $[\nabla A_*](y)h := \nabla[A_*h](y)$ for all $h \in Y$. In particular, if $Y = \mathbb{R}^m$, and $[A\mu]_i = \langle \psi | \mu \rangle$ as in Lemma 3.8, we have $A_*h(y) = \sum_{i=1}^m \psi_i(y)h_i$ and $[\nabla A_*](y)h = \sum_{i=1}^m \nabla\psi_i(y)h_i$. Hence, if each $\nabla\psi_i$ is $L_{\nabla\psi}$-Lipschitz,

$$\begin{aligned}\|\nabla A_*(x) - \nabla A_*(y)\| &= \sup_{\|h\|_2=1} \left\| \sum_{i=1}^m (\nabla\psi_i(x) - \nabla\psi_i(y))h_i \right\| \\ &\leq \sup_{\|h\|_2=1} \sum_{i=1}^m L_{\nabla\psi} \max\{\chi_{\operatorname{supp}\psi_i}(x), \chi_{\operatorname{supp}\psi_i}(y)\}|h_i| \leq \sqrt{2N_\psi} L_{\nabla\psi},\end{aligned}$$

where N_ψ, the maximum number of ψ_i supported at any single point, is as defined in Lemma 3.8. Thus, ∇A_* is $\sqrt{2N_\psi}L_{\nabla\psi}$-Lipschitz. If each ψ_i each L_ψ-Lipschitz, a similar calculation establishes that A_* is $\sqrt{2N_\psi}L_\psi$-Lipschitz. If each $\|\psi_i\|_\infty \leq M_\psi$, then by Jensen's inequality, also

$$\begin{aligned}\|A\| &= \sup_{\|\mu\|=1} \|A\mu\|_Y = \sup_{\|\mu\|=1} \sqrt{\sum_{i=1}^m \left(\int \psi_i(x)\, d\mu(x) \right)^2} \leq \sup_{\|\mu\|=1} \sqrt{\sum_{i=1}^m \|\mu\| \int \psi_i(x)^2\, d|\mu|(x)} \\ &\leq \sup_{\|\mu\|=1} \sqrt{N_\psi M_\psi^2 \|\mu\|^2} = \sqrt{N_\psi} M_\psi.\end{aligned}$$

Thus, $y \mapsto \nabla[A_*A](y)$ is $\sqrt{2}N_\psi M_\psi L_{\nabla\psi}$-Lipschitz, which is helpful with Example 5.13.

As $\nabla v^k(x) = \nabla A_*(x)[A\mu^k - b]$, we also deduce that

$$\nabla v^k \quad \text{is} \quad \sqrt{2N_\psi} L_{\nabla\psi} \|A\mu^k - b\|_Y\text{-Lipschitz.}$$

We do not necessarily know a bound on the final term, but if μ^k is bounded, as in Section 5.4, we can update it adaptively, or we can make an educated guess, such as μ^k being a better solution than

zero:

$$\frac{1}{2}\|A\mu^k - b\|_Y^2 \leq^? \frac{1}{2}\|b\|_Y^2 + G(0).$$

Thus, we could take $\ell_F^{(1)} = \sqrt{2N_\psi}L_{\nabla\psi}\sqrt{\|b\|_Y^2 + 2G(0)}$, and enforce (4.2d), (5.2a) and (5.2b) by reducing the transport γ^{k+1} if necessary, if this guess was wrong.

Finally, instead of directly enforcing (5.14c), we can often ensure it through the rest of (5.14).

Lemma 5.17. *Suppose* (5.14a) *holds with* ε^{k+1} *tightened to* $\bar{\varepsilon}^{k+1} \leq \min\{\varepsilon^{k+1}, c(\varepsilon^{k+1})^2/\|\gamma^{k+1}\|\}$ *for some* $c > 0$. *Also suppose that* $\nabla\tilde{\varepsilon}^{k+1}$ *is* $L/\|\gamma^{k+1}\|$*-Lipschitz for some* $L \geq 0$. *Then* (5.14c) *holds for any* $C' \geq 2(1 + \operatorname{diam}\Omega\sqrt{L/c})$.

Proof. We may assume that $\|\gamma^{k+1}\| > 0$, since otherwise there is nothing to prove. Set $g := \tilde{\varepsilon}^{k+1}$ and let $y \in \pi_\#^1 \operatorname{supp}\gamma^{k+1}$. Observe from (5.14a) that $g(y) \leq \bar{\varepsilon}^{k+1}$ and $g(y+h) \geq -\bar{\varepsilon}^{k+1}$ for any $h \in \mathbb{R}^n$ with $y + h \in \Omega$. Combining these bounds with the descent inequality yields

$$\langle \nabla g(y), -h\rangle = [g(y+h) - \langle \nabla g(y), h\rangle - g(y)] + g(y) - g(y+h) \leq \frac{L}{2\|\gamma^{k+1}\|}\|h\|^2 + 2\bar{\varepsilon}^{k+1}.$$

Taking $\|h\| = \zeta\varepsilon^{k+1}$ for some $c > 0$, we obtain

$$\|\nabla g(y)\| = \sup_{\|h\|=\zeta\varepsilon^{k+1}} \frac{\langle \nabla g(y), -h\rangle}{\|h\|} \leq \sup_{\|h\|=\zeta\varepsilon^{k+1}} \left(\frac{L}{2}\|h\| + \frac{2c(\varepsilon^{k+1})^2}{\|h\|}\right)\frac{1}{\|\gamma^{k+1}\|} = \left(\frac{L\zeta}{2} + \frac{2c}{\zeta}\right)\frac{\varepsilon^{k+1}}{\|\gamma^{k+1}\|}.$$

The right-hand side is minimised by $\zeta = 2\sqrt{c/L}$, yielding

$$\int \|\nabla g(y)\|\, d|\gamma^{k+1}|(x,y) \leq 2\sqrt{L/c}\varepsilon^{k+1}. \tag{5.18}$$

Thus, by the definition of the Bregman divergence, (5.14c) holds when $C' \geq 2(1 + \operatorname{diam}\Omega\sqrt{L/c})$. □

For Radon norm regularisation, we can remove the implicit requirement that ∇w^{k+1} be Lipschitz. Subject to bounds on μ^k and γ^{k+1}, the remaining Lipschitz requirement on $\nabla\check{v}^k$, follows from Example 5.16 for our typical example of a data term, while the Lipschitz requirement on ω^{k+1} holds, for example, when $E_{\mathscr{D}}$ with $\mathscr{D} = \rho*$ for a Lipschitz differentiable kernel ρ.

Corollary 5.18. *Let* $G = \alpha\|\cdot\|_{\mathscr{M}} + \delta_{\geq 0}$ *for some* $\alpha > 0$. *Suppose* (5.14a) *holds with* ε^{k+1} *tightened to* $\bar{\varepsilon}^{k+1} \leq \min\{\varepsilon^{k+1}, c(\varepsilon^{k+1})^2/\|\gamma^{k+1}\|\}$ *for some* $c > 0$. *Also suppose that* $\nabla[\tau\check{v}^k + \omega^{k+1}]$ *is* $L/\|\gamma^{k+1}\|$*-Lipschitz for some* $L \geq 0$, *and that* $\operatorname{supp}\pi_\#^1\gamma^{k+1} \subset \operatorname{supp}\mu^{k+1}$. *Then* (5.14c) *holds for any* $C' \geq 2(1 + \operatorname{diam}\Omega\sqrt{L/c})$.

Proof. Note that $\nabla w^{k+1}(y) = 0$ for $y \in \operatorname{supp}\mu^{k+1}$. Let $g = \tau[\check{v}^k + \alpha] + \omega^{k+1}$ and also observe that (5.14a) shows that

$$\text{on } \Omega \quad -\bar{\varepsilon}^{k+1} \leq g \leq \bar{\varepsilon}^{k+1} \quad \text{on } \operatorname{supp}\mu^{k+1}. \tag{5.19}$$

In particular, $g(y) \leq \bar{\varepsilon}^{k+1}$ and $g(y+h) \geq -\bar{\varepsilon}^{k+1}$ when $y \in \operatorname{supp}\pi_\#^1\gamma^{k+1} \subset \operatorname{supp}\mu^{k+1}$. Now we follow the proof of Lemma 5.17 until (5.18), where we use that $\nabla g(y) = \nabla\bar{\varepsilon}^{k+1}(y)$ for $y \in \operatorname{supp}\mu^{k+1}$ before continuing. □

For $G = \alpha\|\cdot\|_{\mathscr{M}}$, we can use a similar argument if $\operatorname{supp}\mu^{k+1}$ is finite, constructing, e.g., by partition of unity an explicit Lipschitz function $a \in \partial G(\mu^{k+1})$ with $a = \pm\alpha$ and $\nabla a = 0$ on $\operatorname{supp}(\mu^{k+1})^\pm$. The factor L will then tend to infinity as $\min\{\|x^+ - x^-\| \mid x^\pm \in \operatorname{supp}(\mu^{k+1})^\pm\}$ goes to zero.

Algorithm 3 Forward-backward for Radon norm regularisation of non-negative measures (μFB)

Require: Regularisation parameter $\alpha > 0$ and pre-differentiable $F : \mathcal{M}(\Omega) \to \mathbb{R}$, as well as a self-adjoint and positive semi-definite $\mathscr{D} \in \mathbb{L}(\mathcal{M}(\Omega); C_0^1(\Omega))$.

1: Choose tolerances $\{\varepsilon^{k+1}\}_{k\in\mathbb{N}} \subset (0,\infty)$ and subproblem fractional tolerance $\kappa \in (0,1)$.
2: Choose $C_{\mathrm{cur}}, C_{\mathrm{con}}, \ell_F, \ell_r > 0$ for which (5.2a), (5.2b) and (5.15a) (resp. (4.2d) and (5.16a)) might hold.
▷ *See Remark 5.15 and Example 5.12 (resp. Lemma 5.11 for mere subdifferential convergence).*
3: Choose step length parameters $\theta, \tau > 0$ satisfying Assumption 5.2 (ii) (resp. 4.5 (i)).
4: Pick an initial iterate $\mu^0 \in \mathscr{X}(\Omega)$, and set $\gamma^k = 0$.
5: **for** $k \in \mathbb{N}$ **do**
6: $v^k := F'(\mu^k)$
7: $\sum_{j=1}^m \gamma_j \delta_{x_j} := \mu^k$. ▷ *Decompose previous measure*
8: $y_j := x_j - \theta\tau \operatorname{sign} a_j \nabla v^k(x_j)$ for $j = 1, \dots, m$. ▷ *Satisfy (4.2b).*
9: $\gamma := \sum_{j=1}^m \gamma_j \delta_{(x_j, y_j)}$
10: **if** the following were not ensured on Line 2 **then**
11: Reduce the $|\gamma_j|$ to satisfy (5.2a), (5.2b) and (5.15a) (resp. (4.2d) and (5.16a)).
▷ *Curvature and curvature transport error control.*
12: **end if**
13: **repeat**
14: $\mu := \breve{\mu} := \mu^k + (\pi^1_\# - \pi^0_\#)\gamma$
15: $\breve{v} := F'(\breve{\mu})$
16: $\bar{\varepsilon}^{k+1} := \min\{\varepsilon^{k+1}, c(\varepsilon^{k+1})^2 / \|\gamma^{k+1}\|\}$. ▷ *Tighten tolerances for Corollary 5.18.*
17: $\mu := \textsc{insert\_and\_adjust}(\breve{\mu}, \breve{v}, \alpha, \tau, \kappa\bar{\varepsilon}^{k+1}, \mathscr{D})$. ▷ *Solve (5.14a) using Algorithm 1.*
18: For all $j = 1, \dots, m$ with $\mu(y_j) = 0$, set $\gamma_j = 0$. ▷ *Ensure $w^{k+1}(y) = 0$ on $\operatorname{supp}\gamma$*
19: Decrease the weights γ_j in γ to ensure (5.15b) (resp. (5.16b)) and $\|\gamma\| \to 0$ in this inner loop.
▷ *Convexity transport error control. See Example 5.13 with 5.14 and 5.16.*
20: **until** Lines 18 and 19 made no change
21: $\mu^{k+1} := \mu$ and $\gamma^{k+1} := \mu$ pruning zero-weight components.
22: **end for**

5.4 POSITIVITY-CONSTRAINED RADON-NORM REGULARISER WITH PARTICLE-TO-WAVE MARGINAL TERM

With now refine Example 4.1, taking

$$G(\mu) = \alpha\|\mu\|_{\mathcal{M}} + \delta_{\geq 0}(\mu) \quad \text{as well as} \quad E(\mu,\nu) = \frac{1}{2}\|\mu-\nu\|^2_{\mathscr{D}} \tag{5.20a}$$

for $\mathscr{D}$ defined for all $\mu \in \mathcal{M}(\Omega)$ following [35] by

$$\mathscr{D}\mu = \rho * \mu \quad \text{for a symmetric and positive semi-definite } 0 \not\equiv \rho \in C_0^1(\mathbb{R}^n) \cap L^2(\mathbb{R}^n). \tag{5.20b}$$

We recall that ρ is symmetric when $\rho(-x) = \rho(x)$, and positive semi-definite when the Fourier transform $\mathscr{F}[\rho] \geq 0$ [12, 35]. On a closed domain Ω, this guarantees that $\mathscr{D} \in \mathbb{L}(\mathcal{M}(\Omega); C_0^1(\mathbb{R}^n))$ and that it is self-adjoint and positive semi-definite [35, Lemmas 2.4 and 2.1]. Examples of kernels ρ are provided in [35, Section 3].

With Example 4.1 as the starting point, but aiming for the stronger convergence claims of Corollary 5.8 and Theorem 5.7, we need to satisfy the relevant conditions of Table 1. This results in Algorithm 3. It also indicates as the "resp." variant a precise version of Example 4.1. To prove convergence, we start with a bound for Algorithm 1.

Lemma 5.19. *The result μ of Algorithm 1 satisfies $\|\mu\|_{\mathcal{M}} \leq \kappa\varepsilon + \theta^{-2}\|\eta\|_\infty$.*

Proof. The mass of μ is determined by the finite-dimensional weight optimisation problem on Line 6 of Algorithm 1. By the definition of the subdifferential, and the accuracy condition, the inexact solution satisfies $f(\vec{\beta}) = f(\vec{\beta}) - f(0) \leq \inf_{g \in \partial f(\vec{\beta})} \langle g, \vec{\beta} \rangle \leq \kappa\varepsilon \|\vec{\beta}\|_1 / (1 + \|\vec{\beta}\|_1) \leq \kappa\varepsilon$. Hence, $\langle \vec{\eta}, \vec{\beta} \rangle + \theta \|\vec{\beta}\|_1 \leq \kappa\varepsilon$, from where Hölder's and Young's inequalities give $\frac{\theta}{2}\|\mu\|_{\mathcal{M}} = \frac{\theta}{2}\|\vec{\beta}\|_1 \leq \kappa\varepsilon + \frac{1}{2\theta}\|\vec{\eta}\|_\infty = \kappa\varepsilon + \frac{1}{2\theta}\|\eta\|_\infty$. □

Theorem 5.20. *With G and E as in (5.20), suppose Assumption 4.3 (iii) and (iv) and Assumption 5.2 (ii) hold, as well as*

(i) $\inf F =: i_F > -\infty$, *and*
(ii) F' *is Lipschitz,*
(iii) the families $\{\nabla F'(\mu) \mid \mu \in A\} \subset C_0(\Omega; \mathbb{R}^n)$ *are uniformly Lipschitz on bounded sets* A.

Let $\{(\mu^{k+1}, \gamma^{k+1})\}_{k \in \mathbb{N}}$ *be generated by Algorithm 3 for an initial* $\mu^1 \in \mathcal{M}(\Omega)$. *If we employ the "resp." variant of the algorithm, then the subdifferential convergence claim of Theorem 4.11 holds.*

If we employ the main variant of the algorithm, and F is, moreover, convex, then the ergodic function value convergence claim of Theorem 5.7 holds for any $\bar{\mu} \in \mathcal{M}(\Omega)$.

In addition to the previous assumptions, if on Line 2 or Line 11 we also guarantee (4.2d), and on Lines 11 and 19 we also guarantee $C_{\mathrm{con}}\varepsilon^{k+1} \geq (k+1)\check{r}_{\mathrm{con}}^{k+1}$ *and* $C_{\mathrm{cur}}\varepsilon^{k+1} \geq (k+1)\check{r}_{\mathrm{cur}}^{k+1}$ *(see Lemma 5.10), then the non-ergodic function value convergence claim of Corollary 5.8 holds.*

Proof. We first prove that the loop on Line 13 of Algorithm 3 terminates. Clearly, Line 18 can only make a change a finite number of times, so will not stop the loop from terminating. Moreover, since $\|\gamma\|$ is decreasing in the loop, $\check{\mu}^k$ stays bounded in the loop, and ensures due to (ii) that also $\eta := \tau\check{v}^k - \mathscr{D}\check{\mu}^k$ stays bounded. Lemma 5.19 shows that also $\|\mu\|_{\mathcal{M}}$ is bounded within the loop. Thus, (iii) guarantees that $B_{\omega - \tau[F'(\mu) - \check{v}]}$ is bounded between iterations of the loop. Now Line 19 guarantees $\|\gamma\| \to 0$ within the loop. Consequently, the condition in this step will eventually be satisfied, and the loop will terminate.

We then observe that Assumption 4.3 holds in their entirety by the choice of G and E. Likewise, when F is convex, Assumption 5.2 holds. By Lemma 5.5, so does Assumption 5.4 after we use Lemma 5.1 to construct for all $k \in \mathbb{N}$ some $\lambda^{k+1} \in \Lambda_{01}(\gamma^{k+1})$ with $\Lambda^+(\lambda^k)$ for k ≥ 1. Our claims now follow from Theorems 4.11 and 5.7 and Corollary 5.8 if we verify the conditions of Table 1:

Variables: We require the basic condition (4.2a) on the spaces of variables. There we have $\omega^{k+1} \in C_0^1(\Omega)$ by the choice of the kernel $\rho \in C_0^1(\mathbb{R}^n)$. We also need $w^{k+1} \in \partial G(\mu^{k+1}) \cap C_0'(\Omega, \pi^1_\# \gamma^{k+1})$. In Algorithm 3, w^{k+1} is implicitly[4] constructed on Line 17. It ensures (5.14a) for some $\tilde{w}^{k+1} \in \partial G(\mu^{k+1})$ with the tightened tolerance $\kappa\varepsilon^{k+1}$. Since for the factor $\kappa \in (0,1)$, and μ^{k+1} has finite support, we can always construct a smoothed $w^{k+1} \in \partial G(\mu^{k+1}) \cap C_0'(\Omega, \pi^1_\# \gamma^{k+1})$ that still satisfies (5.14a), i.e., (4.2c).

Transport: The transport condition (4.2b) is ensured by Lines 8 and 9.

Curvature: The curvature conditions (4.2d), (5.2a) and (5.2b), as needed by the different claims, are guaranteed by Lines 2 and 11.

Marginal: By [35, Lemma 4.2], Algorithm 1 used on Line 17 terminates and proves (5.14a), i.e., (4.2c) (with the tightened tolerances $\bar{\varepsilon}^{k+1}$). We next inductively verify the indirect conditions (5.14c) and (5.14b) with the remainder conditions.

Remainder: Depending on the claim, we need to verify (4.2e), (5.2c) or, (5.2d). We do this by induction. So suppose one of these conditions holds with (5.14c) and (5.14b) for $k = 0, \ldots, N-1$. Trivially, this is the case if $N = 0$, so let $N \geq 1$. Then Corollary 5.9 provides the bound $\|\mu^N\| \leq m_\mu$, independent of N. Due to the construction on Line 9, also $\|\gamma^N\| \leq m_\mu$. Thus, (5.14a), already verified above, implies

[4] For an explicit construction of $\tilde{w}^{k+1}$, which is only in $C_0(\Omega)$, see the proof of [35, Lemma 4.1]. Given this, to form $w^{k+1} \in \partial G(\mu^{k+1}) \cap C_0'(\Omega, \pi^1_\# \gamma^{k+1})$, we can take small balls around the points of $\pi^1_\# \gamma^{k+1}$, form a smooth partition of unity, and use it to glue indicator functions to $\tilde{w}^{k+1}$.

(5.14b) for $k = N$. The relevant conditions (5.16a) and (5.16b) or (5.15a) and (5.15b) for $k = N$ are ensured by Lines 11 and 19, as are $C_{\mathrm{con}}\varepsilon^{k+1} \geq (k+1)\check{r}_{\mathrm{con}}^{k+1}$ and $C_{\mathrm{cur}}\varepsilon^{k+1} \geq (k+1)\check{r}_{\mathrm{cur}}^{k+1}$ for the non-ergodic claim. Therefore, to use Lemma 5.10 to verify (5.2c), it remains to verify (5.14c) for $k = N$. Indeed, it follows from (iii), that $\nabla\check{v}^{N-1} = \nabla F'(\mu^{N-1} + (\pi_\#^1 - \pi_\#^0)\gamma^N)$ is Lipschitz with some factor $L_{\nabla F'}$ independent of N. Likewise $\omega^N = \mathscr{D}(\mu^N - \mu^{N-1} - (\pi_\#^1 - \pi_\#^0)\gamma^N)$ is Lipschitz with some factor $L_{\mathscr{D}}$ independent of N. Line 18 guarantees $\operatorname{supp}\pi_\#^1\gamma^N \subset \operatorname{supp}\mu^N$. Thus, Corollary 5.18 verifies (5.14c) through (5.14a) holding with the tightened tolerances $\bar{\varepsilon}^{k+1}$. Now Lemma 5.10 proves for $k = N$ the relevant one of (4.2e), (5.2c), or (5.2b). □

Remark 5.21. Replacing Line 17 by Algorithm 2, the subdifferential convergence claim can be verified similarly for the Radon-squared proximal term.

Example 5.22. If $F(\mu) = \frac{1}{2}\|A\mu - b\|^2$ for some $A \in \mathbb{L}(\mathscr{M}(\Omega); Y)$ and a Hilbert space Y, both (i) of Theorem 5.20 obviously holds. If A has the structure of Example 5.16, then also (iii) due to the Lipschitz estimates of ∇A_* therein.

6 EXTENSIONS

We now sketch the extension of the work in Sections 4 and 5 to product spaces and primal-dual methods. We follow the approach of [19, Section 4] in treating the optimality conditions

$$0 \in H(u) := \bar{F}(u) + \bar{G}(u) + \bar{\Xi}u, \tag{6.1}$$

where $\bar{F} : U \to \mathbb{R}$ is convex and Fréchet differentiable on a normed space U, $\bar{G} : U \to \overline{\mathbb{R}}$ is convex, proper and lower semicontinuous, and $\bar{\Xi} \in \mathbb{L}(U; U_*)$ for a predual space U_* of U, is skew-symmetric: it has the pre-adjoint $\bar{\Xi}_* = -\bar{\Xi}$. (In [19], F is not required to be convex, but here, for simplicity, we make that restriction.)

We consider the product structure $U = \mathscr{M}(\Omega) \times Q$, which may or may not agree with a primal-dual product structure $U = Z \times Y$. We assume the normed space Q to be equipped with $\mathscr{Q} \in \mathbb{L}(Q; Q_*)$ (where Q_* is either a predual space or the dual space) and work with the norm $\|\cdot\|_{\mathscr{Q}}$. Moreover, we assume that $\bar{G}$ is separable:

$$\bar{G}(u) = G(\mu) + J(q) \quad \text{where} \quad u = (\mu, q) \in U. \tag{6.2}$$

We first formulate the general algorithm in Section 6.1, and then sketch its convergence in Section 6.2 through an extension of Lemma 5.6. In Section 6.3 we verify the abstract convergence conditions for a more specific primal-dual method.

6.1 ALGORITHM

With $U = \mathscr{M}(\Omega) \times Q$ and $U_* = C_0(\Omega) \times Q_*$ as above, we write $P_{\mathscr{M}_*} \in \mathbb{L}(U_*; C_0(\Omega))$ and $P_{Q_*} \in \mathbb{L}(U_*; Q_*)$ for the projection operators from U_* to $C_0(\Omega)$ and Q_*. Then $(P_{\mathscr{M}_*})^* \in \mathbb{L}(\mathscr{M}; U)$ with $(P_{\mathscr{M}_*})^*\mu = (\mu, 0)$. We also extend the curvature as

$$\mathscr{K}_{\bar{F}}(u, \gamma) = \int_{\Omega^2} B_{\bar{F}'(u)}(x, y) d\gamma(x, y) \quad \text{for} \quad u = (\mu, q) \in U.$$

Algorithm 4 General forward-backward type algorithm with a measure subspace

Require: Setting of Section 6.

1: Follow Algorithm 3 with the following changes:
2: On Lines 6 and 15, replace F' by $P_{\mathscr{M}_*}[\bar{F}'(\cdot, q^k) + \bar{\Xi}(\cdot, q^k)]$.
3: Before Line 6 or after Line 21, as necessary, find q^{k+1} by solving

$$0 \in \tau[P_{Q_*}\bar{F}'(\check{v}^k, q^k) + \partial J(q^{k+1}) + P_{Q_*}\bar{\Xi}(\mu^{k+1}, q^{k+1})] + P_{Q_*}\mathscr{U}(\mu^{k+1} - \mu^k, q^{k+1} - q^k)$$

for the final value of $\check{v}^k = \check{v}$ from the loop on Line 13.

We assume to be given a preconditioning operator $\mathscr{U} \in \mathbb{L}(U; U_*)$ that satisfies

$$P_{\mathscr{M}_*}[\tau\bar{\Xi} + \mathscr{U}] = 0 \quad \text{and} \quad P_{\mathscr{M}_*}\mathscr{U}(P_{\mathscr{M}_*})^* = 0. \tag{6.3}$$

This condition prevents $\mathscr{U}$ from operating on $\mathscr{M}(\Omega)$, where steps are instead determined by $V_{c_2,E}$. With $u^k = (\mu^k, q^k) \in \mathscr{M}(\Omega) \times Q$, to update μ^{k+1} and γ^{k+1}, we solve the conditions of Table 1 with the redefinitions

$$v^k := P_{\mathscr{M}_*}[\bar{F}'(\mu^k, q^k) + \bar{\Xi}(\mu^k, q^k)], \quad \text{and} \quad \check{v}^k := P_{\mathscr{M}_*}[\bar{F}'(\check{\mu}^k, q^k) + \bar{\Xi}(\check{\mu}^k, q^k)], \tag{6.4a}$$

also replacing $\mathscr{K}_F(\mu^k, \gamma)$ by $\mathscr{K}_{\bar{F}}(u^k, \gamma)$. For updating q^{k+1}, we solve

$$0 \in \tau[P_{Q_*}F'(\check{\mu}^k, q^k) + \partial J(q^{k+1}) + P_{Q_*}\bar{\Xi}(\mu^{k+1}, q^{k+1})] + P_{Q_*}\mathscr{U}(\mu^{k+1} - \mu^k, q^{k+1} - q^k). \tag{6.4b}$$

For $G(\mu) = \alpha\|\cdot\|_{\mathscr{M}} + \delta_{\geq 0}(\mu)$ and $E = E_{\mathscr{D}}$, this results in Algorithm 4.

Example 6.1 (Sliding primal-dual proximal splitting). With Z and Y Hilbert spaces, for convex and Fréchet differentiable $F : \mathscr{M}(\Omega) \times Z \to \mathbb{R}$, convex, proper, and lower semicontinuous $G : \mathscr{M}(\Omega) \to \overline{\mathbb{R}}$, $R : Z \to \overline{\mathbb{R}}$ and $H : Y \to \overline{\mathbb{R}}$, as well as $K = (K_\mu, K_z) \in \mathbb{L}(\mathscr{M}(\Omega) \times Z; Y)$, consider the problem

$$\begin{aligned}\min_{\mu \in \mathscr{M}(\Omega),\, z \in Z} \; & F(\mu, z) + G(\mu) + R(z) + H(K(\mu, z)) \\ &= \min_{\mu \in \mathscr{M}(\Omega),\, z \in Z} \sup_{y \in Y} F(\mu, z) + G(\mu) + R(z) + \langle K(\mu, z) | y \rangle - H^*(y).\end{aligned}$$

For $U = \mathscr{M}(\Omega) \times Z \times Y$, $Q = Z \times Y$, and $u = (\mu, z, y)$, we take

$$\bar{G}(u) = G(\mu) + R(z) + H^*(u), \quad \bar{F}(u) = F(\mu, z), \quad \text{and} \quad \bar{\Xi} = \begin{pmatrix} 0 & K^* \\ -K & 0 \end{pmatrix}.$$

For step lengths $\sigma_p, \sigma_d > 0$ that satisfy $\sigma_p\sigma_d\|K\|^2 \leq 1$, take also

$$\mathscr{U} = \tau \begin{pmatrix} 0 & 0 & -K_\mu^* \\ 0 & \sigma_p^{-1}\,\mathrm{Id} & -K_z^* \\ -K_\mu & -K_z & \sigma_d^{-1}\,\mathrm{Id} \end{pmatrix} \quad \text{so that} \quad \tau\Xi + \mathscr{U} = \tau \begin{pmatrix} 0 & 0 & 0 \\ 0 & \sigma_p^{-1}\,\mathrm{Id} & 0 \\ -2K_\mu & -2K_z & \sigma_d^{-1}\,\mathrm{Id}\,. \end{pmatrix}.$$

Clearly (6.3) holds, as does the separability (6.2) with $J(z, y) = R(z) + H^*(y)$.

Denoting the Fréchet derivative of F with respect to μ by $F^{(\mu)}$, (6.4a) now sets

$$v^k = F^{(\mu)}(\mu^k, z^k) + K_\mu^* y^k \quad \text{and} \quad \check{v}^k = F^{(\mu)}(\check{\mu}^k, z^{k+1}) + K_\mu^* y^k, \tag{6.5}$$

while (6.4b) is a single step of standard primal-dual proximal splitting,

$$\begin{cases} z^{k+1} := \operatorname{prox}_{\sigma_p R}(z^k - \sigma_p \nabla_z F(\breve{\mu}^k, z^k) - \sigma_p K_z^* y^k), \\ y^{k+1} := \operatorname{prox}_{\sigma_d H^*}(y^k + \sigma_d K(2\mu^{k+1} - \mu^k, 2z^{k+1} - z^k)). \end{cases}$$

When $G(\mu) = \alpha\|\cdot\|_{\mathscr{M}} + \delta_{\geq 0}(\mu)$, this is also Line 3 of Algorithm 4.

We will return to step length parameter choices in Lemma 6.7.

The above primal-dual method only uses the transported measure $\breve{\mu}^k$ in the smooth component F. Nonsmooth data terms, such as the impulse noise experiments of [35], do not benefit from transport.

6.2 CONVERGENCE

We now sketch extensions of Lemma 5.6 and Theorem 6.5. For this we define the *Lagrangian gap functional*

$$\mathscr{G}(u; \bar{u}) := [\bar{F} + \bar{G}](x) - [\bar{F} + \bar{G}](\bar{u}) - \langle \bar{\Xi} u | \bar{u} \rangle_{U^*,U} \quad (u, \bar{u} \in U).$$

For $\bar{u}$ a solution to (6.1), we have $\mathscr{G}(u; \bar{u}) \geq 0$ [17, Lemma 11.1][5]. We do not here consider an extension of Theorem 4.11, as our focus is on primal-dual methods, and we, and as far as we know, nobody, knows how to prove such a result for them.[6]

For $u = (\mu, q)$ and $\tilde{u} = (\nu, \tilde{q})$, writing $\breve{u} := (\mu + (\pi_\#^1 - \pi_\#^0)\gamma, q)$, we extend (5.3) and (5.4) as

$$V_k^{i,2}(u, \tilde{u}; \lambda) := \bar{V}^{i,2}_{\theta^{-1}c_2,E}(\mu, \nu; \lambda) + \frac{1}{2}\|u - \tilde{u}\|_{\mathscr{U}}^2 - \tau \mathscr{K}_{\bar{F}}(u, \pi_\#^{i,2}\lambda) \quad \text{and} \tag{6.6a}$$

$$V_k^{0,1}(u, \tilde{u}; \lambda) := \bar{V}^{0,1}_{\theta^{-1}c_2,E}(\mu, \nu; \lambda) + \frac{1}{2}\|u - \tilde{u}\|_{\mathscr{U}}^2 - \tau[B_{\bar{F}}(\breve{u}, \tilde{u}) + \mathscr{K}_{\bar{F}}(u, \pi_\#^{0,1}\lambda) + \ell_r \pi_\#^{0,1}|\lambda|(c_2)]. \tag{6.6b}$$

We collect our assumptions in the following adaptation of Assumptions 4.3, 5.2 and 5.4. Parts (iv) and (v) with (5.2a) and (5.2b) ensure bounds analogous to Lemma 5.3 on $V_k^{i,2}$ and $V_k^{0,1}$.

Assumption 6.2. With $U = \mathscr{M}(\Omega) \times Q$ for a normed space Q, we assume that

(i) $E : \mathscr{M}(\Omega) \times \mathscr{M}(\Omega) \to [0, \infty]$ satisfies (3.2) with $E(\nu, \nu) = 0$ for all $\nu \in \mathscr{M}(\Omega)$.
(ii) $\bar{G} : U \to \overline{\mathbb{R}}$ is convex, proper, and lower semicontinuous and satisfies the separability (6.2).
(iii) $\mathscr{U} \in \mathbb{L}(U; U_*)$ is positive semi-definite and self-adjoint, $\Xi \in \mathbb{L}(U; U_*)$ is skew-symmetric, and they satisfy (6.3).
(iv) $\bar{F} : U \to \mathbb{R}$ is convex and pre-differentiable, $P_{\mathscr{M}_*} F'(u) \in C_0^1(\Omega)$ for all $u \in U$. For some $\ell, L \geq 0$, for all $\mu, \nu \in \mathscr{M}(\Omega)$; $q, p \in Q$, for all $\gamma \in \mathscr{M}(\Omega^2)$, we have

$$V_{\ell c_2, LE}(\mu, \nu; \gamma) + \frac{1}{2\tau}\|(\mu, q) - (\nu, p)\|_{\mathscr{U}}^2 \geq B_{\bar{F}}((\mu + (\pi_\#^1 - \pi_\#^0)\gamma, q), (\nu, p)).$$

(v) The step lengths $\tau, \theta > 0$, the parameters ℓ and L, and the parameters ℓ_F and ℓ_r from (4.2) and (5.2), satisfy $\tau L \leq 1$ and $\theta\tau[\ell + \ell_r + \ell_F] \leq 1$.
(vi) The tolerances $\{\varepsilon^{k+1}\}_{k\in\mathbb{N}}$ satisfy $\lim_{N\to\infty} \frac{1}{N}\sum_{k=0}^{N-1} \varepsilon^{k+1} = 0$.
(vii) For all $k \in \mathbb{N}$, for given $u^{k+1} \in \mathscr{M}(\Omega)$, $\gamma^{k+1} \in \mathscr{M}(\Omega^2)$, and $\lambda^{k+1} \in \Lambda_{01}(\gamma^{k+1})$, the choice of $\gamma^{k+2} \in \mathscr{M}(\Omega^2)$ and $\lambda^{k+2} \in \Lambda_{01}(\gamma^{k+2}) \cap \Lambda^+(\lambda^{k+1})$ satisfies $V_{k+1}^{1,2}(u^{k+1}, \bar{u}; \lambda^{k+1}) \geq V_{k+2}^{0,2}(u^{k+1}, \bar{u}; \lambda^{k+2})$ for all $\bar{u} \in U$.

Remark 6.3 (No Radon marginal term). Due to the requirement that E satisfy (3.2), the convergence results of this section do not apply to $E = E_{\mathscr{D}}$. Forgoing this requirement, we could extend Theorem 4.11 to product spaces, but would not be able to treat primal-dual methods, which are our focus here.

[5] The lemma is written in Hilbert spaces, but the proof remains unchanged in general normed spaces.

[6] For an incomplete result, see [19].

The next result extends the first part of Lemma 5.6.

Lemma 6.4. *Suppose Assumption 6.2 (i) to (iv) hold, and that* $k \in \mathbb{N}$ *and* $(\mu^k, q^k, \gamma^{k+1}) \in \mathcal{M}(\Omega) \times Q \times \mathcal{M}(\Omega^2)$ *are given. If* (4.2a)–(4.2b), *and* (6.4) *hold, then for any* $u = (\mu, q) \in \mathcal{M}(\Omega) \times Q$, *and any* $\lambda \in \Lambda_{01}(\gamma^{k+1})$,

$$\tau \mathcal{G}(u^{k+1}; u) \geq V_k^{1,2}(u^{k+1}, u; \lambda) - V_k^{0,2}(u^k, u; \lambda) + V_k^{0,1}(u^k, u^{k+1}; \lambda) - r^{k+1}(\lambda, \mu). \tag{6.7}$$

Proof. Let $p^{k+1} \in \partial J(q^{k+1})$ be the element for which (6.4b) is satisfied. Then $(w^{k+1}, p^{k+1}) \in \partial \bar{G}(u^{k+1})$. Since $\langle \bar{\Xi} u^{k+1} | u^{k+1} \rangle = 0$, minding Assumption 6.2 (ii) and (iv), using the subdifferentiability of G and the Bregman three-point identity of Lemma A.1 on $\bar{F}$, we estimate

$$\begin{aligned} 0 \geq \mathcal{G}(u^{k+1}; u) &+ \langle P_{Q_*} \bar{F}'(\check{u}^k) + p^{k+1} + P_{Q_*} \bar{\Xi} u^{k+1} | q - q^{k+1} \rangle_{Q_*, Q} \\ &+ \langle P_{\mathcal{M}_*} F'(\check{u}^k) + w^{k+1} + P_{\mathcal{M}_*} \bar{\Xi} u^{k+1} | \mu - \mu^{k+1} \rangle_{C_0(\Omega), \mathcal{M}(\Omega)} + B_{\bar{F}}(\check{u}^k, u) - B_{\bar{F}}(\check{u}^k, u^{k+1}). \end{aligned} \tag{6.8}$$

Here $B_{\bar{F}}(\check{u}^k, u) \geq 0$ by the convexity of $\bar{F}$. Due to (6.3) and (6.4a), we have

$$\begin{aligned} \tau P_{\mathcal{M}_*}[F'(\check{u}^k) + \bar{\Xi} u^{k+1}] = P_{\mathcal{M}_*}[\tau F'(\check{u}^k) + \tau \bar{\Xi} \check{u}^k - \mathcal{U}(u^{k+1} - \check{u}^k)] &= \tau \check{v}^k - P_{\mathcal{M}_*} \mathcal{U}(u^{k+1} - \check{u}^k) \\ &= \tau \check{v}^k - P_{\mathcal{M}_*} \mathcal{U}(u^{k+1} - u^k). \end{aligned} \tag{6.9}$$

Likewise, by (6.4b),

$$\tau[P_{Q_*} \bar{F}'(\check{u}^k) + p^{k+1} + P_{Q_*} \bar{\Xi} u^{k+1}] = -P_{Q_*} \mathcal{U}(u^{k+1} - u^k). \tag{6.10}$$

Multiplying (6.8) by τ and using (6.9) and (6.10), we now get

$$0 \geq \tau \mathcal{G}(u^{k+1}; u) + \tau \langle \check{v}^k + w^{k+1} | \mu - \mu^{k+1} \rangle_{C_0(\Omega), \mathcal{M}(\Omega)} - \langle \mathcal{U}(u^{k+1} - u^k) | u - u^{k+1} \rangle_{X_*, X} - B_{\tau \bar{F}}(\check{u}^k, u^{k+1}). \tag{6.11}$$

Using $\lambda \in \Lambda_{01}(\gamma^{k+1})$, (4.2b) and Assumption 6.2 (i) with Theorem 3.1, we prove, exactly as (5.7) in the proof of Lemma 5.6, that

$$\begin{aligned} 0 \geq \bar{V}^{1,2}_{\theta^{-1} c_2, E}(\mu^{k+1}, \mu; \lambda) &- \bar{V}^{0,2}_{\theta^{-1} c_2, E}(\mu^k, \mu; \lambda) + \bar{V}^{0,1}_{\theta^{-1} c_2, E}(\mu^k, \mu^{k+1}; \lambda) \\ &+ \tau[\mathcal{K}_{\bar{F}}(u^k, \pi_\#^{0,2} \lambda) - \mathcal{K}_{\bar{F}}(u^k, \pi_\#^{0,1} \lambda) - \mathcal{K}_{\bar{F}}(u^{k+1}, \pi_\#^{1,2} \lambda)] \\ &- \tau \langle \check{v}^k + w^{k+1} | \mu - \mu^{k+1} \rangle - \tau \ell_r \pi_\#^{0,1} |\lambda|(c_2) - r^{k+1}(\mu, \lambda). \end{aligned}$$

Using this, the Pythagoras' identity in (6.11)[7], and the definitions (6.6) in (6.11), we obtain the claim. □

Now, similarly to Theorem 5.7, we obtain an ergodic convergence result.

Theorem 6.5 (Ergodic gap convergence). *Suppose Assumption 6.2 holds. Generate* $\{u^{k+1} = (\mu^{k+1}, \gamma^{k+1})\}_{k \in \mathbb{N}}$ *through the satisfaction of* (6.4) *with the "Ergodic* $O(1/N)$*" conditions of Table 1. for an initial* $u^0 = (\mu^0, q^0) \in \mathcal{M}(\Omega) \times Q$. *Pick* $N \in \mathbb{N}$ *and* $\lambda^1 \in \Lambda_{01}(\gamma^1)$. *Then for* $\tilde{u}^N := \frac{1}{N} \sum_{k=0}^{N-1} u^{k+1}$ *and any* $\bar{u} = (\bar{\mu}, \bar{q})$, *we have*

$$\mathcal{G}(\tilde{u}^N; \bar{u}) \leq \frac{1}{\tau N} V_0^{0,2}(u^0, \bar{u}; \lambda^1) + \frac{C_{\mathcal{K}}}{N} + \frac{C_{\bar{\mu}}}{\tau N} \sum_{k=0}^{N-1} \varepsilon^{k+1}. \tag{6.12}$$

In particular, if $\bar{x}$ *solves* (6.1), *then* $\mathcal{G}(\tilde{u}^N; \bar{u}) \to 0$ *at the rate* $O(1/N)$.

Sketch of proof. The proof is analogous to Theorem 5.7. We apply Lemma 6.4, bounding the distance terms similarly to Lemma 5.3 with Assumption 6.2 (iv) and (v), (4.2d) and (5.2b). Therefore, using (5.2c), the rebalancing Assumption 6.2 (vii), and telescoping, we obtain (6.12). For the final claim, we use (6.12) and the fact that $\mathcal{G}(\,\cdot\,; \bar{u}) \geq 0$ for $\bar{u}$ solving (6.1). □

Remark 6.6. If $\bar{\Xi} = 0$, i.e., we consider forward-backward splitting in product spaces, we can also extend Corollary 5.8 to show the non-ergodic convergence $\mathcal{G}(u^N; \bar{u}) \to 0$.

[7] See [35] or [19] for the version with an operator in a normed space.

6.3 PRIMAL-DUAL STEP LENGTHS

Let us return to Example 6.1 and verify Assumption 6.2. The conditions (ii) and (iii) hold by the assumptions and constructions in Example 6.1, (i) and (vii) depend on the specific choice of the energy E for the measure space updates—they hold by Example 4.4 and the proof of Lemma 5.5 if $E(\mu,\nu) = \frac{1}{2}\|\mu-\nu\|_{\mathscr{D}}$ for a self-adjoint $\mathscr{D} \in \mathbb{L}(\mathscr{M}(\Omega); C_0^1(\Omega))$. The condition (vi) merely controls the tolerances. So it remains to treat the step length and smoothness conditions (iv) and (v). We prove these through a simpler smoothness condition, which, with the help of Young's inequality, can be verified for $F(\mu,z) = \frac{1}{2}\|A\mu + z - b\|^2$, similarly to Example 4.4.

Lemma 6.7. *Assume the setup of Example 6.1 with $\bar{F} : U \to \mathbb{R}$ convex and pre-differentiable with $P_{\mathscr{M}_*}F'(u) \in C_0^1(\Omega)$ for all $u \in U$. Take $E(\mu,\nu) = \frac{1}{2}\|\mu-\nu\|_{\mathscr{D}}^2$ for a self-adjoint positive semi-definite $\mathscr{D} \in \mathbb{L}(\mathscr{M}(\Omega); C_0(\Omega))$, satisfying $K_\mu^* K_\mu \leq M\mathscr{D}$ for some $M > 0$, Suppose for some $\ell_0, L_0, L_z \geq 0$ that*

$$B_F((\mu + (\pi_\#^1 - \pi_\#^0)\gamma, z), (\nu, w)) \leq V_{\ell_0 c_2, L_0 E}(\mu, \nu; \gamma) + \frac{L_z}{2}\|z-w\|_Z^2$$

for all $\mu, \nu \in \mathscr{M}(\Omega)$; $\gamma \in \mathscr{M}(\Omega^2)$; and $(z,w) \in Z$. Let $\hat{\ell} := \ell_0 + \ell_F + \ell_r$. Then Assumption 6.2 (iv) and (v) hold when

$$0 < \beta := \sigma_p\sigma_d\|K_z\|^2/(1-\sigma_p L_z) < 1, \qquad \text{(determines } \sigma_d, \sigma_p > 0\text{)} \tag{6.13a}$$

$$\tau\sigma_d M < (1-\tau L_0)(1-\beta) \qquad \text{(determines } \tau > 0\text{)} \tag{6.13b}$$

$$\frac{\tau\theta\sigma_d(1-\tau L_0)}{(1-\beta)(1-\tau L_0) - \tau\sigma_d M}\|K_\mu(\pi_\#^1 - \pi_\#^0)\gamma\|_Y^2 \leq 2(1-\tau\theta\hat{\ell})\int c_2\, d\gamma \qquad \text{(determines } \theta > 0\text{).} \tag{6.13c}$$

Proof. By several applications of Young's inequality (compare [17, Lemma 9.12]), for any $x = (\mu, z, y)$, whenever $0 < \beta := \sigma_d\sigma_p\|K_z\|^2/(1-\sigma_p L_z) < 1$, we have

$$\begin{aligned}\frac{1}{\tau}\|u\|_{\mathscr{U}}^2 - L_z\|z\|_Z^2 &\geq (\sigma_p^{-1} - L_z)\|z\|_Z^2 - \frac{\sigma_d}{1-\beta}\|K_\mu\tilde{\mu}\|_Y^2 - \frac{1-\beta}{\sigma_d}\|y\|_Y^2 + \frac{1}{\sigma_d}\|y\|_Y^2 - \frac{\sigma_d}{\beta}\|K_z z\|_Y^2 - \frac{\beta}{\sigma_d}\|y\|_Y^2 \\ &= (\sigma_p^{-1} - L_z)\|z\|^2 - \frac{\sigma_d}{\beta}\|K_z z\|_Y^2 - \frac{\sigma_d}{1-\beta}\|K_\mu\mu\|_Y^2 \geq -\frac{\sigma_d}{1-\beta}\|K_\mu\mu\|_Y^2.\end{aligned}$$

For $\mu = \mu_1 - \mu_0$, for any $\alpha > 0$, we have

$$\|K_\mu\mu\|_Y^2 \leq (1+\alpha)\|K_z(\mu_1 - \mu_0 - (\pi_\#^1 - \pi_\#^0)\gamma)\|_Y^2 + (1+\alpha^{-1})\|K_z(\pi_\#^1 - \pi_\#^0)\gamma\|_Y^2.$$

Applying these two inequalities in the smoothness inequality in Assumption 6.2 (iv), we see the latter to hold if, for all $\mu, \nu \in \mathscr{M}(\Omega)$ and $\gamma \in \mathscr{M}(\Omega^2)$, we have by Young's inequality that

$$\begin{aligned}\int_{\Omega^2} (\ell - \ell_0)c_2(x,y)\, d|\gamma|(x,y) + (L-L_0)E(\mu - \pi_\#^0\gamma, \nu - \pi_\#^1\gamma) \\ \geq \frac{\sigma_d(1+\alpha)}{2(1-\beta)}\|K_z(\nu - \mu - (\pi_\#^1 - \pi_\#^0)\gamma)\|_Y^2 + \frac{\sigma_d(1+\alpha^{-1})}{2(1-\beta)}\|K_z(\pi_\#^1 - \pi_\#^0)\gamma\|_Y^2.\end{aligned} \tag{6.14}$$

With $E(\mu,\nu) = \frac{1}{2}\|\mu-\nu\|_{\mathscr{D}}^2$ and $K_\mu^* K_\mu \leq M\mathscr{D}$, we can take $\alpha = (1-\beta)(L-L_0)/(\sigma_d M) - 1$ if this is positive. We take maximal $L = 1/\tau$ and $\ell = 1/(\theta\tau) - (\ell_F + \ell_r)$ that satisfy Assumption 6.2 (v). Then $\alpha = (1-\beta)(1-\tau L_0)/(\tau\sigma_d M) - 1$. Now (6.14) and our bounds on α and β hold by (6.13). □

Remark 6.8. The conditions (6.13a) and (6.13b) can be rewritten as

$$\tau\sigma_d M\frac{1-\sigma_p L_z}{1-\tau L} + \sigma_p L_z + \sigma_p\sigma_d\|K_z\|^2 < 1 \quad \text{with} \quad 1 > \sigma_p L_z \quad \text{and} \quad 1 > \tau L.$$

The first one is the standard step length condition for the PDPS on (z, y) alone, together with the first term that controls the marginal step length $\tau > 0$ for μ. If there is no z component (and hence no σ_p), the condition is simply $\tau\sigma_d M + \tau L < 1$, and we can take $\beta = 0$ in the final (6.13c), which determines the transport step length. We can then, in appropriate cases, use the transport smoothness estimates of Lemma 3.8, together with the boundedness of iterates and γ^{k+1} (compare Corollary 5.9) to verify (6.13c). If $K_\mu = 0$, as in our numerical experiments, it merely requires $1 \geq \theta\tau[\ell_0 + \ell_F + \ell_r]$.

7 NUMERICAL EXPERIENCE

We now numerically treat the problem (1.1), following the setup of [35]. Moreover, to demonstrate auxiliary parametrisation and the sliding PDPS of Example 6.1, we consider the "biased" problem with TV-regularisation,

$$\min_{\mu\in\mathcal{M}(\Omega),\, z\in Z} \frac{1}{2}\|A\mu + z - b\|^2 + \alpha\|\mu\| + \delta_{\geq 0}(\mu) + \lambda\|\nabla_h z\|_{2,1}, \tag{7.1}$$

where ∇_h describe a discretised gradient operator from 1D-signals $Z = \mathbb{R}^n$ to $\mathbb{R}^n$ or from 2D images $Z = \mathbb{R}^{n_1 n_2}$ to vector fields in $\mathbb{L}(\mathbb{R}^{n_1 n_2}; \mathbb{R}^2)$, and $\|\cdot\|_{2,1}$ is a sum over 2-norms over each component. The operator A is as for (1.1). The idea is that we observe on a sensor grid (a camera) the image b that has the spikes μ superimposed over an unknown background image z.

We first briefly describe the algorithm parametrisation in Section 7.1. We then describe the sample problems that we solve in Section 7.2. We finish with a report and discussion on the performance in Section 7.3.

7.1 ALGORITHM IMPLEMENTATION AND PARAMETRISATION

For the basic problem (1.1) we implemented Algorithm 3 (sFB), the algorithms from [35] (μFB and μPDPS) as well as the "fully corrective" conditional gradient method of [28, Algorithm 2] (FWf). Moreover, we include results for variants of the forward-backward methods with a Radon-norm-squared proximal term (radon^{2}FB and radon^{2}sFB). Our Rust implementations are available on Zenodo [36]. The implementation also includes the inertial μFISTA and the "relaxed" conditional gradient method of [8, Algorithm 5.1], but as these were not, correspondingly, better than μFB and FWf in the experiments of [35], we do not include the results here. For the "biased" problem (7.1), we implemented the sliding PDPS of Example 6.1 (sPDPS), as well as its non-sliding variant (fPDPS). We also implemented their variants with a Radon-norm-squared proximal term (radon^{2}sPDPS and radon^{2}fPDPS), although our theory does not show them to be convergent.

Several details of the implementation, such as the branch-and-bound method for the insertion Algorithm 1, are documented in [35]. Conditional gradient methods crucially depend on merging heuristics to keep the number of spikes computationally manageable. These are also documented in [35]. Contrary to the experiments in [35], on some of the experiments, we now needed to also enable merging heuristics for μPDPS to avoid accumulating a high number of spikes once the method starts to be very near a solution. The situation is the same with the new (non-sliding) radon^{2}FB and radon^{2}fPDPS, with which we use the same objective function decrease based merging as with FWf. This is possible with the weak convergence guarantees of Theorem 4.11, but not with Theorem 5.7 and Corollary 5.8. We do not use or need merging heuristics with the other variants of our methods, except to clean up after the final step.

Most methods use semismooth Newton (SSN) for the finite-dimensional subproblems of Algorithm 1, occasionally falling back to forward-backward splitting if the Newton system is too ill-conditioned. For the Radon norm proximal term the subproblem of Algorithm 2 is significantly more complex, so, due to the significant effort of deriving and implementing semismooth Newton methods, we have,

Table 2: Algorithm parametrisations for the basic experiments. Empty fields are not applicable to the algorithm in question. For merging, "i:ρ" indicates that weighted interpolation is used to form a new spike to replace the merged spikes, with ρ the candidate search radius, while "m:ρ" indicates that merging moves mass between the merged spikes. The transport tolerance multiplier C_{con} is explained in Example 5.13.

(a) 1D "fast" experiment

Method	τ_0	θ_0	σ_0	merging	inner method	c_{con}
μFB	0.99			no	SSN	
μPDPS	5.0		0.198	i:0.01	SSN	
FWf				i:0.01	SSN	
sFB	0.99	0.9		no	SSN	100.0
radon^{2}FB	0.99			m:0.01	PDPS	
radon^{2}sFB	0.99	0.9		m:0.01	PDPS	1000.0

(b) 2D "fast" experiment

Method	τ_0	θ_0	σ_0	merging	inner method	c_{con}
μFB	0.99			no	SSN	
μPDPS	5.0		0.198	i:0.01	SSN	
FWf				i:0.01	SSN	
sFB	0.99	0.9		no	SSN	100.0
radon^{2}FB	0.99			m:0.01	PDPS	
radon^{2}sFB	0.99	0.9		m:0.01	PDPS	1000.0

so far, only implemented the PDPS. This is reflected in the reported higher iteration counts for the inner problem. Even then, our formulation depends on the proximal map of $\vec{\beta} \mapsto \frac{1}{2}\|\vec{\beta} - \vec{\alpha}\|_1 + \tau\alpha\|\vec{\beta}\|_1$. Formulating an algorithm for this requires some effort, but proceeds along similar lines as sorting algorithms for simplex projection (see, e.g., [2]) or the constrained positive ℓ_1 regularisation algorithm of [32, Lemma D.2]. We provide details in the internal documentation of our implementation [36].

We always take the initial iterate $\mu^0 = 0$. For (7.1), also initial $z^0 = 0$ and the dual iterate $y^0 = 0$. Based on trial and error, we take the tolerance sequence $\varepsilon_k = 0.5\tau\alpha/(1 + 0.2k)^{1.4}$, where k is the iteration number. This choice balances between fast initial convergence and not slowing down later iterations too much via excessive accuracy requirements. Moreover, as a bootstrap heuristic, on the first 10 iterations, we insert in Algorithm 1 at most one point irrespective of the tolerance ε_k. This does not affect convergence, as the convergence theory can be applied starting from any fixed iteration number. We indicate in Tables 2 and 3 the step length and other parameters that differ between each algorithm. For the step length parameters we indicate value relative to the maximal value, e.g., for μFBwe take $\tau = \tau_0/L$ where $\tau \in (0, 1)$ is indicated in the table, and L satisfies $A_*A \leq L\mathscr{D}$.

7.2 EXPERIMENTS

We use the squared data term $F(\mu) = \frac{1}{2}\|A\mu - b\|^2$ in both $\Omega = [0, 1]$ and $\Omega = [0, 1]^2$. For the forward operator, following the construction in [35, Theorem 3.3], we take $A \in \mathbb{L}(\mathscr{M}(\Omega); \mathbb{R}^n)$ defined by $[A\mu]_i = \mu(\theta_i * \psi)$, $(i = 1, \ldots, n)$, where each sensor i on a uniform grid with centres z_i has the field-of-view $\theta_i(x) = \chi_{[-r,r]^n}(x - z_i)$. In $[0, 1]$ we use 100, and in $[0, 1]^2$ we use 16×16 equally spaced sensors

Table 3: Algorithm parametrisations for the "biased" experiments. For merging, "m:ρ" indicates that merging moves mass between the merged spikes, with ρ the candidate search radius. The transport tolerance multiplier c_{con} operate similarly to Example 5.13, taking into account the changed structure of v^k and $\check{v}^k$.

(a) 1D "biased" experiment

Method	τ_0	θ_0	$\sigma_{p,0}$	$\sigma_{d,0}$	merging	inner method	c_{con}
sPDPS	0.99	0.9	0.99	0.05	no	SSN	100.0
fPDPS	0.99		0.99	0.05	no	SSN	
radon^{2}sPDPS	0.99	0.3	0.99	0.05	m:0.01	PDPS	1000.0
radon^{2}fPDPS	0.99		0.99	0.05	m:0.01	PDPS	

(b) 2D "biased" experiment

Method	τ_0	θ_0	$\sigma_{p,0}$	$\sigma_{d,0}$	merging	inner method	c_{con}
sPDPS	0.99	0.9	0.99	0.05	no	SSN	100.0
fPDPS	0.99		0.99	0.05	no	SSN	
radon^{2}sPDPS	0.99	0.3	0.99	0.15	m:0.01	PDPS	10000.0
radon^{2}fPDPS	0.99		0.99	0.15	m:0.01	PDPS	

Table 4: Experiment noise levels and regularisation parameters. Empty fields are not applicable to the experiment in question.

Experiment	std.dev.	SNR (dB)	α	λ
1D "fast"	0.2	4.83	0.06	
2D "fast"	0.15	3.80	0.12	
1D "biased"	0.1	20.26	0.2	0.02
2D "biased"	0.15	10.65	0.06	0.005

with r the sensor spacing times 0.4.[8] For the physical spread ψ, we consider the "fast" (compactly supported, differentiable, third-order polynomial) spread of [35, Example 3.4]. We do not consider the "cut Gaussian" spread, as this is not differentiable.[9] We take the kernel ρ for the particle-to-wave operator $\mathscr{D} = \rho *$ from the same example. The standard deviations and other parameters of the spread are as in the numerical experiments of [35]. To generate the synthetic measurement data b, we apply A to a ground-truth measure $\hat{\mu}$ with four spikes of distinct magnitudes. For the "biased" problem (7.1) we additionally add to this the ground-truth bias $\hat{z}$ formed by summing the weighted indicator functions of two intervals (1D) or Euclidean balls (2D). The details can be found in our implementation [36]. Then we add to each sensor reading independent Gaussian noise. The noise level (standard deviation and resulting SSNR) and corresponding regularisation parameters, found by trial-and-error, are listed in Table 4 for each experiment.

[8] The small number of sensors along each axis in 2D is for visualisation purposes: quadrupling sensor count to a grid of 32×32, only doubles CPU time, indicating good scaling behaviour of our implementation.

[9] Our implementation [36] includes a dynamic estimation workaround to the lack of Lipschitz gradient estimates of the dual variables, but the performance of the sliding methods is unpredictable due to the unsatisfied assumption.

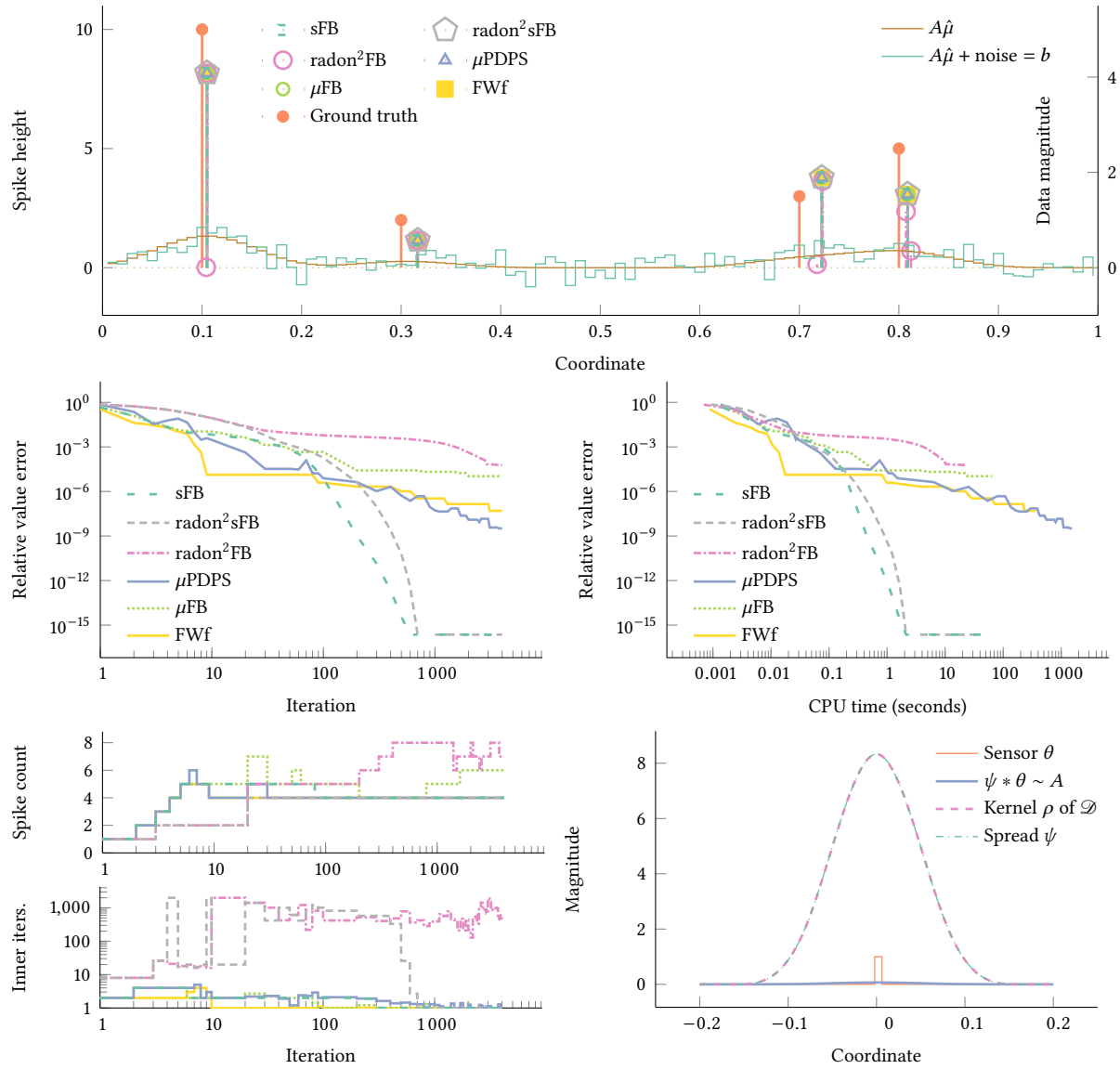

Figure 1: Reconstructions and performance on 1D problem with "fast" spread. **Top:** reconstruction and original data. The measurement data magnitude scale is on the right, spike magnitude on the left. **Middle:** Function value in terms of iteration count (left) and CPU time (right). **Bottom:** spike evolution, inner iteration count (left), and kernels (right). The thick lines indicate the spike count, and the thinner and dimmer lines the inner iteration count.

7.3 RESULTS

We ran the experiments on a 2020 MacBook Air M1 with 16GB of memory. We take advantage of the 4 high performance CPU cores of the 8-core machine by using 4 parallel computational threads to calculate A^*z and for the branch-and-bound optimisation. We report the performance of all the algorithms applicable to each experiment in the corresponding Figures 1 to 4. Each of the figures depicts the spread ψ, kernel ρ, and sensor θ involved in A and $\mathscr{D}$. They also depict the noisy and noise-free data, the ground-truth measure $\hat{\mu}$, and the algorithmic reconstructions. For (7.1) also the ground-truth bias $\hat{z}$ and its reconstruction by sPDPS is included. The reconstructions or optimal solutions to the problems (4.1) and (7.1) cannot be expected to equal $\hat{\mu}$ and $\hat{z}$ due to noise and ill-conditioning of the

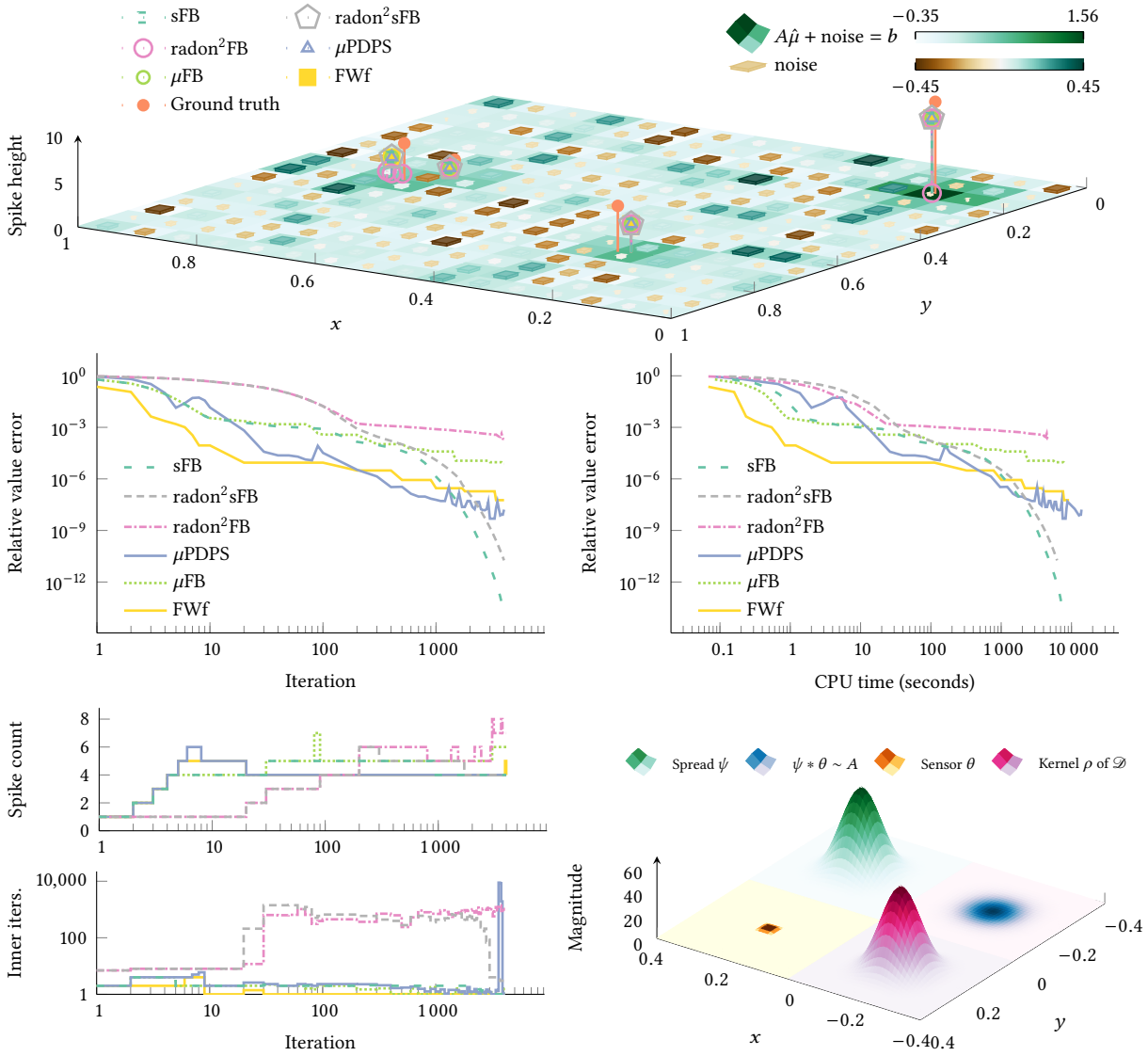

Figure 2: Reconstructions and performance on 2D problem with "fast" spread. **Top:** reconstruction and original data. The area of the top surface of the boxes is proportional the noise level of the underlying sensor, and their colour the sign of the noise. **Middle:** Function value in terms of iteration count and CPU time. **Bottom:** spike and inner iteration count evolution, and kernels.

inverse problem $A\mu + z = b$.

Each of the figures plots against both the iteration count and the spent CPU time, the relative error of the function value,

$$e^k = \frac{v(x^k) - v(x_{\min})}{v(x^0) - v(x_{\min})},$$

where v is the objective of (1.1) and $x^k = \mu^k$ (or (7.1) and $x^k = (\mu^k, z^k)$), and $v(x_{\min})$ is the minimum objective value observed over all algorithms. (Since the relative error is zero for such an iterate, it will not be shown in the logarithmic plots.) The plots are logarithmic on both axes, and we sample the reported values logarithmically only on iterations $1, 2, \ldots, 10, 20, \ldots, 100, 200, \ldots$. We limit the number of iterations to 4000. The CPU time is a sum over the time spent by each computational thread, so

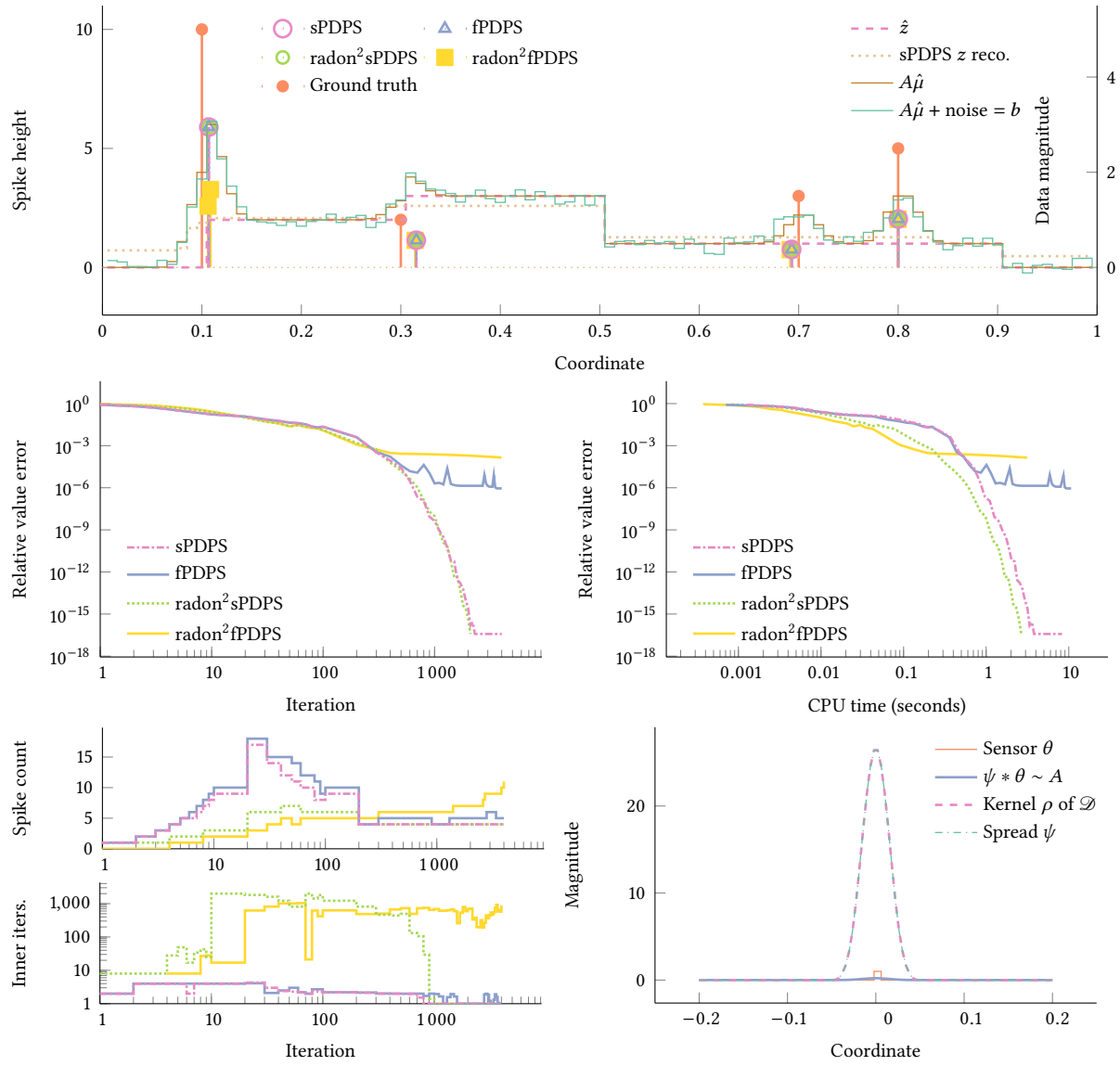

Figure 3: Reconstructions and performance on 1D problem with "fast" spread and TV-regularised bias. **Top:** reconstruction and original data. The measurement data magnitude scale is on the right, spike magnitude on the left. **Middle:** Function value in terms of iteration count (left) and CPU time (right). **Bottom:** spike evolution, inner iteration count (left), and kernels (right). The thick lines indicate the spike count, and the thinner and dimmer lines the inner iteration count.

several times the clock time requirement.

The figures also indicate the spike count evolution, and the number of iterations needed to solve the finite-dimensional subproblems. The subproblem iteration counts are averages over the corresponding period.

7.4 COMPARISON

Studying Figures 1 and 2, we notice the sliding $\mathscr{D}$-(semi)norm proximal term sFB as well as the Radon-norm proximal term radon^{2}sFB to exhibit significant performance advantages over the comparison

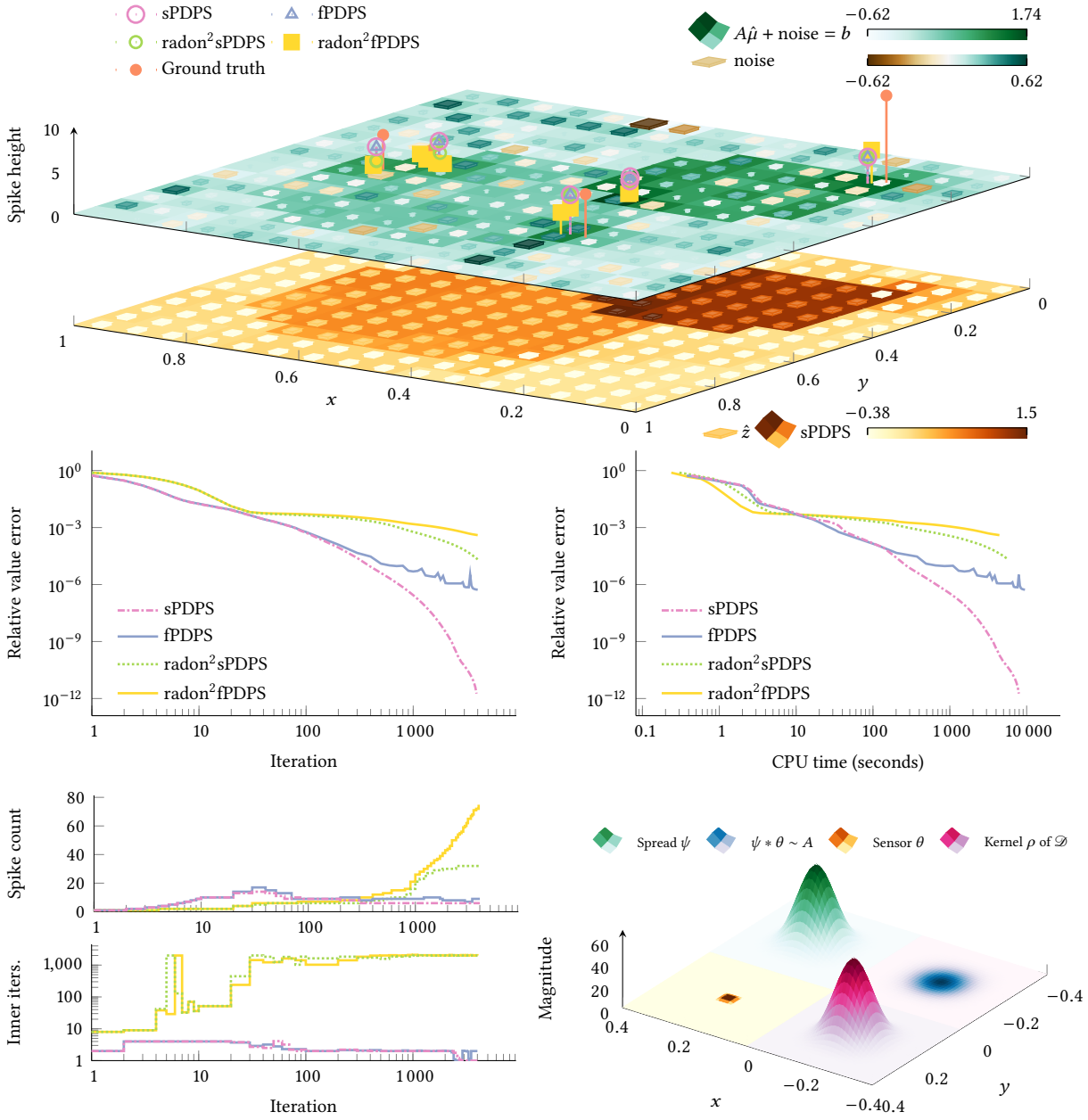

Figure 4: Reconstructions and performance on 2D problem with "fast" spread and TV-regularised bias. **Top:** reconstruction and original data. The upper layer displays the data, noise level, and spike reconstructions, while the lower layer displays the bias and its reconstruction by sPDPS. The area of the top surface of the boxes that indicate the noise level is proportional the noise level of the underlying sensor, and their colour the sign of the noise. **Middle:** Function value in terms of iteration count and CPU time. **Bottom:** spike and inner iteration count evolution, and kernels.

methods. The $\mathcal{D}$-(semi)norm is generally faster, as is to be expected from both being more tailored to the problem, as well as from the worse convergence guarantees that we are able to obtain. Likewise, for the biased problem, Figures 3 and 4 indicate the sliding sPDPS to overperform the other methods variants, with the Radon-norm proximal term exhibiting noticeably slower convergence than the

$\mathscr{D}$-norm. Close to a solution, the Radon norm proximal term also start to exhibit spike accumulation, with merging heuristics not always being effective. This is to be expected, as our theory does not show the convergence of PDPS based on the Radon-squared proximal term.

7.5 CONCLUSION

Overall, the numerical results confirm our intuition that proximal terms based on metrisation of the weak-$*$ topology—in particular, conventional optimal transport distances, but also particle-to-wave or MMD norms–improve the performance of optimisation methods in measure spaces. Lipschitz properties of the data term F, can, however, be more easily verified for proximal terms based on the Radon norm, as well as for conditional gradient methods. An advantage of our approach, compared to the latter, is its flexibility, allowing easily to treat more complex problems through primal-dual methods, product spaces, and nonconvex objectives.

REFERENCES

[1] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, Birkhäuser, second edition, 2008.

[2] S. Angerhausen, *Stochastic Primal-Dual Proximal Splitting Method for Risk-Averse Optimal Control of PDEs*, PhD thesis, University Duisburg–Essen, 2022, doi:10.17185/duepublico/78165.

[3] J. D. Benamou, Numerical resolution of an "unbalanced" mass transport problem, *ESAIM: Mathematical Modelling and Numerical Analysis* 37 (2003), 851–868, doi:10.1051/m2an:2003058.

[4] R. Bergmann, R. Herzog, D. Tenbrick, and J. Vidal-Núñez, Fenchel duality for convex optimization and a primal dual algorithm on Riemannian manifolds, 2019, arXiv:1908.02022.

[5] L. Blank and C. Rupprecht, An extension of the projected gradient method to a Banach space setting with application in structural topology optimization, *SIAM Journal on Control And Optimization* 55 (2017), 1481–1499, doi:10.1137/16m1092301.

[6] K. Bredies, M. Carioni, S. Fanzon, and F. Romero, A generalized conditional gradient method for dynamic inverse problems with optimal transport regularization, 2020, arXiv:2012.11706.

[7] K. Bredies, M. Carioni, S. Fanzon, and D. Walter, Linear convergence of accelerated generalized conditional gradient methods, 2021, arXiv:2110.06756.

[8] K. Bredies and H. K. Pikkarainen, Inverse problems in spaces of measures, *ESAIM: Control, Optimization and Calculus of Variations* 19 (2013), 190–218, doi:10.1051/cocv/2011205.

[9] E. J. Candès and C. Fernandez-Granda, Towards a mathematical theory of super-resolution, *Communications on Pure and Applied Mathematics* 67 (2014), 906–956, doi:10.1002/cpa.21455.

[10] E. Casas, C. Clason, and K. Kunisch, Approximation of elliptic control problems in measure spaces with sparse solutions, *SIAM Journal on Control And Optimization* 50 (2012), 1735–1752, doi:10.1137/110843216.

[11] E. Casas, C. Clason, and K. Kunisch, Parabolic control problems in measure spaces with sparse solutions, *SIAM Journal on Optimization* 51 (2013), 28–63, doi:10.1137/120872395.

[12] W. Cheney and W. Light, *A Course in Approximation Theory*, volume 101 of Graduate Studies in Mathematics, American Mathematical Society, 2000.

[13] L. Chizat, Sparse optimization on measures with over-parameterized gradient descent, *Mathematical Programming* (2021), doi:10.1007/s10107-021-01636-z.

[14] L. Chizat and F. Bach, On the global convergence of gradient descent for over-parameterized models using optimal transport, in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), volume 31, Curran Associates, Inc., 2018, https://proceedings.neurips.cc/paper/2018/file/a1afc58c6ca9540d057299ec3016d726-Paper.pdf.

[15] L. Chizat, G. Peyré, B. Schmitzer, and F. X. Vialard, An Interpolating distance between optimal transport and Fisher–Rao metrics, *Foundations of Computational Mathematics* 18 (2016), 1–44, doi:10.1007/s10208-016-9331-y.

[16] L. Chizat, G. Peyré, B. Schmitzer, and F. X. Vialard, Unbalanced optimal transport: Dynamic and Kantorovich formulations, *Journal of Functional Analysis* 274 (2018), 3090–3123, doi:10.1016/j.jfa.2018.03.008.

[17] C. Clason and T. Valkonen, Introduction to Nonsmooth Analysis and Optimization, 2024, arXiv:2001.00216, https://tuomov.iki.fi/m/nonsmoothbook.pdf. textbook, submitted.

[18] Q. Denoyelle, V. Duval, G. Peyré, and E. Soubies, The sliding Frank–Wolfe algorithm and its application to super-resolution microscopy, *Inverse Problems* (2019), doi:10.1088/1361-6420/ab2a29.

[19] N. Dizon and T. Valkonen, Differential estimates for fast first-order multilevel nonconvex optimisation, 2024, arXiv:2412.01481, https://tuomov.iki.fi/m/tracking.pdf. submitted.

[20] V. Duval and G. Peyré, Sparse regularization on thin grids I: the Lasso, *Inverse Problems* 33 (2017), 055008, doi:10.1088/1361-6420/aa5e12.

[21] A. Flinth, F. de Gournay, and P. Weiss, On the linear convergence rates of exchange and continuous methods for total variation minimization, *Mathematical Programming* 190 (2020), 221–257, doi:10.1007/s10107-020-01530-0.

[22] I. Fonseca and G. Leoni, *Modern methods in the calculus of variations: L^p spaces*, Springer, 2007.

[23] E. Gladin, P. Dvurechensky, A. Mielke, and J. J. Zhu, Interaction-Force Transport Gradient Flows, 2024, arXiv:2405.17075.

[24] S. Kondratyev, L. Monsaingeon, D. Vorotnikov, et al., A new optimal transport distance on the space of finite Radon measures, *Advances in Differential Equations* 21 (2016), 1117–1164, arXiv:1505.07746.

[25] F. Lauster and D. R. Luke, Convergence of proximal splitting algorithms in CAT(κ) spaces and beyond, *Fixed Point Theory and Algorithms for Sciences and Engineering* 2021 (2021), doi:10.1186/s13663-021-00698-0.

[26] M. Liero, A. Mielke, and G. Savaré, Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures, *Inventiones mathematicae* 211 (2017), 969–1117, doi:10.1007/s00222-017-0759-8.

[27] J. Lindberg, Mathematical concepts of optical superresolution, *Journal of Optics* 14 (2012), 083001, doi:10.1088/2040-8978/14/8/083001.

[28] K. Pieper and D. Walter, Linear convergence of accelerated conditional gradient algorithms in spaces of measures, *ESAIM: Control, Optimization and Calculus of Variations* 27 (2021), 38, doi:10.1051/cocv/2021042, arXiv:1904.09218.

[29] F. Santambrogio, *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, Progress in Nonlinear Differential Equations and Their Applications, Springer, 2015.

[30] B. Schmitzer and B. Wirth, Dynamic models of Wasserstein-1-type unbalanced transport, *ESAIM: Control, Optimization and Calculus of Variations* 25 (2019), 23, doi:10.1051/cocv/2018017.

[31] B. Schmitzer and B. Wirth, A Framework for Wasserstein-1-Type Metrics, *Journal of Convex Analysis* (2019), 353–396, arXiv:1701.01945.

[32] E. Suonperä and T. Valkonen, General single-loop methods for bilevel parameter learning, 2024, arXiv:2408.08123, https://tuomov.iki.fi/m/bilevel_general.pdf. submitted.

[33] T. Séjourné, G. Peyré, and F. X. Vialard, Unbalanced Optimal Transport, from theory to numerics, in *Numerical Control: Part B*, Elsevier, 2023, 407–471, doi:10.1016/bs.hna.2022.11.003.

[34] C. Udrişte, *Convex functions and optimization methods on Riemannian manifolds*, volume 297 of Mathematics and its Applications, Kluwer Academic Publishers, Dordrecht, 1994, doi:10.1007/978-94-015-8390-9.

[35] T. Valkonen, Proximal methods for point source localisation, *Journal of Nonsmooth Analysis and Optimization* 4 (2023), 10433, doi:10.46298/jnsao-2023-10433, arXiv:2212.02991, https://tuomov.iki.fi/m/pointsourceo.pdf.

[36] T. Valkonen, Proximal methods for point source localisation: the implementation, Software on Zenodo, written in Rust, 2025, doi:10.5281/zenodo.14884773.

APPENDIX A AUXILIARY LEMMAS

We recall a fundamental three-point identity of Bregman divergences.

Lemma A.1. *Let $J : \mathscr{M}(\Omega) \to \overline{\mathbb{R}}$ be convex, proper, and weakly-$*$ lower semicontinuous. Take $\nu_0, \nu_1, \nu_2 \in \mathscr{M}(\Omega)$, as well as $\omega_0 \in \partial J(\nu_0)$ and $\omega_1 \in \partial J(\nu_1)$. Then*

$$B_J^{\omega_0}(\nu_0, \nu_2) - B_J^{\omega_0}(\nu_0, \nu_1) = B_J^{\omega_1}(\nu_1, \nu_2) + \langle \omega_1 - \omega_0 | \nu_2 - \nu_1 \rangle.$$

Proof. We have

$$\begin{aligned} B_J^{\omega_0}(\nu_0, \nu_2) - B_J^{\omega_0}(\nu_0, \nu_1) &= J(\nu_2) - J(\nu_1) - \langle \omega_0 | \nu_2 - \nu_1 \rangle \\ &= J(\nu_2) - J(\nu_1) - \langle \omega_1 | \nu_2 - \nu_1 \rangle + \langle \omega_1 - \omega_0 | \nu_2 - \nu_1 \rangle \\ &= B_J^{\omega_1}(\nu_1, \nu_2) + \langle \omega_1 - \omega_0 | \nu_2 - \nu_1 \rangle. \end{aligned}$$

□