A Stochastic Bregman Primal-Dual Splitting Algorithm for Composite Optimization

Antonio Silveti-Falls, Cesare Molinari, Jalal Fadili

A Stochastic Bregman Primal-Dual Splitting Algorithm for Composite Optimization

Antonio Silveti-Falls* Cesare Molinari † Jalal Fadili‡

Abstract. We study a stochastic first order primal-dual method for solving convex-concave saddle point problems over real reflexive Banach spaces using Bregman divergences and relative smoothness assumptions, in which we allow for stochastic error in the computation of gradient terms within the algorithm. We show ergodic convergence in expectation of the Lagrangian optimality gap with a rate of $O(1/k)$ and that every almost sure weak cluster point of the ergodic sequence is a saddle point in expectation under mild assumptions. Under slightly stricter assumptions, we show almost sure weak convergence of the pointwise iterates to a saddle point. Under a relative strong convexity assumption on the objective functions and a total convexity assumption on the entropies of the Bregman divergences, we establish almost sure strong convergence of the pointwise iterates to a saddle point. Our framework is general and does not need strong convexity of the entropies inducing the Bregman divergences in the algorithm. Numerical applications are considered including entropically regularized Wasserstein barycenter problems and regularized inverse problems on the simplex.

Key words. Bregman divergence; primal-dual splitting; noneuclidean splitting; saddle point problems; first order algorithms; convergence rates; relative smoothness; total convexity; Banach space.

AMS subject classifications. 49J52, 65K05, 65K10.

1 Introduction

1.1 Problem Statement and Algorithm

The goal is to solve the following primal-dual, or saddle point, problem over the real reflexive Banach spaces $\mathcal{X}_p$ and $\mathcal{X}_d$, where the subscript p refers to primal and d to dual:

$$\min_{x\in\mathcal{X}_p}\max_{\mu\in\mathcal{X}_d}\mathcal{L}(x,\mu) \tag{$\mathscr{P}.\mathscr{D}.$}$$

where

$$\mathcal{L}(x,\mu)\overset{\text{def}}{=}f(x)+g(x)+\iota_{\mathcal{C}_p}(x)+\langle Tx,\ \mu\rangle-h^*(\mu)-l^*(\mu)-\iota_{\mathcal{C}_d}(\mu) \tag{1.1}$$

is the Lagrangian functional and $\iota_{\mathcal{C}_p}$ and $\iota_{\mathcal{C}_d}$ are the indicator functions of the convex constraint sets $\mathcal{C}_p$ and $\mathcal{C}_d$, respectively, and $T:\mathcal{X}_p\to\mathcal{X}_d^*$ is a linear mapping. This composite problem formulation is general enough

*Toulouse School of Economics, University of Toulouse, France. E-mail: Tony.S.Falls@gmail.com. This work was done while the author was at GREYC-ENSICAEN.

†Istituto Italiano di Tecnologia, Italy. E-mail: cecio.molinari@gmail.com

‡Normandie Université, ENSICAEN, UNICAEN, CNRS, GREYC, France. E-mail: Jalal.Fadili@ensicaen.fr.

to encompass in particular the problem settings of [16, 20] as well as image processing problems involving infimal convolution based regularizers accounting for multiple orders of smoothness, see e.g., [7, 32, 28].

We denote the primal and dual problems as

$$\min_{x\in\mathcal{C}_p}\left\{f(x)+g(x)+\left(\left(h\underset{\mathcal{C}_d}{\square}l\right)\circ T\right)(x)\right\} \tag{$\mathscr{P}$}$$

$$\min_{\mu\in\mathcal{C}_d}\left\{h^*(\mu)+l^*(\mu)+\left(\left(f^*\underset{\mathcal{C}_p}{\square}g^*\right)\circ(-T^*)\right)(\mu)\right\} \tag{$\mathscr{D}$}$$

where $\left(f^*\underset{\mathcal{C}_p}{\square}g^*\right)^* \overset{\text{def}}{=} f+g+\iota_{\mathcal{C}_p}$, using $*$ to denote the Fenchel conjugate, and similarly for $\underset{\mathcal{C}_d}{\square}$. In the case in which $\mathcal{C}_p$ and $\mathcal{C}_d$ are trivial constraints, i.e., the entire spaces $\mathcal{X}_p$ and $\mathcal{X}_d$, the corresponding primal and dual problems related to ($\mathscr{P}.\mathscr{D}.$) are

$$\min_{x\in\mathcal{X}_p}\left\{f(x)+g(x)+\left((h\square l)\circ T\right)(x)\right\}$$
$$\min_{\mu\in\mathcal{X}_d}\left\{h^*(\mu)+l^*(\mu)+\left((f^*\square g^*)\circ(-T^*)\right)(\mu)\right\}$$

where $\square$ recovers the classical *infimal convolution* defined by $f\square g(v)=\inf_{w\in\mathcal{X}_p}\left(f(w)+g(v-w)\right)$. The set of solutions for ($\mathscr{P}$) and ($\mathscr{D}$) are written as

$$\begin{aligned}\boldsymbol{S}_{\mathscr{P}} &\overset{\text{def}}{=} \operatorname*{argmin}_{x\in\mathcal{C}_p}\left\{\max_{\mu\in\mathcal{C}_d}\left\{f(x)+g(x)+\langle Tx,\mu\rangle-h^*(\mu)-l^*(\mu)\right\}\right\}\\ \boldsymbol{S}_{\mathscr{D}} &\overset{\text{def}}{=} \operatorname*{argmax}_{\mu\in\mathcal{C}_d}\left\{\min_{x\in\mathcal{C}_p}\left\{f(x)+g(x)+\langle Tx,\mu\rangle-h^*(\mu)-l^*(\mu)\right\}\right\}.\end{aligned} \tag{1.2}$$

The set of saddle points for the Lagrangian defined in (1.1) is denoted

$$\boldsymbol{S}\overset{\text{def}}{=}\left\{(x^\star,\mu^\star)\in\mathcal{X}_p\times\mathcal{X}_d:\quad\forall(x,\mu)\in\mathcal{X}_p\times\mathcal{X}_d,\quad\mathcal{L}(x^\star,\mu)\leq\mathcal{L}(x^\star,\mu^\star)\leq\mathcal{L}(x,\mu^\star)\right\}$$

which obeys the inclusion $\boldsymbol{S}\subset\boldsymbol{S}_{\mathscr{P}}\times\boldsymbol{S}_{\mathscr{D}}$.

Given a real reflexive Banach space $\mathcal{X}$, we denote by $\Gamma_0(\mathcal{X})$ the space of proper convex lower semicontinuous functions from $\mathcal{X}$ to $\mathbb{R}\cup\{+\infty\}$. For a subset $\mathcal{C}$ of a Banach space, $\operatorname{int}\mathcal{C}$ denotes its interior. We suppose the following standing hypotheses on the problem, which we collectively denote by **(H)**:

(H)
- **($\mathbf{H}_1$)** The Banach spaces $\mathcal{X}_p$ and $\mathcal{X}_d$ are real and reflexive, while $\mathcal{C}_p\subset\mathcal{X}_p$ and $\mathcal{C}_d\subset\mathcal{X}_d$ are nonempty convex closed subsets.
- **($\mathbf{H}_2$)** The functions f and g belong to $\Gamma_0(\mathcal{X}_p)$ while l and h belong to $\Gamma_0(\mathcal{X}_d)$, with $\mathcal{C}_p\subset\operatorname{dom}(f)$ (resp. $\mathcal{C}_d\subset\operatorname{dom}(h^*)$) and f (resp. h^*) is differentiable on $\operatorname{int}\mathcal{C}_p$ (resp. $\operatorname{int}\mathcal{C}_d$).
- **($\mathbf{H}_3$)** $\mathcal{C}_p\cap\operatorname{dom}(g)\neq\emptyset$ and $\mathcal{C}_d\cap\operatorname{dom}(l^*)\neq\emptyset$.
- **($\mathbf{H}_4$)** The operator $T:\mathcal{X}_p\to\mathcal{X}_d^*$ is linear and continuous.
- **($\mathbf{H}_5$)** The set of saddle points $\boldsymbol{S}$ for ($\mathscr{P}.\mathscr{D}.$) is nonempty.

It is well-known that $\boldsymbol{S}$ is non-empty under suitable domain qualification conditions.

Before introducing the method, we recall the definition of Bregman divergence which will be key to our algorithm and to the theoretical analysis of convergence.

Definition 1.1 (Bregman divergence). Given a function $\phi : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$, often referred to as the entropy, differentiable on $\operatorname{int} \operatorname{dom}(\phi)$, its Bregman divergence is defined by

$$D_\phi(x, y) \overset{\text{def}}{=} \begin{cases} \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y\rangle & \text{if } x \in \operatorname{dom}(\phi) \text{ and } y \in \operatorname{int} \operatorname{dom}(\phi), \\ +\infty & \text{else.} \end{cases}$$

Notice that if ϕ belongs to $\Gamma_0(\mathcal{X})$ then D_ϕ is always nonnegative by the subdifferential inequality.

The Stochastic Bregman Primal-Dual Splitting algorithm, SBPD for short, is presented in Algorithm 1. We introduce two entropy functions, ϕ_p and ϕ_d, and we denote by D_p and D_d their Bregman divergences, respectively. We further consider the possibility of some stochastic error in the computation of the gradients[1] ∇f and ∇h^* which we will denote for $\nabla f(x_k)$ as δ_k^p and for $\nabla h^*(\mu_k)$ as δ_k^d.

Algorithm 1: Stochastic Bregman Primal-Dual Splitting (SBPD).

for $k = 0, 1, \ldots$ **do**

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathcal{C}_p} \left\{ g(x) + \langle \nabla f(x_k) + \delta_k^p, x\rangle + \langle Tx, \widetilde{\mu}_k\rangle + \frac{1}{\lambda_k} D_p(x, x_k) \right\}$$

$$\mu_{k+1} = \operatorname*{argmin}_{\mu \in \mathcal{C}_d} \left\{ l^*(\mu) + \langle \nabla h^*(\mu_k) + \delta_k^d, \mu\rangle - \langle T\widetilde{x}_k, \mu\rangle + \frac{1}{\nu_k} D_d(\mu, \mu_k) \right\}$$

where $\widetilde{\mu}_k = \mu_k$ and $\widetilde{x}_k = 2x_{k+1} - x_k$.

In the deterministic setting for the primal update, i.e., $\delta_k^p = 0$ for each $k \in \mathbb{N}$, the first step of the algorithm can be re-written in the following way:

$$\begin{aligned} x_{k+1} &= \operatorname*{argmin}_{x \in \mathcal{C}_p} \left\{ g(x) + f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \langle Tx, \widetilde{\mu}_k\rangle + \frac{1}{\lambda_k} D_p(x, x_k) \right\} \\ &= (\nabla\phi_p + \lambda_k \partial g)^{-1} \left(\nabla\phi_p - \lambda_k \nabla\left(f(\cdot) + \langle T\cdot, \widetilde{\mu}_k\rangle\right)\right)(x_k) \\ &= (\nabla\phi_p + \lambda_k \partial g)^{-1} \left(\nabla\phi_p(x_k) - \lambda_k \nabla f(x_k) - \lambda_k T^* \widetilde{\mu}_k\right). \end{aligned}$$

Analogously, if $\delta_k^d = 0$ for all $k \in \mathbb{N}$,

$$\mu_{k+1} = (\nabla\phi_d + \nu_k \partial l^*)^{-1} \left(\nabla\phi_d(\mu_k) - \nu_k \nabla h^*(\mu_k) + \nu_k T\widetilde{x}_k\right).$$

A priori, the mappings $(\nabla\phi_p + \lambda_k\partial g)^{-1}$ and $(\nabla\phi_d + \nu_k\partial l^*)^{-1}$, sometimes referred to as D-prox mappings, may be empty, may not be single-valued, or may not map $\operatorname{int} \operatorname{dom}(\phi_p)$ (resp. $\operatorname{int} \operatorname{dom}(\phi_d)$) to $\operatorname{int} \operatorname{dom}(\phi_p)$ (resp. $\operatorname{int} \operatorname{dom}(\phi_d)$). In light of this, we will only consider ϕ_p and ϕ_d for which these mappings are well-defined and map from $\operatorname{int} \operatorname{dom}(\phi_p)$ to $\operatorname{int} \operatorname{dom}(\phi_p)$ and the analog for ϕ_d (see $(\mathbf{A}_1)$). In Section 3.1, we will elaborate on the class of Legendre functions on a real reflexive Banach space given in [4, Definition 2.2] which will help us to ensure that the D-prox mappings are well-defined.

[1] The addition of stochastic error in the computation of D-prox operators associated to g or l^*, while interesting, is problematic for the algorithm in the sense that the monotone inclusions may no longer hold and the iterates themselves might not remain in the interior of the domain as desired.

1.2 Contribution and Prior Work

The idea of using primal-dual methods to solve convex-concave saddle point problems has been around since the 1960s, e.g., [41], [50], [36], or [40]. For an introduction into the use of primal-dual methods in convex optimization, we refer the reader to [35]. More recently, without being exhaustive, there were the notable works [19], [14], [21], [56], and [16] which examined problems quite similar to the one posed here using first order primal-dual methods.

In particular, [16] studied ($\mathscr{P}.\mathscr{D}.$) using D-prox mappings, i.e., proximal mappings where the euclidean energy has been replaced by a suitable Bregman divergence, under the assumption that f and h^* are Lipschitz-smooth Γ_0 functions and that the entropies ϕ_p and ϕ_d are strongly convex. They show ergodic convergence of the Lagrangian optimality gap with a rate of $O\left(1/k\right)$ under mild assumptions and also faster rates, e.g., $O\left(1/k^2\right)$ and linear convergence, under stricter assumptions involving strong convexity. We generalize their results by relaxing the Lipschitz-smooth assumption to a relative smoothness assumption, by analyzing the totally convex and relatively strongly convex case, by introducing stochastic error to the algorithm, and by showing almost sure weak convergence of the pointwise iterates themselves. Additionally, the recent work [33] studied a variant of the problem considered in [16] focused on semidefinite programming with D-prox mappings and an adaptive step size. As in [16], they assume that the entropies inducing the Bregman divergences are strongly convex, in contrast to our work. The authors in [20] proposed a Bregman primal-dual method that iteratively constructs the best Bregman approximation to an arbitrary point from the Kuhn-Tucker set of a composite monotone inclusion in real reflexive Banach spaces, and for which they established strong convergence of the iterates. When specialized to structured minimization, their framework covers ($\mathscr{P}.\mathscr{D}.$) but without the smooth parts nor infimal-convolutions or the constraint sets $\mathcal{C}_p$ and $\mathcal{C}_d$. Moreover, their algorithm necessitates a complicated Bergman projection step and they do not consider stochastic versions.

Generalizations of [16] involving inexactness already exist in the form of [48] and [15], however, [48] only considers determinstic inexactness and proximal operators computed in the euclidean sense, i.e., with entropy equal to the euclidean energy, and requires Lipschitz-smoothness. It's worth noting that the inexactness considered in their paper allows for the inexact computation of the proximal operators, in contrast to our work. While Algorithm 1 allows for inexactness, in the form of stochastic error, it is only allowable in the computation of gradient terms. The paper [15] allows for a very particular kind of stochastic error in which one randomly samples a set of indices at each iteration in an arbitrary but fixed way, i.e., according to some fixed distribution. However, the stochastic error we consider in the present paper is more general while encompassing the previous cases, although with less sharp results if the noise is not well behaved.

Another related work is that of [30] which generalizes the problem considered in [16] by allowing for a nonlinear coupling $\Phi\left(x,\mu\right)$ in ($\mathscr{P}.\mathscr{D}.$) instead of $\left\langle Tx,\mu\right\rangle$, although they maintain essentially the same Lipschitz-smoothness assumptions as in [16] translated to $\Phi\left(x,\mu\right)$. They are able to show a $O\left(1/k\right)$ convergence rate for the ergodic Lagrangian optimality gap under mild assumptions and an accelerated rate $O\left(1/k^2\right)$ when g in ($\mathscr{P}.\mathscr{D}.$) is strongly convex with another assumption on the coupling $\Phi\left(x,\mu\right)$.

The notion of relative smoothness is key to the analysis of differentiable but not Lipschitz-smooth optimization problems. The earliest reference to this notion can be found in an economics paper [6] where it is used to address a problem in game theory involving fisher markets. Later it was parallelly developed for Bregman Forward-Backward splitting in [31] and then in [39] (see also [42, 10]), and coined relative smoothness in [39]. This idea allows one to apply arguments involving descent lemmas which are normally relegated to Lipschitz-smooth problems and it has been extended, for instance to define relative Lipschitz-continuity in [37], in [38] for the stochastic generalized conditional gradient, and to define a generalized curvature constant for the generalized conditional gradient algorithm in [54]. The analogous idea of relative strong convexity, while noted before in [16], was not explored in detail; here we analyze our algorithm under such assumptions

in combination with total convexity of the entropies.

To our knowledge, our paper is the first to analyze ($\mathscr{P}.\mathscr{D}.$) under a relative smoothness condition with D-prox mappings. Additionally, we are the first to include stochastic error in the computation of the gradient terms for ($\mathscr{P}.\mathscr{D}.$) under these assumptions.

1.3 Paper Organization

The rest of the paper is divided into four sections. In Section 2, we recall some basic definitions to make precise all the notions used in the paper along with some useful elementary results regarding sequences of random variables.

In Section 3, we make explicit all the assumptions **($\mathbf{A}_1$)-($\mathbf{A}_{11}$)** we will use on the objective functions, entropies, step sizes, etc. We go on to establish the main estimation of Lemma 3.10 under **(H)** and **($\mathbf{A}_1$)-($\mathbf{A}_3$)** that will be used in the convergence analysis of the ergodic, pointwise, and relatively strongly convex cases. The key idea is to utilize the descent lemma given by relative smoothness along with the usual inequalities for Γ_0 functions to estimate the optimality gap $\mathcal{L}(x_k, \mu) - \mathcal{L}(x, \mu_k)$ in terms of the Bregman divergences induced by the entropies ϕ_p and ϕ_d. The proof of the estimation here is similar in spirit to the proof of the main estimation in [14], with the main difference being that we are unable to use Young's inequality to deal with the coupling terms, which we handle using **($\mathbf{A}_3$)**. There are also some lemmas involving **(H)** and **($\mathbf{A}_1$)-($\mathbf{A}_5$)** regarding the stochastic error, culminating in a summability result for the sequences $(\mathbb{E}[\langle \Delta_k, w - w_{k+1}\rangle])_{k\in\mathbb{N}}$ and $(\mathbb{E}[\langle \Delta_k, w - w_{k+1}\rangle \mid \mathscr{S}_k])_{k\in\mathbb{N}}$ which appear in the convergence analysis.

In Section 4, we use the estimation developed in Section 3 along with **(H)** and **($\mathbf{A}_1$)-($\mathbf{A}_{11}$)** regarding the entropies ϕ_p and ϕ_d and the regularity of their induced Bregman divergences to show convergence of the algorithm; first convergence of the expectation of the Lagrangian optimality gap for the ergodic iterates under **(H)** and **($\mathbf{A}_1$)-($\mathbf{A}_5$)** and then almost sure weak convergence of the pointwise iterates under **(H)** and **($\mathbf{A}_1$)-($\mathbf{A}_{10}$)**. Finally, we examine the case where **($\mathbf{A}_{11}$)** holds, i.e., there is relative strong convexity of the objective functions with respect to the entropies, and total convexity of the entropies themselves. For the ergodic analysis, denote by (x_∞, μ_∞) an almost sure weak sequential cluster point of the ergodic primal-dual sequence $((\bar{x}_k, \bar{\mu}_k))_{k\in\mathbb{N}}$. Then we show that its expectation, namely $\mathbb{E}[(x_\infty, \mu_\infty)]$, is a saddle point. We prove also, for every x and μ, convergence of the expectation of the Lagrangian optimality gap $\mathbb{E}[\mathcal{L}(\bar{x}_k, \mu) - \mathcal{L}(x, \bar{\mu}_k)]$ with a rate of $O(1/k)$. For the pointwise analysis, we begin by showing an almost sure asymptotic regularity result for the primal-dual sequence $(w_k)_{k\in\mathbb{N}}$. With this, we are then able to adapt the well known Opial's lemma (see [43]) to the Bregman primal-dual setting to establish almost sure weak convergence of the primal-dual sequence $(w_k)_{k\in\mathbb{N}}$ to a saddle point $w^\star$. In the final part of this section, we establish almost sure strong convergence of the primal-dual sequence $(w_k)_{k\in\mathbb{N}}$ to a saddle point $w^\star$ under **($\mathbf{A}_{11}$)** and total convexity of the entropies.

Lastly, in Section 5, we explore potential applications of the algorithm and demonstrate numerically its effectiveness when applied to two different problems. The first is a simple linear inverse problem on the simplex with total variation regularization, which we examine in the deterministic and stochastic case. The second is an application in optimal transport involving the entropically regularized Wasserstein distance and inverse problems. There is also a discussion of other possible applications of the algorithm to entropic Wasserstein barycenter problems.

2 Notations and preliminary facts

2.1 Basic notation

Given a real reflexive Banach space $\mathcal{X}$, we denote by $\mathcal{X}^*$ its topological dual and by $\langle u, x\rangle$ the duality pairing for $x \in \mathcal{X}$ and $u \in \mathcal{X}^*$. The norm on $\mathcal{X}$ is denoted $\|\cdot\|_{\mathcal{X}}$. The symbols $\rightharpoonup$ and $\to$ denote respectively weak and strong convergence. The set of weak sequential cluster points of a sequence $(x_k)_{k\in\mathbb{N}}$ in $\mathcal{X}$ is defined as

$$\mathfrak{W}\left[(x_k)_{k\in\mathbb{N}}\right] \stackrel{\text{def}}{=} \left\{x \in \mathcal{X} : \exists \left(x_{k_j}\right)_{j\in\mathbb{N}}, x_{k_j} \rightharpoonup x\right\}. \tag{2.1}$$

For a function $f \in \Gamma_0(\mathcal{X})$, $\partial f : \mathcal{X} \to 2^{\mathcal{X}}$ is its subdifferential operator. When referring to the differentiability or the gradient of a function $f : \mathcal{X} \to \mathbb{R}$, it is meant in the sense of Gâteaux. For a non-empty closed convex set $\mathcal{C} \subset \mathcal{X}$, $N_{\mathcal{C}}(x)$ is the normal cone of $\mathcal{C}$ at $x \in \mathcal{C}$. $\operatorname{int} \mathcal{C}$ and $\overline{\mathcal{C}}$ denote the interior and the closure of a set $\mathcal{C} \subset \mathcal{X}$.

2.2 Bregman divergence notation

We denote by D, without subscript, the Bregman divergence associated to $\phi(x,\mu) \stackrel{\text{def}}{=} \phi_p(x)+\phi_d(\mu)$; namely, given $w_i \stackrel{\text{def}}{=} (x_i, \mu_i)$ with $(x_i, \mu_i) \in \mathcal{X}_p \times \mathcal{X}_d$ for $i \in \{1, 2\}$,

$$D(w_1, w_2) \stackrel{\text{def}}{=} D_p(x_1, x_2) + D_d(\mu_1, \mu_2).$$

We proceed with some notions about regularity of functions.

Definition 2.1 (Legendre function). The function ϕ is called a Legendre function if $\partial\phi$ is both locally bounded and single-valued on its domain, $(\partial\phi)^{-1}$ is locally bounded on its domain, and ϕ is strictly convex on every convex subset of $\operatorname{dom}(\partial\phi)$.

Definition 2.2 (Relative smoothness). Given a function $\phi : \mathcal{X} \to \mathbb{R}\cup\{+\infty\}$ differentiable on $\operatorname{int}\operatorname{dom}(\phi)$, we say that the function $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is L-smooth with respect to ϕ if it is differentiable on $\operatorname{int}\operatorname{dom}(\phi)$ and $L\phi - f$ is convex on $\operatorname{int}\operatorname{dom}(\phi)$; namely, if for every $x, y \in \operatorname{int}\operatorname{dom}(\phi)$

$$D_f(x, y) \le LD_\phi(x, y).$$

Remark 2.3. The relative smoothness property, used notably in [31], [42] and [39], implies the following fact which can be interpreted as a "generalized descent lemma": for every $x, y \in \operatorname{int}\operatorname{dom}(\phi)$,

$$f(x) \le f(y) + \langle \nabla f(y), x - y\rangle + LD_\phi(x, y). \tag{2.2}$$

When ϕ is the euclidean square norm, or energy, relative smoothness is equivalent to Lipschitz-smoothness, i.e., Lipschitz-continuity of the gradient of f.

Definition 2.4 (Relative strong convexity). Given a function $\phi : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ differentiable on $\operatorname{int}\operatorname{dom}(\phi)$, and a non-empty closed convex set $\mathcal{C} \subset \operatorname{dom}(f) \cap \operatorname{dom}(\phi)$, we say that $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is m-strongly convex on $\mathcal{C}$ with respect to ϕ if for every $x \in \mathcal{C}$ and $y \in \operatorname{int}\operatorname{dom}(\phi) \cap \operatorname{dom}(\partial f) \cap \mathcal{C}$

$$f(x) - f(y) - \langle u,\, y - x\rangle \ge mD_\phi(x, y), \quad \text{for all} \quad u \in \partial f(y).$$

Note that the idea of relative strong convexity can be found in a footnote of [16] but it was not explored further. With these definitions, we use the following notation to improve readability

$$\begin{aligned}\left(\frac{1}{\Lambda_k} - L\right) D\left(w_1, w_2\right) &\stackrel{\text{def}}{=} \left(\frac{1}{\lambda_k} - L_p\right) D_p\left(x_1, x_2\right) + \left(\frac{1}{\nu_k} - L_d\right) D_d\left(\mu_1, \mu_2\right) \\ \left(\frac{1}{\Lambda_\infty} - L\right) D\left(w_1, w_2\right) &\stackrel{\text{def}}{=} \left(\frac{1}{\lambda_\infty} - L_p\right) D_p\left(x_1, x_2\right) + \left(\frac{1}{\nu_\infty} - L_d\right) D_d\left(\mu_1, \mu_2\right) \\ M(w_1, w_2) &\stackrel{\text{def}}{=} \langle T(x_1 - x_2), \mu_1 - \mu_2 \rangle\end{aligned} \tag{2.3}$$

where λ_k and ν_k are the step-sizes in Algorithm 1, and L_p and L_d are the constants introduced in $(\mathbf{A}_1)$.

2.3 Probabilistic notation and preliminaries

We denote by $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space with set of events Ω, σ-algebra $\mathcal{F}$, and probability measure $\mathbb{P}$. Throughout, we assume that any real reflexive Banach space $\mathcal{X}$ is endowed with its Borel σ-algebra, $\mathcal{B}(\mathcal{X})$. Formally, we define the stochastic primal and dual errors at iteration k as δ_k^p and δ_k^d, i.e., δ_k^p and δ_k^d are measurable functions from Ω to $\mathcal{X}_p^*$ and $\mathcal{X}_d^*$ with their respective Borel σ-algebras. When it makes sense, we will also denote the combined error as Δ_k in the same way that we use w_k, e.g.,

$$\langle \Delta_k, w - w_k \rangle \stackrel{\text{def}}{=} \langle \delta_k^p, x - x_k \rangle + \langle \delta_k^d, \mu - \mu_k \rangle.$$

We denote a filtration on $(\Omega, \mathcal{F}, \mathbb{P})$ by $\mathfrak{F} \stackrel{\text{def}}{=} (\mathscr{F}_k)_{k \in \mathbb{N}}$ where $\mathscr{F}_k$ is a sub-σ-algebra satisfying, for each $k \in \mathbb{N}$, $\mathscr{F}_k \subset \mathscr{F}_{k+1} \subset \mathcal{F}$. Furthermore, given a set of random variables $\{a_0, \ldots, a_n\}$ we denote by $\sigma(a_0, \ldots, a_n)$ the σ-algebra generated by $a_0, \ldots, a_n$. Finally, an expression (P) is said to hold $(\mathbb{P}\text{-a.s.})$ if

$$\mathbb{P}\left(\{\omega \in \Omega : (P) \text{ holds}\}\right) = 1.$$

Using the above notation, we denote the canonical filtration associated to the iterates of the algorithm as $\mathfrak{S} \stackrel{\text{def}}{=} (\mathscr{S}_k)_{k \in \mathbb{N}}$ with, for all $k \in \mathbb{N}$,

$$\mathscr{S}_k \stackrel{\text{def}}{=} \sigma\left\{(x_0, \mu_0), (x_1, \mu_1), \ldots, (x_k, \mu_k)\right\}$$

such that all iterates up to (x_k, μ_k) are completely determined by $\mathscr{S}_k$.

For the remainder of the paper, all equalities and inequalities involving random quantities should be understood as holding $(\mathbb{P}\text{-a.s.})$ even if it is not explicitly written.

Definition 2.5. Given a filtration $\mathfrak{F}$, we denote by $\ell_+(\mathfrak{F})$ the set of sequences of $[0, +\infty[$-valued random variables $(a_k)_{k \in \mathbb{N}}$ such that, for each $k \in \mathbb{N}$, a_k is $\mathscr{F}_k$ measurable. Then, we also define the following set of summable random variables,

$$\ell_+^1(\mathfrak{F}) \stackrel{\text{def}}{=} \left\{ (a_k)_{k \in \mathbb{N}} \in \ell_+(\mathfrak{F}) : \sum_{k \in \mathbb{N}} a_k < +\infty \ (\mathbb{P}\text{-a.s.}) \right\}.$$

The set of non-negative summable sequences is denoted ℓ_+^1.

The following probabilistic results will be useful in the convergence analysis of Algorithm 1.

Lemma 2.6 (Robbins-Siegmund, [49, Theorem 1]). *Given a filtration $\mathfrak{F}$ and the sequences of real-valued random variables $(r_k)_{k\in\mathbb{N}} \in \ell_+(\mathfrak{F})$, $(a_k)_{k\in\mathbb{N}} \in \ell_+(\mathfrak{F})$, and $(z_k)_{k\in\mathbb{N}} \in \ell_+^1(\mathfrak{F})$ satisfying, for each $k \in \mathbb{N}$,*

$$\mathbb{E}[r_{k+1} \mid \mathscr{F}_k] - r_k \leq -a_k + z_k \ (\mathbb{P}\text{-a.s.})$$

it holds that $(a_k)_{k\in\mathbb{N}} \in \ell_+^1(\mathfrak{F})$ and $(r_k)_{k\in\mathbb{N}}$ converges $(\mathbb{P}$-a.s.$)$ to a random variable with value in $[0, +\infty[$.

Remark 2.7. In the deterministic case, Lemma 2.6 reduces to the following statement. Let $(a_k)_{k\in\mathbb{N}} \in \ell_+, (r_k)_{k\in\mathbb{N}} \in \ell_+$ and $(z_k)_{k\in\mathbb{N}} \in \ell_+^1$ such that, for each $k \in \mathbb{N}$,

$$r_{k+1} - r_k \leq -a_k + z_k.$$

Then $(a_k)_{k\in\mathbb{N}} \in \ell_+^1$ and $(r_k)_{k\in\mathbb{N}}$ converges to $r \in [0, +\infty[$. This result is [47, Lemma 2, page 44].

Lemma 2.8. *If $(x_n)_{k\in\mathbb{N}}$ is a sequence of $\mathcal{X}$-valued random variables such that $\left(\mathbb{E}\left(\|x_k\|_{\mathcal{X}}^q\right)\right)_{k\in\mathbb{N}} \in \ell_+^1$ for some $q \in]0, +\infty[$, then $x_k \to 0$ almost surely.*

Proof. For every $\varepsilon > 0$, by Markov's inequality,

$$\sum_{n=0}^{N} \mathbb{P}\left(\|x_n\|_{\mathcal{X}}^q \geq \varepsilon\right) \leq \frac{1}{\varepsilon} \sum_{n=0}^{N} \mathbb{E}\left(\|x_n\|_{\mathcal{X}}^q\right). \tag{2.4}$$

Taking the limit for $N \to +\infty$ and using the assumption $\left(\mathbb{E}\left(\|x_k\|_{\mathcal{X}}^q\right)\right)_{k\in\mathbb{N}} \in \ell_+^1$, we get that, for every $\varepsilon > 0$, it holds $\mathbb{P}\left(\|x_n\|_{\mathcal{X}}^q \geq \varepsilon\right)$ belongs to ℓ_+^1. As a consequence of the Borel-Cantelli Lemma, $\|x_n\|_{\mathcal{X}}^q \to 0$ almost surely whence the claim follows. □

3 Main assumptions and estimations

3.1 Main assumptions

We first state our assumptions and then remark on their motivations and common examples where they hold. Note that for several results, only a subset of these assumptions are needed; we will comment ont this hereafter. For brevity, throughout the remainder of the paper we employ the following notation

$$\mathcal{U}_p \stackrel{\text{def}}{=} \operatorname{int}\operatorname{dom}(\phi_p) \cap \operatorname{dom}(\partial g) \qquad \tilde{\mathcal{U}}_p \stackrel{\text{def}}{=} \operatorname{dom}(\phi_p) \cap \operatorname{dom}(\partial g)$$
$$\mathcal{U}_d \stackrel{\text{def}}{=} \operatorname{int}\operatorname{dom}(\phi_d) \cap \operatorname{dom}(\partial l^*) \qquad \tilde{\mathcal{U}}_d \stackrel{\text{def}}{=} \operatorname{dom}(\phi_d) \cap \operatorname{dom}(\partial l^*).$$

- **($\mathbf{A}_1$)** The entropies ϕ_p and ϕ_d belong to $\Gamma_0(\mathcal{X}_p)$ and $\Gamma_0(\mathcal{X}_d)$ with $\operatorname{dom}(\phi_p) \times \operatorname{dom}(\phi_d) = \mathcal{C}_p \times \mathcal{C}_d$ and with f and h^* being L_p and L_d - smooth wrt ϕ_p and ϕ_d, respectively (see Definition 2.2). The D-prox mappings $(\nabla\phi_p + \lambda_k\partial g)^{-1}$ and $(\nabla\phi_d + \nu_k\partial l^*)^{-1}$ are well-defined (i.e., nonempty and single-valued) maps from $\operatorname{int}\operatorname{dom}(\phi_p)$ and $\operatorname{int}\operatorname{dom}(\phi_d)$ to $\operatorname{int}\operatorname{dom}(\phi_p)$ and $\operatorname{int}\operatorname{dom}(\phi_d)$, respectively.
- **($\mathbf{A}_2$)** The step size sequences $(\lambda_k)_{k\in\mathbb{N}}$ and $(\nu_k)_{k\in\mathbb{N}}$ are positive, nondecreasing, and bounded above with their limits denoted $\lim_{k\to\infty} \lambda_k = \lambda_\infty$ and $\lim_{k\to\infty} \nu_k = \nu_\infty$.
- **($\mathbf{A}_3$)** The step sizes satisfy **($\mathbf{A}_2$)** and one of the following holds:
  - **(I)** there is a function $d : (\mathcal{X}_p \times \mathcal{X}_d)^2 \to \mathbb{R}_+$ and $\varepsilon \geq 0$ such that

$$\inf_{\substack{w_1 \in \tilde{\mathcal{U}}_p \times \tilde{\mathcal{U}}_d,\ w_2 \in \mathcal{U}_p \times \mathcal{U}_d; \\ w_1 \neq w_2}} \frac{\left(\frac{1}{\Lambda_\infty} - L\right) D(w_1, w_2) - M(w_1, w_2)}{d(w_1, w_2)} \geq \varepsilon; \tag{3.1}$$

(II) the above holds with $\varepsilon > 0$.

($\mathbf{A}_4$) The error sequence $(\Delta_k)_{k\in\mathbb{N}}$ is unbiased conditioned on the filtration $\mathfrak{S}$, i.e., for each $k \in \mathbb{N}$,

$$\mathbb{E}\left[\delta_k^p \mid \mathscr{S}_k\right] = \mathbb{E}\left[\delta_k^d \mid \mathscr{S}_k\right] = 0.$$

($\mathbf{A}_5$) One of the following holds:

(I) for each $k \in \mathbb{N}$, the stochastic errors δ_k^p and δ_k^d are zero almost surely;

(II) the following sequences satisfy

$$\left(\mathbb{E}\left[\left\|\delta_k^p\right\|_{\mathcal{X}_p^*} \mid \mathscr{S}_k\right]\right)_{k\in\mathbb{N}} \in \ell_+^1(\mathfrak{S}) \text{ and } \left(\mathbb{E}\left[\left\|\delta_k^p\right\|_{\mathcal{X}_p^*}\right]\right)_{k\in\mathbb{N}} \in \ell_+^1$$
$$\left(\mathbb{E}\left[\left\|\delta_k^d\right\|_{\mathcal{X}_d^*} \mid \mathscr{S}_k\right]\right)_{k\in\mathbb{N}} \in \ell_+^1(\mathfrak{S}) \text{ and } \left(\mathbb{E}\left[\left\|\delta_k^d\right\|_{\mathcal{X}_d^*}\right]\right)_{k\in\mathbb{N}} \in \ell_+^1$$

and the sets $\mathcal{U}_p$ and $\mathcal{U}_d$ are bounded, i.e., $0 < \operatorname{diam}_{\mathcal{U}_p} < +\infty$ and the same for $\operatorname{diam}_{\mathcal{U}_d}$;

(III) the entropies ϕ_p and ϕ_d are strongly convex with respect to $\|\cdot\|_{\mathcal{X}_p}^2$ and $\|\cdot\|_{\mathcal{X}_d}^2$ with moduli m_p and m_d, respectively. Additionally, the step sizes $(\lambda_k)_{k\in\mathbb{N}}$ and $(\nu_k)_{k\in\mathbb{N}}$ satisfy **($\mathbf{A}_2$)** with

$$\nu_\infty \lambda_\infty < \frac{m_p m_d}{\|T\|_{p\to d^*}^2},$$

where $\|\cdot\|_{p\to d^*}$ is the standard operator norm between $\mathcal{X}_p$ and $\mathcal{X}_d^*$ and the following sequences satisfy

$$\mathbb{E}\left[\left\|\delta_k^p\right\|_{\mathcal{X}_p^*}^2 \mid \mathscr{S}_k\right] \in \ell_+^1(\mathfrak{S}) \text{ and } \mathbb{E}\left[\left\|\delta_k^p\right\|_{\mathcal{X}_p^*}^2\right] \in \ell_+^1.$$

($\mathbf{A}_6$) For the function d used in (3.1) and all bounded sequences $(v_k)_{k\in\mathbb{N}}$ and $(z_k)_{k\in\mathbb{N}}$ in $\operatorname{int}\operatorname{dom}(\phi) \stackrel{\text{def}}{=} \operatorname{int}\operatorname{dom}(\phi_p) \times \operatorname{int}\operatorname{dom}(\phi_d)$

$$d(v_k, z_k) \to 0 \quad \Rightarrow \quad v_k - z_k \to 0. \tag{3.2}$$

($\mathbf{A}_7$) For every $w \stackrel{\text{def}}{=} (x, \mu) \in \operatorname{int}\operatorname{dom}(\phi)$, at least one of $D(w, \cdot)$ or $d(w, \cdot)$ is coercive.

($\mathbf{A}_8$) For any bounded sequence $(w_k)_{k\in\mathbb{N}}$ with $w_k \in \operatorname{int}\operatorname{dom}\phi$ for each $k \in \mathbb{N}$, if $w_{k+1} - w_k \to 0$, then

$$\nabla\phi_p(x_{k+1}) - \nabla\phi_p(x_k) \to 0 \text{ and } \nabla f(x_{k+1}) - \nabla f(x_k) \to 0;$$
$$\nabla\phi_d(\mu_{k+1}) - \nabla\phi_d(\mu_k) \to 0 \text{ and } \nabla h^*(\mu_{k+1}) - \nabla h^*(\mu_k) \to 0.$$

($\mathbf{A}_9$) For any sequence $(w_k)_{k\in\mathbb{N}}$ with $w_k \in \operatorname{int}\operatorname{dom}\phi$, for each $k \in \mathbb{N}$, if $w_k \rightharpoonup w_\infty$, then

$$\nabla\phi(w_k) \rightharpoonup \nabla\phi(w_\infty).$$

($\mathbf{A}_{10}$) For an arbitrary sequence $(w_k)_{k\in\mathbb{N}} \in \mathcal{X}_p \times \mathcal{X}_d$, if $w_k \rightharpoonup 0$, then

$$\langle Tx_k, \mu_k\rangle \to 0.$$

($\mathbf{A}_{11}$) **(I)** At least one of the functions f or g is relatively strongly convex on $\mathcal{C}_p \cap \operatorname{dom}(g)$ wrt an entropy $\psi_p : \mathcal{X}_p \to \mathbb{R} \cup \{+\infty\}$ with constant m_f or m_g, respectively (see Definition 2.4). The entropy ψ_p satisfies $\operatorname{dom}(\phi_p) \subseteq \operatorname{dom}(\psi_p)$.

(II) At least one of the functions h^* or l^* is relatively strongly convex on $\mathcal{C}_d \cap \operatorname{dom}(l^*)$ wrt an entropy $\psi_d : \mathcal{X}_d \to \mathbb{R} \cup \{+\infty\}$ with constant m_{h^*} or m_{l^*}, respectively (see Definition 2.4). The entropy ψ_d satisfies $\operatorname{dom}(\phi_d) \subseteq \operatorname{dom}(\psi_d)$.

Analogously to (2.3), we will also use the shorhand notation using the relative strong convexity constants and entropies from $(\mathbf{A}_{11})$:

$$\begin{aligned} m_{(f,h^*)} D'(w_1, w_2) &\stackrel{\text{def}}{=} m_f D_{\psi_p}(x_1, x_2) + m_{h^*} D_{\psi_d}(\mu_1, \mu_2) \\ m_{(g,l^*)} D'(w_1, w_2) &\stackrel{\text{def}}{=} m_g D_{\psi_p}(x_1, x_2) + m_{l^*} D_{\psi_d}(\mu_1, \mu_2). \end{aligned} \tag{3.3}$$

Remark 3.1 ($(\mathbf{A}_1)$ and $(\mathbf{A}_2)$). There are several, technical characterizations of sufficient conditions that ensure the latter half of $(\mathbf{A}_1)$ holds. Classical examples start by assuming that the spaces are reflexive and that ϕ_p and ϕ_d are Legendre functions and then add assumptions depending on the space being considered; see for instance, the comprehesinve treatment in [4, Section 3]. Notice that we do not require ϕ_p and ϕ_d to be Legendre in general, that is indeed incompatible with $(\mathbf{A}_9)$ if the limit point is on the boundary. In practice, the latter half of $(\mathbf{A}_1)$ is required only for the existence and uniqueness of the sequence generated by the algorithm and is not used explicitly elsewhere in the convergence analysis. For $(\mathbf{A}_2)$, it is sufficient to take the step sizes $(\lambda_k)_{k\in\mathbb{N}}$ and $(\nu_k)_{k\in\mathbb{N}}$ to simply be constant.

Remark 3.2 ($(\mathbf{A}_3)$). The infimum in $(\mathbf{A}_3)$ is taken with $w_1 \in \tilde{\mathcal{U}}_p \times \tilde{\mathcal{U}}_d$ and $w_2 \in \mathcal{U}_p \times \mathcal{U}_d$ because, a priori, a solution $w^\star$ may lie in the boundary of $\tilde{\mathcal{U}}_p \times \tilde{\mathcal{U}}_d$ even if the iterates $(w_k)_{k\in\mathbb{N}}$ themselves remain in $\mathcal{U}_p \times \mathcal{U}_d$ due to $(\mathbf{A}_1)$. Since the Bregman divergence is still well defined when the first argument (but not the second) is in $\operatorname{dom}(\phi) \setminus \operatorname{int}\operatorname{dom}(\phi)$, there is no issue with taking the infimum over this set. Observe that $(\mathbf{A}_3)$ also entails that, for every $w_1 \in \tilde{\mathcal{U}}_p \times \tilde{\mathcal{U}}_d$ and $w_2 \in \mathcal{U}_p \times \mathcal{U}_d$, for each $k \in \mathbb{N}$,

$$\frac{1}{\Lambda_k} D(w_1, w_2) - M(w_1, w_2) \geq L D(w_1, w_2) + \varepsilon d(w_1, w_2) \geq 0. \tag{3.4}$$

Example 3.3. Suppose that $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is a convex nondecreasing function with φ^* its positive conjugate and γ a finite coercive gauge with domain $\mathbb{R}_+(\mathcal{U}_p - \mathcal{U}_p) \subset \mathcal{X}_d$ (in the Minkowski sense) and polar γ°. Assume that the quantities defined by

$$\|T\|_{D_p} \stackrel{\text{def}}{=} \sup_{x_1, x_2 \in \mathcal{U}_p;\ x_1 \neq x_2} \frac{\varphi(\gamma(T(x_2 - x_1)))}{D_p(x_1, x_2)} \quad \text{and} \quad \|I\|_{D_d} \stackrel{\text{def}}{=} \sup_{\mu_1, \mu_2 \in \mathcal{U}_d;\ \mu_1 \neq \mu_2} \frac{\varphi^*(\gamma^\circ(\mu_2 - \mu_1))}{D_d(\mu_1, \mu_2)}$$

are finite. We use the notation $\|\cdot\|_{D_p}$ and $\|\cdot\|_{D_p}$, but notice that they may not be norms. If, moreover, we suppose that the step sizes verify, for each $k \in \mathbb{N}$, for some $\varepsilon_k \geq 0$,

$$\left(\frac{1}{\lambda_k} - L_p\right) \geq \|T\|_{D_p} + \varepsilon_k \ \text{ and } \ \left(\frac{1}{\nu_k} - L_d\right) \geq \|I\|_{D_d} + \varepsilon_k, \tag{3.5}$$

then $(\mathbf{A}_3)$ is satisfied with $d(w_1, w_2) = D(w_1, w_2)$. Indeed, for any pair $w_1, w_2 \in \mathcal{U}_p \times \mathcal{U}_d$, we have, for

each $k \in \mathbb{N}$,

$$\begin{aligned}
&\left(\frac{1}{\Lambda_k} - L\right) D\left(w_1, w_2\right) - M\left(w_1, w_2\right) \\
= \;& \left(\frac{1}{\lambda_k} - L_p\right) D_p\left(x_1, x_2\right) + \left(\frac{1}{\nu_k} - L_d\right) D_d\left(\mu_1, \mu_2\right) - \langle T(x_1 - x_2), \mu_1 - \mu_2 \rangle \\
\geq \;& \|T\|_{D_p} D_p\left(x_1, x_2\right) + \|I\|_{D_d} D_d\left(\mu_1, \mu_2\right) - \gamma\left(T(x_1 - x_2)\right) \gamma^\circ\left(\mu_1 - \mu_2\right) + \varepsilon_k D\left(w_1, w_2\right) \\
\geq \;& \varphi(\gamma\left(T(x_1 - x_2)\right)) + \varphi^*(\gamma^\circ\left(\mu_1 - \mu_2\right)) - \varphi(\gamma\left(T(x_1 - x_2)\right)) - \varphi^*(\gamma^\circ\left(\mu_1 - \mu_2\right)) + \varepsilon_k D\left(w_1, w_2\right) \\
= \;& \varepsilon_k D\left(w_1, w_2\right).
\end{aligned} \tag{3.6}$$

Note that in this example we have taken the action of T on the primal variables into the definition of $\|\cdot\|_{D_p}$. It is equally possible, and sometimes desirable, to define things such that the action of the adjoint T^* on the dual variables is incorporated into $\|\cdot\|_{D_d}$ instead, which can change the values (and consequently step sizes) in a non-Hilbertian setting.

Remark 3.4 ($\mathbf{A_4}$) and ($\mathbf{A_5}$). Notice that, using Lemma 2.8, ($\mathbf{A_4}$) and ($\mathbf{A_5}$) (in any case) imply that $\left(\delta_k^p\right)_{k\in\mathbb{N}}$ and $\left(\delta_k^d\right)_{k\in\mathbb{N}}$ converge strongly (with respect to $\|\cdot\|_{\mathcal{X}_p^*}$ and $\|\cdot\|_{\mathcal{X}_d^*}$ respectively) to zero a.s. and that, furthermore, for any fixed $w \in \mathcal{C}_p \times \mathcal{C}_d$, $\left(\mathbb{E}\left[\langle \Delta_k, w - w_{k+1}\rangle\right]\right)_{k\in\mathbb{N}} \in \ell_+^1$ and $\left(\mathbb{E}\left[\langle \Delta_k, w - w_{k+1}\rangle \mid \mathscr{S}_k\right]\right)_{k\in\mathbb{N}} \in \ell_+^1\left(\mathfrak{S}\right)$ (see Lemma 3.14 for details). In ($\mathbf{A_5}$)(III), the norms $\|\cdot\|_{\mathcal{X}_p}$ and $\|\cdot\|_{\mathcal{X}_d}$ can be replaced with arbitrary norms as long as ϕ_p and ϕ_d are strongly convex with respect to their square. The different cases for ($\mathbf{A_5}$) can be mixed for the primal and dual, e.g., one can take ($\mathbf{A_5}$)(III) for the primal but have ($\mathbf{A_5}$)(II) for the dual; the current presentation simply for convenience.

Remark 3.5 ($\mathbf{A_6}$). In the case where $d\left(x, y\right)$ is the Bregman divergence induced by the Shannon-Boltzman entropy, the Hellinger entropy, the fractional-power entropy, the Fermi-Dirac entropy, or the energy/euclidean entropy, ($\mathbf{A_6}$) holds (see [31, Remark 4].

More generally, when $d = D_\zeta$ for some entropy ζ which is Legendre, we have from [4, Example 4.10] that ($\mathbf{A_6}$) is satisfied whenever one of the following holds

- ζ is uniformly convex on bounded sets;
- $\mathcal{X}_p \times \mathcal{X}_d$ is finite dimensional, $\mathrm{dom}\left(\zeta\right)$ is closed, and $\zeta\,|_{\mathrm{dom}(\zeta)}$ is strictly convex and continuous.

Thus, if $\zeta = \phi$, with ϕ Legendre, we require only $\mathrm{dom}\left(\phi\right)$ to be closed if $\mathcal{X}_p \times \mathcal{X}_d$ is finite dimensional.

Remark 3.6 ($\mathbf{A_7}$). Sufficient conditions for ($\mathbf{A_7}$) to hold for Legendre functions in real reflexive Banach spaces are given in [3, Lemma 7.3(viii) & (ix)]. For example, $D_p\left(x, \cdot\right)$ is coercive if ϕ_p is supercoercive and $x \in \mathrm{int\,dom}\left(\phi_p\right)$; alternatively, if $\mathcal{X}_p$ is finite-dimensional, $\mathrm{dom}\left(\phi_p^*\right)$ is open, and $x \in \mathrm{int\,dom}\left(\phi_p\right)$. Similar conditions hold for ϕ_d.

Remark 3.7 ($\mathbf{A_{10}}$). Assumption ($\mathbf{A_{10}}$) is very mild and holds when the operator T (or T^*) is for instance compact.

Remark 3.8 ((A$_8$), (A$_9$)). Assumptions (A$_8$), (A$_9$) and (A$_{10}$) are required only for the pointwise weak convergence of the iterates, namely in Section 4.3. (A$_8$) and (A$_9$) have been previously assumed by other authors to prove weak convergence of the iterates for the Bregman Forward-Backward algorithm on a real reflexive Banach space; see [42, 10]. In particular, (A$_9$) is a weak sequential continuity assumption on the gradients of the entropies, while (A$_8$) can be obtained for instance from norm-to-norm uniform continuity on bounded sets of $\nabla\phi_p$, $\nabla\phi_d$, ∇f, and ∇h^*. A typical example where these assumptions hold is when $\mathcal{X}_p$ is the ℓ_q space[2], $q \in]1, +\infty[$, and $\phi_p = \|\cdot\|_{\ell_q}^q / q$, in which case $\nabla\phi_p$ is the duality mapping on ℓ_q. The latter is known in this case to be weakly continuous [9] and norm-to-norm uniformly continuous on every bounded subset of ℓ^q [18]. However, if the duality mapping is replaced with the normalized duality mapping, i.e., $\phi_p = \|\cdot\|_{\ell_q}^2 / 2$, then (A$_9$) fails unless $q = 2$ (i.e., Hilbertian setting) while (A$_8$) still holds for ϕ_p; see [57].

On the other hand, (A$_8$) is satisfied when $\mathcal{X}_p \times \mathcal{X}_d$ is finite dimensional. Indeed, in finite dimension not only do strong and weak convergence coincide but also $\nabla\phi_p$, $\nabla\phi_d$, ∇f, and ∇h^* are all continuous on the interior of their domains by [51, Corollary 9.20] since $\phi_p, f \in \Gamma_0(\mathcal{X}_p)$ and $\phi_d, h^* \in \Gamma_0(\mathcal{X}_d)$. Again, (A$_9$) is more subtle even in finite dimension since Legenderness of the entropy entails that if an interior sequence converges to a point on the boundary of the domain of the entropy, the sequence of gradients will diverge.

We finish this section by providing an infinite-dimensional example where all asumptions hereabove are verified.

Example 3.9. We give an example of an infinite-dimensional Banach space $\mathcal{X}_p$ and an entropy ϕ_p for which assumptions (A$_1$), (A$_8$) and (A$_9$) both hold. Consider $\mathcal{H}$ an infinite-dimensional Hilbert space and $\mathcal{V}$ a finite-dimensional Banach space, with respective norms $\|\cdot\|_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{V}}$, and define $\mathcal{X}_p = \mathcal{H} \times \mathcal{V}$ to be the Banach space with norm $\|(h,v)\|_{\mathcal{X}_p} = \sqrt{\|h\|_{\mathcal{H}}^2 + \|v\|_{\mathcal{V}}^2}$. Let $\mathcal{C}_p = \mathcal{X}_p$, we can pick the entropy $\phi_p(x) = \frac{1}{2}\|x\|_{\mathcal{X}_p}^2$ whose gradient is given, for $x = (h,v)$, by $\nabla\phi_p(x) = h + J_{\mathcal{V}} v$ where $J_{\mathcal{V}}$ is the normalized duality mapping for $\mathcal{V}$. Then, if we assume that $\mathcal{V}$ is a smooth and rotund space as in [3, Lemma 6.2], $\mathcal{X}_p$ will be a smooth and rotund space and we will have that ϕ_p is Legendre, i.e., (A$_1$) will be satisfied. Since $\mathcal{V}$ is finite-dimensional and $\mathcal{X}_p$ is open, (A$_9$) is satisfied for $\nabla\phi_p$. Indeed, the limit point x^∞ cannot lie on the boundary since the boundary is empty while $\mathcal{V}$ being finite-dimensional guarantees the continuity of $J_{\mathcal{V}}$.

3.2 Main estimations

The following results constitute the main estimations that will be used in the convergence analysis of Algorithm 1.

Lemma 3.10. *Recall the notation of (2.3). Assume that (H) and (A$_1$)-(A$_3$) hold, then we have the following energy estimation. For every $w \overset{\text{def}}{=} (x,\mu) \in (\mathcal{C}_p \times \mathcal{C}_d) \cap (\mathrm{dom}(g) \times \mathrm{dom}(l^*))$, for each $k \in \mathbb{N}$,*

$$\mathcal{L}(x_{k+1},\mu) - \mathcal{L}(x,\mu_{k+1}) + \left[\frac{1}{\Lambda_{k+1}} D(w,w_{k+1}) - M(w,w_{k+1})\right] + \langle w_{k+1} - w, \Delta_k\rangle + \varepsilon d(w_{k+1},w_k) \\ \leq \left[\frac{1}{\Lambda_k} D(w,w_k) - M(w,w_k)\right]. \quad (3.7)$$

If, moreover, (A$_{11}$)(I) and (A$_{11}$)(II) hold, we have (using the notation of (3.3)) for every $w \overset{\text{def}}{=} (x,\mu) \in$

[2] We focus on the primal space $\mathcal{X}_p$ but the same reasoning applies to $\mathcal{X}_d$.

$(\mathcal{C}_p \times \mathcal{C}_d) \cap (\mathrm{dom}(g) \times \mathrm{dom}(l^*))$, *for each* $k \in \mathbb{N}$,

$$\begin{aligned}\mathcal{L}(x_{k+1}, \mu) - \mathcal{L}(x, \mu_{k+1}) + \left[\frac{1}{\Lambda_{k+1}} D(w, w_{k+1}) - M(w, w_{k+1})\right] + \langle w_{k+1} - w, \Delta_k \rangle + \varepsilon d(w_{k+1}, w_k) \\ \leq \left[\frac{1}{\Lambda_k} D(w, w_k) - M(w, w_k)\right] - m_{(f,h^*)} D'(w, w_k) - m_{(g,l^*)} D'(w, w_{k+1}). \quad (3.8)\end{aligned}$$

Proof. We will prove claim (3.8) since (3.7) is a special case of it when $m_{(f,h^*)} = m_{(g,l^*)} = 0$. For all $k \in \mathbb{N}$, the following holds by the definitions of x_{k+1} and μ_{k+1} in Algorithm 1,

$$\begin{aligned}&\frac{1}{\lambda_k}(\nabla\phi_p(x_k) - \nabla\phi_p(x_{k+1})) - \nabla f(x_k) - \delta_k^p - T^*\widetilde{\mu}_k \in \partial g(x_{k+1}) \\ &\frac{1}{\nu_k}(\nabla\phi_d(\mu_k) - \nabla\phi_d(\mu_{k+1})) - \nabla h^*(\mu_k) - \delta_k^d + T\widetilde{x}_k \in \partial l^*(\mu_{k+1}).\end{aligned} \quad (3.9)$$

Observe that by assumptions **(H)** and **($\mathbf{A}_{11}$)(I)**, we have $\mathcal{C}_p \cap \mathrm{dom}(g) = \mathrm{dom}(\phi_p) \cap \mathrm{dom}(g) \subset \mathrm{dom}(\psi_p) \cap \mathrm{dom}(g)$. Morover, using also that $\mathrm{dom}(\partial g) \subset \mathrm{dom}(g)$, we have $\forall k \in \mathbb{N}$, $x_k \in \mathrm{int\,dom}(\phi_p) \cap \mathrm{dom}(\partial g) = \mathrm{int\,dom}(\phi_p) \cap \mathrm{dom}(\partial g) \cap \mathcal{C}_p \cap \mathrm{dom}(g) \subset \mathrm{int\,dom}(\psi_p) \cap \mathrm{dom}(\partial g) \cap \mathcal{C}_p \cap \mathrm{dom}(g)$. A similar reasoning is also valid replacing $(\mathcal{C}_p, g, \phi_p, \psi_p)$ with their dual counterparts $(\mathcal{C}_d, l^*, \phi_d, \psi_d)$ and invoking **($\mathbf{A}_{11}$)(II)**. We are then in position to apply the relative strong convexity inequality of Definition 2.4, which holds at any $(x, \mu) \in (\mathcal{C}_p \times \mathcal{C}_d) \cap (\mathrm{dom}(g) \times \mathrm{dom}(l^*))$ and (x_{k+1}, μ_{k+1}), hence giving

$$\begin{aligned}g(x) &\geq g(x_{k+1}) + \langle u, x - x_{k+1} \rangle + m_g D_{\psi_p}(x, x_{k+1}) \\ l^*(\mu) &\geq l(\mu_{k+1}) + \langle v, \mu - \mu_{k+1} \rangle + m_{l^*} D_{\psi_d}(\mu, \mu_{k+1})\end{aligned} \quad (3.10)$$

for any $u \in \partial g(x_{k+1})$ and $v \in \partial l^*(\mu_{k+1})$. Combining (3.9) and (3.10) and applying the three-point identity for Bregman divergences [17, Lemma 3.1], we have

$$\begin{aligned}D_p(x, x_k) \geq \lambda_k \Big(g(x_{k+1}) - g(x) + \langle \nabla f(x_k) + \delta_k^p,\ x_{k+1} - x \rangle + \langle T(x_{k+1} - x),\ \widetilde{\mu}_k \rangle \Big) \\ + m_g \lambda_k D_{\psi_p}(x, x_{k+1}) + D_p(x, x_{k+1}) + D_p(x_{k+1}, x_k); \\ D_d(\mu, \mu_k) \geq \nu_k \Big(l^*(\mu_{k+1}) - l^*(\mu) + \langle \nabla h^*(\mu_k) + \delta_k^d,\ \mu_{k+1} - \mu \rangle - \langle T\widetilde{x}_k,\ \mu_{k+1} - \mu \rangle \Big) \\ + m_{l^*} \nu_k D_{\psi_d}(\mu, \mu_{k+1}) + D_d(\mu, \mu_{k+1}) + D_d(\mu_{k+1}, \mu_k).\end{aligned} \quad (3.11)$$

Moreover, from the relative smoothness assumed in **($\mathbf{A}_1$)** and the consequent generalized descent lemma (2.2), we have, for each $k \in \mathbb{N}$,

$$\begin{aligned}f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k),\ x_{k+1} - x_k \rangle + L_p D_p(x_{k+1}, x_k) \\ h^*(\mu_{k+1}) &\leq h^*(\mu_k) + \langle \nabla h^*(\mu_k),\ \mu_{k+1} - \mu_k \rangle + L_d D_d(\mu_{k+1}, \mu_k).\end{aligned} \quad (3.12)$$

To apply the relative strong convexity inequality to f and h^*, we again check the required qualification conditions of Definition 2.4. First, from **(H)** and **($\mathbf{A}_{11}$)(I)**, $\mathcal{C}_p \cap \mathrm{dom}(g) \subset \mathcal{C}_p = \mathrm{dom}(f) \cap \mathrm{dom}(\phi_p) \subset \mathrm{dom}(f) \cap \mathrm{dom}(\psi_p)$. In addition, $\forall k \in \mathbb{N}$, $x_k \in \mathrm{int\,dom}(\phi_p) \subset \mathrm{int\,dom}(\psi_p)$. Since f is differentiable on $\mathrm{int\,dom}(\phi_p)$, we have $\mathrm{int\,dom}(\phi_p) \subset \mathrm{dom}(\partial f)$, i.e, $x_k \in \mathrm{int\,dom}(\psi_p) \cap \mathrm{dom}(\partial f)$. We have also argued above that $x_k \in \mathrm{int\,dom}(\phi_p) \cap \mathrm{dom}(\partial g) \subset \mathcal{C}_p \cap \mathrm{dom}(g)$, and thus $x_k \in \mathrm{int\,dom}(\psi_p) \cap \mathrm{dom}(\partial f) \cap \mathcal{C}_p \cap \mathrm{dom}(g)$ as required to apply the relative strong convexity inequality of f at any $x \in \mathcal{C}_p \cap \mathrm{dom}(g)$ and

x_{k+1}. The same reasoning remains valid replacing $(\mathcal{C}_p, f, g, \phi_p, \psi_p)$ with $(\mathcal{C}_d, h^*, l^*, \phi_d, \psi_d)$ and invoking **($\mathbf{A}_{11}$)(II)**. We then have for any $(x, \mu) \in (\mathcal{C}_p \times \mathcal{C}_d) \cap (\mathrm{dom}(g) \times \mathrm{dom}(l^*))$, for each $k \in \mathbb{N}$,

$$\begin{aligned} f(x) &\geq f(x_k) + \langle \nabla f(x_k),\ x - x_k \rangle + m_f D_{\psi_p}(x, x_k) \\ h^*(\mu) &\geq h^*(\mu_k) + \langle \nabla h^*(\mu_k),\ \mu - \mu_k \rangle + m_{h^*} D_{\psi_d}(\mu, x_k)\,. \end{aligned} \tag{3.13}$$

Summing (3.12) and (3.13), we obtain, for each $(x, \mu) \in \mathcal{C}_p \times \mathcal{C}_d$, for each $k \in \mathbb{N}$,

$$\begin{aligned} f(x_{k+1}) &\leq f(x) + \langle \nabla f(x_k),\ x_{k+1} - x \rangle + L_p D_p(x_{k+1}, x_k) - m_f D_{\psi_p}(x, x_k) \\ h^*(\mu_{k+1}) &\leq h^*(\mu) + \langle \nabla h^*(\mu_k),\ \mu_{k+1} - \mu \rangle + L_d D_d(\mu_{k+1}, \mu_k) - m_{h^*} D_{\psi_d}(\mu, \mu_k)\,. \end{aligned}$$

Summing the latter with (3.11), we obtain

$$\begin{aligned} &\lambda_k \left[f(x_{k+1}) + g(x_{k+1}) - f(x) - g(x) + \langle T(x_{k+1} - x),\ \widetilde{\mu}_k \rangle + \left\langle x_{k+1} - x, \delta_k^p \right\rangle \right] + D_p(x, x_{k+1}) \\ &\quad + (1 - L_p \lambda_k) D_p(x_{k+1}, x_k) \leq D_p(x, x_k) - m_f \lambda_k D_{\psi_p}(x, x_k) - m_g \lambda_k D_{\psi_p}(x, x_{k+1})\,; \end{aligned}$$

$$\begin{aligned} &\nu_k \left[h^*(\mu_{k+1}) + l^*(\mu_{k+1}) - h^*(\mu) - l^*(\mu) - \langle T\widetilde{x}_k,\ \mu_{k+1} - \mu \rangle + \left\langle \mu_{k+1} - \mu, \delta_k^d \right\rangle \right] + D_d(\mu, \mu_{k+1}) \\ &\quad + (1 - L_d \nu_k) D_d(\mu_{k+1}, \mu_k) \leq D_d(\mu, \mu_k) - m_{h^*} \nu_k D_{\psi_d}(\mu, \mu_k) - m_{l^*} \nu_k D_{\psi_d}(\mu, \mu_{k+1})\,. \end{aligned}$$

Recall the notations of (2.3), (3.3), and that

$$\langle w_1 - w_2,\ \Delta_k \rangle \overset{\text{def}}{=} \langle x_1 - x_2,\ \delta_k^p \rangle + \langle \mu_1 - \mu_2,\ \delta_k^d \rangle,$$

then, for each $(x, \mu) \in \mathcal{C}_p \times \mathcal{C}_d$, for each $k \in \mathbb{N}$,

$$\begin{aligned} &\mathcal{L}(x_{k+1}, \mu) - \mathcal{L}(x, \mu_{k+1}) + \langle T(x_{k+1} - x), \widetilde{\mu}_k \rangle - \langle T\widetilde{x}_k, \mu_{k+1} - \mu \rangle + \langle w_{k+1} - w, \Delta_k \rangle \\ &\qquad + \frac{1}{\Lambda_k} D(w, w_{k+1}) - \frac{1}{\Lambda_k} D(w, w_k) + \left(\frac{1}{\Lambda_k} - L \right) D(w_{k+1}, w_k) \\ &\qquad\qquad \leq \langle Tx_{k+1}, \mu \rangle - \langle Tx, \mu_{k+1} \rangle - m_{(g,l^*)} D'(w, w_{k+1}) - m_{(f,h^*)} D'(w, w_k)\,. \end{aligned}$$

Rearranging the terms, we arrive at

$$\begin{aligned} &\mathcal{L}(x_{k+1}, \mu) - \mathcal{L}(x, \mu_{k+1}) + \frac{1}{\Lambda_k} D(w, w_{k+1}) - \frac{1}{\Lambda_k} D(w, w_k) + \left(\frac{1}{\Lambda_k} - L \right) D(w_{k+1}, w_k) + \langle w_{k+1} - w, \Delta_k \rangle \\ &\leq \langle Tx_{k+1}, \mu - \widetilde{\mu}_k \rangle + \langle T(\widetilde{x}_k - x), \mu_{k+1} \rangle + \langle Tx, \widetilde{\mu}_k \rangle - \langle T\widetilde{x}_k, \mu \rangle - m_{(g,l^*)} D'(w, w_{k+1}) - m_{(f,h^*)} D'(w, w_k) \\ &= \langle T(x_{k+1} - x), \mu - \widetilde{\mu}_k \rangle + \langle T(\widetilde{x}_k - x), \mu_{k+1} - \mu \rangle - m_{(g,l^*)} D'(w, w_{k+1}) - m_{(f,h^*)} D'(w, w_k)\,. \end{aligned}$$

Now we use that $\widetilde{x}_k = 2x_{k+1} - x_k$ and $\widetilde{\mu}_k = \mu_k$, to obtain

$$\begin{aligned} &\mathcal{L}(x_{k+1}, \mu) - \mathcal{L}(x, \mu_{k+1}) + \frac{1}{\Lambda_k} D(w, w_{k+1}) - \frac{1}{\Lambda_k} D(w, w_k) + \left(\frac{1}{\Lambda_k} - L \right) D(w_{k+1}, w_k) + \langle w_{k+1} - w, \Delta_k \rangle \\ &\leq \langle T(x_{k+1} - x), \mu - \mu_k \rangle + \langle T(x_{k+1} - x), \mu_{k+1} - \mu \rangle + \langle T(x_{k+1} - x_k), \mu_{k+1} - \mu \rangle - m_{(g,l^*)} D'(w, w_{k+1}) \\ &\quad - m_{(f,h^*)} D'(w, w_k) \\ &= \Big(\langle T(x_{k+1} - x_k), \mu_{k+1} - \mu_k \rangle + \langle T(x - x_{k+1}), \mu - \mu_{k+1} \rangle - \langle T(x - x_k), \mu - \mu_k \rangle \Big) - m_{(g,l^*)} D'(w, w_{k+1}) \\ &\quad - m_{(f,h^*)} D'(w, w_k)\,. \end{aligned}$$

Equivalently, recalling that $M(w_1, w_2) \overset{\text{def}}{=} \langle T(x_1 - x_2), \mu_1 - \mu_2 \rangle$, we get

$$\begin{aligned}\mathcal{L}(x_{k+1}, \mu) - \mathcal{L}(x, \mu_{k+1}) + \langle w_{k+1} - w, \Delta_k \rangle + \left[\frac{1}{\Lambda_k} D(w, w_{k+1}) - M(w, w_{k+1})\right] \\ - \left[\frac{1}{\Lambda_k} D(w, w_k) - M(w, w_k)\right] + \left[\left(\frac{1}{\Lambda_k} - L\right) D(w_{k+1}, w_k) - M(w_{k+1}, w_k)\right] \\ \leq -m_{(g,l^*)} D'(w, w_{k+1}) - m_{(f,h^*)} D'(w, w_k). \end{aligned} \tag{3.14}$$

Recall that, by $(\mathbf{A}_2)$, $(\lambda_k)_{k\in\mathbb{N}}$ and $(\nu_k)_{k\in\mathbb{N}}$ are nondecreasing sequences, and thus

$$\frac{1}{\Lambda_{k+1}} D(w, w_{k+1}) \leq \frac{1}{\Lambda_k} D(w, w_{k+1}). \tag{3.15}$$

Finally, combining (3.14) with (3.15) and $(\mathbf{A}_3)$ applied at the points w_{k+1} and w_k, we get (3.8). □

Lemma 3.11. *Assume* **(H)** *and* $(\mathbf{A}_1)$*-*$(\mathbf{A}_3)$ *hold and, for each* $k \in \mathbb{N}$*, denote by* $\widehat{w}_{k+1}$ *the exact update of the algorithm, i.e.,*

$$\widehat{w}_{k+1} = \begin{pmatrix} \widehat{x}_{k+1} \\ \widehat{\mu}_{k+1} \end{pmatrix} = \begin{pmatrix} (\nabla\phi_p + \lambda_k \partial g)^{-1} (\nabla\phi_p(x_k) - \lambda_k (\nabla f(x_k)) - \lambda_k T^* \mu_k) \\ (\nabla\phi_d + \nu_k \partial l^*)^{-1} (\nabla\phi_d(\mu_k) - \nu_k (\nabla h^*(\mu_k)) + \nu_k T(2\widehat{x}_{k+1} - x_k)) \end{pmatrix}. \tag{3.16}$$

Then, the following holds, for each $k \in \mathbb{N}$*,*

$$\langle \Delta_k, \widehat{w}_{k+1} - w_{k+1} \rangle \geq \frac{1}{\Lambda_k} \left(D(\widehat{w}_{k+1}, w_{k+1}) + D(w_{k+1}, \widehat{w}_{k+1})\right) - 2M(\widehat{w}_{k+1}, w_{k+1}) \geq 0. \tag{3.17}$$

Proof. By design of the algorithm, the following monotone inclusions hold, for each $k \in \mathbb{N}$,

$$\begin{aligned} \nabla\phi_p(x_k) - \lambda_k (\nabla f(x_k) - T^* \mu_k) - \nabla\phi_p(\widehat{x}_{k+1}) \in \lambda_k \partial g(\widehat{x}_{k+1}) \\ \nabla\phi_p(x_k) - \lambda_k \left(\nabla f(x_k) + \delta_k^p - T^* \mu_k\right) - \nabla\phi_p(x_{k+1}) \in \lambda_k \partial g(x_{k+1}). \end{aligned} \tag{3.18}$$

and similarly for the dual

$$\begin{aligned} \nabla\phi_d(\mu_k) - \nu_k (\nabla h^*(\mu_k) + T(2\widehat{x}_{k+1} - x_k)) - \nabla\phi_d(\widehat{\mu}_{k+1}) \in \nu_k \partial l^*(\widehat{\mu}_{k+1}) \\ \nabla\phi_d(\mu_k) - \nu_k \left(\nabla h^*(\mu_k) + \delta_k^d + T(2x_{k+1} - x_k)\right) - \nabla\phi_d(\mu_{k+1}) \in \nu_k \partial l^*(\mu_{k+1}). \end{aligned} \tag{3.19}$$

By monotonicity of the operators ∂l^* and ∂g combined with (3.19) and (3.18), we then have, for each $k \in \mathbb{N}$,

$$\begin{aligned} \langle \widehat{x}_{k+1} - x_{k+1}, \delta_k^p \lambda_k - \nabla\phi_p(\widehat{x}_{k+1}) + \nabla\phi_p(x_{k+1}) \rangle \geq 0 \\ \left\langle \widehat{\mu}_{k+1} - \mu_{k+1}, \delta_k^d \nu_k - \nabla\phi_d(\widehat{\mu}_{k+1}) + \nabla\phi_d(\mu_{k+1}) + 2\nu_k T(\widehat{x}_{k+1} - x_{k+1}) \right\rangle \geq 0. \end{aligned} \tag{3.20}$$

We can rewrite the above using Definition 1.1 to have, for each $k \in \mathbb{N}$,

$$\begin{aligned} \langle \widehat{x}_{k+1} - x_{k+1}, \delta_k^p \rangle \geq \frac{1}{\lambda_k} \left(D_p(\widehat{x}_{k+1}, x_{k+1}) + D_p(x_{k+1}, \widehat{x}_{k+1})\right) \\ \left\langle \widehat{\mu}_{k+1} - \mu_{k+1}, \delta_k^d \right\rangle \geq \frac{1}{\nu_k} \left(D_d(\widehat{\mu}_{k+1}, \mu_{k+1}) + D_d(\mu_{k+1}, \widehat{\mu}_{k+1})\right) - 2\langle T(\widehat{x}_{k+1} - x_{k+1}), \widehat{\mu}_{k+1} - \mu_{k+1} \rangle. \end{aligned} \tag{3.21}$$

Adding the above inequalities together gives, for each $k \in \mathbb{N}$,

$$\langle \Delta_k, \widehat{w}_{k+1} - w_{k+1} \rangle \geq \frac{1}{\Lambda_k} \left(D\left(\widehat{w}_{k+1}, w_{k+1}\right) + D\left(w_{k+1}, \widehat{w}_{k+1}\right) \right) - 2M\left(\widehat{w}_{k+1}, w_{k+1}\right). \tag{3.22}$$

Using ($\mathbf{A}_3$) and (3.4), and the fact that M is symmetric wrt its arguments, for each $k \in \mathbb{N}$,

$$\frac{1}{\Lambda_k} \left(D\left(\widehat{w}_{k+1}, w_{k+1}\right) + D\left(w_{k+1}, \widehat{w}_{k+1}\right) \right) - 2M\left(\widehat{w}_{k+1}, w_{k+1}\right) \geq 0.$$

□

Lemma 3.12. *Assume* (H), ($\mathbf{A}_1$)-($\mathbf{A}_3$), *and* ($\mathbf{A}_5$)(III) *all hold. One can choose* $a > 0$ *so that, for each* $k \in \mathbb{N}$,

$$\frac{m_p}{\lambda_k} - \frac{\|T\|_{p \to d^*}^2}{a} > 0 \ \text{ and } \ \frac{m_d}{\nu_k} - a > 0$$

and the following holds, for each $k \in \mathbb{N}$,

$$\langle \Delta_k, \widehat{w}_{k+1} - w_{k+1} \rangle \leq \left(\frac{m_p}{\lambda_k} - \frac{\|T\|_{p \to d^*}^2}{a} \right)^{-1} \left\| \delta_k^p \right\|_{\mathcal{X}_p^*}^2 + \left(\frac{m_d}{\nu_k} - a \right)^{-1} \left\| \delta_k^d \right\|_{\mathcal{X}_d^*}^2.$$

Proof. It follows from the strong convexity of ϕ_p and ϕ_d given by ($\mathbf{A}_5$)(III) that, for each $k \in \mathbb{N}$,

$$\begin{aligned} \frac{1}{\Lambda_k} \left(D\left(w_{k+1}, \widehat{w}_{k+1}\right) + D\left(\widehat{w}_{k+1}, w_{k+1}\right) \right) &= \frac{1}{\Lambda_k} \left\langle \nabla \phi\left(w_{k+1}\right) - \nabla \phi\left(\widehat{w}_{k+1}\right), w_{k+1} - \widehat{w}_{k+1} \right\rangle \\ &\geq \frac{m_p}{\lambda_k} \left\| \widehat{x}_{k+1} - x_{k+1} \right\|_{\mathcal{X}_p}^2 + \frac{m_d}{\nu_k} \left\| \widehat{\mu}_{k+1} - \mu_{k+1} \right\|_{\mathcal{X}_d}^2. \end{aligned} \tag{3.23}$$

Substituting this result into Lemma 3.11 (3.17) and applying Young's inequality with $a > 0$ we get, for each $k \in \mathbb{N}$,

$$\begin{aligned} &\langle \Delta_k, \widehat{w}_{k+1} - w_{k+1} \rangle \\ &\geq \frac{m_p}{\lambda_k} \left\| \widehat{x}_{k+1} - x_{k+1} \right\|_{\mathcal{X}_p}^2 + \frac{m_d}{\nu_k} \left\| \widehat{\mu}_{k+1} - \mu_{k+1} \right\|_{\mathcal{X}_d}^2 - 2M\left(\widehat{w}_{k+1}, w_{k+1}\right) \\ &= \frac{m_p}{\lambda_k} \left\| \widehat{x}_{k+1} - x_{k+1} \right\|_{\mathcal{X}_p}^2 + \frac{m_d}{\nu_k} \left\| \widehat{\mu}_{k+1} - \mu_{k+1} \right\|_{\mathcal{X}_d}^2 - 2\left(\left\langle T\left(\widehat{x}_{k+1} - x_{k+1}\right), \widehat{\mu}_{k+1} - \mu_{k+1} \right\rangle \right) \\ &\geq \frac{m_p}{\lambda_k} \left\| \widehat{x}_{k+1} - x_{k+1} \right\|_{\mathcal{X}_p}^2 + \frac{m_d}{\nu_k} \left\| \widehat{\mu}_{k+1} - \mu_{k+1} \right\|_{\mathcal{X}_d}^2 - \frac{\|T\|_{p \to d^*}^2}{a} \left\| \widehat{x}_{k+1} - x_{k+1} \right\|_{\mathcal{X}_p}^2 - a \left\| \widehat{\mu}_{k+1} - \mu_{k+1} \right\|_{\mathcal{X}_d}^2 \\ &= \left(\frac{m_p}{\lambda_k} - \frac{\|T\|_{p \to d^*}^2}{a} \right) \left\| \widehat{x}_{k+1} - x_{k+1} \right\|_{\mathcal{X}_p}^2 + \left(\frac{m_d}{\nu_k} - a \right) \left\| \widehat{\mu}_{k+1} - \mu_{k+1} \right\|_{\mathcal{X}_d}^2. \end{aligned} \tag{3.24}$$

Then, since the step size sequences $(\lambda_k)_{k \in \mathbb{N}}$ and $(\nu_k)_{k \in \mathbb{N}}$ are bounded and nondecreasing by ($\mathbf{A}_2$), and furthemore by ($\mathbf{A}_5$)(III) are chosen small enough to satisfy

$$\nu_\infty \lambda_\infty < \frac{m_p m_d}{\|T\|_{p \to d^*}^2},$$

one can choose $a > 0$ so that

$$\frac{m_p}{\lambda_\infty} - \frac{\|T\|^2_{p\to d^*}}{a} > 0 \text{ and } \frac{m_d}{\nu_\infty} - a > 0$$

and, by extension under **(A$_2$)**, for each $k \in \mathbb{N}$,

$$\frac{m_p}{\lambda_k} - \frac{\|T\|^2_{p\to d^*}}{a} > 0 \text{ and } \frac{m_d}{\nu_k} - a > 0.$$

Finally, we apply Young's inequality twice to the following to find, for each $k \in \mathbb{N}$,

$$\begin{aligned}
\langle \Delta_k, \widehat{w}_{k+1} - w_{k+1}\rangle &= \left\langle \delta_k^p, \widehat{x}_{k+1} - x_{k+1}\right\rangle + \left\langle \delta_k^d, \widehat{\mu}_{k+1} - \mu_{k+1}\right\rangle \\
&\leq \frac{1}{2}\left(\frac{m_p}{\lambda_k} - \frac{\|T\|^2_{p\to d^*}}{a}\right)^{-1} \left\|\delta_k^p\right\|^2_{\mathcal{X}_p^*} + \frac{1}{2}\left(\frac{m_p}{\lambda_k} - \frac{\|T\|^2_{p\to d^*}}{a}\right) \|\widehat{x}_{k+1} - x_{k+1}\|^2_{\mathcal{X}_p} \\
&\quad + \frac{1}{2}\left(\frac{m_d}{\nu_k} - a\right)^{-1} \left\|\delta_k^d\right\|^2_{\mathcal{X}_d^*} + \frac{1}{2}\left(\frac{m_d}{\nu_k} - a\right) \|\widehat{\mu}_{k+1} - \mu_{k+1}\|^2_{\mathcal{X}_d} \\
&\leq \frac{1}{2}\left(\frac{m_p}{\lambda_k} - \frac{\|T\|^2_{p\to d^*}}{a}\right)^{-1} \left\|\delta_k^p\right\|^2_{\mathcal{X}_p^*} + \frac{1}{2}\left(\frac{m_d}{\nu_k} - a\right)^{-1} \left\|\delta_k^d\right\|^2_{\mathcal{X}_d^*} \\
&\quad + \frac{1}{2}\langle \Delta_k, \widehat{w}_{k+1} - w_{k+1}\rangle
\end{aligned}$$

and the desired claim follows. □

Remark 3.13. In Lemma 3.12, one can instead choose to use $\|T^*\|^2_{d\to p^*}$ to have, for each $k \in \mathbb{N}$,

$$\langle \Delta_k, \widehat{w}_{k+1} - w_{k+1}\rangle \leq \left(\frac{m_p}{\lambda_k} - \frac{1}{a}\right) \left\|\delta_k^p\right\|^2_{\mathcal{X}_p^*} + \left(\frac{m_d}{\nu_k} - a\|T^*\|^2_{d\to p^*}\right) \left\|\delta_k^d\right\|^2_{\mathcal{X}_d^*}$$

if there is asymmtery in the size of m_p and m_d.

In the event that only ϕ_p is strongly convex with respect to $\|\cdot\|^2_{\mathcal{X}_p}$ but the analog does not hold for ϕ_d, we can make the following argument. Take (3.21) from Lemma 3.11 and use strong convexity, to get

$$\left\langle \widehat{x}_{k+1} - x_{k+1}, \delta_k^p\right\rangle \geq \frac{1}{\lambda_k}\left(D_p\left(\widehat{x}_{k+1}, x_{k+1}\right) + D_p\left(x_{k+1}, \widehat{x}_{k+1}\right)\right) \geq \frac{m_p}{\lambda_k}\|\widehat{x}_{k+1} - x_{k+1}\|^2_{\mathcal{X}_p}$$

and so, by Cauchy-Schwarz,

$$\|\widehat{x}_{k+1} - x_{k+1}\|_{\mathcal{X}_p} \leq \frac{\lambda_k}{m_p}\left\|\delta_k^p\right\|_{\mathcal{X}_p^*}.$$

Then, using again Cauchy-Schwarz and the previous inequality,

$$\left\langle \delta_k^p, \widehat{x}_{k+1} - x_{k+1}\right\rangle \leq \left\|\delta_k^p\right\|_{\mathcal{X}_p^*} \|\widehat{x}_{k+1} - x_{k+1}\|_{\mathcal{X}_p} \leq \frac{\lambda_k}{m_p}\left\|\delta_k^p\right\|^2_{\mathcal{X}_p^*}$$

without the restriction on λ_∞ and ν_∞ imposed in Lemma 3.12 because we no longer to need control the term $2M\left(\widehat{w}_{k+1}, w_{k+1}\right)$. This term, $2M\left(\widehat{w}_{k+1}, w_{k+1}\right)$, is a result of the way we have defined $\widehat{\mu}_{k+1}$ to depend on $\widehat{x}_{k+1}$, which is necessary to keep $\widehat{w}_{k+1}$ deterministic conditioned on the filtration $\mathscr{S}_k$. Thus, if only one of the entropies can be chosen to be strongly convex, one is inclined to formulate the problem in such a way that the primal problem has the strongly convex entropy, and to deal with the dual problem using **(A$_5$)(I)** or **(A$_5$)(II)** for the dual.

Lemma 3.14. *Under* **(H)** *and* **(A$_1$)**-**(A$_5$)**, *the following sequences satisfy, for any fixed* $w \in \mathcal{C}_p \times \mathcal{C}_d$,

$$(\mathbb{E}\left[\langle \Delta_k, w - w_{k+1}\rangle \mid \mathscr{S}_k\right])_{k\in\mathbb{N}} \in \ell^1_+(\mathfrak{S}) \quad \text{and} \quad (\mathbb{E}\left[\langle \Delta_k, w - w_{k+1}\rangle\right])_{k\in\mathbb{N}} \in \ell^1_+.$$

Proof. The assumption **(A$_5$)** has three cases with the first, **(A$_5$)(I)**, corresponding to the deterministic setting, i.e., the lemma holds trivially. For both of the following two cases we note that, by Lemma 3.11, for each $k \in \mathbb{N}$, for any fixed $w \in \mathcal{C}_p \times \mathcal{C}_d$,

$$\begin{aligned}\mathbb{E}\left[\langle \Delta_k, w - w_{k+1}\rangle \mid \mathscr{S}_k\right] &= \mathbb{E}\left[\langle \Delta_k, w - \widehat{w}_{k+1}\rangle + \langle \Delta_k, \widehat{w}_{k+1} - w_{k+1}\rangle \mid \mathscr{S}_k\right] \\ &= \mathbb{E}\left[\langle \Delta_k, \widehat{w}_{k+1} - w_{k+1}\rangle \mid \mathscr{S}_k\right] \geq 0\end{aligned} \tag{3.25}$$

since, due to **(A$_4$)**, Δ_k is unbiased conditioned on the filtration $\mathscr{S}_k$. By the law of total expectation applied to the above, it follows that, for each $k \in \mathbb{N}$, for any fixed $w \in \mathcal{C}_p \times \mathcal{C}_d$,

$$\mathbb{E}\left[\langle \Delta_k, w - w_{k+1}\rangle\right] = \mathbb{E}\left[\langle \Delta_k, \widehat{w}_{k+1} - w_{k+1}\rangle\right] \geq 0$$

and thus the following sequences satisfy, for any fixed $w \in \mathcal{C}_p \times \mathcal{C}_d$,

$$(\mathbb{E}\left[\langle \Delta_k, w - w_{k+1}\rangle \mid \mathscr{S}_k\right])_{k\in\mathbb{N}} \in \ell_+(\mathfrak{S}) \quad \text{and} \quad (\mathbb{E}\left[\langle \Delta_k, w - w_{k+1}\rangle\right])_{k\in\mathbb{N}} \in \ell_+.$$

Now assume that **(A$_5$)(II)** holds, recall that, for each $k \in \mathbb{N}$,

$$\langle \Delta_k, \widehat{w}_{k+1} - w_{k+_1}\rangle \overset{\text{def}}{=} \left\langle \delta^p_k, \widehat{x}_{k+1} - x_{k+1}\right\rangle + \left\langle \delta^d_k, \widehat{\mu}_{k+1} - \mu_{k+1}\right\rangle.$$

By **(A$_5$)(II)**, the sets $\mathcal{U}_p$ and $\mathcal{U}_d$ are bounded and thus have finite diameters, $\mathrm{diam}_{\mathcal{U}_p}$ and $\mathrm{diam}_{\mathcal{U}_d}$ respectively. Furthermore, by **(A$_1$)** and the definition of the updates in the algorithm, the exact update $\widehat{w}_{k+1}$ will remain in $\mathcal{U}_p \times \mathcal{U}_d$ for all $k \in \mathbb{N}$. Then, for each $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\left[\left\langle \delta^p_k, \widehat{x}_{k+1} - x_{k+1}\right\rangle \mid \mathscr{S}_k\right] &\leq \mathbb{E}\left[\left\|\delta^p_k\right\|_{\mathcal{X}_p^*} \left\|\widehat{x}_{k+1} - x_{k+1}\right\|_{\mathcal{X}_p} \mid \mathscr{S}_k\right] \leq \mathrm{diam}_{\mathcal{U}_p}\mathbb{E}\left[\left\|\delta^p_k\right\|_{\mathcal{X}_p^*} \mid \mathscr{S}_k\right]; \\ \mathbb{E}\left[\left\langle \delta^d_k, \widehat{\mu}_{k+1} - \mu_{k+1}\right\rangle \mid \mathscr{S}_k\right] &\leq \mathbb{E}\left[\left\|\delta^d_k\right\|_{\mathcal{X}_d^*} \left\|\widehat{\mu}_{k+1} - \mu_{k+1}\right\|_{\mathcal{X}_d} \mid \mathscr{S}_k\right] \leq \mathrm{diam}_{\mathcal{U}_d}\mathbb{E}\left[\left\|\delta^d_k\right\|_{\mathcal{X}_d^*} \mid \mathscr{S}_k\right].\end{aligned}$$

Since $\left(\mathbb{E}\left[\left\|\delta^p_k\right\|_{\mathcal{X}_p^*} \mid \mathscr{S}_k\right]\right)_{k\in\mathbb{N}} \in \ell^1_+(\mathfrak{S})$ and $\left(\mathbb{E}\left[\left\|\delta^d_k\right\|_{\mathcal{X}_d^*} \mid \mathscr{S}_k\right]\right)_{k\in\mathbb{N}} \in \ell^1_+(\mathfrak{S})$ by **(A$_5$)(II)**, and noting (3.25), it holds that, for any fixed $w \in \mathcal{C}_p \times \mathcal{C}_d$,

$$(\mathbb{E}\left[\langle \Delta_k, w - w_{k+1}\rangle \mid \mathscr{S}_k\right])_{k\in\mathbb{N}} \in \ell^1_+(\mathfrak{S}).$$

Using the same argument with the law of total expectation together with the fact that $\left(\mathbb{E}\left[\left\|\delta^p_k\right\|_{\mathcal{X}_p^*}\right]\right)_{k\in\mathbb{N}} \in \ell^1_+$ and $\left(\mathbb{E}\left[\left\|\delta^d_k\right\|_{\mathcal{X}_d^*}\right]\right)_{k\in\mathbb{N}} \in \ell^1_+$ by **(A$_5$)(II)**, it then follows that, for any fixed $w \in \mathcal{C}_p \times \mathcal{C}_d$,

$$(\mathbb{E}\left[\langle \Delta_k, w - w_{k+1}\rangle\right])_{k\in\mathbb{N}} \in \ell^1_+.$$

Finally, in the case of **(A$_5$)(III)**, we assume that the entropies ϕ_p and ϕ_d are strongly convex with respect to $\|\cdot\|^2_{\mathcal{X}_p}$ and $\|\cdot\|^2_{\mathcal{X}_d}$ respectively. Using Lemma 3.12 and taking expectation conditioned on $\mathscr{S}_k$, we have, for each $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\left[\langle \Delta_k, \widehat{w}_{k+1} - w_{k+1}\rangle \mid \mathscr{S}_k\right] &\leq \left(\frac{m_p}{\lambda_k} - \frac{\|T\|^2_{p\to d^*}}{a}\right)^{-1} \mathbb{E}\left[\left\|\delta^p_k\right\|^2_{\mathcal{X}_p^*} \mid \mathscr{S}_k\right] + \left(\frac{m_d}{\nu_k} - a\right)^{-1} \mathbb{E}\left[\left\|\delta^d_k\right\|^2_{\mathcal{X}_d^*} \mid \mathscr{S}_k\right] \\ &\leq \left(\frac{m_p}{\lambda_\infty} - \frac{\|T\|^2_{p\to d^*}}{a}\right)^{-1} \mathbb{E}\left[\left\|\delta^p_k\right\|^2_{\mathcal{X}_p^*} \mid \mathscr{S}_k\right] + \left(\frac{m_d}{\nu_\infty} - a\right)^{-1} \mathbb{E}\left[\left\|\delta^d_k\right\|^2_{\mathcal{X}_d^*} \mid \mathscr{S}_k\right].\end{aligned}$$

Thus by the summability assumption of $(\mathbf{A}_5)$**(III)**, we have

$$\left(\mathbb{E}\left[\left\|\delta_k^p\right\|_{\mathcal{X}_p^*}^2 \mid \mathscr{S}_k\right]\right)_{k\in\mathbb{N}} \in \ell_+^1(\mathfrak{S}) \quad \text{and} \quad \left(\mathbb{E}\left[\left\|\delta_k^d\right\|_{\mathcal{X}_d^*}^2 \mid \mathscr{S}_k\right]\right)_{k\in\mathbb{N}} \in \ell_+^1(\mathfrak{S})$$

and so, for any fixed $w \in \mathcal{C}_p \times \mathcal{C}_d$,

$$\left(\mathbb{E}\left[\langle \Delta_k, w - w_{k+1}\rangle \mid \mathscr{S}_k\right]\right)_{k\in\mathbb{N}} \in \ell_+^1(\mathfrak{S}).$$

Similarly, taking Lemma 3.12 with total expectation and the summability assumption of $(\mathbf{A}_5)$**(III)** yields, for any fixed $w \in \mathcal{C}_p \times \mathcal{C}_d$,

$$\left(\mathbb{E}\left[\langle \Delta_k, w - w_{k+1}\rangle\right]\right)_{k\in\mathbb{N}} \in \ell_+^1.$$

□

4 Convergence Analysis

4.1 Ergodic Convergence

Define, for each $k \in \mathbb{N}$, the *ergodic* iterates $\bar{x}_k \overset{\text{def}}{=} \frac{1}{k}\sum_{i=1}^{k} x_i$ and $\bar{\mu}_k \overset{\text{def}}{=} \frac{1}{k}\sum_{i=1}^{k} \mu_i$. The following theorem characterizes the convergence of the algorithm for the Lagrangian optimality gap evaluated at the ergodic iterates; later results on pointwise convergence will also imply ergodic convergence.[3]

Theorem 4.1. *Let* **(H)** *and* $(\mathbf{A}_1)$*-*$(\mathbf{A}_4)$ *hold. Then we have the following convergence rate to a noise-dominated regime: for each $k \in \mathbb{N}$, for every $(x, \mu) \in (\mathcal{C}_p \times \mathcal{C}_d) \cap (\mathrm{dom}(g) \times \mathrm{dom}(l^*))$,*

$$\mathbb{E}\left[\mathcal{L}(\bar{x}_k, \mu) - \mathcal{L}(x, \bar{\mu}_k)\right] \leq \frac{\frac{1}{\Lambda_0} D(w, w_0) - M(w, w_0)}{k} + \frac{\sum_{i=0}^{k-1} \mathbb{E}\left[\langle \Delta_i, w - w_{i+1}\rangle\right]}{k}. \tag{4.1}$$

In particular, if also $(\mathbf{A}_5)$ *holds, every almost sure weak sequential cluster point of $(\bar{w}_k)_{k\in\mathbb{N}}$ is optimal in mean; if $\bar{w}_{k_j} \rightharpoonup w_\infty$ almost surely, then $\mathbb{E}(w_\infty)$ is a saddle point for the Lagrangian.*

Proof. Let $w \overset{\text{def}}{=} (x, \mu) \in (\mathcal{C}_p \times \mathcal{C}_d) \cap (\mathrm{dom}(g) \times \mathrm{dom}(l^*))$. Beginning with Lemma 3.10, taking the the total expectation of (3.7) and summing up from 0 to $k-1$, discarding positive terms on the left hand side, we have

$$\sum_{i=0}^{k-1} \mathbb{E}\left[\mathcal{L}(x_{i+1}, \mu) - \mathcal{L}(x, \mu_{i+1})\right] \leq \frac{1}{\Lambda_0} D(w, w_0) - M(w, w_0) + \sum_{i=0}^{k-1} \mathbb{E}\left[\langle \Delta_i, w - w_{i+1}\rangle\right]. \tag{4.2}$$

Notice that $\sum_{i=0}^{k-1} \mathbb{E}\left[\langle \Delta_i, w - w_{i+1}\rangle\right]$ is nonnegative by $(\mathbf{A}_4)$ and Lemma 3.11. Using Jensen's inequality with the convex-concave function $\mathcal{L}$, we have (4.1).

[3] By "ergodic convergence", we mean convergence of the Lagrangian optimality gap evaluated at the ergodic iterates; not any ergodic averaging of the Lagrangian values themselves.

Now, assuming also ($\mathbf{A}_5$), let $(\bar{x}_{k_j}, \bar{\mu}_{k_j}) \rightharpoonup (x_\infty, \mu_\infty)$ almost surely. First note that, by Lemma 3.14,

$$\sum_{i=0}^{\infty} \mathbb{E}\left[\langle \Delta_i, w - w_{i+1}\rangle\right] < +\infty.$$

Then, for every $(x, \mu) \in (\mathcal{C}_p \times \mathcal{C}_d) \cap (\mathrm{dom}(g) \times \mathrm{dom}(l^*))$,

$$\begin{aligned} \mathcal{L}\left(\mathbb{E}\left(x_\infty\right), \mu\right) - \mathcal{L}\left(x, \mathbb{E}\left(\mu_\infty\right)\right) &\leq \mathbb{E}\left[\mathcal{L}\left(x_\infty, \mu\right) - \mathcal{L}\left(x, \mu_\infty\right)\right] \\ &\leq \mathbb{E}\left[\liminf_{j\to\infty} \left[\mathcal{L}\left(\bar{x}_{k_j}, \mu\right) - \mathcal{L}\left(x, \bar{\mu}_{k_j}\right)\right]\right] \\ &\leq \liminf_{j\to\infty} \mathbb{E}\left[\mathcal{L}\left(\bar{x}_{k_j}, \mu\right) - \mathcal{L}\left(x, \bar{\mu}_{k_j}\right)\right] \\ &\leq 0, \end{aligned} \tag{4.3}$$

where we used Jensen's inequality, weak lower semicontinuity of $\mathcal{L}$, Fatou's Lemma and (4.1) with ($\mathbf{A}_5$) and Lemma 3.14. Inequality (4.3) trivially holds outside $(\mathcal{C}_p \times \mathcal{C}_d) \cap (\mathrm{dom}(g) \times \mathrm{dom}(l^*))$, and so holds for any $(x, \mu) \in \mathcal{X}_p \times \mathcal{X}_d$, whence we get that $(\mathbb{E}\left(x_\infty\right), \mathbb{E}\left(\mu_\infty\right))$ is a saddle point for $\mathcal{L}$. □

Remark 4.2. The term $k^{-1}\sum_{i=0}^{k-1}\mathbb{E}\left[\langle \Delta_i, w - w_{i+1}\rangle\right]$ in Theorem 4.1 is an averaging of the noise which dictates the radius of the noise-dominated region in some sense. For example, if we assume that there exists a constant $c \geq 0$ such that $\mathbb{E}\left[\langle \Delta_i, w - w_{i+1}\rangle\right] \leq c$ for all $i \in \mathbb{N}$ and for all $w \in \mathcal{X}_p \times \mathcal{X}_d$, then we have

$$\frac{\sum_{i=0}^{k-1}\mathbb{E}\left[\langle \Delta_i, w - w_{i+1}\rangle\right]}{k} \leq c$$

for all $k \in \mathbb{N}$, i.e., the radius of the noise-dominated region in Theorem 4.1 is at most c.

Remark 4.3. Consider the algorithm in the deterministic case, then choose $(x, \mu) = (x^\star, \mu^\star)$ for some saddle point $(x^\star, \mu^\star)$ in (4.1). In this case, the constant in the rate of convergence, $\frac{1}{\Lambda_0} D\left(w^\star, w_0\right) - M\left(w^\star, w_0\right)$, is given in terms of the Bregman divergence, in contrast to methods like [14] which have constants in terms of the Euclidean norm. With this change in the geometry, the dependence of the constant on the dimension of the problem can be greatly reduced, even from linear to logarithmic dependence for some problems and appropriately chosen entropies.

4.2 Asymptotic Regularity

Theorem 4.4. *Let* **(H)**, ($\mathbf{A}_1$), ($\mathbf{A}_2$), ($\mathbf{A}_3$)**(II)**, ($\mathbf{A}_4$), ($\mathbf{A}_5$), *and* ($\mathbf{A}_6$) *hold. Then the primal-dual sequence* $(x_k, \mu_k)_{k\in\mathbb{N}}$ *is almost surely asymptotically regular, meaning that* $x_{k+1} - x_k \to 0$ *and* $\mu_{k+1} - \mu_k \to 0$ *almost surely.*

Proof. Use again (3.7) in Lemma 3.10 with w equal to a saddle point $w^\star \in \boldsymbol{S}$ and take the total expectation to get, for each $k \in \mathbb{N}$,

$$\begin{aligned} &\mathbb{E}\left[\mathcal{L}\left(x_{k+1}, \mu^\star\right) - \mathcal{L}\left(x^\star, \mu_{k+1}\right)\right] + \mathbb{E}\left[\frac{1}{\Lambda_{k+1}} D\left(w^\star, w_{k+1}\right) - M\left(w^\star, w_{k+1}\right)\right] \\ &\quad + \varepsilon\mathbb{E}\left[d\left(w_{k+1}, w_k\right)\right] \leq \mathbb{E}\left[\frac{1}{\Lambda_k} D\left(w^\star, w_k\right) - M\left(w^\star, w_k\right)\right] + \mathbb{E}\left[\langle \Delta_k, w^\star - w_{k+1}\rangle\right]. \end{aligned} \tag{4.4}$$

By the definition of saddle point in (1.1), it holds, for each $k \in \mathbb{N}$,

$$\mathcal{L}\left(x_k, \mu^\star\right) - \mathcal{L}\left(x^\star, \mu_k\right) \geq 0$$

and so, from Lemma 2.7 with $(\mathbf{A}_4)$, $(\mathbf{A}_5)$, Lemma 3.14, and $(\mathbf{A}_3)$**(II)**,

$$\mathbb{E}\left[d\left(w_{k+1}, w_k\right)\right] \in \ell_+^1.$$

By Lemma 2.8, $d\left(w_{k+1}, w_k\right) \to 0$ almost surely. In view of $(\mathbf{A}_6)$, we get that, almost surely,

$$w_{k+1} - w_k \to 0, \tag{4.5}$$

i.e., the primal-dual sequence $\left(w_k\right)_{k\in\mathbb{N}}$ is almost surely asymptotically regular. □

4.3 Pointwise Convergence

The main result of this section is related to the pointwise weak convergence of the primal-dual sequence $\left(x_k, \mu_k\right)_{k\in\mathbb{N}}$ to a saddle point. These results require the stronger assumptions $(\mathbf{A}_8)$-$(\mathbf{A}_{10})$, although they are verified in many situations (see the discussion in Remark 3.8 and example thereafter). We will also impose the following conditions, which are only necessary for this particular section in the stochastic case and can be dropped for the deterministic case or the other sections.

(PW$_1$) $\mathcal{X}_p$ and $\mathcal{X}_d$ are separable.

(PW$_2$) The Bregman divergence D satisfies the following property. Let $\tilde{\Omega}$ be a full-measure subset of Ω ($\tilde{\Omega} \in \mathcal{F}$ with $\mathbb{P}(\tilde{\Omega}) = 1$). Let $w^\star \in \boldsymbol{S}$ and $\left(s_n\right)_{n\in\mathbb{N}} \subset \boldsymbol{S}$ such that $s_n \to w^\star$. If, for every $n \in \mathbb{N}$ and for every $\omega \in \tilde{\Omega}$,

$$\lim_{k\to\infty} \Lambda_k^{-1} D\left(s_n, w_k\left(\omega\right)\right) - M\left(s_n, w_k\left(\omega\right)\right) = r_{s_n}\left(\omega\right) \in [0, +\infty[,$$

then there exists a $[0, +\infty[$-valued random variable $r_{w^\star}$ such that, for any $\omega \in \tilde{\Omega}$,

$$\lim_{k\to\infty} \Lambda_k^{-1} D\left(w^\star, w_k\left(\omega\right)\right) - M\left(w^\star, w_k\left(\omega\right)\right) = r_{w^\star}\left(\omega\right).$$

Proposition 4.5. *Let* **(H)**, $(\mathbf{A}_1)$, $(\mathbf{A}_2)$, $(\mathbf{A}_3)$**(I)** *or* **(II)**, *and* $(\mathbf{A}_4)$-$(\mathbf{A}_8)$ *hold. Then* $\left(\left(x_k, \mu_k\right)\right)_{k\in\mathbb{N}}$ *is almost surely bounded and, recalling the notation of* (2.1) *and* (1.2), $\mathfrak{W}\left[\left(w_k\right)_{k\in\mathbb{N}}\right] \subset \boldsymbol{S}$ ($\mathbb{P}$-*a.s.*).

Proof. Evaluating Lemma 3.10 at a saddle point $w = w^\star \in \boldsymbol{S}$ and taking expectation conditioned on the filtration $\mathcal{S}_k$, we get, for each $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}\left(x_{k+1}, \mu^\star\right) - \mathcal{L}\left(x^\star, \mu_{k+1}\right) \mid \mathcal{S}_k\right] + \mathbb{E}\left[\frac{1}{\Lambda_{k+1}} D\left(w^\star, w_{k+1}\right) - M\left(w^\star, w_{k+1}\right) \mid \mathcal{S}_k\right] \\ + \varepsilon\mathbb{E}\left[d\left(w_{k+1}, w_k\right) \mid \mathcal{S}_k\right] \leq \left[\frac{1}{\Lambda_k} D\left(w^\star, w_k\right) - M\left(w^\star, w_k\right)\right] + \mathbb{E}\left[\left\langle \Delta_k, w^\star - w_{k+1}\right\rangle \mid \mathcal{S}_k\right].\end{aligned}$$

Then, by $(\mathbf{A}_4)$, $(\mathbf{A}_5)$, Lemma 3.14, and Lemma 2.6, $\left(\Lambda_k^{-1} D\left(w^\star, w_k\right) - M\left(w^\star, w_k\right)\right)_{k\in\mathbb{N}}$ is almost surely convergent to some $r \in [0, +\infty[$. In particular, from $(\mathbf{A}_3)$ and (3.4), both $\left(D\left(w^\star, w_k\right)\right)_{k\in\mathbb{N}}$ and $\left(d\left(w^\star, w_k\right)\right)_{k\in\mathbb{N}}$ are almost surely bounded and the coercivity condition $(\mathbf{A}_7)$ entails that the sequence $\left(w_k\right)_{k\in\mathbb{N}}$ is almost surely bounded in $\operatorname{int}\operatorname{dom}\left(\phi\right)$. Since $\mathcal{X}_p$ and $\mathcal{X}_d$ are reflexive, $\mathfrak{W}\left[\left(w_k\right)_{k\in\mathbb{N}}\right] \neq \emptyset$ almost surely. Let $w_\infty = \left(x_\infty, \mu_\infty\right)$ be an almost sure weak sequential cluster point of $\left(w_k\right)_{k\in\mathbb{N}}$, i.e., there is a subsequence

$(w_{k_i})_{i\in\mathbb{N}}$ such that $w_{k_i} \rightharpoonup w_\infty$ almost surely. The updates of Algorithm 1 are equivalent to the following monotone inclusions,

$$\begin{pmatrix} \frac{\nabla\phi_p(x_{k_i}) - \nabla\phi_p(x_{k_{i+1}})}{\lambda_k} + (\nabla f(x_{k_{i+1}}) - \nabla f(x_{k_i}) - \delta_{k_i}^p) + T^*(\mu_{k_{i+1}} - \mu_{k_i}) \\ \frac{\nabla\phi_d(\mu_{k_i}) - \nabla\phi_d(\mu_{k_{i+1}})}{\nu_k} + (\nabla h^*(\mu_{k_{i+1}}) - \nabla h^*(\mu_{k_i}) - \delta_{k_i}^d) + T(x_{k_{i+1}} - x_{k_i}) \end{pmatrix}$$
$$\in \begin{pmatrix} \partial g + \nabla f & 0 \\ 0 & \partial l^* + \nabla h^* \end{pmatrix} \begin{pmatrix} x_{k_{i+1}} \\ \mu_{k_{i+1}} \end{pmatrix} + \begin{pmatrix} 0 & T^* \\ -T & 0 \end{pmatrix} \begin{pmatrix} x_{k_{i+1}} \\ \mu_{k_{i+1}} \end{pmatrix}. \quad (4.6)$$

Since $(w_{k_i})_{i\in\mathbb{N}}$ lies in $\operatorname{int}\mathcal{C}_p \times \operatorname{int}\mathcal{C}_d$, we have $N_{\mathcal{C}_p}(x_{k_{i+1}}) = 0$ and $N_{\mathcal{C}_d}(\mu_{k_{i+1}}) = 0$. This together with [58, Theorem 2.4.2(viii)] implies

$$\begin{pmatrix} \partial g + \nabla f & 0 \\ 0 & \partial l^* + \nabla h^* \end{pmatrix} \begin{pmatrix} x_{k_{i+1}} \\ \mu_{k_{i+1}} \end{pmatrix} + \begin{pmatrix} 0 & T^* \\ -T & 0 \end{pmatrix} \begin{pmatrix} x_{k_{i+1}} \\ \mu_{k_{i+1}} \end{pmatrix} \subset$$
$$\begin{pmatrix} \partial(g + f + \iota_{\mathcal{C}_p}) & 0 \\ 0 & \partial(l^* + h^* + \iota_{\mathcal{C}_d}) \end{pmatrix} \begin{pmatrix} x_{k_{i+1}} \\ \mu_{k_{i+1}} \end{pmatrix} + \begin{pmatrix} 0 & T^* \\ -T & 0 \end{pmatrix} \begin{pmatrix} x_{k_{i+1}} \\ \mu_{k_{i+1}} \end{pmatrix}. \quad (4.7)$$

The first operator on the right hand side of (4.7) is maximal monotone thanks to **(H)** and [58, Theorem 3.1.11]. The second operator is a skew-symmetric linear operator which is then maximal monotone with full domain by [55, Section 17]. By [55, Theorem 24.1(a)], we deduce that the operator in the right hand side of (4.7) is maximal monotone. Hence its graph is sequentially closed in the weak-strong topology by [8, Lemma 1.2]. Recall that, by **($\mathbf{A}_4$)**, **($\mathbf{A}_5$)**, and Remark 3.4, $\left(\delta_k^p\right)_{k\in\mathbb{N}}$ and $\left(\delta_k^d\right)_{k\in\mathbb{N}}$ converge strongly to zero almost surely. From Theorem 4.4 and the fact that $w_{k_i} \rightharpoonup w_\infty$, we have also that $((x_{k_i+1}, \mu_{k_i+1}))_{i\in\mathbb{N}}$ converges weakly to (x_∞, μ_∞) almost surely. In addition, by **(H)**, T is linear (and bounded) which, combined with Theorem 4.4, yields

$$T(x_{k_{i+1}} - x_{k_i}) \to 0 \text{ and } T^*(\mu_{k_{i+1}} - \mu_{k_i}) \to 0$$

almost surely. From **($\mathbf{A}_8$)** combined with Theorem 4.4, we deduce that, almost surely,

$$\nabla\phi_p(x_{k_{i+1}}) - \nabla\phi_p(x_{k_i}) \to 0 \text{ and } \nabla f(x_{k_{i+1}}) - \nabla f(x_{k_i}) \to 0$$
$$\nabla\phi_d(\mu_{k_{i+1}}) - \nabla\phi_d(\mu_{k_i}) \to 0 \text{ and } \nabla h^*(\mu_{k_{i+1}}) - \nabla h^*(\mu_{k_i}) \to 0.$$

Now since both $(\lambda_k)_{k\in\mathbb{N}}$ and $(\nu_k)_{k\in\mathbb{N}}$ are bounded away fro zero by **($\mathbf{A}_2$)**, we have shown that, almost surely, the left hand side of (4.6) converge strongly. Hence, by weak-strong sequential closedness of the graph of the operator in (4.7) we have shown above, we get

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \begin{pmatrix} \partial(g + f + \iota_{\mathcal{C}_p}) & T^* \\ -T & \partial(l^* + h^* + \iota_{\mathcal{C}_d}) \end{pmatrix} \begin{pmatrix} x_\infty \\ \mu_\infty \end{pmatrix},$$

holds almost surely, whence it follows that each weak sequential cluster point of $(w_k)_{k\in\mathbb{N}}$ is a saddle point almost surely. □

The significance of the following proposition is in the order of the quantifiers; it guarantees that there exists a full-measure set $\tilde{\Omega}$ for which the conclusion holds for every solution $w^\star$.

Proposition 4.6. *Let* **(H)**, **($\mathbf{A}_1$)**, **($\mathbf{A}_2$)**, **($\mathbf{A}_3$)(I)** *or* **(II)**, *and* **($\mathbf{A}_4$)**-**($\mathbf{A}_8$)** *hold as well as* **($\mathbf{PW}_1$)** *and* **($\mathbf{PW}_2$)**. *Then, there exists* $\tilde{\Omega} \in \mathcal{F}$ *such that* $\mathbb{P}\left(\tilde{\Omega}\right) = 1$ *and, for every* $w^\star \in \boldsymbol{S}$ *and for every* $\omega \in \tilde{\Omega}$, *the sequence*

$$\left(\Lambda_k^{-1} D\left(w^\star, w_k\left(\omega\right)\right) - M\left(w^\star, w_k\left(\omega\right)\right)\right)_{k\in\mathbb{N}}$$

converges with limit in $[0, +\infty[$.

Proof. By $(\mathbf{PW}_1)$, there exists a countable set S such that $\overline{S} = \boldsymbol{S}$. Once again, as in the proof of Proposition 4.5, for every $w^\star \in \boldsymbol{S}$ there exist $\Omega_{w^\star} \in \mathcal{F}$ such that $\mathbb{P}(\Omega_{w^\star}) = 1$ and, for every $\omega \in \Omega_{w^\star}$, it holds

$$\lim_{k\to\infty} \Lambda_k^{-1} D(w^\star, w_k(\omega)) - M(w^\star, w_k(\omega)) = r^\star(\omega) \in [0, +\infty[.$$

Let $\tilde{\Omega} = \bigcap_{s\in S} \Omega_s$ and notice that $\mathbb{P}\left(\tilde{\Omega}\right) = 1$ since, by countability of S, we have

$$\mathbb{P}\left(\tilde{\Omega}\right) = 1 - \mathbb{P}\left(\tilde{\Omega}^c\right) = 1 - \mathbb{P}\left(\bigcup_{s\in S} \Omega_s^c\right) \geq 1 - \sum_{s\in S} \mathbb{P}(\Omega_s^c) = 1.$$

Fix a particular $w^\star \in \boldsymbol{S}$; since $\overline{S} = \boldsymbol{S}$, there exists a sequence $(s_n)_{n\in\mathbb{N}}$ in S such that $s_n \to w^\star$. At the same time, for each $n \in \mathbb{N}$, there exists r_n, a $[0, +\infty[$-valued random variable such that, for each $\omega \in \tilde{\Omega}$,

$$\lim_{k\to\infty} \Lambda_k^{-1} D(s_n, w_k(\omega)) - M(s_n, w_k(\omega)) = r_n(\omega) \in [0, +\infty[.$$

Applying now $(\mathbf{PW}_2)$, we find that, for any $\omega \in \tilde{\Omega}$,

$$\lim_{k\to\infty} \Lambda_k^{-1} D(w^\star, w_k(\omega)) - M(w^\star, w_k(\omega)) = r_{w^\star}(\omega) \in [0, \infty[.$$

□

Theorem 4.7. *Let* $\mathbf{(H)}$, $\mathbf{(A_1)}$, $\mathbf{(A_2)}$, *and* $\mathbf{(A_4)}$-$\mathbf{(A_{10})}$ *hold as well as* $\mathbf{(PW_1)}$ *and* $\mathbf{(PW_2)}$. *Suppose also that one of the following holds:*

(i) $\boldsymbol{S}$ *is a singleton.*

(ii) $\mathbf{(A_3)(I)}$, $\boldsymbol{S} \subset \operatorname{int} \mathcal{C}_p \times \operatorname{int} \mathcal{C}_d$ *and* ϕ_p *and* ϕ_d *are Legendre.*

(iii) $\mathbf{(A_3)(II)}$, *and* $d(w^1, w^2) = 0 \Rightarrow w^1 = w^2$.

Then, there exists $\bar{w}$, *an* $\boldsymbol{S}$*-valued random variable, such that* $(w_k)_{k\in\mathbb{N}} \rightharpoonup \bar{w}$ $(\mathbb{P}$*-a.s.*$)$.

Proof. We use a standard reasoning inspired by Opial's lemma (see [43]). We recall the notation of (2.1) for the set of weak cluster points of a sequence. By Proposition 4.5, there exists $\Omega' \in \mathcal{F}$ with $\mathbb{P}(\Omega') = 1$ such that, for any $\omega \in \Omega'$, the following holds

$$\mathfrak{W}[(w_k(\omega))] \subset \boldsymbol{S}$$

and the sequence $(w_k(\omega))_{k\in\mathbb{N}}$ is bounded, and thus $\mathfrak{W}[(w_k(\omega))] \neq \emptyset$ since the spaces are reflexive. Furthermore, by Proposition 4.6, there exists $\Omega'' \in \mathcal{F}$ with $\mathbb{P}(\Omega'') = 1$ such that, for any $\omega \in \Omega''$, for any $w^\star \in \boldsymbol{S}$, it holds

$$\lim_{k\to\infty} \Lambda_k^{-1} D(w^\star, w_k(\omega)) - M(w^\star, w_k(\omega)) = r_{w^\star}(\omega) \in [0, +\infty[.$$

Let $\tilde{\Omega} = \Omega' \cap \Omega''$, for any $\omega \in \tilde{\Omega}$ we let $w^1(\omega) \in \mathfrak{W}\left[(w_k(\omega))_{k\in\mathbb{N}}\right]$ and $w^2(\omega) \in \mathfrak{W}\left[(w_k(\omega))_{k\in\mathbb{N}}\right]$ be two weak sequential cluster points of $(w_k(\omega))_{k\in\mathbb{N}}$, i.e., there exists two subsequences $(w_{k_i}(\omega))_{i\in\mathbb{N}}$ and $\left(w_{k_j}(\omega)\right)_{j\in\mathbb{N}}$ such that $w_{k_i}(\omega) \rightharpoonup w^1(\omega)$ and $w_{k_j}(\omega) \rightharpoonup w^2(\omega)$ almost surely. Since $\mathfrak{W}\left[(w_k(\omega))_{k\in\mathbb{N}}\right] \subset \boldsymbol{S}$, $w^1(\omega)$ and $w^2(\omega)$ are saddle points.Thus, there exist $r_{w^1}(\omega), r_{w^2}(\omega) \in [0, +\infty[$ such that,

$$\lim_{k\to\infty} \left(\Lambda_k^{-1} D(w^1(\omega), w_k(\omega)) - M(w^1(\omega), w_k(\omega))\right) = r_{w^1}(\omega)$$

and

$$\lim_{k\to\infty}\left(\Lambda_k^{-1}D(w^2(\omega),w_k(\omega))-M(w^2(\omega),w_k(\omega))\right)=r_{w^2}(\omega).$$

Using the three point identity, we have, for each $i\in\mathbb{N}$,

$$\begin{aligned}
&\Lambda_{k_i}^{-1}D(w^1(\omega),w_{k_i}(\omega))-M(w^1(\omega),w_{k_i}(\omega))-\Lambda_{k_i}^{-1}D(w^2(\omega),w_{k_i}(\omega))+M(w^2(\omega),w_{k_i}(\omega))\\
&=\Lambda_{k_i}^{-1}\left(D(w^1(\omega),w_{k_i}(\omega))-D(w^2(\omega),w_{k_i}(\omega))\right)-\left(M(w^1(\omega),w_{k_i}(\omega))-M(w^2(\omega),w_{k_i}(\omega))\right)\\
&=\Lambda_{k_i}^{-1}\left(D(w^1(\omega),w^2(\omega))-\left\langle\nabla\phi(w_{k_i}(\omega))-\nabla\phi(w^2(\omega)),w^1(\omega)-w^2(\omega)\right\rangle\right)\\
&\qquad-\left(M(w^1(\omega),w_{k_i}(\omega))-M(w^2(\omega),w_{k_i}(\omega))\right).
\end{aligned}\tag{4.8}$$

Recall that, by $(\mathbf{A}_2)$, both $(\lambda_k)_{k\in\mathbb{N}}$ and $(\nu_k)_{k\in\mathbb{N}}$ are nondecreasing and bounded above with limits λ_∞ and ν_∞, respectively. We denote $\Lambda_\infty\overset{\text{def}}{=}(\lambda_\infty,\nu_\infty)$. Then, recalling $(\mathbf{A}_9)$ and $(\mathbf{A}_{10})$ and passing to the limit in (4.8) we get

$$\begin{aligned}
r_{w^1}(\omega)-r_{w^2}(\omega)&=\Lambda_\infty^{-1}\left(D(w^1(\omega),w^2(\omega))-\left\langle\nabla\phi(w^1(\omega))-\nabla\phi(w^2(\omega)),w^1(\omega)-w^2(\omega)\right\rangle\right)\\
&\qquad+M(w^2(\omega),w^1(\omega))\\
&=\Lambda_\infty^{-1}\left(D(w^1(\omega),w^2(\omega))-D(w^1(\omega),w^2(\omega))-D(w^2(\omega),w^1(\omega))\right)\\
&\qquad+M(w^2(\omega),w^1(\omega))\\
&=-\Lambda_\infty^{-1}D(w^2(\omega),w^1(\omega))+M(w^2(\omega),w^1(\omega)).
\end{aligned}$$

Repeating this argument, replacing $w_{k_i}(\omega)$ by $w_{k_j}(\omega)$ above, we furthermore have

$$r_{w^1}(\omega)-r_{w^2}(\omega)=\Lambda_\infty^{-1}D(w^1(\omega),w^2(\omega))-M(w^1(\omega),w^2(\omega)),$$

which shows that

$$\left[\Lambda_\infty^{-1}D(w^1(\omega),w^2(\omega))-M(w^1(\omega),w^2(\omega))\right]+\left[\Lambda_\infty^{-1}D(w^2(\omega),w^1(\omega))-M(w^2(\omega),w^1(\omega))\right]=0.$$

By $(\mathbf{A}_3)$ and (3.4), we arrive at

$$L\left[D\left(w^1(\omega),w^2(\omega)\right)+D\left(w^2(\omega),w^1(\omega)\right)\right]+\varepsilon\left[d\left(w^1(\omega),w^2(\omega)\right)+d\left(w^2(\omega),w^1(\omega)\right)\right]=0,$$

or equivalently, in view of $(\mathbf{A}_3)$,

$$\begin{aligned}
D\left(w^1(\omega),w^2(\omega)\right)+D\left(w^2(\omega),w^1(\omega)\right)&=0\ \text{ and}\\
\varepsilon\left[d\left(w^1(\omega),w^2(\omega)\right)+d\left(w^2(\omega),w^1(\omega)\right)\right]&=0.
\end{aligned}\tag{4.9}$$

To complete the proof, it remains to show that $w^1(\omega)=w^2(\omega)$ for all $\omega\in\tilde{\Omega}$, from which we can conclude that $w^1=w^2$ ($\mathbb{P}$-a.s.) since $\mathbb{P}\left(\tilde{\Omega}\right)=1$.

(i) Thanks to Proposition 4.5, we have $\mathfrak{W}\left[(w_k(\omega))\right]\subset\boldsymbol{S}=\{\bar{w}(\omega)\}$.

(ii) In this case, we have $\boldsymbol{S}\subset\operatorname{int}\operatorname{dom}(\phi_p)\times\operatorname{int}\operatorname{dom}(\phi_p)$ thanks to $(\mathbf{A}_1)$. Thus, in view of Proposition 4.5, $w^i(\omega)\in\mathfrak{W}\left[(w_k(\omega))_{k\in\mathbb{N}}\right]\subset\operatorname{int}\operatorname{dom}(\phi_p)\times\operatorname{int}\operatorname{dom}(\phi_p)$ for $i=1,2$. Moreover, (4.9) gives

$$D\left(w^1(\omega),w^2(\omega)\right)+D\left(w^2(\omega),w^1(\omega)\right)=\langle\nabla\phi(w^1(\omega))-\nabla\phi(w^2(\omega)),\,w^1(\omega)-w^2(\omega)\rangle=0.$$

Unless $w^1(\omega)=w^2(\omega)$, this is in contradiction with strict monotonicity of $\nabla\phi$ on $\operatorname{int}\operatorname{dom}(\phi_p)\times\operatorname{int}\operatorname{dom}(\phi_p)$ since ϕ_p and ϕ_d are Legendre.

(iii) Under these assumptions, $\varepsilon > 0$ and (4.9) entails

$$d\left(w^1(\omega), w^2(\omega)\right) = 0,$$

whence we conclude $w^1(\omega) = w^2(\omega)$ by the assumption on d.

□

Remark 4.8. The assumptions (i) and (ii) and corresponding conclusions in Theorem 4.7, can be separated to cover either the primal or the dual variables. For instance, assumption (i) can be weakened to the set $\{x : (x, \mu) \in \boldsymbol{S}\}$ is a singleton (rather than the entire set $\boldsymbol{S}$) then we will retain weak convergence of the primal iterates $(x_k)_{k\in\mathbb{N}}$ to the solution to the primal problem. Similarly, if (ii) holds only for the primal or dual, we will nevertheless retain weak convergence of the primal or dual iterates, respectively. We will not elaborate more on this for the sake of brevity and space limitation.

4.4 Strong Convergence under Relative Strong Convexity

In this part we assume that either f, g, or both are relatively strongly convex (see Definition 2.4) with respect to ψ_p with constant m_f, m_g, or $m_f + m_g$, respectively, as in $(\mathbf{A}_{11})$. For brevity, we analyze only the primal case but all of the analogous convergence results will hold for the dual case by making the corresponding assumptions on h^*, l^*, and ψ_d, as in $(\mathbf{A}_{11})$. In addition, if the assumptions made here on the primal functions and entropies hold for the corresponding dual functions and entropies, we will have convergence results for the whole primal-dual sequence $(w_k)_{k\in\mathbb{N}}$.

Central to our arguments are the concepts of total convexity and sequential consistency which provide an elegant framework relating convergence in terms of the Bregman divergence and convergence in terms of the ambient norm of the space. We will assume that ψ_p is sequentially consistent and totally convex, which we now go on to define. The following definitions come from [12] although earlier notions of total convexity and its modulus exist.

Definition 4.9. Define, for all $x \in \operatorname{int}\operatorname{dom}(\psi_p)$ and $t \in [0, \infty[$,

$$\Theta_{\psi_p}(x, t) \stackrel{\text{def}}{=} \inf\left\{D_{\psi_p}(x', x) : \|x - x'\|_{\mathcal{X}_p} = t\right\}.$$

The function Θ is called the modulus of total convexity and it is clearly nondecreasing in t (see [12, Page 18]). We call a function ψ_p totally convex at a point $x \in \operatorname{int}\operatorname{dom}(\psi_p)$ iff $\Theta_{\psi_p}(x, t) > 0$ for any $t > 0$. We say the function ψ_p is totally convex on a subset $X \subseteq \operatorname{int}\operatorname{dom}(\psi_p)$ iff it is totally convex for each $x \in X$.

Total convexity is a sort of generalization of strict convexity to functions defined on Banach spaces. Indeed, for finite-dimensional spaces, strict convexity and total convexity are equivalent for functions with full domain [12, Proposition 1.2.6]. Examples of totally convex functions include the Shannon-Boltzmann entropy, the Hellinger entropy, the Fermi-Dirac entropy, the energy/euclidean entropy, and any strongly convex function as well.

Definition 4.10. A function ψ_p is called sequentially consistent on a subset $X \subseteq \operatorname{int}\operatorname{dom}(\psi_p)$ iff for any bounded subset $V \subseteq X$, for any $t > 0$, we have

$$\inf_{x\in V} \Theta_{\psi_p}(x, t) > 0.$$

Lemma 4.11. *Let $(x^\star, \mu^\star) \in \boldsymbol{S}$ be a saddle point. Assume* **(H)**, $(\mathbf{A}_1)$-$(\mathbf{A}_5)$, *and* $(\mathbf{A}_{11})$**(I)**. *Then $D_{\psi_p}(x^\star, x_k) \to 0$ almost surely. Similarly, if* $(\mathbf{A}_{11})$**(II)** *holds, then $D_{\psi_d}(\mu^\star, \mu_k) \to 0$ almost surely.*

Proof. Under ($\mathbf{A}_{11}$)(I), evaluating (3.8) in Lemma 3.10 at a saddle point $w = w^\star \in \boldsymbol{S}$ we have, for each $k \in \mathbb{N}$,

$$\frac{1}{\Lambda_k} D\left(w^\star, w_k\right) - \frac{1}{\Lambda_{k+1}} D\left(w^\star, w_{k+1}\right) - M\left(w^\star, w_k\right) + M\left(w^\star, w_{k+1}\right) + \left\langle \Delta_k, w^\star - w_{k+1} \right\rangle \\ \geq m_g D_{\psi_p}\left(x^\star, x_{k+1}\right) + m_f D_{\psi_p}\left(x^\star, x_k\right).$$

We now break the proof into two cases based on whether $m_g > 0$ or $m_f > 0$, starting with $m_f > 0$. Taking the expectation conditioned on the filtration, we have, for each $k \in \mathbb{N}$,

$$m_f D_{\psi_p}\left(x^\star, x_k\right) \leq \frac{1}{\Lambda_k} D\left(w^\star, w_k\right) - \frac{1}{\Lambda_{k+1}} \mathbb{E}\left[D\left(w^\star, w_{k+1}\right) \mid \mathscr{S}_k\right] - M\left(w^\star, w_k\right) \\ + \mathbb{E}\left[M\left(w^\star, w_{k+1}\right) \mid \mathscr{S}_k\right] + \mathbb{E}\left[\left\langle \Delta_k, w^\star - w_{k+1} \right\rangle \mid \mathscr{S}_k\right]. \quad (4.10)$$

Applying Lemma 2.6 to (4.10) along with the assumption that $m_f > 0$, ($\mathbf{A}_4$), and ($\mathbf{A}_5$) with Lemma 3.14, we find that $\left(D_{\psi_p}\left(x^\star, x_k\right)\right)_{k\in\mathbb{N}} \in \ell_+^1(\mathfrak{S})$ and, in particular, $D_{\psi_p}\left(x^\star, x_k\right) \to 0$ almost surely.

Now, assuming $m_g > 0$ gives, for each $k \in \mathbb{N}$,

$$m_g D_{\psi_p}\left(x^\star, x_{k+1}\right) \leq \frac{1}{\Lambda_k} D\left(w^\star, w_k\right) - \frac{1}{\Lambda_{k+1}} D\left(w^\star, w_{k+1}\right) - M\left(w^\star, w_k\right) + M\left(w^\star, w_{k+1}\right) + \left\langle \Delta_k, w^\star - w_{k+1} \right\rangle.$$

Taking the expectation then leads to, for each $k \in \mathbb{N}$,

$$m_g \mathbb{E}\left[D_{\psi_p}\left(x^\star, x_{k+1}\right)\right] \leq \frac{1}{\Lambda_k} \mathbb{E}\left[D\left(w^\star, w_k\right)\right] - \frac{1}{\Lambda_{k+1}} \mathbb{E}\left[D\left(w^\star, w_{k+1}\right)\right] - \mathbb{E}\left[M\left(w^\star, w_k\right)\right] + \mathbb{E}\left[M\left(w^\star, w_{k+1}\right)\right] \\ + \mathbb{E}\left[\left\langle \Delta_k, w^\star - w_{k+1} \right\rangle\right].$$

Then, by Remark 2.7 with the assumption $m_g > 0$, ($\mathbf{A}_5$) and Lemma 3.14, we have that $\left(\mathbb{E}\left[D_{\psi_p}\left(x^\star, x_k\right)\right]\right)_{k\in\mathbb{N}} \in \ell_+^1$ and so, by Lemma 2.8, we have that $D_{\psi_p}\left(x^\star, x_k\right) \to 0$ almost surely. □

Theorem 4.12. *Assume* (H), ($\mathbf{A}_1$)-($\mathbf{A}_5$), *and* ($\mathbf{A}_{11}$)(I) *hold, that ψ_p is sequentially consistent on $\mathcal{U}_p$, and assume $x^\star$ is the unique solution to the primal problem (i.e., $\boldsymbol{S}_{\mathscr{P}} = \{x^\star\}$). Then, if the sublevel sets of $D_{\psi_p}\left(x^\star, \cdot\right)$ are bounded, the sequence $\left(x_k\right)_{k\in\mathbb{N}}$ converges strongly to the solution $x^\star$ almost surely. Furthermore, if* ($\mathbf{A}_{11}$)(II) *holds, $\mu^\star$ is the unique solution to the dual problem, ψ_d is sequentially consistent on $\mathcal{U}_d$, and the sublevel sets of $D_{\psi_d}\left(\mu^\star, \cdot\right)$ are bounded, then almost surely, the sequence $\left(w_k\right)_{k\in\mathbb{N}}$ converges strongly to the saddle point $w^\star$.*

Proof. Under these assumptions, Lemma 4.11 ensures $D_{\psi_p}\left(x^\star, x_k\right) \to 0$ almost surely. The sublevel sets of $D_{\psi_p}\left(x^\star, \cdot\right)$ are bounded and thus the sequence $\left(x_k\right)_{k\in\mathbb{N}}$ is bounded. Since $\left(x_k\right)_{k\in\mathbb{N}}$ also remains in $\mathcal{U}_p$ by ($\mathbf{A}_1$), there exists $U_p \subseteq \mathcal{U}_p$ a bounded set such that, for each $k \in \mathbb{N}$, $x_k \in U_p$. Since ψ_p is sequentially consistent on $\mathcal{U}_p$, we have, for any $t > 0$,

$$\inf_{x\in U_p} \Theta_{\psi_p}(x, t) > 0.$$

Assume now that $\left(x_k\right)_{k\in\mathbb{N}}$ does not converge strongly to $x^\star$. Then there exists a subsequence $\left(x_{k_j}\right)_{j\in\mathbb{N}}$ and $\epsilon > 0$ such that for all $j \in \mathbb{N}$ it holds,

$$\left\|x_{k_j} - x^\star\right\|_{\mathcal{X}_p} > \epsilon.$$

Since $\left(x_{k_j}\right)_{j\in\mathbb{N}}$ is a subsequence of $(x_k)_{k\in\mathbb{N}}$, $\left(D_{\psi_p}\left(x^\star, x_{k_j}\right)\right)_{j\in\mathbb{N}}$ is a subsequence of $\left(D_{\psi_p}\left(x^\star, x_k\right)\right)_{k\in\mathbb{N}}$ and so its limit is 0. Since ψ_p is sequentially consistent and $\left\|x_{k_j} - x^\star\right\| > \epsilon$, the following is true: for any $j \in \mathbb{N}$,

$$D_{\psi_p}\left(x^\star, x_{k_j}\right) \geq \Theta_{\psi_p}\left(x_{k_j}, \left\|x_{k_j} - x^\star\right\|_{\mathcal{X}_p}\right) \geq \Theta_{\psi_p}\left(x_{k_j}, \epsilon\right) \geq \inf_{x\in\mathcal{U}_p} \Theta_{\psi_p}\left(x, \epsilon\right) > 0, \tag{4.11}$$

which contradicts the fact that $\lim_{j\to\infty} D_{\psi_p}\left(x^\star, x_{k_j}\right) = 0$ since the positive lower bound $\inf_{x\in\mathcal{U}_p} \Theta_{\psi_p}\left(x, \epsilon\right)$ does not depend on j. Thus such a subsequence $\left(x_{k_j}\right)_{j\in\mathbb{N}}$ cannot exist and the desired claim follows.

Repeating this argument for the dual gives convergence of $(\mu_k)_{k\in\mathbb{N}}$ to the solution of the dual problem $\mu^\star$ and thus, if ($\mathbf{A}_{11}$) holds for the primal and the dual, we have that $(w_k)_{k\in\mathbb{N}}$ converges to the saddle point $w^\star$. □

Remark 4.13. The assumption that the sublevel sets of the the Bregman divergence be bounded, used in Theorem 4.12, holds for a wide class of entropies which includes the Shannon-Boltzmann entropy, the Hellinger entropy, the Fermi-Dirac entropy, the fractional power entropy, and energy/euclidean entropy (see [31, Remark 4]).

Remark 4.14. In the statement of Theorem 4.12, uniqueness of the solution $x^\star$ is assumed only for clarity of presentation. Indeed, without the assumption the same argument used in the proof works for every solution $x^\star$; and this implies that the solution to the primal problem under our considerations must be unique, as the sequence x_k converges to any solution taken. We do not have a more direct proof for uniqueness of the solution in the general setting of Theorem 4.12, but we point at Proposition A.2 where we show a direct proof of uniqueness under the assumption that there exists a saddle point $w^\star = (x^\star, \mu^\star)$ with $x^\star \in \mathcal{U}_p$.

5 Applications and Numerical Experiments

We examine two applications that satisfy our assumptions for Theorem 4.1. The following results will be useful throughout the applications section, particularly when it comes to satisfying ($\mathbf{A}_3$). In the rest of the section, $\|\cdot\|_q$, $q \in [1, +\infty]$, will stand for the ℓ^q norm on $\mathbb{R}^n$. $\mathcal{B}_r^q$ is the ℓ^q ball of radius $r > 0$.

We begin with a famous result, Pinsker's inequality, which shows that the Kullback-Leibler divergence is strongly convex on the simplex wrt the ℓ^1 norm.

Lemma 5.1 (Pinsker's Inequality [46]). *Let $x, y \in \Sigma^n \overset{\text{def}}{=} \left\{u \in \mathbb{R}^n : u \geq 0, u^T\mathbb{1} = 1\right\}$ and let K be the Shannon-Boltzmann entropy: $K(x) = \sum_{i=1}^n x_i \log(x_i)$ on $\mathbb{R}^n_+$ with the convention that $0\log 0 = 0$. Then it holds*

$$\frac{1}{2}\|x - y\|_1^2 \leq D_K(x, y).$$

Lemma 5.2. *Let $T : \mathbb{R}^n \to \mathbb{R}^m$, $\mathcal{C}_p = \mathbb{R}^n_+$, $\mathcal{C}_d = \mathbb{R}^n$, $g(x) = \iota_{\{1\}}\left(x^T\mathbb{1}\right)$ and $l \in \Gamma_0(\mathbb{R}^m)$. Choose $\phi_p(x) = \sum_{i=1}^n x_i \log(x_i)$ on $\mathbb{R}^n_+$ with $0\log 0 = 0$, and $\phi_d(\mu) = \frac{1}{2}\|\mu\|_2^2$. Let $\gamma > 0$. Then ($\mathbf{A}_3$) and ($\mathbf{A}_6$) are satisfied with $\tilde{\mathcal{U}}_p = \Sigma^n$, $\mathcal{U}_p = \operatorname{ri}\Sigma^n$, $\mathcal{U}_d = \tilde{\mathcal{U}}_d = \operatorname{dom}(\partial l^*)$,*

$$\lambda_\infty < \frac{1}{L_p + \gamma\|T\|_2^2} \text{ and } \nu_\infty < \frac{1}{L_d + \gamma^{-1}},$$

$\varepsilon = \frac{1}{2}$ *and*

$$d(w_1, w_2) = \left(\frac{1}{\lambda_\infty} - L_p - \gamma \|T\|_2^2\right) \|x_1 - x_2\|_1^2 + \left(\frac{1}{\nu_\infty} - L_d - \frac{1}{\gamma}\right) \|\mu_1 - \mu_2\|_2^2 .$$

In the above, ri denotes the relative interior.

Proof. The expressions of $\tilde{\mathcal{U}}_p, \mathcal{U}_p, \tilde{\mathcal{U}}_d$ and $\mathcal{U}_d$ are immediate. By definition (see (2.3)), for any $w \in \tilde{\mathcal{U}}_p \times \tilde{\mathcal{U}}_d$ and $w' \in \mathcal{U}_p \times \mathcal{U}_d$, we have

$$\left(\frac{1}{\Lambda_\infty} - L\right) D(w, w') - M(w, w') = \left(\frac{1}{\lambda_\infty} - L_p\right) D_p(x, x') + \left(\frac{1}{\nu_\infty} - L_d\right) \frac{1}{2} \|\mu - \mu'\|_2^2 - \langle T(x - x'), \mu - \mu' \rangle .$$

Using Lemma 5.1, it holds for any $x \in \tilde{\mathcal{U}}_p$ and $x' \in \mathcal{U}_p$,

$$D_p(x, x') \geq \frac{1}{2} \|x - x'\|_1^2 .$$

By Young's inequality, for any $\gamma > 0$, we also have

$$\begin{aligned} -\langle T(x - x'), \mu - \mu' \rangle &\geq -\frac{\gamma}{2} \|T(x - x')\|_2^2 - \frac{1}{2\gamma} \|\mu - \mu'\|_2^2 \geq -\frac{\gamma}{2} \|T\|_2^2 \|(x - x')\|_2^2 - \frac{1}{2\gamma} \|\mu - \mu'\|_2^2 \\ &\geq -\frac{\gamma}{2} \|T\|_2^2 \|(x - x')\|_1^2 - \frac{1}{2\gamma} \|\mu - \mu'\|_2^2 . \end{aligned}$$

Combining the two, we find, for any $w \in \tilde{\mathcal{U}}_p \times \tilde{\mathcal{U}}_d$ and $w' \in \mathcal{U}_p \times \mathcal{U}_d$

$$\left(\frac{1}{\Lambda_\infty} - L\right) D(w, w') - M(w, w') \geq \frac{1}{2} \left[\left(\frac{1}{\lambda_\infty} - L_p - \gamma \|T\|_2^2\right) \|x - x'\|_1^2 + \left(\frac{1}{\nu_\infty} - L_d - \frac{1}{\gamma}\right) \|\mu - \mu'\|_2^2 \right]$$

which gives (3.1). Checking ($\mathbf{A}_6$) is immediate. □

5.1 Linear Inverse Problems on the Simplex

In [16], the problem of least squares regression on the simplex was considered as an application of the Chambolle-Pock algorithm. A natural extension for Algorithm 1 is to replace the euclidean norm with the Kullback-Leibler divergence. The Kullback-Leibler divergence is not Lipschitz-smooth and so the Chambolle-Pock algorithm of [14] and [16] cannot be applied, although [16] does allow one to use an entropy in computing the D-proximal mapping associated to g.

Consider the problem,

$$\min_{\substack{x \in \mathbb{R}_+^n \\ x^T \mathbb{1} = 1}} D_K(Ax, b) + \beta \|Bx\|_1 \tag{5.1}$$

where $A \in \mathbb{R}_+^{m\times n}$ is a matrix which does not contain any rows which are identically 0, $b \in \mathbb{R}_{++}^m$, K is the Shannon-Boltzmann entropy with the convention that $0\log 0 = 0$,

$$K(x) = \sum_{i=1}^{n} x_i \log(x_i), \quad \text{with} \quad \operatorname{dom}(K) = \mathbb{R}_+^n,$$

and $B : \mathbb{R}^n \to \mathbb{R}^{n-1}$ is the linear operator given by

$$Bx = \begin{pmatrix} x_2 - x_1 \\ \vdots \\ x_n - x_{n-1} \end{pmatrix}.$$

It is known that the term $\|Bx\|_1$ in (5.1) is intended to promote piecewise-constant solutions [52]. Rewriting (5.1), the associated saddle point problem is given by,

$$\min_{x\in\mathbb{R}_+^n} \max_{\mu\in\mathbb{R}^{n-1}} D_K(Ax, b) + \iota_{\{1\}}\left(x^T\mathbb{1}\right) + \langle Bx, \mu\rangle - \iota_{\mathcal{B}_\beta^\infty}(\mu).$$

Problem 5.1 can be put in a form solvable using the primal-dual algorithm of [14]. But, in addition to working over higher dimensional spaces, this algorithm does not exploit the geometry underlying the problem hence requiring computing (euclidean) prox mappings which are computationally more demanding.

We can apply Algorithm 1 with the following choices,

$$f(x) = D_K(Ax, b), \quad g(x) = \iota_{\{1\}}\left(x^T\mathbb{1}\right), \quad T = B, \quad h^* \equiv 0, \quad l^*(\mu) = \iota_{\mathcal{B}_\beta^\infty}(\mu),$$
$$\mathcal{C}_p = \mathbb{R}_+^n, \text{ and } \mathcal{C}_d = \mathbb{R}^{n-1}.$$

We choose ϕ_p and ϕ_d (with the same convention $0\log 0 = 0$) to be

$$\phi_p(x) = \sum_{i=1}^{n} x_i \log(x_i) \text{ and } \phi_d(\mu) = \frac{1}{2}\|\mu\|_2^2$$

which induces the divergences D_p and D_d

$$D_p(x, x') = \sum_{i=1}^{n} x_i \log\left(\frac{x_i}{x_i'}\right) - x_i + x_i' \text{ and } D_d(\mu, \mu') = \frac{1}{2}\left\|\mu - \mu'\right\|_2^2.$$

This gives us the following D-prox operator for our problem,

$$\operatorname{prox}_{\lambda_k g}^{D_p}(x) \overset{\text{def}}{=} \operatorname*{argmin}_{u\in\mathcal{C}_p} \{\lambda_k g(u) + D_p(u, x)\} = \operatorname*{argmin}_{\substack{u\in\mathbb{R}_+^n \\ u^T\mathbb{1}=1}} \{D_p(u, x)\} = \left(\frac{\exp(x_i)}{\sum_{j=1}^{n}\exp(x_j)}\right)_{i=1}^{n}.$$

The main hypothesis **(H)** is clearly satisfied in this problem. In order to satisfy **(A_1)**, we must find a constant $L_p > 0$ such that $L_p\phi_p(x) - f(x)$ is convex for all $x \in \operatorname{int}(\operatorname{dom}\phi_p) = \mathbb{R}_{++}^n$. This is precisely what is shown in [31, Lemma 8], which we include here for clarity.

Lemma 5.3. *Let* $\phi_p(x) = \sum_{i=1}^{n} x_i \log(x_i)$, $f(x) = D_K(Ax, b)$, *and* $A \in \mathbb{R}_+^{m\times n}$ *such that none of the rows of* A *are completely* 0. *Then, for any* L_p *such that*

$$L_p \geq \max_{1\leq j\leq m} \left(\sum_{i=1}^{n} A_{i,j} \right),$$

$L_p\phi_p - f$ *is convex on* $\mathbb{R}^n_{++}$.

Proof. See [31, Lemma 8] □

It remains to choose step sizes $(\lambda_k)_{k\in\mathbb{N}}$ and $(\nu_k)_{k\in\mathbb{N}}$ such that $(\mathbf{A}_2)$ and $(\mathbf{A}_3)$ are satisfied, for which we refer to Lemma 5.2.

Remark 5.4. Notice that the constant $\gamma > 0$ in Lemma 5.2 is arbitrary. For the experiments, we took $\gamma = \|B\|_2^{-1}$ to have symmetric step sizes,

$$\lambda_k = \frac{1}{L_p + \|B\|_2} \quad \text{and} \quad \nu_k = \frac{1}{L_d + \|B\|_2}$$

since $L_d = 0$ in this problem.

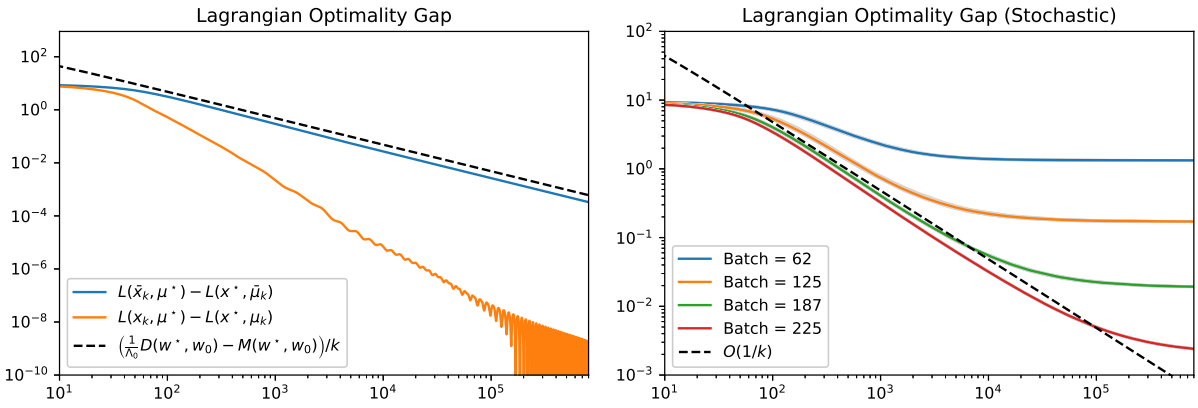

Figure 1: (Left) Lagrangian optimality gap, computed for both the pointwise iterates in orange and the ergodic iterates in blue, for Algorithm 1 applied deterministically to the linear inverse problem on the simplex in dimension $n = 250$. (Right) The average Lagrangian optimality gap for the ergodic iterates of Algorithm 1 applied stochastically for various batch sizes, showing convergence to a noise dominated region as predicted by Theorem 4.1. The colored lines are the average Lagrangian values for 20 runs of the algorithm, with individual runs displayed in light gray. The $O(1/k)$ theoretical rate is the same in both plots, as given in Theorem 4.1.

We now apply Algorithm 1 to solve (5.1) using the step size and entropy choices discussed above. We take $n = 250$ and $m = 250$, generate A with $a_{i,j} \in [0.01, 1.01]$ uniformly i.i.d., and generate b with entries uniformly i.i.d. in $]0, 1]$. We initialize with $x_0 = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ and $\mu_0 = 0$ with the constant step sizes $\lambda_k = \frac{1}{L_p + \|B\|_2}$ and $\nu_k = \frac{1}{\|B\|_2}$. We also consider $D_K(Ax, b) = \sum_{i=1}^{m} (Ax)_i \log\left(\frac{(Ax)_i}{b_i}\right)$ as a finite-sum

for which we can sample $f_i(x) = (Ax)_i \log\left(\frac{(Ax)_i}{b_i}\right)$ in batches, uniformly, when computing the gradient. Theorem 4.1 ensures convergence of the Lagrangian optimality gap in the deterministic setting for the ergodic iterates, and convergence in expectation to a noise-dominated region for stochastic sampling if the error is bounded in expectation as discussed in Remark 4.2. In Lemma A.1 in the appendix, we prove that this is indeed the case.

On the left of Figure 1, the Lagrangian optimality gap is presented for both the ergodic and pointwise iterates in the deterministic case. We show the same gaps for the stochastic version of the algorithm with batch sampling in Figure 1 on the right. To plot these gaps, we first run the deterministic version of the algorithm for a high number of iterations to find an (approximate) saddle point $(x^\star, \mu^\star) \in \boldsymbol{S}$ and then rerun the algorithm for 80% of the number of inital iterations, computing the gap at each iteration. For the stochastic version, we run the algorithm 20 times for each batch size and then plot the average of the gap for the ergodic iterates over these 20 runs in color, with individual runs represented in light gray. Clearly as the batch size increases, the radius of the noise-dominated region shrinks.

5.2 Variational problems with the entropic Wasserstein distance

Consider the optimal transport problem between two discrete measures, ρ and θ, defined on two metric spaces $\mathcal{X}$ and $\mathcal{Y}$. Let $C \in \mathbb{R}^{n\times m}$ be the ground cost on $\mathcal{X}\times\mathcal{Y}$. The cost C is typically application-dependent, and reflects some prior knowledge on the data to be processed. We regularize the optimal transport problem by subtracting in the objective the entropy of the transport plan π,

$$E(\pi) = -\sum_{i=1}^{n}\sum_{j=1}^{m} \pi_{i,j}\log(\pi_{i,j}).$$

The idea of regularizing the optimal transport problem by including the entropy of the transport plan π is not new. It was popularized by [22] and then explored, for example, in [23] for computing entropic Wasserstein barycenters, in [44] for approximating entropic Wasserstein gradient flows, in [24] for variational Wasserstein problems, in [25], etc. For $\gamma > 0$, the entropic regularization of the Kantorovich formulation of optimal transport can be written as the convex optimization problem

$$W_\gamma(\rho,\theta) \stackrel{\text{def}}{=} \inf_{\pi\in\Pi(\rho,\theta)} \left\{ \langle C,\pi\rangle + \gamma\sum_{i=1}^{n}\sum_{j=1}^{m}\pi_{i,j}\log(\pi_{i,j}) = \gamma\sum_{i=1}^{n}\sum_{j=1}^{m}\pi_{i,j}\log\left(\frac{\pi_{i,j}}{\xi_{i,j}}\right)\right\}, \tag{5.2}$$

where $\Pi(\rho,\theta) \stackrel{\text{def}}{=} \left\{\pi\in\mathbb{R}_+^{n\times m} : \pi\mathbb{1} = \rho, \pi^T\mathbb{1} = \theta\right\}$ is the so-called transportation polytope and $\xi_{i,j} \stackrel{\text{def}}{=} \exp(\frac{-C_{i,j}}{\gamma})$ is the Gibbs Kernel. When $\mathcal{X} = \mathcal{Y}$, $\gamma = 0$ and $C = d^p$, where d is a distance on $\mathcal{X}$, then $W_0^{1/p}$ is the well-known p-Wasserstein distance.

We consider solving the following variational problem over discrete measures, i.e., vectors in the simplex $\Sigma^n \stackrel{\text{def}}{=} \left\{x : x\geq 0, x^T\mathbb{1} = 1\right\}$,

$$\min_{\rho\in\Sigma^n} W_\gamma(F\rho,\theta) + J\circ B(\rho), \tag{5.3}$$

where $J\in\Gamma_0(\mathbb{R}^p)$, $F:\Sigma^n\to\Sigma^m$ and $B:\mathbb{R}^n\to\mathbb{R}^p$ are both linear operators. Seen as a matrix, F is typically column-stochastic while $\rho\in\Sigma^n$ is a discrete measure over the metric space $\mathcal{X}$ and $\theta\in\Sigma^m$ is the fixed observed discrete measure over the metric space $\mathcal{Y}$.

Problem (5.3) is a natural way to solve inverse problems on discrete measures where one assumes that

$$\theta \approx F\rho_0,$$

where ρ_0 is an unknown discrete measure over $\mathcal{Y}$ to recover from the observed θ. When $F = \mathrm{Id}$ and $\gamma = 0$, (5.3) is closely related to computing the Wasserstein gradient flow (aka JKO flow [34]) of $J \circ B$. The JKO flow was first studied in [34] as it relates to the Fokker-Planck equation before being generalized (cf. [2], [53]). Entropic regularization, i.e., with $\gamma > 0$, was studied in [44] to compute Wasserstein gradient flows over spaces of probability distributions with the topology induced by the Wasserstein metric.

Applying Fenchel-Rockafellar duality to (5.2) (see [45, Proposition 2.4] for the unregularized case and [23, Section 5.1] for the entropic case), it is straightforward to see that problem (5.3) reads also

$$\min_{\rho \in \Sigma^n} \sup_{\tau \in \mathbb{R}^m, \eta \in \mathbb{R}^m} \langle \tau, F\rho \rangle + \langle \eta, \theta \rangle - \gamma \sum_{j=1}^{m} \sum_{i=1}^{m} \exp\left(\frac{\tau_i + \eta_j - C_{i,j}}{\gamma} \right) + J \circ B(\rho). \tag{5.4}$$

Taking the supremum over η, one can easily show that (see also [29, Proposition 2.1]),

$$\min_{\rho \in \Sigma^n} \sup_{\tau \in \mathbb{R}^m} \langle \tau, F\rho \rangle - \gamma \sum_{j=1}^{m} \theta_j \log\left(\sum_{i=1}^{m} \exp\left(\frac{\tau_i - C_{i,j}}{\gamma} \right) \right) + J \circ B(\rho). \tag{5.5}$$

Remark 5.5. Observe in (5.5) that the smooth term in τ (excluding the inner product $\langle \tau, F\rho \rangle$) is actually a log-sum-exp smooth approximation of the $\max$ function, which would appear naturally when marginalizing with respect to η in the case $\gamma = 0$.

Now, dualizing on J, we finally get that (5.3) is equivalent to

$$\min_{\rho \in \mathbb{R}^n_+} \sup_{\tau \in \mathbb{R}^m, \zeta \in \mathbb{R}^p} \iota_{\{1\}}(\rho^T \mathbb{1}) + \langle (\tau, \zeta), (F\rho, B\rho) \rangle - \gamma \sum_{j=1}^{m} \theta_j \log\left(\sum_{i=1}^{m} \exp\left(\frac{\tau_i - C_{i,j}}{\gamma} \right) \right) - J^*(\zeta). \tag{5.6}$$

The problem in (5.6) is a saddle point problem which can be solved with Algorithm 1 by taking

$$\mathcal{C}_p = \mathbb{R}^n_+, \quad \mathcal{C}_d = \mathbb{R}^{m+p}, \quad T\rho = (F\rho, B\rho), \quad f(\rho) = 0, \quad g(\rho) = \iota_{\{1\}}(\rho^T \mathbb{1}),$$
$$l^*(\mu) = l^*(\zeta) = J^*(\zeta), \text{ and } h^*(\mu) = h^*(\tau) = \gamma \sum_{j=1}^{m} \theta_j \log\left(\sum_{i=1}^{m} \exp\left(\frac{\tau_i - C_{i,j}}{\gamma} \right) \right).$$

The natural choice for the entropies is, again,

$$\phi_p(x) = \sum_{i=1}^{n} x_i \log(x_i) \text{ and } \phi_d(\mu) = \frac{1}{2} \|\mu\|_2^2.$$

Lemma 5.6. *The function $h^*(\mu)$ is L_d Lipschitz-smooth for $L_d \geq \gamma^{-1} \sum_{j=1}^{m} \theta_j = \gamma^{-1}$.*

Proof. The log-sum-exp function (with temperature constant γ),

$$\mathrm{LSE}_\gamma(x) \stackrel{\text{def}}{=} \gamma \log\left(\sum_{i=1}^{n} \exp\left(\frac{x_i}{\gamma} \right) \right),$$

is C^2 and convex on $\mathbb{R}^n$ (see [27, Lemma 4], [51, Example 2.16, page 48]) and thus so is $h^*(\tau, \zeta)$. The gradient, $\nabla \mathrm{LSE}_\gamma(x)$, is given, component-wise, for each $k \in \{1, \ldots, n\}$ by

$$(\sigma_\gamma(x))^{(k)} = \frac{\exp(x_k/\gamma)}{\sum_{i=1}^{n} \exp(x_i/\gamma)}.$$

The function $\sigma_\gamma(x)$ is called the softmax function with temperature constant γ and is Lipschitz-continuous in the euclidean norm with Lipschitz constant γ^{-1} (see [27, Proposition 4]). Thus, to see that the function h^* is Lipschitz-smooth, denote the jth column of C as $C_{\cdot,j}$ and notice

$$h^*(\mu) = h^*(\tau) = \sum_{j=1}^m \theta_j \mathrm{LSE}_\gamma(\tau - C_{\cdot,j}) \implies \nabla h^*(\mu) = \nabla h^*(\tau) = \sum_{j=1}^m \theta_j \sigma_\gamma(\tau - C_{\cdot,j}).$$

With this we write,

$$\begin{aligned}
\left\|\nabla h^*(\mu) - \nabla h^*(\mu')\right\|_2 &= \left\|\sum_{j=1}^m \theta_j \left(\sigma_\gamma(\tau - C_{\cdot,j}) - \sigma_\gamma(\tau' - C_{\cdot,j})\right)\right\|_2 \\
&\leq \left(\sum_{j=1}^m \theta_j\right) \left\|\sigma_\gamma(\tau - C_{\cdot,j}) - \sigma_\gamma(\tau' - C_{\cdot,j})\right\|_2 \\
&\leq \gamma^{-1} \left(\sum_{j=1}^m \theta_j\right) \left\|\tau - \tau'\right\|_2
\end{aligned}$$

and the desired claim follows. □

It is clear that **(H)** holds in this setting. It remains to find suitable step sizes $(\lambda_k)_{k\in\mathbb{N}}$ and $(\nu_k)_{k\in\mathbb{N}}$ to satisfy **($\mathbf{A}_2$)** and **($\mathbf{A}_3$)**. Since the entropies here are exactly the same as in the linear inverse problem on the simplex, we refer again to Lemma 5.2. With these step sizes, we consider a one-dimensional instance of the problem with $n = 108$, $C_{i,j} = \frac{1}{2}\|i-j\|_2^2$, F a convolution operator with kernel $\exp\left(-\frac{1}{1-t^2}\right)$ for $t \in]-,1,1[$ and 0 otherwise, $J \circ B$ the total variation [52], and $\theta \approx F\rho_0$ our observation of $F\rho_0$ is corrupted by Dirichlet distributed noise. We take $x_0 = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ and $\mu_0 = 0$. The results are displayed in Figure 2.

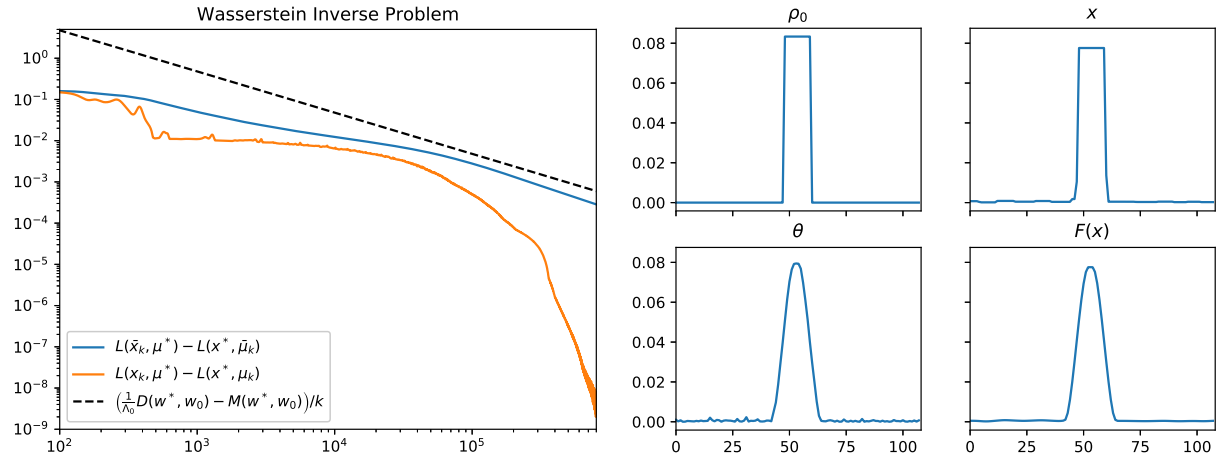

Figure 2: (Left) Ergodic and pointwise convergence profiles for Algorithm 1 applied to the Wasserstein inverse problem with entropic regularization parameter $\gamma = 1$, total variation regularization parameter $\beta = 1$, and $n = 108$. (Right) The ground truth measure ρ_0, the recovered measure x, the corrupted observation θ, and the image Fx of the recovered measure x.

Remark 5.7. A chief advantage of (5.6), in contrast to optimizing with respect to the transport plan π, is the significant difference in computational complexity, since the former is operating over $n+m+p$ variables only rather than nm. Indeed, one can rewrite problem (5.3) as

$$\min_{\substack{\rho\in\Sigma^n\\ \pi^T\mathbb{1}=\theta\\ \pi\mathbb{1}=F\rho\\ \pi\in\mathbb{R}_+^{n\times m}}} D_K(\pi,\xi)+J\circ B(\rho)=\min_{\substack{\rho\in\mathbb{R}^n\\ \pi\in\mathbb{R}^{n\times m}}} g(\rho,\pi)+f\circ L(\rho,\pi)$$

where $g(\rho,\pi)\overset{\text{def}}{=}\iota_{\mathbb{R}_+^n}(\rho)+\iota_{\mathbb{R}_+^{n\times m}}(\pi)+D_K(\pi,\xi)$, L is a linear operator defined as

$$L(\rho,\pi)\overset{\text{def}}{=}\begin{pmatrix} B\rho\\ \rho^T\mathbb{1}\\ -F\rho+\pi\mathbb{1}\\ \pi^T\mathbb{1}\end{pmatrix}$$

and $f(s,t,r,u)\overset{\text{def}}{=}J(s)+\iota_{\{1\}\times\{0\}\times\{\theta\}}(t,r,u)$. This formulation is solvable using the Chambolle-Pock algorithm of [14] but, in addition to working with much more variables, over higher dimensional spaces, does not exploit the geometryof the simplex, and requires computing prox mappings which are computationally more demanding. The prox operator associated to $D_K(\pi,\xi)+\iota_{\mathbb{R}_+^{n\times m}}(\pi)$ will require the Lambert W function which is a special function (see [26] for more). Even starting from the semidualized form (5.6) will require either sorting or incrteasing the number of dual variables if euclidean splitting methods like in [14] are applied.

Remark 5.8. Although we considered here only a simple Wasserstein inverse problem involving a single observed measure, Algorithm 1 and our problem framework readily extend to more complicated settings such as computing the Wasserstein barycenter of indirectly observed measures. Wasserstein barycenter problems were first introduced in [1] without entropic regularization of the Wasserstein distance. Later, the use of entropic regularization of the Wasserstein distance to speed up computation of barycenters was put forth in [23], however the barycenter itself was not regularized; such developments would come later, e.g., [13], [5], etc, and even then the problems considered did not include the possibility of observing the image of the measure θ under a linear operator F rather than observing the measure θ itself.

Let $q\in\mathbb{N}$ and consider q reference measures $\theta^i\in\mathbb{R}^{n_i}$ with $n_i\in\mathbb{N}$ for each $1\leq i\leq q$, each having been observed through some linear operator $F^i:\mathbb{R}^n\to\mathbb{R}^{n_i}$ applied to an unknown discrete measure $\rho^i\in\Sigma^n$, i.e., $\theta^i\approx F^i\rho^i$. Let $\alpha\in\Sigma^q$. Then we can write the regularized Wasserstein barycenter problem as

$$\min_{\rho\in\Sigma^n}\sum_{k=1}^{q}\alpha_q W_{\gamma_k}\left(F^k\rho,\theta^k\right)+\sum_{r=1}^{q'}J^r\circ B^r(\rho)$$

which is equivalent to the following,

$$\min_{\rho\in\Sigma^n}\sup_{\substack{\tau^1\in\mathbb{R}^{n_1},\ldots,\tau^q\in\mathbb{R}^{n_q}\\ \zeta^1\in\mathbb{R}^{m_1},\ldots,\zeta^{q'}\in\mathbb{R}^{m_{q'}}}}\sum_{k=1}^{q}\left[\langle\alpha_k\tau_k,F_k\rho\rangle-\alpha_k\gamma_k\sum_{j=1}^{n_k}\theta_j^k\log\left(\sum_{i=1}^{n_k}\exp\left(\frac{\tau_i^k-C_{i,j}^k}{\gamma_k}\right)\right)\right]$$
$$+\sum_{r=1}^{q'}\left[\langle\zeta^r,B^r\rho\rangle-(J^r)^*(\zeta^r)\right].$$

This formulation of the problem can be solved with with Algorithm 1 by taking

$$\mathcal{C}_p = \mathbb{R}^n_+, \quad \mathcal{C}_d = \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_q} \times \mathbb{R}^{m_1} \times \cdots \times \mathbb{R}^{m_{q'}}, \quad f(\rho) = 0, \quad g(\rho) = \iota_{\{\sum_{i=1}^n \rho_i = 1\}}(\rho),$$

$$l^*(\mu) = l^*(\zeta_1, \ldots, \zeta_{q'}) = \sum_{l=1}^{q'} J_l^*(\zeta_l), \text{ and}$$

$$h^*(\mu) = h^*(\tau^1, \ldots, \tau^q) = \sum_{k=1}^{q} \alpha_k \gamma_k \sum_{j=1}^{n_k} \theta_j^k \log\left(\sum_{i=1}^{n_k} \exp\left(\frac{\tau_i^k - C_{i,j}^k}{\gamma_k}\right)\right),$$

with the same entropy choices as we took for (5.6).

Remark 5.9. Consider the same setup as in the previous remark with $(\theta^1, \ldots, \theta^q)$ and let $\beta \in \mathbb{R}_+$. Another interesting formulation of the regularized Wasserstein barycenter problem that can be solved using Algorithm 1 is the following

$$\min_{\rho \in \Sigma^n} \min_{\rho_1, \ldots, \rho_q \in \Sigma^n} \sum_{i=1}^{q} \left(W_{\gamma_i}(\theta^i, F^i \rho_i) + J \circ A(\rho^i)\right) + \beta \sum_{i=1}^{q} \alpha_i W_{\gamma_i}(\rho_i, \rho).$$

This problem is simultaneously solving the Wasserstein inverse problem for each observed measure θ^i while also finding a barycenter ρ among the proposed solutions ρ^i of the Wasserstein inverse problems.

Acknowledgements

ASF was supported by the ERC Consolidated grant NORIA and by the Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant number FA9550-19-1-7026. JF was partly supported by Institut Universitaire de France. CM was supported by Project MONOMADS funded by Conseil Régional de Normandie.

A Appendix

Lemma A.1. *For each $k \in \mathbb{N}$, for a fixed batch size $0 < q < m$, let $B_k \subset \{1, \ldots, m\}$ be the batch of q indices sampled at iteration k and define $B_k^c \overset{\text{def}}{=} \{1, \ldots, m\} \setminus B_k$. Consider the error term induced by the batch sampling:*

$$\delta_k^p = -\nabla\left(\sum_{i \in B_k^c} (Ax)_i \log\left(\frac{(Ax)_i}{b_i}\right)\right).$$

If the entries of A are positive then $\left(\mathbb{E}\left[\left\langle \delta_k^p, x - x_{k+1}\right\rangle\right]\right)_{k \in \mathbb{N}}$ is bounded for all $x \in \Sigma^n$.

Proof. By Cauchy-Schwarz we have

$$\mathbb{E}\left[\left\langle \delta_k^p, x - x_{k+1}\right\rangle\right] \leq \mathbb{E}\left[\left\|\delta_k^p\right\| \|x - x_{k+1}\|\right] \leq \operatorname{diam}_{\Sigma^n} \mathbb{E}\left[\left\|\delta_k^p\right\|\right]$$

and so it suffices to bound $\mathbb{E}\left[\left\|\delta_k^p\right\|\right]$ for all $k \in \mathbb{N}$. Rather than bound the expectation itself, we will provide a coarse bound which holds deterministically for an arbitrary batch. For any batch $B \subset \{1, \ldots, m\}$ of size

$0 < q < m$, define $\tilde{A} \in \mathbb{R}_{++}^{(m-q)\times n}$ to be the matrix composed of rows which were not sampled in the batch B and similarly for $\tilde{b} \in \mathbb{R}_{++}^{m-q}$,

$$\tilde{A} \overset{\text{def}}{=} \left[A_{i,\cdot}\right]_{i\in B^c} \qquad \text{and} \qquad \tilde{b} \overset{\text{def}}{=} \left(b_i\right)_{i\in B^c}.$$

Using component wise $\log$ and division we have, for all $k \in \mathbb{N}$,

$$\begin{aligned}\left\|\nabla\left(\sum_{i\in B^c}\left(Ax_k\right)_i \log\left(\frac{\left(Ax_k\right)_i}{b_i}\right)\right)\right\| &= \left\|\tilde{A}^T \log\left(\frac{\tilde{A}x_k}{\tilde{b}}\right)\right\| \leq \left\|\tilde{A}\right\| \left\|\log\left(\frac{\tilde{A}x_k}{\tilde{b}}\right)\right\| \\ &\leq \left\|\tilde{A}\right\| \left(\left\|\log\left(\tilde{A}x_k\right)\right\| + \left\|\log\left(\tilde{b}\right)\right\|\right)\end{aligned}$$

As A and b are fixed from the problem data, $\left\|\tilde{A}\right\|$ and $\left\|\log\left(\tilde{b}\right)\right\|$ are bounded. All that remains is to bound $\left\|\log\left(\tilde{A}x_k\right)\right\|$, for which we first recall that $x_k \in \Sigma^n \cap \mathbb{R}_{++}^n$ for all $k \in \mathbb{N}$ by design of the algorithm and the choice of ϕ_p. Let $\underline{a} = \min_{i,j} A_{i,j} > 0$ and $\overline{a} = \max_{i,j} A_{i,j}$, then for each $i \in \{1, \ldots, m-q\}$

$$\log\left(\left\langle \tilde{A}_{i,\cdot}, x_k\right\rangle\right) \leq \log\left(\left\|\tilde{A}_{i,\cdot}\right\|_\infty \|x_k\|_1\right) = \log\left(\left\|\tilde{A}_{i,\cdot}\right\|_\infty\right) \leq \log\left(\overline{a}\right)$$

as well as

$$\log\left(\left\langle \tilde{A}_{i,\cdot}, x_k\right\rangle\right) \geq \log\left(\min_j \tilde{A}_{i,j}\right) \geq \log\left(\underline{a}\right)$$

so that the components of $\log\left(\tilde{A}x\right)$ are contained in a ball of radius $\max\left\{\left|\log\left(\underline{a}\right)\right|, \left|\log\left(\overline{a}\right)\right|\}$, which is finite since the entries of A are positive and A is fixed. Thus $\left\|\log\left(\tilde{A}x\right)\right\|$ is bounded and the proof is complete. □

Proposition A.2. *Assume* (H), ($\mathbf{A}_1$), *and* ($\mathbf{A}_{11}$) *hold and that ψ_p is totally convex and sequentially consistent on $\mathcal{U}_p$. Moreover, suppose that $\boldsymbol{S}_{\mathscr{P}} \cap \mathcal{U}_p \neq \emptyset$; meaning that there exists at least one solution $x^\star$ of ($\mathscr{P}$) in $\mathcal{U}_p$. Then, there exists a unique solution to the primal problem (i.e., $\boldsymbol{S}_{\mathscr{P}} = \{x^\star\}$).*

Proof. First notice that, from ($\mathbf{A}_{11}$), $\operatorname{dom}\left(\phi_p\right) \subseteq \operatorname{dom}\left(\psi_p\right)$ and so $\mathcal{U}_p \subseteq \operatorname{int}\operatorname{dom}\left(\phi_p\right) \subseteq \operatorname{int}\operatorname{dom}\left(\psi_p\right)$. As ψ_p is sequentially consistent on $\mathcal{U}_p$, we have that, for any bounded subset $V \subseteq \mathcal{U}_p$ and for any $t > 0$,

$$\inf_{x\in V} \Theta_{\psi_p}\left(x, t\right) > 0.$$

Suppose for instance that ($\mathbf{A}_{11}$) holds specifically with f relatively strongly convex with respect to ψ_p. Then, by Definition 2.4, for any $x, y \in \operatorname{int}\operatorname{dom}\left(\psi_p\right)$,

$$m_f D_{\psi_p}\left(x, y\right) \leq D_f\left(x, y\right).$$

Then, for any bounded subset $V \subseteq \mathcal{U}_p$ and for any $t > 0$,

$$0 < m_f \inf_{x\in V} \Theta_{\psi_p}\left(x, t\right) \leq \inf_{x\in V} \Theta_f\left(x, t\right),$$

and so f is sequentially consistent on $\mathcal{U}_p$. Recall from [11, Proposition page 50-51] that sequential consistency of f on the set $\mathcal{U}_p$ implies uniform convexity of f on any bounded subset $V \subseteq \mathcal{U}_p$. Denote by $x^\star$ a point in $\boldsymbol{S}_{\mathscr{P}} \cap \mathcal{U}_p$. As $x^\star \in \mathcal{U}_p$, that is an open set, there is a ball $\mathcal{B}_\tau(x^\star)$, for $\tau > 0$, which is bounded and contained in $\mathcal{U}_p$. In particular, f is uniformly convex on $\mathcal{B}_\tau(x^\star)$ and so is the objective function in ($\mathscr{P}$). Suppose by contradiction that there exists another solution $\bar{x} \in \boldsymbol{S}_{\mathscr{P}}$ with $\bar{x} \neq x^\star$. By convexity, the segment connecting $x^\star$ and $\bar{x}$ is contained in $\boldsymbol{S}_{\mathscr{P}}$. Then, the intersection of this segment with $\mathcal{B}_\tau(x^\star)$ has more than one element and is contained both in $\boldsymbol{S}_{\mathscr{P}} \cap \mathcal{B}_\tau(x^\star)$. This is a contradiction with the uniform convexity of the objective function in ($\mathscr{P}$) on $\mathcal{B}_\tau(x^\star)$.

□

References

[1] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[2] L. Ambrosio, N. Gigli, and G. Savare. *Gradient Flows*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2008.

[3] Heinz H. Bauschke, Jonathan M. Borwein, and Patrick L. Combettes. Essential smoothness, essential strict convexity, and legendre functions in banach spaces. *Communications in Contemporary Mathematics*, 03(04):615–647, 2001.

[4] Heinz H. Bauschke, Jonathan M. Borwein, and Patrick L. Combettes. Bregman monotone optimization algorithms. *SIAM Journal on Control and Optimization*, 42(2):596–636, 2003.

[5] Jérémie Bigot, Elsa Cazelles, and Nicolas Papadakis. Data-driven regularization of wasserstein barycenters with an application to multivariate density registration. *Information and Inference: A Journal of the IMA*, 8(4):719–755, 2019.

[6] Benjamin Birnbaum, Nikhil R Devanur, and Lin Xiao. Distributed algorithms via gradient descent for fisher markets. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 127–136, 2011.

[7] Kristian Bredies, Karl Kunisch, and Thomas Pock. Total generalized variation. *SIAM Journal on Imaging Sciences*, 3(3):492–526, 2010.

[8] F. E. Browder. Multi-valued monotone nonlinear mappings and duality mappings in banach spaces. *Trans. Amer. Math. Soc.*, 118:338–351, 1965.

[9] F. E. Browder. Fixed point theorems for nonlinear semicontractive mappings in banach spaces. *Arch. Rational Mech. Anal.*, 21:259–269, 1966.

[10] M. N. Bùi and P. L. Combettes. Bregman forward-backward operator splitting. *Set-Valued and Variational Analysis*, 29(3):583–603, 2021.

[11] Dan Butnariu, Alfredo Iusem, and Constantin Zalinescu. On uniform convexity, total convexity and convergence of the proximal point and outer bregman projection algorithms in banach spaces. *Journal of Convex Analysis*, 10, 01 2003.

[12] Dan Butnariu and Alfredo N Iusem. *Totally convex functions for fixed points computation and infinite dimensional optimization*, volume 40. Springer Science & Business Media, 2000.

[13] Elsa Cazelles, Jérémie Bigot, and Nicolas Papadakis. Regularized barycenters in the wasserstein space. In *International Conference on Geometric Science of Information*, pages 83–90. Springer, 2017.

[14] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[15] Antonin Chambolle, Matthias J. Ehrhardt, Peter Richtárik, and Carola-Bibiane Schönlieb. Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. *SIAM Journal on Optimization*, 28(4):2783–2808, 2018.

[16] Antonin Chambolle and Thomas Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1-2):253–287, 2016.

[17] Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

[18] I. Cioranescu. *Geometry of Banach spaces, DualityMappings and Nonlinear Problems*. Kluwer, Dordrecht, 1990.

[19] Patrick L. Combettes and Jean-Christophe Pesquet. Primal-dual splitting algorithm for solving inclusions with mixtures of composite, lipschitzian, and parallel-sum type monotone operators. *Set-Valued and Variational Analysis*, 20(2):307–330, Jun 2012.

[20] P.L. Combettes and Q. Nguyen. Solving composite monotone inclusions in reflexive banach spaces by constructing best bregman approximations from their kuhn-tucker set. *Journal of Convex Analysis*, 23:481–510, 05 2016.

[21] L. Condat. A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, pages 1–20, 2012.

[22] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[23] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. *Journal of Machine Learning Research*, 2014.

[24] Marco Cuturi and Gabriel Peyré. A smoothed dual approach for variational wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016.

[25] Marco Cuturi and Gabriel Peyré. Semidual regularized optimal transport. *SIAM Review*, 60(4), Jan 2018.

[26] Mireille El Gheche, Jean-Christophe Pesquet, and Joumana Farah. A proximal approach for optimization problems involving kullback divergences. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5984–5988. IEEE, 2013.

[27] Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning, 2017.

[28] Yiming Gao and Kristian Bredies. Infimal convolution of oscillation total generalized variation for the recovery of images with structured texture. *SIAM Journal on Imaging Sciences*, 11(3):2021–2063, 2018.

[29] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems 29*, pages 3440–3448. Curran Associates, Inc., 2016.

[30] Erfan Yazdandoost Hamedani and Necdet Serhat Aybat. A primal-dual algorithm for general convex-concave saddle point problems. *arXiv preprint arXiv:1803.01401*, 2018.

[31] J. Bolte H.H. Bauschke and M. Teboulle. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Math. Oper. Res.*, 42(2):330–348, 2017.

[32] Martin Holler and Karl Kunisch. On infimal convolution of tv-type functionals and applications to video and image reconstruction. *SIAM Journal on Imaging Sciences*, 7(4):2258–2300, 2014.

[33] Xin Jiang and Lieven Vandenberghe. Bregman primal–dual first-order method and application to sparse semidefinite programming. *arXiv preprint*, 2021.

[34] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

[35] N. Komodakis and J. Pesquet. Playing with duality: An overview of recent primal-dual approaches for solving large-scale optimization problems. *IEEE Signal Processing Magazine*, 32(6):31–54, 2015.

[36] VN Lebedev and NT Tynjanskiı. Duality theory of concave-convex games. In *Soviet Math. Dokl*, volume 8, pages 752–756, 1967.

[37] Haihao Lu. "relative continuity" for non-lipschitz nonsmooth convex optimization using stochastic (or deterministic) mirror descent. *INFORMS Journal on Optimization*, 1(4):288–303, 2019.

[38] Haihao Lu and Robert M. Freund. Generalized stochastic frank–wolfe algorithm with stochastic "substitute" gradient for structured convex optimization. *Mathematical Programming*, Mar 2020.

[39] Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

[40] L. McLinden. An extension of fenchel's duality theorem to saddle functions and dual minimax problems. *Pacific J. Math.*, 50(1):135–158, 1974.

[41] J.-J. Moreau. Théorèmes "inf-sup,". *C. R. Acad. Sci. Paris Sér. A Math.*, 258:2720–2722, 1964.

[42] Q. V. Nguyen. Forward-backward splitting with bregman distances. *Vietnam J. Math.*, 45(519–539), 2017.

[43] Z. Opial. Weak convergence of the sequence of successive approximations for nonexpansive mappings. *Bulletin of the American Mathematical Society*, 73(4):591–597, 1967.

[44] Gabriel Peyré. Entropic approximation of wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4):2323–2351, 2015.

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

[46] Mark Semenovich Pinsker. The information stability of gaussian random variables and processes (in russian)). In *Doklady Akademii Nauk*, volume 133, pages 28–30. Russian Academy of Sciences, 1960.

[47] B. T. Polyak. *Introduction to optimization*. Optimization Software, 1987.

[48] Julian Rasch and Chambolle Antonin. Inexact first-order primal–dual algorithms. *Computational Optimization and Applications*, 76(2):381–430, 2020.

[49] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In Jagdish S. Rustagi, editor, *Optimizing Methods in Statistics*, pages 233 – 257. Academic Press, 1971.

[50] R. T. Rockafellar. Minimax theorems and conjugate saddle-functions. *Mathematica Scandinavica*, 14(2):151–173, 1964.

[51] R. T. Rockafellar and R. Wets. *Variational analysis*, volume 317. Springer Verlag, 1998.

[52] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.

[53] Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017.

[54] Antonio Silveti-Falls, Cesare Molinari, and Jalal Fadili. Generalized conditional gradient with augmented lagrangian for composite minimization. *SIAM Journal on Optimization*, 30(4):2687–2725, 2020.

[55] S. Simons. *From Hahn-Banach to Monotonicity*, volume 1693 of *Lecture Notes in Math.* Springer-Verlag, New York, 2008.

[56] B. C. Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, pages 1–15, 2011.

[57] Hong Xu, Tae-Hwa Kim, and Ximing Yin. Weak continuity of the normalized duality map. *Journal of Nonlinear and Convex Analysis*, 15(3):595–604, 2014.

[58] C. Zălinescu. *Convex Analysis in General Vector Spaces*. World Scientific Publishing, River Edge, NJ, 2002.