
Fast Stochastic Bregman Gradient Methods: Sharp Analysis and Variance Reduction

Radu-Alexandru Dragomir^{*12} Mathieu Even^{*23} Hadrien Hendrikx^{*23}

Abstract

We study the problem of minimizing a relatively-smooth convex function using stochastic Bregman gradient methods. We first prove the convergence of Bregman Stochastic Gradient Descent (BSGD) to a region that depends on the noise (magnitude of the gradients) at the optimum. In particular, BSGD with a constant step-size converges to the exact minimizer when this noise is zero (*interpolation* setting, in which the data is fit perfectly). Otherwise, when the objective has a finite sum structure, we show that variance reduction can be used to counter the effect of noise. In particular, fast convergence to the exact minimizer can be obtained under additional regularity assumptions on the Bregman reference function. We illustrate the effectiveness of our approach on two key applications of relative smoothness: tomographic reconstruction with Poisson noise and statistical preconditioning for distributed optimization.

1. Introduction

We are interested in solving the minimization problem

$$\min_{x \in C} f(x), \text{ where } f(x) = \mathbb{E}_\xi [f_\xi(x)], \quad (1)$$

where C is a closed convex subset of \mathbb{R}^d and f_ξ are differentiable convex functions. These problems typically arise in machine learning when performing (empirical) risk minimization, in which case f_ξ is for instance a loss function for some random sample ξ . Problem (1) is also encountered in signal processing applications such as image deblurring or tomographic reconstruction inverse problems, in which the

goal is to recover an unknown signal from a large number of noisy observations. First-order methods are often very efficient for solving problems such as (1), but computing a gradient ∇f might be very expensive for large-scale problems (large number of components f_ξ), and even impossible in the case of true risk minimization (infinite number of f_ξ). In this case, stochastic gradient methods have proven to be particularly effective thanks to their low cost per iteration. The simplest one, Stochastic Gradient Descent (SGD), consists in updating x_t as

$$x_{t+1} = \arg \min_{x \in C} \left\{ \eta_t g_t^\top x + \frac{1}{2} \|x - x_t\|^2 \right\}$$

where g_t is a gradient estimate such that $\mathbb{E}[g_t] = \nabla f(x_t)$. In our case, a natural choice would be $g_t = \nabla f_{\xi_t}(x_t)$ for some ξ_t . The choice of the step size η_t is crucial for obtaining good performances and is typically related to the smoothness of f with respect to the Euclidean norm.

Beyond simply adapting the step size, a powerful generalization of SGD consists in refining the geometry and performing instead Bregman gradient (a.k.a mirror) steps as

$$x_{t+1} = \arg \min_{x \in C} \left\{ \eta_t g_t^\top x + D_h(x, x_t) \right\}, \quad (2)$$

where the Euclidean distance has been replaced by the Bregman divergence with respect to a reference function h , which writes:

$$D_h(x, y) = h(x) - h(y) - \nabla h(y)^\top (x - y), \quad (3)$$

for all $x \in \text{dom } h, y \in \text{int dom } h$. We make the following blanket assumptions on h throughout the article, which guarantee well-posedness of the update (2).

Assumption 1. *The function $h : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ is twice continuously differentiable and strictly convex on $\text{int } C$. Moreover, for every $y \in \mathbb{R}^d$, the problem*

$$\min_{x \in C} h(x) - x^\top y$$

has a unique solution, which lies in $\text{int } C$.

The standard SGD algorithm corresponds to the case where $h = \frac{1}{2} \|\cdot\|^2$. However, a different choice of h might better fit

^{*}Equal contribution ¹Université Toulouse 1 Capitole ²D.I. Ecole Normale Supérieure, CRNS, PSL University, Paris ³INRIA Paris. Correspondence to: Radu-Alexandru Dragomir <radu-alexandru.dragomir@inria.fr>, Mathieu Even <mathieu.even@ens.psl.eu>, Hadrien Hendrikx <hadrien.hendrikx@inria.fr>.

the geometry of the set C and the curvature of the function, allowing the algorithm to take larger steps in directions where the objective gradient changes slowly. This choice is guided by the notion of relative smoothness and strong convexity, introduced in Bauschke et al. (2017); Lu et al. (2018). Instead of the squared Euclidean norm for standard smoothness, *relative* regularity is measured with respect to the reference function h .

Definition 1. *The function f is said to be L -relatively smooth and μ -relatively strongly convex with respect to h if it is differentiable and for all $x, y \in \text{int dom } h$,*

$$\mu D_h(x, y) \leq D_f(x, y) \leq L D_h(x, y). \quad (4)$$

where D_f is defined similarly to (3). Note that if $\mu = 0$, the left-hand side inequality reduces to assuming convexity of f . Similarly, if $h = \frac{1}{2} \|\cdot\|^2$, then $D_h(x, y) = \frac{1}{2} \|x - y\|^2$, and the usual notions of smoothness and strong convexity are recovered. If both functions are two times differentiable, Equation (4) can be turned into an equivalent condition on the Hessians: $\mu \nabla^2 h(x) \preceq \nabla^2 f(x) \preceq L \nabla^2 h(x)$. Throughout the article, we will generally write $\mu_{f/h}$ and $L_{f/h}$ to insist on the relative aspect.

Writing the optimality conditions for the minimization problem of Equation (2), we obtain the following equivalent iteration, which is in the alternative Mirror Descent form (Nemirovsky and Yudin, 1983):

$$\nabla h(x_{t+1}) = \nabla h(x_t) - \eta_t g_t. \quad (5)$$

Although these updates have a closed-form solution for many choices of the reference function h , they may be harder to perform than standard gradient steps, since they require solving the subproblem defined in (2). Yet, this may be worth doing in some cases to reduce the overall iteration complexity, if the resulting majorization in (4) is much tighter than with the Euclidean distance. Let us list some applications of relative regularity:

Problems with unbounded curvature. Some problems have singularities at some boundary points in C where the Hessian grows arbitrarily large. In this situation, smoothness with respect to the Euclidean norm does not hold globally, and standard gradient methods become inefficient as they necessitate excessively small step sizes or costly line search procedures. A typical example arises in inverse problems with Poisson noise, which are used in particular for image deblurring (Bertero et al., 2009) or tomographic reconstruction (Kak and Slaney, 2001). In this case, the objective function involves the Kullback-Leibler divergence, which becomes singular as one of its arguments approaches 0. However, by choosing the reference function $h(x) = -\sum_{i=1}^d \log(x^{(i)})$, one can show that relative smoothness holds globally (Bauschke et al., 2017). For

more examples, see Lu et al. (2018); Bolte et al. (2018); Nesterov (2019); Mishchenko (2019).

Distributed optimization. When h approximates f in the sense of (4), Bregman methods can be used to speed up convergence by performing non-uniform preconditioning (Shamir et al., 2014; Reddi et al., 2016; Yuan and Li, 2020; Hendrikx et al., 2020b). Typically, h is chosen as the objective function on a smaller portion of the dataset of size n_{prec} (e.g., the dataset of the server), which improves the conditioning by a factor of up to n_{prec} compared to Euclidean methods, while naturally taking advantage of an eventually small effective dimension of the dataset (Even and Massoulié, 2021). In this case, forming the gradient g_t requires communication with the workers (where most of the data is held), and is thus expensive. Although the updates may not have a simple expression, the inner problem of Equation (2) can be solved locally at the server without additional communications. Therefore, Bregman methods allow to drastically reduce the communication cost by reducing the overall iteration complexity.

Despite these applications, there are still many gaps in our understanding of convergence guarantees of Bregman gradient methods. In particular, most existing results focus on the deterministic case $g_t = \nabla f(x_t)$, or do not leverage the relative regularity assumptions.

Contributions and outline

In this work, we develop convergence theorems for Bregman SGD in the relatively-smooth setting. Our variance condition depends on the magnitude of the stochastic gradients at the optimum, and can thus be much smaller than the one used in Hanzely et al. (2018), in particular for overparametrized models (which verify the interpolation condition that all stochastic gradients are equal to 0 at the optimum). Our analysis relies on the Bregman generalization of a few technical lemmas such as the celebrated $\|a + b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$ inequality (Lemma 2) or the co-coercivity inequality (Lemma 3), which we believe to be of independent interest.

Then, we show that variance-reduction techniques, which are widely used to accelerate traditional Euclidean stochastic methods when the objective has a finite-sum structure (Schmidt et al., 2013; Johnson and Zhang, 2013; Defazio et al., 2014; Allen-Zhu, 2017), can be adapted to the Bregman setting. Although this generally requires stronger regularity assumptions (such as global smoothness of h and Lipschitz continuity of $\nabla^2 h^*$), we show that the asymptotic rate of convergence solely depends on relative regularity constants. The same type of results (asymptotic speedup under additional smoothness assumptions) is observed when applying Nesterov-type acceleration to Bregman gradient methods (Hanzely et al., 2018; Dragomir et al., 2021; Hen-

driks et al., 2020b). We provide a summary of the rates proven in this paper in the appendix.

We start by discussing the related work in Section 2. Then, Section 3 presents the results for stochastic gradient descent, along with the main technical lemmas. Section 4 develops a Bregman version of the standard SAGA algorithm (Defazio et al., 2014). Finally, Section 5 illustrates the efficiency of the proposed methods on several applications, including Poisson inverse problems, tomographic reconstruction and distributed optimization.

2. Related work

The Bregman gradient method was first introduced as the *Mirror Descent* scheme¹ (Nemirovsky and Yudin, 1983; Beck and Teboulle, 2003) for minimizing convex nonsmooth functions, and enjoyed notable success in online learning (Bubeck, 2011). More recently, the introduction of relative smoothness (Bauschke et al., 2017; Lu et al., 2018; Bolte et al., 2018) has also brought interest in applying Bregman methods to differentiable objectives. This condition guides the choice of a well-suited reference function h which can greatly improve efficiency over standard gradient descent. While the vanilla Bregman descent method yields the same convergence rate as the Euclidean counterpart, subsequent work has focused on obtaining better rates with acceleration schemes (Hanzely et al., 2018). However, lower bounds show that the rates for relatively smooth optimization cannot be accelerated in general (Dragomir et al., 2021), and that additional regularity assumptions are needed. Similar notions of relative regularity have also been investigated for non-differentiable functions, such as relative continuity (Lu, 2019; Antonakopoulos et al., 2019). Zhou et al. (2020) also study non-differentiable functions, but in the online setting and without relative continuity.

Stochastic optimization methods, and in particular SGD, are very efficient when the number of samples is high (Bottou, 2012) and are often referred to as “the workhorse of machine learning”. The problem with SGD is that, in general, it only converges to a neighbourhood of the optimum unless a diminishing step-size is used. Variance reduction can be used to counter this problem, and many variance-reduced methods have been developed, such as SAG (Schmidt et al., 2013), SDCA (Shalev-Shwartz and Zhang, 2013; Shalev-Shwartz, 2016), SVRG (Johnson and

Zhang, 2013) or SAGA (Defazio et al., 2014).

Surprisingly, stochastic Bregman gradients algorithms have received less attention. Hanzely and Richtárik (2018); Gao et al. (2020); Hendriks et al. (2020a) study Bregman coordinate descent methods, and Zhang and He (2018) study the non-convex non-smooth setting. Antonakopoulos et al. (2020) study stochastic algorithms for online optimization, under Riemann-Lipschitz continuity. In contrast, our work focuses on Bregman SGD for relatively-smooth objectives. Hanzely and Richtárik (2018) study the same setting and obtain comparable convergence rates, but with a much looser notion of variance, which we discuss more in details in the next section. This is problematic since their bound on the variance is thus proportional to the magnitude of the gradients along the trajectory, and may thus be very large when far from the optimum if f is strongly convex. In contrast, our definition of variance leverages the stochastic gradients at the optimum, which allows us to obtain significant results without bounded gradients and in the interpolation regime (zero gradients at the optimum). In particular, our analysis can be seen as a Bregman generalization of the analysis from Gower et al. (2019). Davis et al. (2018) also analyze a similar setting, but again with more restrictive assumptions on the noise and boundedness of the gradients. Besides, to the best of our knowledge, variance reduction for Bregman stochastic methods was only studied in Shi et al. (2017) in the context of stochastic saddle-point optimization, but without leveraging relative regularity assumptions like we do in this work.

When preparing the final version of this paper, we became aware of the recent work by (Latafat et al., 2021), who also study Bregman stochastic algorithms for finite-sum minimization. However, the setting they consider is more restricted, as the convergence rates depend on the strong convexity modulus of h as well as the Euclidean smoothness constant of f .

3. Bregman Stochastic Gradient Descent

3.1. Preliminaries

We start by introducing a few technical lemmas, which are Bregman analogs to well-known Euclidean results, and which are at the heart of our analysis. All missing proofs can be found in Appendix A.

Recall that the conjugate h^* is defined for $y \in \mathbb{R}^d$ as $h^*(y) = \sup_{x \in \mathbb{R}^d} x^\top y - h(x)$. In particular, under Assumption 1, h^* is convex and differentiable on \mathbb{R}^d (Bauschke and Combettes, 2011, Cor. 18.12), and $\nabla h^*(\nabla h(y)) = y$ for $y \in \text{int } C$, which implies the following result:

Lemma 1 (Duality). *For $x, y \in \text{int dom } h$, we have $D_h(x, y) = D_{h^*}(\nabla h(y), \nabla h(x))$.*

¹Note that *Mirror Descent* and *Bregman Gradient* refer to the same algorithm, but that *Mirror Descent* is typically used when f is non-smooth, or in the online optimization community, whereas *Bregman Gradient* is generally preferred when using the relative smoothness assumption. Yet, both names are valid and there are exceptions, for instance Hanzely and Richtárik (2018) use the *Mirror Descent* terminology although they assume relative smoothness.

See, e.g., [Bauschke and Borwein \(1997, Thm 3.7.\)](#) for the proof. Using duality, we prove the following key lemma:

Lemma 2. *Let x^+ be such that $\nabla h(x^+) = \nabla h(x) - g$, and similarly define x_1^+ and x_2^+ from g_1 and g_2 . Then, if $g = \frac{g_1 + g_2}{2}$, we obtain:*

$$D_h(x, x^+) \leq \frac{1}{2} [D_h(x, x_1^+) + D_h(x, x_2^+)].$$

Lemma 2 can be adapted for any $g = (1 - \alpha)g_1 + \alpha g_2$ with $\alpha \in [0, 1]$. In the Euclidean case $h = \|\cdot\|^2$, we recover $\|\frac{g_1 + g_2}{2}\|^2 \leq \frac{1}{2} (\|g_1\|^2 + \|g_2\|^2)$. We now generalize the cocoercivity of the gradients ([Nesterov, 2003, Eq. 2.1.7](#)) to the relatively smooth case:

Lemma 3 (Bregman Cocoercivity). *If a convex function f is relatively L -smooth w.r.t to h , then for any $\eta \leq \frac{1}{L}$,*

$$D_f(x, y) \geq \frac{1}{\eta} D_{h^*}(\nabla h(x) - \eta(\nabla f(x) - \nabla f(y)), \nabla h(x))$$

3.2. Variance definition

We start by specifying two assumptions on the structure of the noise. Note that we use a constant step-size $\eta > 0$ throughout this section for simplicity, but similar results hold with decreasing step-sizes. We denote $x^* = \arg \min_x f(x)$ the minimizer of f and $\|x\|_H^2 = x^\top H x$ for a positive definite operator H and $x \in \mathbb{R}^d$.

Assumption 2. *The stochastic gradients g_t are such that $g_t = \nabla f_{\xi_t}(x_t)$, with $\mathbb{E}_{\xi_t}[f_{\xi_t}] = f$ and f_{ξ_t} is convex and $L_{f/h}$ -relatively smooth with respect to h for all ξ_t . Besides, there exists a constant $\sigma^2 \geq 0$ such that:*

$$\begin{aligned} \sigma^2 &\geq \frac{1}{2\eta^2} \mathbb{E}_{\xi_t} [D_{h^*}(\nabla h(x_t) - 2\eta \nabla f_{\xi_t}(x^*), \nabla h(x_t))] \\ &= \mathbb{E}_{\xi_t} [\|\nabla f_{\xi_t}(x^*)\|_{\nabla^2 h^*(z_t)}^2], \end{aligned}$$

for some $z_t \in [\nabla h(x_t) - 2\eta \nabla f_{\xi_t}(x^*), \nabla h(x_t)]$.

The assumption that the stochastic gradients are actual gradients of stochastic functions which are themselves smooth with respect to h is rather natural, as already discussed in the introduction. It is at the heart of variance reduction in the finite sum setting (though the sum does not need to be finite in the case of Assumption 2), and is in particular verified when solving (Empirical) Risk minimization problems.

Yet, it prevents the analysis from applying to coordinate descent methods for instance, in which $g_t = \nabla_i f(x_t)$, with $i \in \{1, \dots, d\}$. However, in this case, the extra structure can also be leveraged to obtain similar results ([Hanzely and Richtárik, 2018; Hendriks et al., 2020a; Gao et al., 2020](#)).

For the variance, Assumption 2 is a Bregman adaptation of the usual variance at the optimum definition used

for instance in [Bach and Moulines \(2011\); Gower et al. \(2019\)](#). Note that if h^* is μ_h -strongly convex with respect to the Euclidean norm, then the assumption is verified for instance when the variance is bounded in ℓ_2 norm, as $\|\nabla f_{\xi_t}(x^*)\|_{\nabla^2 h^*(z_t)}^2 \leq \mu_h^{-1} \|\nabla f_{\xi_t}(x^*)\|^2$ (we used the fact that if h is μ_h -strongly convex, then h^* is $1/\mu_h$ -smooth, see e.g., [Kakade et al. \(2009\)](#)).

We now compare our noise assumption with ([Hanzely and Richtárik, 2018, Assumption 5.1.](#)), which writes:

$$\frac{1}{\eta_t} \mathbb{E}_{\xi_t} [(\nabla f(x_t) - \nabla f_{\xi_t}(x_t))^\top (x_{t+1} - \bar{x}_{t+1})] \leq \sigma^2, \quad (6)$$

for $t \geq 0$, where g_t is the stochastic gradient estimate and \bar{x}_{t+1} is the output of the (theoretical) Bregman gradient step taken with the true gradient, that is, $\nabla h(\bar{x}_{t+1}) = \nabla h(x_t) - \eta_t \nabla f(x_t)$. Thus, their condition can be written:

$$\frac{1}{\eta_t^2} \mathbb{E}_{\xi_t} [D_h(x_{t+1}, \bar{x}_{t+1}) + D_h(\bar{x}_{t+1}, x_{t+1})] \leq \sigma^2,$$

so that σ^2 bounds at each step the distance (in the Bregman sense) between x_{t+1} and \bar{x}_{t+1} , the point that would be obtained by the expected (deterministic) gradient update. To illustrate why our assumption is weaker, let us consider the case where h is μ_h -strongly convex. In this setting, a sufficient condition for (6) to hold is that

$$\frac{1}{\mu_h} \mathbb{E}_{\xi_t} [\|\nabla f(x_t) - \nabla f_{\xi_t}(x_t)\|^2] \leq \sigma^2, \quad (7)$$

while a sufficient condition for our variance definition to hold is (using that $\nabla f(x^*) = 0$):

$$\frac{1}{\mu_h} \mathbb{E}_{\xi_t} [\|\nabla f(x^*) - \nabla f_{\xi_t}(x^*)\|^2] \leq \sigma^2, \quad (8)$$

which only depends on the magnitude of the gradients at the optimum instead of the variance along the full trajectory since x_t is replaced by x^* . In particular, in the *interpolation setting* where $\nabla f_{\xi}(x^*) = 0$ for every ξ , $\sigma^2 = 0$ with our condition. Besides, if f is strongly convex then the norm of its gradients increases when far from the optimum, and so one needs to restrict x_t to a compact set of \mathbb{R}^d for a condition such as (7) to hold. In contrast, the condition from (8) can hold globally without further assumptions.

3.3. Convergence results

We now prove the actual convergence theorems for Bregman SGD. To avoid notation clutter, we generally omit with respect to which variable expectations are taken when clear from the context.

Theorem 1. *If f is $L_{f/h}$ -smooth and $\mu_{f/h}$ -strongly convex relative to h with $\mu_{f/h} > 0$, and Assumptions 1 and 2 hold,*

then for $\eta \leq 1/(2L_{f/h})$, the iterates produced by Bregman stochastic gradient (2) satisfy

$$\mathbb{E}[D_h(x^*, x_t)] \leq (1 - \eta\mu_{f/h})^t D_h(x^*, x_0) + \eta \frac{\sigma^2}{\mu_{f/h}}. \quad (9)$$

Note that since we are in a Bregman setting, convergence is measured in terms of $D_h(x^*, x_t)$, the distance between x^* and x_t in the metric induced by h . If h is μ_h -strongly convex, then $D_h(x^*, x_t) \geq \frac{\mu_h}{2} \|x_t - x^*\|^2$ and convergence in ℓ_2 distance is recovered.

Proof. By using Lemma 4 from Appendix A, we obtain:

$$\begin{aligned} \mathbb{E}_{\xi_t}[D_h(x^*, x_{t+1})] &= D_h(x^*, x_t) - \eta D_f(x^*, x_t) \\ &\quad - \eta D_f(x_t, x^*) + \mathbb{E}_{\xi_t}[D_h(x_t, x_{t+1})]. \end{aligned} \quad (10)$$

Using Lemma 2, the last term can be bounded as $D_h(x_t, x_{t+1}) \leq \frac{1}{2}[D_1 + D_2]$. We use Lemma 3 (Bregman co-coercivity) to write:

$$\begin{aligned} D_1 &= D_{h^*}(\nabla h(x_t) - 2\eta[\nabla f_{\xi_t}(x_t) - \nabla f_{\xi_t}(x^*)], \nabla h(x_t)) \\ &\leq 2\eta D_{f_{\xi_t}}(x_t, x^*), \end{aligned}$$

so that $\mathbb{E}_{\xi_t}[D_1/2] \leq \eta D_f(x_t, x^*)$. Similarly,

$$D_2 = D_{h^*}(\nabla h(x_t) - 2\eta \nabla f_{\xi_t}(x^*), \nabla h(x_t)), \quad (11)$$

so that $\mathbb{E}_{\xi_t}[D_2/2] \leq \eta^2 \sigma_t^2$. Thus, using the relative strong convexity of f to bound the $D_f(x^*, x_t)$ term, we obtain:

$$\mathbb{E}_{\xi_t} D_h(x^*, x_{t+1}) \leq (1 - \eta\mu_{f/h}) D_h(x^*, x_t) + \eta^2 \sigma^2, \quad (12)$$

which yields the desired result. \square

Remark 1 (Interpolation). *In the interpolation setting (when $\nabla f_{\xi_t}(x^*) = 0$ for all ξ_t), we have that $\sigma^2 = 0$. Theorem 1 thus proves linear convergence in this case. For instance, when solving objectives of the form $D_{\text{KL}}(Ax, b)$ (which has applications in optimal transport (Mishchenko, 2019)) or $D_{\text{KL}}(b, Ax)$ (which has application in deblurring or tomographic reconstruction), then the variance as defined in Hanzely and Richtárik (2018) may be unbounded, whereas the variance as we define it is equal to 0 if there exists z such that $Az = b$.*

When f is convex ($\mu_{f/h} = 0$), Theorem 1 can be adapted to obtain a $1/T$ decrease of the error up to a noise region.

Theorem 2 (Convex case). *Under the same assumptions as Theorem 1, if $\mu = 0$, then*

$$\mathbb{E} \left[\frac{1}{T} \sum_{t=0}^T D_f(x^*, x_t) \right] \leq \frac{D_h(x^*, x_0)}{\eta T} + \eta \sigma^2 \quad (13)$$

Contrary to the Euclidean case, we do not obtain a guarantee on the average iterate in general. This is because the bound is on the average of $D_f(x^*, x_t)$ instead of $D_f(x_t, x^*)$, and Bregman divergences are not necessarily convex in their second argument (except for the Euclidean distance and Kullback-Leibler divergence). Therefore, the final bound is obtained on $\min_t D_f(x^*, x_t)$, meaning that there is at least one x_t such that this is true. Note that the nice properties regarding interpolation still hold in this setting.

Proof. We start from Lemma 4 and bound the $D_h(x_t, x_{t+1})$ in the same way as when $\mu > 0$, which yields:

$$\eta D_f(x^*, x_t) = D_h(x^*, x_t) - \mathbb{E}_{\xi_t}[D_h(x^*, x_{t+1})] + \eta^2 \sigma^2.$$

Averaging over t and dividing by η leads to (13). \square

4. Variance reduction

We have shown in the previous section that BSGD enjoys guarantees that are similar to that of its Euclidean counterpart, although the notion of variance needs to be adapted. We show in this section that it is also possible to apply variance reduction to accelerate convergence. To this end, we solve for $n \in \mathbb{N}^*$ and some convex functions f_i :

$$\min_{x \in C} f(x) := \frac{1}{n} \sum_{i=1}^n f_i(x). \quad (14)$$

The difference with Section 3 is that we now assume that f is a finite sum, which is required for variance reduction. We also assume that the minimizer x^* belongs to $\text{int } C$, so that $\nabla f(x^*) = 0$. The case where x^* lies on the border of C is more delicate, as h might not be differentiable there (e.g., the log-barrier); this would require an involved technical analysis which we leave for future work.

To solve Problem (14), we consider Algorithm 1, which is a Bregman adaptation of the SAGA algorithm (Defazio et al., 2014). Following its Euclidean counterpart, Algorithm 1 stores the stochastic gradients computed at each iteration, and reuses them to estimate the full gradient. Therefore, only one stochastic gradient needs to be computed at each iteration, thus drastically reducing the iteration cost compared to batch gradient descent, at the expense of using more memory. Note that the stochastic updates are unbiased since $\mathbb{E}_i[g_t] = \nabla f(x_t)$, and at the optimum (when $x_t = \phi_i = x^*$ for all i), $g_t = \nabla f(x^*) = 0$ so the variance at the optimum is 0 (contrary to SGD). We now study the convergence guarantees of Algorithm 1 in more details.

4.1. Convergence Results

For analyzing the Bregman-SAGA scheme, we first need to introduce, in addition to relative smoothness, an assumption on the regularity of D_h .

Algorithm 1 Bregman-SAGA($(\eta_t)_{t \geq 0}, x_0$)

- 1: $\phi_i = x_0$ for $i = 1, \dots, n$
- 2: **for** $t = 0, 1, 2, \dots$ **do**
- 3: Pick $i_t \in \{1, \dots, n\}$ uniformly at random
- 4: $g_t = \nabla f_{i_t}(x_t) - \nabla f_{i_t}(\phi_{i_t}^t) + \frac{1}{n} \sum_{j=1}^n \nabla f_j(\phi_j^t)$
- 5: $x_{t+1} = \arg \min_x \{ \eta_t g_t^\top x + D_h(x, x_t) \}$
- 6: $\phi_{i_t}^{t+1} = x_{t+1}$, and store $\nabla f_{i_t}(\phi_{i_t}^{t+1})$.
- 7: $\phi_j^{t+1} = \phi_j^t$ for $j \neq i_t$.
- 8: **end for**

Assumption 3. For all $i \in \{1, \dots, n\}$, f_i is $L_{f/h}$ relatively smooth w.r.t. h , and f is $\mu_{f/h}$ relatively strongly convex w.r.t. h . Moreover, there exists a gain function G such that for any $x, y, v \in \mathbb{R}^d$ and $\lambda \in [-1, 1]$,

$$D_{h^*}(x + \lambda v, x) \leq G(x, y, v) \lambda^2 D_{h^*}(y + v, y).$$

Such structural assumptions appear to be essential for analyzing Bregman-type methods that use information provided by gradients of past iterates. The function G models the fact that the Bregman divergence $D_{h^*}(x + v, x)$ is not homogeneous nor invariant to translation in x in general (except for the Euclidean case where it is equal to $\|v\|^2/2$). Note that such difficulties are also encountered for obtaining accelerated rates with inertial variants of Bregman descent, where similar assumptions are needed (Hanzely et al., 2018). This seems unavoidable, as suggested by the lower bound in Dragomir et al. (2021).

Although the gain function G is relatively abstract at this point, it plays a key role in defining the step-size, and convergence guarantees similar those of Euclidean SAGA can be obtained provided G can be chosen small enough. We first state the general Theorem 3 (convergence proof for Algorithm 1), and then detail how G can be bounded in several interesting cases.

For $t \geq 0$ and step-sizes $\eta_t > 0$, define $H_t = \frac{1}{n} \sum_{i=1}^n D_{f_i}(\phi_i^t, x^*)$, and the potential ψ_t as follows:

$$\psi_t = \frac{1}{\eta_t} D_h(x^*, x_t) + \frac{n}{2} H_t. \quad (15)$$

First note that by convexity of h and of the f_i , $\psi_t \geq 0$ for all t . Our goal in this section is to show that $\{\psi_t\}_{t \geq 0}$ converges to 0 at a given speed. Indeed, since $D_h(x^*, x_t) \leq \psi_t$, this implies (as in Section 3) that x_t converges to x^* at the same rate. To ease notations, we define

$$\bar{\alpha}^t = \frac{1}{n} \sum_{j=1}^n \nabla f_j(\phi_j^t), \text{ and } \bar{\alpha}_i^t = \nabla f_i(\phi_i^t) - \bar{\alpha}^t. \quad (16)$$

Theorem 3. Assume that Algorithm 1 is run with a step size sequence $\{\eta_t\}_{t \geq 0}$ satisfying $\eta_t = 1/(8L_{f/h}G_t)$ for

every $t \geq 0$, with G_t decreasing in t and such that for all $j \in \{1, \dots, n\}$:

$$G_t \geq G \left(\nabla h(x_t), \nabla h(x_t), \frac{1}{L_{f/h}} (\nabla f_j(x_t) - \nabla f_j(x^*)) \right),$$

$$G_t \geq G \left(\nabla h(x_t) - 2\eta_t \bar{\alpha}^t, \nabla h(\phi_j^t), \frac{1}{L_{f/h}} (\nabla f_j(\phi_j^t) - \nabla f_j(x^*)) \right).$$

Then, under Assumptions 1 and 3, the potential ψ_t satisfies

$$\mathbb{E}_{i_t} [\psi_{t+1}] \leq \left(1 - \min \left(\eta_t \mu_{f/h}, \frac{1}{2n} \right) \right) \psi_t, \quad (17)$$

In the convex case ($\mu_{f/h} = 0$), we obtain that

$$\mathbb{E} \left[\frac{1}{4T} \sum_{t=1}^T D_f(x_t, x^*) + H_t \right] \leq \frac{\psi_0}{T}. \quad (18)$$

Proof. Similarly to BSGD, we apply Lemma 4 (Appendix A), which yields

$$\mathbb{E}_{i_t} [D_h(x^*, x_{t+1})] = D_h(x^*, x_t) - \eta_t D_f(x^*, x_t) - \eta_t D_f(x_t, x^*) + \mathbb{E}_{i_t} [D_h(x_t, x_{t+1})]. \quad (19)$$

Lemmas 1 and 2 yield $D_h(x_t, x_{t+1}) \leq (D_1 + D_2)/2$, with

$$D_1 = D_{h^*}(\nabla h(x_t) - 2\eta_t [\nabla f_{i_t}(x_t) - \nabla f_{i_t}(x^*)], \nabla h(x_t)),$$

$$D_2 = D_{h^*}(\nabla h(x_t) - 2\eta_t (\nabla f_{i_t}(x^*) - \bar{\alpha}_i^t), \nabla h(x_t)).$$

Using Assumption 3 together with Lemma 3, we obtain:

$$D_1 \leq 4\eta_t^2 L_{f/h}^2 G_t \times$$

$$D_{h^*}(\nabla h(x_t) - \frac{1}{L_{f/h}} [\nabla f_{i_t}(x_t) - \nabla f_{i_t}(x^*)], \nabla h(x_t))$$

$$\leq 4\eta_t^2 L_{f/h} G_t D_{f_i}(x_t, x^*).$$

To bound the second term, we use Lemma 5 (Pfau, 2013) which is a Bregman version of the bias-variance decomposition. We write $V = 2\eta_t [\nabla f_{i_t}(\phi_{i_t}^t) - \nabla f_{i_t}(x^*)]$, so that $\mathbb{E}_{i_t} [V] = 2\eta_t \bar{\alpha}^t$ and:

$$\mathbb{E}_{i_t} [D_2] = \mathbb{E}_{i_t} [D_{h^*}(\nabla h(x_t) - \mathbb{E}_{i_t} [V] + V, \nabla h(x_t))]$$

$$\leq \mathbb{E}_{i_t} [D_{h^*}(\nabla h(x_t) - \mathbb{E}_{i_t} [V] + V, \nabla h(x_t) - \mathbb{E}_{i_t} [V])]$$

$$\leq 4\eta_t^2 L_{f/h}^2 G_t \mathbb{E}_{i_t} D_{h^*}(\nabla h(\phi_{i_t}^t) - L_{f/h}^{-1} [\nabla f_{i_t}(\phi_{i_t}^t) - \nabla f_{i_t}(x^*)], \nabla h(\phi_{i_t}^t))$$

$$\leq 4\eta_t^2 G_t L_{f/h} \mathbb{E}_{i_t} [D_{f_i}(\phi_{i_t}^t, x^*)]$$

where we used the gain function for translation and rescaling the step size. Following Hofmann et al. (2015), we write:

$$\mathbb{E}_{i_t} [H_{t+1}] = \left(1 - \frac{1}{n} \right) H_t + \frac{1}{n} D_f(x_t, x^*). \quad (20)$$

Therefore, we can use the $-H_t/n$ term to control the excess term from bounding $D_h(x_t, x_{t+1})$. In the end, we obtain:

$$\mathbb{E}_{i_t} [\psi_{t+1} - \psi_t] \leq -D_f(x^*, x_t) - \left(\frac{1}{2} - 2\eta_t L_{f/h} G_t \right) H_t - \left(1 - 2\eta_t L_{f/h} G_t - \frac{1}{2} \right) D_f(x_t, x^*).$$

If we choose $\eta_t \leq 1/(8L_{f/h}G_t)$ then the last term is positive and $1 - 4\eta_t L_{f/h}G_t \geq 1/2$. If $\mu_{f/h} > 0$ then we use the relative strong convexity of f to obtain that the right hand side is proportional to ψ_t , thus leading to a linear convergence rate. Otherwise, we obtain a telescopic sum, leading to the $1/T$ rate of Equation (18). \square

Note that the monotonicity of η_t (through G_t) is a technical condition to ensure that the Lyapunov is non-increasing. Otherwise, ψ_t could blow up even though x_{t+1} is very close to x_t , simply because η_t shrinks. It could be replaced by the condition that η_t does not vary too much (not more than a factor $1 - O(1/n)$), which achieves the same goal. The rest of this section is devoted to shong that non-trivial G_t can be chosen in many cases, thus leading to strong convergence guarantees. In particular, the rate recovers that of Euclidean SAGA in case h is a quadratic form.

Corollary 1. *If $\nabla^2 h$ is constant (h is quadratic), then Assumption 3 is satisfied with $G = 1$, so that*

$$\mathbb{E} [\psi_t] \leq \left(1 - \min \left(\frac{1}{8\kappa_{f/h}}, \frac{1}{2n} \right) \right)^t \psi_0, \quad (21)$$

where $\kappa_{f/h} = L_{f/h}/\mu_{f/h}$ is the relative condition number.

If h is not quadratic, but f^* and h^* are regular with respect to a norm, then strong guarantees can also be obtained:

Corollary 2. *If h^* is μ_h^{-1} -smooth and f^* is L_f^{-1} -strongly convex with respect to a norm $\|\cdot\|^2$, then the stepsize can be chosen constant as $\eta_t = \frac{\mu_h}{8L_f}$, and*

$$\mathbb{E} [\psi_t] \leq \left(1 - \min \left(\frac{\mu_h \mu_{f/h}}{8L_f}, \frac{1}{2n} \right) \right)^t \psi_0. \quad (22)$$

Note that following Kakade et al. (2009), having h^* be μ_h^{-1} -smooth is equivalent to having h be μ_h strongly-convex.

Proof. The proof follows the same step as the proof of Theorem 3, but the translation invariance and homogeneity are obtained by comparison with the norm, instead of using Assumption 3. Thus, we pay a factor μ_h^{-1} when bounding D_{h^*} by the norm, and a factor L_f when bounding the norm by D_{f^*} . It is also possible to directly use Assumption 3, but in this case the L_f factor is replaced by $L_{f/h}L_h$, which is an upper bound on L_f , and may thus be slightly looser. \square

Note that Corollary 1 is actually a consequence of Corollary 2, since $\mu_h = 1$ and $L_f = L_{f/h}$ if D_h is a norm itself. Otherwise, the constant G_t is chosen in a rather pessimistic way, and depends on the difference between directly bounding D_f by D_h (in which case we pay a factor $L_{f/h}$), or going through a norm $\|\cdot\|$ in the middle (in which we case we pay $L_f/\mu_h \geq L_{f/h}$).

As stated at the beginning of this section, one of the problems is that Bregman divergences lack translation invariance and homogeneity. However, as the algorithm converges, one can expect these conditions to hold locally, as $D_{h^*}(x+v, x)$ is approximated by $\frac{1}{2}\|v\|_{\nabla^2 h^*(x^*)}^2$ for small enough v , and x close enough to x^* . This is indeed what happens under enough regularity assumptions on h .

Proposition 1. *If h is L_h -smooth and the Hessian $\nabla^2 h^*$ is M -smooth, then the gain function can be chosen as:*

$$G(x, y, v) = 1 + 2ML_h (\|y - x\| + \|v\|).$$

Note that, even if the regularity conditions of Proposition 1 do not hold globally (such as for problems with unbounded curvature), they are at least valid on every bounded subset of $\text{int } C$, as soon as h is C^3 on $\text{int } C$. We now explicit a possible explicit choice for G_t in this setting.

Corollary 3. *Assume that h is L_h -smooth, μ_h -strongly convex and that the Hessian $\nabla^2 h^*$ is M -smooth. Then, there exists an explicit constant C such that if Algorithm 1 is run with a step size $\eta_t = 1/(8L_{f/h}G_t)$ with G_t decreasing and satisfying*

$$G_t \geq \min \left(\frac{L_{f/h}L_h}{\mu_h}, 1 + C \left(\sum_{j=1}^n \|x_t - \phi_j^t\| + \left\| \sum_{j=1}^n \nabla f_j(\phi_j^t) \right\| \right) \right), \quad (23)$$

then we have the convergence rate

$$\mathbb{E} [\psi_{t+1}] \leq \left(1 - \min \left(\frac{1}{8G_t\kappa_{f/h}}, \frac{1}{2n} \right) \right) \psi_t, \quad (24)$$

where $\lim_{t \rightarrow \infty} G_t = 1$, or, more precisely,

$$\mathbb{E} [G_t] \leq 1 + \mathcal{O} \left(1 - \min \left(\frac{1}{8\kappa_h\kappa_{f/h}}, \frac{1}{2n} \right) \right)^t. \quad (25)$$

The explicit expression for the constant C is provided in Appendix B along with the proof. Although the result involves smoothness constants of h which can be large in the relatively-smooth setting, this dependence disappears asymptotically. Hence, after some time t , which we can roughly estimate using Equation (25), we obtain that $G_t = O(1)$. Thus, we reach the same kind of convergence

rate as in the ideal quadratic case, which depends only on the *relative* condition number $\kappa_{f/h}$, but with more general functions h , and thus possibly much better conditioning. Besides, the order of magnitude required for G_t can be estimated during the optimization process using Equation (23).

4.2. Remarks on adaptivity

Assumption 3 highlights the fact that the key difficulty is purely geometric, and that in general we need to make up for the lack of translation invariance and homogeneity of Bregman divergences. Although Corollary 3 gives a criterion for G_t that can be evaluated throughout training (since the constant C is explicit), several approximations are required to obtain it, and it may be loose overall. Yet, for the theory to hold, it suffices to have η_t small enough such that:

$$\mathbb{E}_{i_t} D_h(x_t, x_{t+1}) \leq \frac{\eta_t}{4} [D_f(x_t, x^*) + \mathbb{E}_{i_t} D_{f_{i_t}}(\phi_{i_t}^t, x^*)].$$

Unfortunately, one would need to know x^* to evaluate such a condition, which is thus hard to use in practice. For the sake of clarity, we have only presented results for Bregman SAGA in this section. Yet, similar results hold for SVRG-style variance reduction, and we present them in Appendix C. An important difference is that in this case, $\phi_i^t = \phi_t$ for all i , and so the last term becomes $\mathbb{E}[D_{f_{i_t}}(\phi_t, x^*) | \mathcal{F}_t] = f(\phi_t) - f(x^*)$ since $\nabla f(x^*) = 0$, so we only need to know $f(x^*)$ (or an estimation of it) in order to compute this criterion. In this case, we don't need to know the relative smoothness constant of the problem and the step-size can be set adaptively, similarly to Barré et al. (2020). Although it may be expensive to compute $D_f(x_t, x^*)$ at each iteration, one can also approximate $f(x_t)$ on the fly, or only update η_t periodically. On a side note, a similar criterion could be used for BSGD (without the second term in this case), in particular for over-parametrized problems for which we know that $f(x^*) = 0$.

5. Experiments

In order to show the effectiveness of our method, we consider the two key settings mentioned in the introduction: problems with unbounded curvature (inverse problems with Poisson noise) and preconditioned distributed optimization. The first setting corresponds to the convex case ($\mu_{f/h} = 0$), whereas the second one corresponds to the relatively strongly convex case ($\mu_{f/h} > 0$). We observe that leveraging stochasticity (and, when needed, variance reduction) drastically improves the performance of Bregman methods in both cases. Additional details on the setting (such as the precise formulation of the objective or the relative smoothness constants) are given in Appendix D.

5.1. Poisson inverse problems

Figure 1(a) considers the minimization problem $\min_{x \in \mathbb{R}_+^d} f(x) = D_{\text{KL}}(b, Ax)$, where $A \in \mathbb{R}_+^{n \times d}$ and $D_{\text{KL}}(u, v) = \sum_{i=1}^n u_i \log(u_i/v_i) - u_i + v_i$ is the Kullback-Leibler divergence. The goal is to recover an unknown signal x_* , observed through the matrix A and corrupted by Poisson noise. This is a fundamental signal processing problem, with applications in astronomy and medicine (see Bertero et al. (2009) for a review). We use the log-barrier reference function, $h(x) = -\sum_i \log x_i$, for which relative smoothness holds with $L_{f/h} = \sum_{i=1}^n b_i/n$ (Bauschke et al., 2017).

We verify experimentally in this section that SGD is fast when the gradients at optimum are zero by first studying a problem where $b = Ax^*$. $A \in \mathbb{R}^{n \times d}$ and $x \in \mathbb{R}^d$ are random (indices sampled uniformly between 0 and 1), with $n = 10000$ and $d = 1000$. We compare the results of the deterministic and stochastic versions of Bregman Gradient descent. We also compare to the Multiplicative Updates (MU) algorithm, also known as Lucy-Richardson or Expectation-Maximization (Shepp and Vardi, 1982), which is a standard baseline for this problem. We observe that BGD is by far the slowest algorithm, but that BSGD is faster than Lucy-Richardson thanks to the stochastic speedup. We also observe that BSGD does not plateau in a noise region and converges to the true solution, which is consistent with Theorem 2. The step-size for BGD and BSGD is chosen as $1/L_{f/h}$, whereas Lucy-Richardson is parameter-free.

Figure 1(b) considers experiments on the tomographic reconstruction problem on the standard Shepp-Logan phantom (Kak and Slaney, 2001). Due to space limitations, the main text mainly describes the results, but the setting details can be found in Appendix D. The step-size given by theory was rather conservative in this case, so we increased it by a factor of 5 for all Bregman algorithms (and even 10 for BGD). Figure 1(b) shows again that stochastic algorithms drastically outperform BGD. Yet, BSGD quickly reaches a plateau because of the noise. On the other hand, BSAGA enjoys variance reduction and fast convergence to the optimum. In this case, BSAGA is on par with MU, the state-of-the-art algorithm for this problem. This is because of the log barrier that allows relative smoothness to hold, but heavily slows down Bregman algorithms when coordinates are close to 0. Yet, these results are encouraging and one may hope for even faster convergence of BSAGA for tomographic reconstruction with a tighter reference function.

5.2. Statistically Preconditioned Distributed Optimization

In this section we consider the problem of solving a distributed optimization problem in which data is distributed

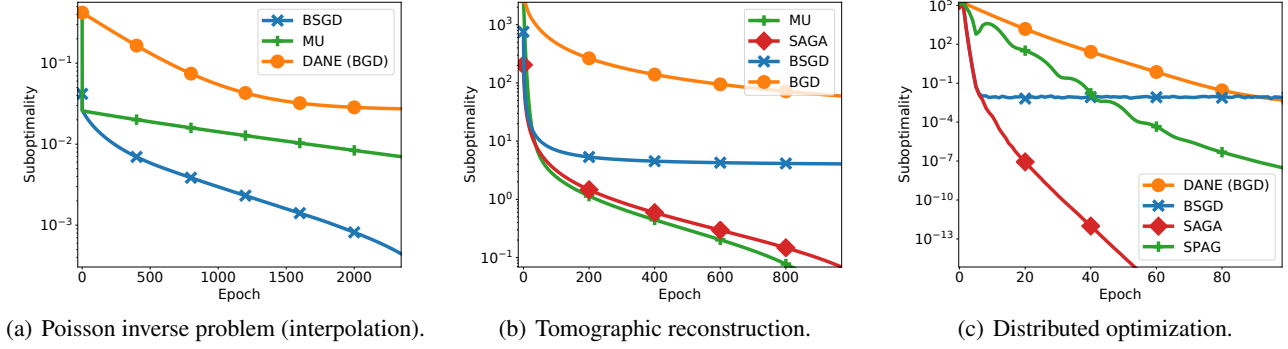


Figure 1. Bregman first-order methods on various applications.

among many workers. We closely follow the setting of Hendrikx et al. (2020a), and solve a logistic regression problem for the RCV1 dataset (Lewis et al., 2004). Function h is taken as the same logistic regression objective as for the global objective f , but on a much smaller dataset of size $n_{\text{prec}} = 1000$ and with an added regularization $c_{\text{prec}} = 10^{-5}$. In this case, BGD corresponds to a widely used variant of DANE (Shamir et al., 2014), in which only the server performs the update. The stochastic updates in BSGD are obtained by subsampling a set of workers at each iteration, so that all the nodes do not have to participate in every iteration. Regularization is taken as $\lambda = 10^{-5}$, and there are $n = 100$ nodes with $N = 1000$ samples each. A fixed learning rate is used, and the best one is selected among $[0.025, 0.05, 0.1, 0.25, 0.5, 1.]$. BGD uses $\eta = 0.5$ while SAGA and BSGD use $\eta = 0.05$. The x-axis represents the total number of communications (or number of passes over the dataset). Note that at each epoch, BGD communicates once with all workers (one round trip for each worker) whereas BSGD and BSAGA communicate n times with one worker sampled uniformly at random each time. Therefore, BSAGA requires much less gradients from the workers to reach a given precision level, yet, it is at the cost of having to solve more local iterations.

Figure 1(c) first shows that BSAGA clearly outperforms BGD. BSGD on the other hand is as fast as BSAGA at the beginning of training, until it hits a variance region at which it saturates. This is consistent with the theory, and is similar to what can be observed in the Euclidean case. An interesting feature is that although the step-size has to be selected smaller than that of gradient descent (which is also the case in the Euclidean setting since f is smoother than the least smooth f_i), choosing a constant step-size is enough to ensure convergence in this case, thus hinting at the fact that the analysis is rather conservative and that G_t does not slow down the algorithm as much as we could have feared when far from the optimum. This is consistent with the results obtained by Hendrikx et al. (2020b) on acceleration.

6. Conclusion

Throughout the paper, we have (i) given tight convergence guarantees for Bregman SGD that allow to accurately describe its behaviour in the interpolation setting, and (ii) introduced and analyzed Bregman analogs to the standard variance-reduced algorithm SAGA. These convergence results require stronger assumptions on the objective than relative smoothness and strong convexity, but we show that fast rates can be obtained nonetheless when h is nicely behaved (quadratic or Lipschitz Hessian). We also prove that these fast rates can be obtained for more general functions h after a transient regime. Besides, we show experimentally that variance reduction greatly accelerates Bregman first-order methods for several key applications, including distributed optimization and tomographic reconstruction. In particular, there does not seem to be a slow transient regime in the applications considered, despite the lack of regularity of the objectives. This need for higher order regularity assumptions but great practical performance is consistent with the results obtained for acceleration in the Bregman setting. Better understanding the transient regime (in which G_t can be high) and finding better reference functions h for the tomographic reconstruction problem are two promising extensions of our work.

Acknowledgements

RD was supported by an AMX fellowship. RD would like to acknowledge support from the Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant number FA9550-19-1-7026/19IOE033 and FA9550-18-1-0226. HH was funded in part by the French government under management of Agence Nationale de la Recherche as part of the ‘‘Investissements d’avenir’’ program, reference ANR-19-P3IA-0001(PRAIRIE 3IA Institute). HH also acknowledges support from the European Research Council (grant SEQUOIA 724063) and from the MSR-INRIA joint centre.

References

- Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *The Journal of Machine Learning Research*, 2017.
- Kimion Antonakopoulos, Elena Veronica Belmega, and Panayotis Mertikopoulos. An adaptive mirror-prox algorithm for variational inequalities with singular operators. In *Advances in Neural Information Processing Systems*, 2019.
- Kimion Antonakopoulos, Elena Veronica Belmega, and Panayotis Mertikopoulos. Online and stochastic optimization beyond lipschitz continuity: A riemannian approach. In *International Conference on Learning Representations*, 2020.
- Francis Bach and Eric Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in neural information processing systems*, 2011.
- Mathieu Barré, Adrien Taylor, and Alexandre d’Aspremont. Complexity guarantees for polyak steps with momentum. In *Conference on Learning Theory*, 2020.
- Heinz Bauschke and Patrick Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. 2011.
- Heinz H. Bauschke and Jonathan M. Borwein. Legendre Functions and the Method of Random Bregman Projections. *Journal of Convex Analysis*, 4(1):27–67, 1997.
- Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A Descent Lemma Beyond Lipschitz Gradient Continuity: First-Order Methods Revisited and Applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.
- Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.
- M Bertero, P Boccaci, G Desidera, and G Vicidomini. Image deblurring with Poisson data: from cells to galaxies. *Inverse Problems*, 25, 2009.
- Jérôme Bolte, Shoham Sabach, Marc Teboulle, and Yakov Vaisbourd. First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3), 2018.
- Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*. Springer, 2012.
- Sébastien Bubeck. Introduction to online optimization. *Lecture Notes*, 2011.
- D. Davis, D. Drusvyatskiy, and Kellie J. MacPhee. Stochastic model-based minimization under high-order growth. *arXiv preprint arXiv:1807.00255*, 2018.
- Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, 2014.
- Radu-Alexandru Dragomir, Adrien Taylor, Alexandre d’Aspremont, and Jérôme Bolte. Optimal complexity and certification of Bregman first-order methods. *Mathematical Programming*, 2021.
- Mathieu Even and Laurent Massoulié. Concentration of non-isotropic random tensors with applications to learning and empirical risk minimization. *arXiv preprint arXiv:2102.04259*, 2021.
- Tianxiang Gao, Songtao Lu, Jia Liu, and Chris Chu. Randomized Bregman coordinate descent methods for non-Lipschitz optimization. *arXiv preprint arXiv:2001.05202*, 2020.
- Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209. PMLR, 2019.
- Filip Hanzely and Peter Richtárik. Fastest rates for stochastic mirror descent methods. *arXiv preprint arXiv:1803.07374*, 2018.
- Filip Hanzely, Peter Richt, and Lin Xiao. Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. *ArXiv preprint arXiv:1808.03045v1*, 2018.
- Hadrien Hendrikx, Francis Bach, and Laurent Massoulié. Dual-free stochastic decentralized optimization with variance reduction. In *Advances in Neural Information Processing Systems*, 2020a.
- Hadrien Hendrikx, Lin Xiao, Sébastien Bubeck, Francis Bach, and Laurent Massoulié. Statistically preconditioned accelerated gradient method for distributed optimization. In *International Conference on Machine Learning*, 2020b.
- Thomas Hofmann, Aurelien Lucchi, Simon Lacoste-Julien, and Brian McWilliams. Variance reduced stochastic gradient descent with neighbors. In *Advances in Neural Information Processing Systems*, 2015.
- Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems*, 2013.

- Avinash C. Kak and Malcolm Slaney. *Principles of computerized tomographic imaging*. SIAM, 2001.
- Sham Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. *Unpublished Manuscript*, <http://ttic.uchicago.edu/shai/papers/KakadeShalevTewari09.pdf>, 2009.
- Puya Latafat, Andreas Themelis, Masoud Ahookhosh, and Panagiotis Patrinos. Bregman finito/miso for nonconvex regularized finite sum minimization without lipschitz gradient continuity. *arXiv preprint arXiv:2102.10312*, 2021.
- David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. RCV1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5 (Apr):361–397, 2004.
- 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.
- Haihao Lu, Robert M Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 2018.
- Konstantin Mishchenko. Sinkhorn algorithm as a special case of stochastic mirror descent. *arXiv preprint arXiv:1909.06918*, 2019.
- Arkadiĭ Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.
- Yuri Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer, 2003.
- Yurii Nesterov. Implementable tensor methods in unconstrained convex optimization. *Mathematical Programming*, 2019.
- David Pfau. A generalized bias-variance decomposition for bregman divergences. *Unpublished Manuscript*, http://davidpfau.com/assets/generalized_bvd_proof.pdf, 2013.
- Sashank J. Reddi, Jakub Konečný, Peter Richtárik, Barnabás Póczós, and Alex Smola. AIDE: Fast and communication efficient distributed optimization. *arXiv preprint arXiv:1608.06879*, 2016.
- Mark Schmidt, Nicolas Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162, 09 2013.
- Shai Shalev-Shwartz. SDCA without duality, regularization, and individual convexity. In *International Conference on Machine Learning*, 2016.
- Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 2013.
- Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate Newton-type method. In *International Conference on Machine Learning*, 2014.
- L. A. Shepp and Y. Vardi. Maximum likelihood reconstruction for emission tomography. *IEEE Transactions on Medical Imaging*, 1(2), 1982.
- Zhan Shi, Xinhua Zhang, and Yaoliang Yu. Bregman divergence for stochastic variance reduction: Saddle-point and adversarial prediction. In *Advances in Neural Information Processing Systems*, 2017.
- Xiao-Tong Yuan and Ping Li. On convergence of distributed approximate Newton methods: Globalization, sharper bounds and beyond. *Journal of Machine Learning Research*, 2020.
- Siqi Zhang and Niao He. On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization. *arXiv preprint arXiv:1806.04781*, 2018.
- Yihan Zhou, Victor Sanches Portella, Mark Schmidt, and Nicholas Harvey. Regret bounds without Lipschitz continuity: Online learning with relative-Lipschitz losses. *Advances in Neural Information Processing Systems*, 33, 2020.

This appendix is organized as follows. We start by providing a detailed comparison of assumptions and convergence rates for our algorithms and related work in Table 6. Then, Section A provides the missing proofs for Bregman SGD, and Section B for the variance-reduced scheme Bregman SAGA. Additionally, we also analyze in Section C another variant based on the SVRG algorithm. Finally, Section D lists additional details for the numerical applications.

Algorithm	Gradient noise	Regularity assumptions	Convergence rate of $D_h(x^*, x_t)$
Gradient Descent	deterministic	$\mu I \preceq \nabla^2 f \preceq LI$	$\mathcal{O}\left(1 - \frac{\mu}{nL}\right)^t$
Bregman Gradient Descent	deterministic	$\mu_{f/h} \nabla^2 h \preceq \nabla^2 f \preceq L_{f/h} \nabla^2 h$	$\mathcal{O}\left(1 - \frac{\mu_{f/h}}{nL_{f/h}}\right)^t$
Stochastic Gradient Descent	variance bounded at x^*	$\mu I \preceq \nabla^2 f_\xi \preceq LI$	$\mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^t + \frac{\sigma^2 L}{\mu}\right)$
Bregman Stochastic Gradient Descent (Theorem 1)	variance bounded at x^*	$\mu_{f/h} \nabla^2 h \preceq \nabla^2 f_\xi \preceq L_{f/h} \nabla^2 h$	$\mathcal{O}\left(\left(1 - \frac{\mu_{f/h}}{L_{f/h}}\right)^t + \frac{\sigma^2 L_{f/h}}{\mu_{f/h}}\right)$
SAGA (Defazio et al., 2014)	finite sum	$\mu I \preceq \nabla^2 f_i \preceq LI$	$\mathcal{O}\left(1 - \min\left(\frac{1}{4n}, \frac{\mu}{3L}\right)\right)^t$
Bregman-SAGA, Corollary 1	finite sum	$\mu_{f/h} \nabla^2 h \preceq \nabla^2 f_i \preceq L_{f/h} \nabla^2 h$, $\nabla^2 h$ constant	$\mathcal{O}\left(1 - \min\left(\frac{1}{2n}, \frac{\mu_{f/h}}{8L_{f/h}}\right)\right)^t$
Bregman-SAGA, Corollary 2	finite sum	$\mu_{f/h} \nabla^2 h \preceq \nabla^2 f_i \preceq L_f I$, $\mu_h I \preceq \nabla^2 h$	$\mathcal{O}\left(1 - \min\left(\frac{1}{2n}, \frac{\mu_{f/h} \mu_h}{8L_f}\right)\right)^t$
Bregman-SAGA, Corollary 3	finite sum	$\mu_{f/h} \nabla^2 h \preceq \nabla^2 f_i \preceq L_{f/h} \nabla^2 h$, $\mu_h I \preceq \nabla^2 h \preceq L_h I$, $\nabla^2 h^*$ is M -smooth	$\mathcal{O}\left(1 - \min\left(\frac{1}{2n}, \frac{\mu_{f/h}}{8G_t L_{f/h}}\right)\right)^t$ with $G_t \rightarrow 1$ as $t \rightarrow \infty$

Figure 2. Summary of convergence rates for standard (stochastic) first-order methods, and their Bregman counterparts in different settings.

A. Missing proofs for Bregman SGD (Section 3)

Lemma 2. Let x^+ be such that $\nabla h(x^+) = \nabla h(x) - g$, and similarly define x_1^+ and x_2^+ from g_1 and g_2 . Then, if $g = \frac{g_1 + g_2}{2}$, we obtain:

$$D_h(x, x^+) \leq \frac{1}{2} [D_h(x, x_1^+) + D_h(x, x_2^+)].$$

Proof. By Lemma 1 (duality), we have:

$$\begin{aligned}
 D_h(x, x^+) &= D_{h^*}(\nabla h(x^+), \nabla h(x)) \\
 &= D_{h^*}(\nabla h(x) - g, \nabla h(x)) \\
 &= D_{h^*}\left(\frac{1}{2}[\nabla h(x) - g_1] + \frac{1}{2}[\nabla h(x) - g_2], \nabla h(x)\right) \\
 &\leq \frac{1}{2} D_{h^*}(\nabla h(x) - g_1, \nabla h(x)) + \frac{1}{2} D_{h^*}(\nabla h(x) - g_2, \nabla h(x)),
 \end{aligned}$$

where the inequality step is obtained by the convexity of the Bregman divergence in its first argument. The final result is obtained by using duality back. \square

Lemma 4. If $\nabla h(x_{t+1}) = \nabla h(x_t) - \eta_t g_t$ with $\mathbb{E}[g_t] = \nabla f(x_t)$, $\nabla f(x^*) = 0$, then:

$$D_h(x^*, x_{t+1}) = D_h(x^*, x_t) - \eta_t D_f(x^*, x_t) - \eta_t D_f(x_t, x^*) + D_h(x_t, x_{t+1}). \quad (26)$$

Note that this descent lemma is an equality, and we can then use standard assumptions to bound the different terms.

Proof. We start by writing $V_t(x) = \eta_t g_t^\top x + D_h(x, x_t)$. Since x_{t+1} is defined as $\arg \min_x V_t(x)$ and by Assumption 1, we have $x_{t+1} \in \text{int } C$ then $\nabla V_t(x_{t+1}) = 0$ and so:

$$V_t(x^*) - V_t(x_{t+1}) = D_{V_t}(x^*, x_{t+1}) = D_h(x^*, x_{t+1}), \quad (27)$$

since $\nabla^2 V_t = \nabla^2 h$. This writes:

$$\eta_t g_t^\top (x^* - x_{t+1}) + D_h(x^*, x_t) - D_h(x_{t+1}, x_t) = D_h(x^*, x_{t+1}). \quad (28)$$

Then, we split the first term as $g_t^\top (x^* - x_{t+1}) = g_t^\top (x^* - x_t) + g_t^\top (x_t - x_{t+1})$. For the first term, we use the fact that $\mathbb{E}[g_t] = \nabla f(x_t)$ and obtain

$$\mathbb{E}[g_t^\top (x^* - x_t)] = -[\nabla f(x^*) - \nabla f(x_t)]^\top (x^* - x_t) = -D_f(x^*, x_t) - D_f(x_t, x^*), \quad (29)$$

For the second term, we write:

$$\begin{aligned} D_h(x_{t+1}, x_t) + D_h(x_t, x_{t+1}) &= \langle \nabla h(x_t) - \nabla h(x_{t+1}), x_t - x_{t+1} \rangle \\ &= \eta_t g_t^\top (x_t - x_{t+1}), \end{aligned}$$

so that

$$\eta_t g_t^\top (x_t - x_{t+1}) - D_h(x_{t+1}, x_t) = D_h(x_t, x_{t+1}). \quad (30)$$

Combining Equations (28), (29) and (30), we obtain:

$$D_h(x^*, x_{t+1}) = D_h(x^*, x_t) - \eta_t D_f(x^*, x_t) - \eta_t D_f(x_t, x^*) + D_h(x_t, x_{t+1}), \quad (31)$$

which finishes the proof. \square

Lemma 3. *If a convex function f is relatively L -smooth w.r.t to h , then for any $\eta \leq \frac{1}{L}$,*

$$D_f(x, y) \geq \frac{1}{\eta} D_{h^*}(\nabla h(x) - \eta(\nabla f(x) - \nabla f(y)), \nabla h(x))$$

Proof. Let $y \in \text{int dom } h$ and consider the function g_y defined by

$$g_y(x) = D_f(x, y) = f(x) - f(y) - \nabla f(y)^\top (x - y)$$

for $x \in C$. g_y is nonnegative, convex and relatively L -smooth with respect to h , since it has the same Hessian than f . Therefore, for $\eta \in (0, \frac{1}{L}]$ the relative smoothness inequality (4) implies that for every $u \in \text{int dom } h$ we have $D_{g_y}(u, x) \leq \frac{1}{\eta} D_h(u, x)$, that is

$$g_y(u) \leq g_y(x) + \nabla g_y(x)^\top (u - x) + \frac{1}{\eta} D_h(u, x) := Q_y(u, x). \quad (32)$$

The right-hand side $Q_y(u, x)$ is a convex function of u and is minimized for a point u^+ such that

$$\nabla h(u^+) - \nabla h(x) + \eta \nabla g_y(x) = 0, \quad (33)$$

and therefore

$$\begin{aligned} 0 &\leq g_y(u^+) \leq Q_y(u^+, x) \\ &= g_y(x) - \frac{1}{\eta} (\nabla h(u^+) - \nabla h(x))^\top (u^+ - x) + \frac{1}{\eta} D_h(u^+, x) \\ &= g_y(x) - \frac{1}{\eta} D_h(x, u^+) \\ &= D_f(x, y) - \frac{1}{\eta} D_{h^*}(\nabla h(u^+), \nabla h(x)) \\ &= D_f(x, y) - \frac{1}{\eta} D_{h^*}(\nabla h(x) - \eta \nabla g_y(x), \nabla h(x)) \end{aligned}$$

and the result follows from the fact that $\nabla g_y(x) = \nabla f(x) - \nabla f(y)$. \square

B. Missing proofs for Variance Reduced methods (Section 4)

B.1. Bregman variance decomposition

First, we use the following Bregman counterpart of a standard variance identity (Pfau, 2013), which we prove for completeness.

Lemma 5 (Bregman variance decomposition). *Let X be a random variable on \mathbb{R}^d . Then for any $u \in \mathbb{R}^d$,*

$$\mathbb{E}[D_{h^*}(X, u)] = D_{h^*}(\mathbb{E}[X], u) + \mathbb{E}[D_{h^*}(X, \mathbb{E}[X])] \quad (34)$$

As a consequence, for any random variable V on \mathbb{R}^d and point $y \in \mathbb{R}^d$ we have

$$\mathbb{E}[D_{h^*}(y + V - \mathbb{E}[V], y - \mathbb{E}[V])] \geq \mathbb{E}[D_{h^*}(y + V - \mathbb{E}[V], y)]. \quad (35)$$

Proof. Denoting $\bar{x} := \mathbb{E}[X]$, We have for $u \in \mathbb{R}^d$

$$\begin{aligned} D_{h^*}(\bar{x}, u) + \mathbb{E}[D_{h^*}(X, \bar{x})] &= h^*(\bar{x}) - h^*(u) - \nabla h^*(u)^\top (\bar{x} - u) + \mathbb{E}[h^*(X) - h^*(\bar{x}) - \nabla h^*(\bar{x})^\top (X - \bar{x})] \\ &= -h^*(u) - \nabla h^*(u)^\top (\bar{x} - u) + \mathbb{E}[h^*(X)] \\ &= \mathbb{E}[h^*(X) - h^*(u) - \nabla h^*(u)^\top (X - u)] \\ &= \mathbb{E}[D_{h^*}(X, u)] \end{aligned}$$

which proves (34). Then, (35) follows from applying it to the point $u = y - \mathbb{E}[V]$ and the random variable $X = y + V - \mathbb{E}[V]$, along with using the nonnegativity of the Bregman divergence $D_{h^*}(\mathbb{E}[X], u)$. \square

B.2. Proof of Theorem 3: generic Bregman-SAGA convergence bound

In this subsection, we give a more detailed proof of Theorem 3, and include derivations that had to be skipped in the main text because of space limitations.

More detailed proof of Theorem 3. Similarly to BSGD, we start by applying Lemma 4 (Appendix A), which yields

$$\mathbb{E}_{i_t}[D_h(x^*, x_{t+1})] = D_h(x^*, x_t) - \eta_t D_f(x^*, x_t) - \eta_t D_f(x_t, x^*) + \mathbb{E}_{i_t}[D_h(x_t, x_{t+1})]. \quad (36)$$

Lemmas 1 and 2 yield $D_h(x_t, x_{t+1}) \leq (D_1 + D_2)/2$, with

$$\begin{aligned} D_1 &= D_{h^*}(\nabla h(x_t) - 2\eta_t [\nabla f_i(x_t) - \nabla f_i(x^*)], \nabla h(x_t)), \\ D_2 &= D_{h^*}(\nabla h(x_t) - 2\eta_t (\nabla f_i(x^*) - \bar{\alpha}_i^t), \nabla h(x_t)). \end{aligned}$$

Using the gain function with the fact that $\eta_t \leq 1/L_{f/h}$ and Lemma 3, we have

$$\begin{aligned} \mathbb{E}_{i_t}[D_1] &= \mathbb{E}_{i_t}[D_{h^*}(\nabla h(x_t) - 2\eta_t (\nabla f_i(x_t) - \nabla f_i(x^*)), \nabla h(x_t))] \\ &\leq 4L_{f/h}^2 \eta_t^2 \mathbb{E}_i \left[G \left(x_t, x_t, \frac{1}{L_{f/h}} (\nabla f_i(x_t) - \nabla f_i(x^*)) \right) D_{h^*} \left[\nabla h(x_t) - \frac{1}{L_{f/h}} (\nabla f_i(x_t) - \nabla f_i(x^*)), \nabla h(x_t) \right] \right] \\ &\leq 4L_{f/h} \eta_t^2 \mathbb{E}_i \left[G \left(x_t, x_t, \frac{1}{L_{f/h}} (\nabla f_i(x_t) - \nabla f_i(x^*)) \right) D_f(x_t, x^*) \right] \\ &\leq 4L_{f/h} \eta_t^2 G_t D_f(x_t, x^*). \end{aligned} \quad (37)$$

Note that we can pull the G_t term out of the expectation over the choice of i since G_t holds for all i . For bounding D_2 ,

Lemma 5 with $V = -2\eta_t(\nabla f_i(x^*) - \nabla f_i(\phi_i^t))$ leads to

$$\begin{aligned}
 \mathbb{E}_{i_t} [D_2] &= \mathbb{E}_i \left[D_{h^*} \left(\nabla h(x_t) - 2\eta_t \left(\nabla f_i(x^*) - \nabla f_i(\phi_i^t) + \frac{1}{n} \sum_{j=1}^n \nabla f_j(\phi_j^t) \right), \nabla h(x_t) \right) \right] \\
 &\leq \mathbb{E}_i \left[D_{h^*} \left(\nabla h(x_t) - 2\eta_t \left(\nabla f_i(x^*) - \nabla f_i(\phi_i^t) + \frac{1}{n} \sum_{j=1}^n \nabla f_j(\phi_j^t) \right), \nabla h(x_t) - \frac{2\eta_t}{n} \sum_{j=1}^n \nabla f_j(\phi_j^t) \right) \right] \\
 &\leq 4\eta_t^2 L_{f/h}^2 \mathbb{E}_i \left[G \left(\nabla h(x_t) - \frac{1}{n} \sum_{j=1}^n \nabla f_j(\phi_j^t), \nabla h(\phi_j^t), \frac{1}{L_{f/h}} (\nabla f_i(\phi_j^t) - \nabla f_i(x^*)) \right) \right. \\
 &\quad \left. D_{h^*} \left[\nabla h(\phi_i^t) - \frac{1}{L_{f/h}} (\nabla f_i(\phi_i^t) - \nabla f_i(x^*)), \nabla h(\phi_i^t) \right] \right] \\
 &\leq 4L_{f/h} \eta_t^2 G_t \mathbb{E}_i [D_{f_i}(\phi_i^t, x^*)].
 \end{aligned} \tag{38}$$

Recall that $H_t = \frac{1}{n} \sum_{j=1}^n D_{f_j}(\phi_j^t, x^*)$. Plugging the expressions for D_1 and D_2 into Equation (36), we obtain:

$$\mathbb{E}_{i_t} [D_h(x^*, x_{t+1})] - D_h(x^*, x_t) \leq -\eta_t D_f(x^*, x_t) - \eta_t D_f(x_t, x^*) + 2L_{f/h} \eta_t^2 G_t [D_f(x_t, x^*) + H_t]. \tag{39}$$

Following Hofmann et al. (2015), we write:

$$\mathbb{E}_{i_t} [H_{t+1}] = \left(1 - \frac{1}{n}\right) H_t + \frac{1}{n} D_f(x_t, x^*), \tag{40}$$

Indeed, $\phi_j^{t+1} = \phi_j^t$ with probability $1 - 1/n$, and $\phi_i^{t+1} = x_t$ with probability $1/n$. Therefore, we can use the $-H_t/n$ term to control the excess term from bounding $D_h(x_t, x_{t+1})$. In the end, using that G_t is decreasing and so η_t is increasing, we obtain the following recursion:

$$\begin{aligned}
 \mathbb{E}_{i_t} [\psi_{t+1}] - \psi_t &= \frac{1}{\eta_{t+1}} D_h(x^*, x_{t+1}) + \frac{n}{2} H_{t+1} - \frac{1}{\eta_t} D_h(x^*, x_t) - \frac{n}{2} H_t \\
 &\leq \frac{1}{\eta_t} (D_h(x^*, x_{t+1}) - D_h(x^*, x_t)) + \frac{n}{2} (H_{t+1} - H_t) \\
 &\leq -D_f(x^*, x_t) - \frac{1}{2} (1 - 4\eta_t L_{f/h} G_t) H_t - \left(1 - 2\eta_t L_{f/h} G_t - \frac{1}{2}\right) D_f(x_t, x^*).
 \end{aligned} \tag{41}$$

If we choose $\eta_t \leq 1/(8L_{f/h} G_t)$ then the last term is positive and $1 - 4\eta_t L_{f/h} G_t \geq 1/2$, so that using the relative strong convexity of f leads to:

$$\begin{aligned}
 \mathbb{E}_{i_t} [\psi_{t+1}] &\leq (\eta_t^{-1} - \mu_{f/h}) D_h(x^*, x_t) + \left(1 - \frac{1}{2n}\right) \frac{n}{2} H_t \\
 &\leq \left(1 - \min\left(\eta_t \mu_{f/h}, \frac{1}{2n}\right)\right) \psi_t.
 \end{aligned}$$

The result can then be obtained by chaining this inequality. If $\mu_{f/h} = 0$ then we start back from Equation (41), use that $D_f(x^*, x_t) \geq 0$ and the same fact that $1 - 4\eta_t L_{f/h} G_t \geq 1/2$ to obtain:

$$\frac{1}{4} [D_f(x_t, x^*) + H_t] \leq \psi_t - \mathbb{E}_{i_t} [\psi_{t+1}].$$

The result is obtained by averaging over T , since the right hand side yields a telescopic sum, leading to the $1/T$ rate of Equation (18). \square

B.3. Lipschitz-Hessian setting

In this section, we add the additional assumption that h is L_h -smooth, and that the Hessian $\nabla^2 h^*$ is M -smooth in the operator norm, that is

$$\|(\nabla^2 h^*(x) - \nabla^2 h^*(y))u\| \leq M\|x - y\|\|u\|$$

for every $x, y, u \in \mathbb{R}^d$.

Proposition 1. *If h is L_h -smooth and the Hessian $\nabla^2 h^*$ is M -smooth, then the gain function can be chosen as:*

$$G(x, y, v) = 1 + 2ML_h(\|y - x\| + \|v\|).$$

Proof of Proposition 1. Writing the divergence in integral form, we have for $x, y, v \in \mathbb{R}^d$ and $\lambda \in [-1, 1]$

$$\begin{aligned} D_{h^*}(x + \lambda v, x) &= \lambda^2 \int_0^1 \int_0^t v^\top \nabla^2 h^*(x + s\lambda v) v \, ds \, dt \\ &\leq \lambda^2 \int_0^1 \int_0^t (v^\top \nabla^2 h^*(y + sv) v + M\|y + sv - x - \lambda sv\|\|v\|^2) \, ds \, dt \\ &\leq \lambda^2 \int_0^1 \int_0^t (v^\top \nabla^2 h^*(y + sv) v + M(\|y - x\| + 2s\|v\|)\|v\|^2) \, ds \, dt \\ &= \lambda^2 (D_{h^*}(y + v, y) + M(\|y - x\| + \|v\|)\|v\|^2). \end{aligned}$$

Using the fact that h is L_h -smooth, h^* is $1/L_h$ -strongly convex and hence $\|v\|^2 \leq 2L_h D_{h^*}(y + v, y)$, leading to

$$D_{h^*}(x + \lambda v, x) \leq \lambda^2 [1 + 2ML_h(\|y - x\| + \|v\|)] D_{h^*}(y + v, y).$$

□

Corollary 3. *Assume that h is L_h -smooth and the Hessian $\nabla^2 h^*$ is M -smooth. Then, there exists an explicit constant C such that if Algorithm 1 is run with a step size $\eta_t = 1/(8L_{f/h}G_t)$ with G_t decreasing in t and satisfying*

$$G_t \geq \min \left(\frac{L_{f/h}L_h}{\mu_h}, 1 + C \left(\sum_{j=1}^n \|x_t - \phi_j^t\| + \left\| \sum_{j=1}^n \nabla f_j(\phi_j^t) \right\| \right) \right), \quad (42)$$

then we have the convergence rate

$$\mathbb{E}_{i_t} [\psi_{t+1}] \leq \left(1 - \min \left(\frac{1}{8G_t\kappa_{f/h}}, \frac{1}{2n} \right) \right) \psi_t, \quad (43)$$

where $\lim_{t \rightarrow \infty} G_t = 1$, or, more precisely,

$$\mathbb{E} [G_t] \leq 1 + \mathcal{O} \left(1 - \min \left(\frac{1}{8\kappa_h\kappa_{f/h}}, \frac{1}{2n} \right) \right)^t. \quad (44)$$

Proof of Corollary 3. Using the gain function from Proposition 1, to satisfy the assumptions of Theorem 3 it is sufficient to choose G_t such that

$$\begin{aligned} G_t &\geq 1 + 2ML_h \left(\frac{1}{L_{f/h}} \|\nabla f_{i_t}(x_t) - \nabla f_{i_t}(x^*)\| + \frac{1}{L_{f/h}} \|\nabla f_{i_t}(\phi_{i_t}^t) - \nabla f_{i_t}(x^*)\| \right. \\ &\quad \left. + \|\nabla h(x_t) - \nabla h(\phi_{i_t}^t)\| - \frac{1}{4nL_{f/h}} \sum_{j=1}^n \|\nabla f_j(\phi_j^t)\| \right). \end{aligned} \quad (45)$$

As the quantities involving $\nabla f_{i_t}(x^*)$ are unknown, we provide an upper estimate. We can proceed in the following way, using the fact that, due to relative regularity, f_i is also smooth with constant $L_hL_{f/h}$, and f is strongly convex with constant

$\mu_h \mu_{f/h}$:

$$\begin{aligned}
 \|\nabla f_{i_t}(x_t) - \nabla f_{i_t}(x^*)\|^2 &\leq 2L_h L_{f/h} D_{f_{i_t}}(x_t, x^*) \\
 &\leq 2L_h L_{f/h} n D_f(x_t, x^*) \\
 &\leq \frac{L_h L_{f/h}}{\mu_h \mu_{f/h}} n \left\| \frac{1}{n} \sum_{j=1}^n \nabla f_j(x_t) \right\|^2 \\
 &\leq \frac{\kappa_f \kappa_{f/h}}{n} \left(\sum_{j=1}^n \|\nabla f_j(x_t) - \nabla f_j(\phi_j^t)\| + \left\| \sum_{j=1}^n \nabla f_j(\phi_j^t) \right\| \right)^2 \\
 &\leq \frac{\kappa_f \kappa_{f/h}}{n} \left(\sum_{j=1}^n L_h L_{f/h} \|x_t - \phi_j^t\| + \left\| \sum_{j=1}^n \nabla f_j(\phi_j^t) \right\| \right)^2.
 \end{aligned}$$

And similarly, we can estimate the second term from

$$\|\nabla f_{i_t}(\phi_{i_t}^t) - \nabla f_{i_t}(x^*)\| \leq \|\nabla f_{i_t}(x_t) - \nabla f_{i_t}(x^*)\| + L_h L_{f/h} \|\phi_{i_t}^t - x_t\|,$$

which leads to the following upper estimate of the RHS of Condition (45):

$$\begin{aligned}
 &1 + 2ML_h \left(\frac{1}{L_{f/h}} \|\nabla f_{i_t}(x_t) - \nabla f_{i_t}(x^*)\| + \frac{1}{L_{f/h}} \|\nabla f_{i_t}(\phi_{i_t}^t) - \nabla f_{i_t}(x^*)\| + \|\nabla h(x_t) - \nabla h(\phi_{i_t}^t)\| - \frac{1}{4nL_{f/h}} \sum_{j=1}^n \nabla f_j(\phi_j^t) \right) \\
 &\leq 1 + 2ML_h \left(\frac{2}{L_{f/h}} \|\nabla f_{i_t}(x_t) - \nabla f_{i_t}(x^*)\| + L_h \|\phi_{i_t}^t - x_t\| + \|\nabla h(x_t) - \nabla h(\phi_{i_t}^t)\| + \frac{1}{4nL_{f/h}} \left\| \sum_{j=1}^n \nabla f_j(\phi_j^t) \right\| \right) \\
 &\leq 1 + 2ML_h \left(2\sqrt{\frac{\kappa_h \kappa_{f/h}}{n}} L_h \sum_{j=1}^n \|x_t - \phi_j^t\| + 2L_h \|\phi_{i_t}^t - x_t\| + \left(\frac{1}{4nL_{f/h}} + \frac{2}{L_{f/h}} \sqrt{\frac{\kappa_h \kappa_{f/h}}{n}} \right) \left\| \sum_{j=1}^n \nabla f_j(\phi_j^t) \right\| \right) \\
 &\leq 1 + C \left(\sum_{j=1}^n \|x_t - \phi_j^t\| + \left\| \sum_{j=1}^n \nabla f_j(\phi_j^t) \right\| \right)
 \end{aligned}$$

where C is defined as

$$C = 2ML_h \max \left(4L_h \left(1 + \sqrt{\frac{\kappa_h \kappa_{f/h}}{n}} \right), \frac{1}{L_{f/h}} \left(\frac{1}{4n} + 2\sqrt{\frac{\kappa_h \kappa_{f/h}}{n}} \right) \right).$$

Now, with such choice of G_t , Theorem 3 applies and the convergence rate (43) holds. It remains to prove the estimate for

Algorithm 2 Bregman-SVRG($(\eta_t)_{t \geq 0}, x_0$)

- 1: $\phi_0 = x_0$, compute and store $\nabla f(\phi_0)$.
 - 2: **for** $t = 0, 1, 2, \dots$ **do**
 - 3: Pick $i_t \in \{1, \dots, n\}$ uniformly at random
 - 4: $g_t = \nabla f_{i_t}(x_t) - \nabla f_{i_t}(\phi_t) + \nabla f(\phi_t)$
 - 5: $x_{t+1} = \arg \min_x \{ \eta_t g_t^\top x + D_h(x, x_t) \}$
 - 6: $\phi_{t+1} = \begin{cases} x_t, \text{ and compute and store } \nabla f(\phi_{t+1}) \text{ with probability } p \\ \phi_t \text{ otherwise.} \end{cases}$
 - 7: **end for**
-

the convergence rate of G_t towards 1. To this end, we show that it is upper bounded by $\mathcal{O}(1 + \psi_t^{1/2})$ since

$$\begin{aligned}
 1 + C \left(\sum_{j=1}^n \|x_t - \phi_j^t\| + \left\| \sum_{j=1}^n \nabla f_j(\phi_j^t) \right\| \right) \\
 \leq 1 + C \left(\sum_{j=1}^n \|x_t - \phi_j^t\| + \left\| \sum_{j=1}^n \nabla f_j(x_t) \right\| + \sum_{j=1}^n \|\nabla f_j(\phi_j^t) - \nabla f_j(x_t)\| \right) \\
 \leq 1 + C \left(\sum_{j=1}^n (1 + L_h L_{f/h}) \|x_t - \phi_j^t\| + n \|\nabla f(x_t)\| \right) \\
 \leq 1 + C \left(\sum_{j=1}^n (1 + L_h L_{f/h}) (\|x_t - x^*\| + \|x^* - \phi_j^t\|) + n L_h L_{f/h} \|x_t - x^*\| \right) \\
 \leq 1 + C \left(n(1 + 2L_h L_{f/h}) \|x_t - x^*\| + \sum_{j=1}^n (1 + L_h L_{f/h}) \|x^* - \phi_j^t\| \right) \\
 \leq 1 + C \left(n(1 + 2L_h L_{f/h}) \sqrt{\frac{2}{\mu_h} D_h(x^*, x_t)} + \sum_{j=1}^n (1 + L_h L_{f/h}) \sqrt{\frac{2}{\mu_h \mu_{f/h}} D_{f_j}(\phi_j^t, x^*)} \right) \\
 = 1 + \mathcal{O} \left(\sqrt{D_h(x^*, x_t)} + \sum_{j=1}^n \sqrt{D_{f_j}(\phi_j^t, x^*)} \right) \\
 = 1 + \mathcal{O} \left(\sqrt{\psi_t} \right)
 \end{aligned} \tag{46}$$

Since we imposed a safeguard such that $G_t \geq \frac{L_{f/h} L_h}{\mu_h}$, the convergence rate of ψ_t is bounded by

$$\mathbb{E}[\psi_t] = \mathcal{O} \left(1 - \min \left(\frac{1}{8\kappa_h \kappa_{f/h}}, \frac{1}{2n} \right) \right)^t$$

as stated by Corollary 2. Indeed, the assumptions are verified as h^* is $1/\mu_h$ -smooth and f^* is $1/L_f$ -strongly convex with $L_f = L_h L_{f/h}$. This worst-case estimate for ψ_t , along with the majorization (46), gives the resulting rate for G_t . \square

C. Bregman SVRG

We consider in this section the convergence guarantees of Bregman SVRG (BSVRG), which is presented in Algorithm 2. We consider the same variant as Hofmann et al. (2015), in which the full gradient used for variance reduction is recomputed

at each step with a small probability p , instead of after a fixed number of steps. We study this variant of BSVRG since it is very closely related to BSAGA. The main difference is that instead of updating ϕ_{i_t} when i_t is picked, the algorithm chooses only one common ϕ_t to perform variance reduction, and this common ϕ_t is updated with probability p at the end of each iteration. Thus, the convergence Theorem for Algorithm 2 closely follows Theorem 3.

Theorem 4. Assume that Algorithm 2 is run with a step size sequence $\{\eta_t\}_{t \geq 0}$ satisfying $\eta_t = 1/(8L_{f/h}G_t)$ for every $t \geq 0$, with G_t decreasing in t and such that for all $j \in \{1, \dots, n\}$:

$$\begin{aligned} G_t &\geq G \left(\nabla h(x_t), \nabla h(x_t), \frac{1}{L_{f/h}} (\nabla f_j(x_t) - \nabla f_j(x^*)) \right), \\ G_t &\geq G \left(\nabla h(x_t) - 2\eta_t \nabla f(\phi_t), \nabla h(\phi_t), \frac{1}{L_{f/h}} (\nabla f_j(\phi_t) - \nabla f_j(x^*)) \right). \end{aligned}$$

Then, under Assumptions 1 and 3, the potential $\psi_t = D_h(x^*, x_t) + \frac{\eta_t}{2p} D_f(\phi_t, x^*)$ satisfies

$$\mathbb{E}_{i_t} [\psi_{t+1}] \leq \left(1 - \min \left(\eta_t \mu_{f/h}, \frac{p}{2} \right) \right) \psi_t, \quad (47)$$

In the convex case ($\mu_{f/h} = 0$), we obtain that

$$\mathbb{E} \left[\frac{1}{4T} \sum_{t=1}^T \eta_t [D_f(x_t, x^*) + D_f(\phi_t, x^*)] \right] \leq \frac{\psi_0}{T}. \quad (48)$$

Proof. As explained before Theorem 4, the only thing that changes between BSAGA and BSVRG is that a global ϕ_t is used instead of separate ϕ_i^t , and that it is update with probability p at the end of each iteration (instead of updating $\phi_{i_t}^t$ at time t for SAGA). Thus, all the derivations performed for BSAGA hold for BSVRG if we replace ϕ_i^t with ϕ_t for all i . The only equation that needs to be adapted is Equation (40), since it relies on the way the ϕ_i^t are updated. Yet, in the case of BSVRG, it writes:

$$\mathbb{E} [H_{t+1}] = (1 - p) H_t + p D_f(x_t, x^*), \quad (49)$$

which is the same as for BSAGA but with p instead of $1/n$. Therefore, the conclusions are unchanged if we replace n by $1/p$ whenever it appears in the bounds. Similar convergence guarantees hold when ϕ_t is updated every fixed number of steps T , but the proof is substantially more involved since Equation (49) does not hold in such a simple form. \square

D. Additional details for the experiments

Due to space limitations, some details of the experimental setting are missing from the main text, and we thus present them in this section. Note that all the experiments presented in this paper run in less than an hour on a standard laptop (and usually much less). Our code is also available in supplementary material.

D.1. Poisson inverse problems

We consider the minimization problem

$$\min_{x \in \mathbb{R}_+^d} f(x) = \frac{1}{n} D_{\text{KL}}(b, Ax) \quad (50)$$

where $D_{\text{KL}}(u, v) = \sum_{i=1}^n u_i \log(u_i/v_i) - u_i + v_i$ is the Kullback-Leibler divergence, and $A \in \mathbb{R}^{n \times d}$ is a typically sparse matrix that models the measurement process. Problem (50) models the maximum likelihood estimation problem when assuming the statistical model

$$b \sim \text{Poisson}(Ax^*)$$

where x^* is the true unknown signal. Inverse problems with Poisson noise arise in various signal processing applications such as astronomy or computerized tomography, see Bertero et al. (2009) and references therein.

As a motivating application of relative smoothness, Bauschke et al. (2017) prove that the Poisson objective f is relatively smooth with respect to the log-barrier reference function

$$h(x) = - \sum_{i=1}^d \log x_i$$

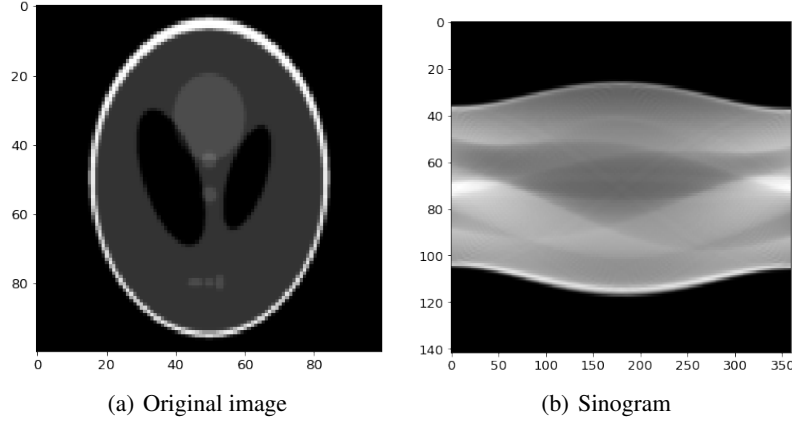


Figure 3. Illustration of the Radon transform on the Shepp-Logan phantom. On the sinogram, each column corresponds to the line integral of the image under different projection angles.

with constant $\sum_{j=1}^n b_j/n$. This constant can be quite conservative when A is a sparse matrix, and so we prove a better estimate by leveraging this structure. For $j \in \{1 \dots n\}$, we denote S_j the support of the j -th column of A , that is

$$S_j := \{i \in \{1 \dots n\} : A_{ij} \neq 0\}.$$

Proposition 2. *The Poisson objective function defined in (50) is relatively L -smooth w.r.t the log-barrier for*

$$L \geq \frac{1}{n} \max_{j \in \{1 \dots d\}} \sum_{i \in S_j} b_i. \quad (51)$$

Proof. Let us denote A_1, \dots, A_n the row vectors of A . We refine the analysis from Bauschke et al. (2017, Lemma 7) and start by writing for $x \in \mathbb{R}_{++}^d, u \in \mathbb{R}^d$

$$u^\top \nabla^2 f(x) u = \frac{1}{n} \sum_{i=1}^n b_i \frac{(A_i^\top u)^2}{(A_i^\top x)^2}.$$

Applying the Jensen inequality to the function $t \mapsto t^2$ and weights $w_{ij} = A_{ij}x_j/(A_i^\top x)$ yields

$$\begin{aligned} d^\top \nabla^2 f(x) d &= \frac{1}{n} \sum_{i=1}^n b_i \left(\sum_{j=1}^d w_{ij} \frac{u_j}{x_j} \right)^2 \\ &\leq \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^d b_i w_{ij} \cdot \frac{u_j^2}{x_j^2} \\ &\leq \frac{1}{n} \sum_{j=1}^d \sum_{i \in S_j} b_i \frac{u_j^2}{x_j^2} \\ &\leq L \sum_{j=1}^d \frac{u_j^2}{x_j^2} \\ &= L u^\top \nabla^2 h(x) u \end{aligned}$$

where we used the fact that $w_{ij} \in [0, 1]$ if $i \in S_j$, and $w_{ij} = 0$ otherwise. □

The relative Lipschitz constant provided by Proposition 2 can be considerably smaller than $\sum_{j=1}^n b_j/n$ when A is sparse, which is the case in practical applications.

For our numerical experiments, we compare full-batch Bregman gradient descent (BGD), Bregman stochastic gradient descent (BSGD), and the Bregman SAGA scheme described in Algorithm 1. We also implement the Multiplicative Update (MU), also known as Lucy-Richardson or Expectation-Maximization (Shepp and Vardi, 1982), which is the standard baseline for Poisson inverse problems.

Synthetic problem in the interpolation setting. In Figure 1(a), we simulate a synthetic problem the gradients at optimum are zero, by choosing $b = Ax^*$ for some random $A \in \mathbb{R}^{n \times d}$ and $x^* \in \mathbb{R}^d$ (indices sampled uniformly between 0 and 1), with $n = 10000$ and $d = 1000$.

Tomographic reconstruction problem. Computerized tomography (Kak and Slaney, 2001) is the task of reconstructing an object from cross-sectional projections, with fundamental applications to medical imaging. We study a classical synthetic toy problem for this task: the Shepp-Logan phantom (Figure 3(a)). In this setting, the observation matrix A corresponds to the discrete *Radon transform*, which is the cross-sectional projection of the original image x along different projection angles $\theta_1, \dots, \theta_n$ (Figure 3(b)). That is, the objective writes

$$f(x) = \frac{1}{n} D_{\text{KL}}(b, Ax) = \frac{1}{n} \sum_{i=1}^n D_{\text{KL}}(b_{\theta_i}, A_{\theta_i} x) \quad (52)$$

where $b_{\theta_i}, A_{\theta_i}$ correspond to the observation and projection matrix along the angle θ_i . For stochastic algorithms, the formulation (52) naturally yields a finite-sum structure: we thus take $f_i(x) = D_{\text{KL}}(b_{\theta_i}, A_{\theta_i} x)$ for $i = 1 \dots n$.

We corrupt the sinogram with Poisson inverse noise, and apply our algorithms. We use $n = 360$ projection angles, and the image dimension is $d = 100^2$. As the matrix A has a sparse structure, we use the relative smoothness constant provided by Proposition 2 for a better estimate. The step-size given by theory was rather conservative in this case, so we increased it by a factor of 5 for all Bregman algorithms (and even 10 for BGD).

D.2. Statistically Preconditioned Distributed Optimization

We detail in this section the setting that was used to obtain Figure 1(c). In particular, we use the following logistic regression objective with quadratic regularization, meaning that the function at node i is:

$$f_i : x \mapsto \frac{1}{N} \sum_{j=1}^N \log \left(1 + \exp(-y_{i,j} x^\top a_j^{(i)}) \right) + \frac{\lambda}{2} \|x\|^2,$$

where $y_{i,j} \in \{-1, 1\}$ is the label associated with $a_j^{(i)}$, the j -th sample of node i . We use a regularization parameter of $\lambda = 10^{-5}$, and the size of the local datasets is equal to $N = 1000$. The local dataset is constructed by shuffling the RCV1 dataset, downloaded from LibSVM, and then assigning a fixed portion to each worker. Then, one node (without loss of generality, node 0) uses its local dataset to construct the preconditioning dataset, so that:

$$h : x \mapsto f_0(x) + \frac{c_{\text{prec}}}{2} \|x\|^2, \quad (53)$$

where $c_{\text{prec}} = 10^{-5}$. Tuning c_{prec} in order to obtain the fastest algorithms is hard in general, as detailed in Hendrikx et al. (2020b) (in which it is denoted as μ). One strategy is to choose c_{prec} of order $1/n_{\text{prec}}$ (in our case $n_{\text{prec}} = N = 1000$), and then decrease it as long as BGD is stable. Our chosen value (10^{-5}) is smaller than that of Hendrikx et al. (2020b) for this problem (10^{-4}), in which they used a rougher $c_{\text{prec}} = c/n_{\text{prec}}$ criterion with varying n_{prec} , and a larger step-size $\eta = 1$ for BGD (which is the same as DANE). Besides, we see that SPAG is slightly unstable in our example, and increasing c_{prec} would help with that. In this case, theory gives that $L_{f/h} \approx 1$. Yet, when $c_{\text{prec}} \approx \lambda$, this step-size usually has to be chosen a bit smaller. Therefore, we choose in our case $\eta = 0.5$ for BGD and SPAG, and $\eta = 0.05$ for BSGD and BGD. Note that there is always a constant factor between the maximum step-size for SAGA and that of BGD, and the difference could further be explained by the difference between the batch condition number (relative smoothness of f) versus the stochastic one (max relative smoothness of the f_i).

We compute the minimum error as the smallest error over all iterations for all algorithms. Then, we subtract it to the running error of an algorithm to get the suboptimality at each step. Following Hendrikx et al. (2020b), local problems are solved using a sparse implementation of SDCA (Shalev-Shwartz, 2016). We warm-start the local problems (initializing on the

solution of the previous one), and perform 10 passes over the preconditioning dataset at each step, or until the norm of the gradient of the inner problem is small enough (10^{-6}). The number of inner passes could be reduced further, but then the algorithms started to converge slightly more slowly. This results in an overall computational overhead for the server, since BSAGA and BSGD require to solve many more inner problems, which are not so cheap to compute. Yet, this overhead only affects the server, and the iteration complexity is much lower, meaning that BSAGA is indeed very efficient to reduce the communication complexity of solving distributed empirical risk minimization problems.