Fast Stochastic Bregman Gradient Methods Sharp Analysis and Variance Reduction

Radu-Alexandru Dragomir\textsuperscript{1,2}, joint work with Mathieu Even\textsuperscript{2} and Hadrien Hendrikx\textsuperscript{2}
May 2021

\textsuperscript{1}Université Toulouse Capitole, \textsuperscript{2} INRIA Paris
Consider the problem

$$\min_{x \in C} f(x) := \mathbb{E}_\xi [f_\xi(x)],$$  \hspace{1cm} (P)$$

where $C \subset \mathbb{R}^d$ is convex and $f_\xi : \mathbb{R}^d \to \mathbb{R}$ are differentiable functions.
Consider the problem

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

where $C \subset \mathbb{R}^d$ is convex and $f_\xi : \mathbb{R}^d \rightarrow \mathbb{R}$ are differentiable functions.

Standard method: (projected) Stochastic Gradient Descent

$$
x_{t+1} = \Pi_C [x_t - \eta_t g_t],
$$

where

$$
\mathbb{E} [g_t] = \nabla f(x_t)
$$

is an unbiased gradient estimate. An equivalent form is

$$
x_{t+1} = \arg \min_{x \in C} \left\{ f(x_t) + g_t^\top (x - x_t) + \frac{1}{2\eta_t} \|x - x_t\|^2 \right\} \quad (SGD)
$$
Stochastic gradient descent

\[\begin{align*}
  x_{t+1} &= \arg\min_{x \in C} \left\{ f(x_t) + g_t^\top (x - x_t) + \frac{1}{2\eta_t} \|x - x_t\|^2 \right\} \\
  \text{(SGD)}
\end{align*}\]

When is this method efficient?

- **noise**: the variance of the gradient estimate \( \mathbb{E} \left[ \|g_t - \nabla f(x_t)\|^2 \right] \) is small,
- **smoothness**: the quadratic model is a good approximation of \( f \).
Stochastic gradient descent

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

(SGD)

When is this method efficient?

- **noise**: the variance of the gradient estimate \( \mathbb{E} [\|g_t - \nabla f(x_t)\|^2] \) is small,
- **smoothness**: the quadratic model is a good approximation of \( f \).

If \( f \) has a \( L \)-Lipschitz continuous gradient, then for every \( \eta \in (0, 1/L] \),

\[
f(x) \leq f(x_t) + \nabla f(x_t)^\top (x - x_t) + \frac{1}{2\eta} \|x - x_t\|^2.
\]

The quadratic model is an upper approximation of \( f \).
Bregman stochastic gradient descent

We can try to find a better model of $f$ by regularizing with a more general Bregman divergence:

$$x_{t+1} = \arg \min_{x \in C} \left\{ f(x_t) + \nabla^\top g_t(x - x_t) + \frac{1}{\eta_t} D_h(x, x_t) \right\} \quad \text{(B-SGD)}$$

where

$$D_h(x, y) = h(x) - h(y) - \nabla h(y)^\top (x - y) \geq 0,$$

is the Bregman divergence induced by some differentiable strictly convex reference function $h$. 

Note: also known as stochastic Mirror Descent.
Bregman stochastic gradient descent

We can try to find a better model of $f$ by regularizing with a more general Bregman divergence:

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

(B-SGD)

where

$$D_h(x, y) = h(x) - h(y) - \nabla h(y)^\top (x - y) \geq 0,$$

is the **Bregman divergence** induced by some differentiable strictly convex reference function $h$.

When is this a good approximation of $f$? When $f$ is **smooth relative** to $h$:

$$f(x) \leq f(x_t) + \nabla f(x_t)^\top (x - x_t) + \frac{1}{\eta} D_h(x, x_t).$$

**Note:** also known as stochastic **Mirror Descent**.
1. Relatively-smooth optimization

2. Bregman stochastic gradient descent

3. Variance reduction for finite sum problems
Relatively-smooth optimization
Bregman divergences

Let \( h : \mathbb{R}^d \to \mathbb{R} \) be a convex reference function, and \( D_h \) its Bregman divergence

\[
D_h(x, y) = h(x) - h(y) - \nabla h(y)^\top (x - y) \geq 0.
\]

Examples:

- **Quadratic** \( h \):
  - \( h(x) = \frac{1}{2} \|x\|^2 \): then \( D_h(x, y) = \frac{1}{2} \|x - y\|^2 \), we recover the Euclidean setting
  - \( h(x) = \frac{1}{2} x^\top Q x \) with \( Q \in S^+_d \): linear preconditioning
Bregman divergences

Let $h : \mathbb{R}^d \rightarrow \mathbb{R}$ be a convex reference function, and $D_h$ its Bregman divergence

$$D_h(x, y) = h(x) - h(y) - \nabla h(y)^\top (x - y) \geq 0.$$ 

Examples:

- **Quadratic $h$:**
  - $h(x) = \frac{1}{2} \|x\|^2$: then $D_h(x, y) = \frac{1}{2} \|x - y\|^2$, we recover the Euclidean setting
  - $h(x) = \frac{1}{2} x^\top Q x$ with $Q \in S^+_{++}$: linear preconditioning

- **Entropy $h(x)$:**

$$x_{t+1} = x_t \cdot \exp[-\eta_t g_t]$$
Bregman divergences

Let $h : \mathbb{R}^d \to \mathbb{R}$ be a convex reference function, and $D_h$ its Bregman divergence

$$D_h(x, y) = h(x) - h(y) - \nabla h(y)^\top (x - y) \geq 0.$$ 

Examples:

- **Quadratic $h$:**
  - $h(x) = \frac{1}{2} \|x\|^2$: then $D_h(x, y) = \frac{1}{2} \|x - y\|^2$, we recover the Euclidean setting
  - $h(x) = \frac{1}{2} x^\top Q x$ with $Q \in S^+_{d^+}$: linear preconditioning

- **Entropy $h(x) = \sum_{i=1}^d x^i \log(x^i) - x^i$, exponential weights algorithm**
  $$x_{t+1} = x_t \cdot \exp[-\eta_t g_t]$$

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

- **Quartic $h(x) = \frac{1}{4} \|x\|^4 + \frac{1}{2} \|x\|^2$**
Relative smoothness

\[ f(x) + \nabla f(x)^\top (u - x) + LD_h(u, x) \]

How to choose the reference function \( h \)?

A natural idea is to require the inner objective of (deterministic) BGD to be a global majorant of the objective function.

Relative smoothness (Bauschke, Bolte, Teboulle 2017)

\( f \) is **L-smooth relative** to the reference function \( h \) if

\[
f(u) \leq f(x) + \nabla f(x)^\top (u - x) + LD_h(u, x) \quad \forall u, x \in C.
\]
Relative smoothness

How to choose the reference function $h$?
A natural idea is to require the inner objective of (deterministic) BGD to be a global majorant of the objective function.

Relative smoothness (Bauschke, Bolte, Teboulle 2017)

$f$ is $L$-smooth relative to the reference function $h$ if

$$f(u) \leq f(x) + \nabla f(x)^\top (u - x) + L D_h(u, x) \quad \forall u, x \in C.$$  

Equivalent to $L h - f$ convex, or, for twice differentiable functions, that

$$\nabla^2 f(x) \preceq L \nabla^2 h(x)$$
Relative smoothness

How to choose the reference function $h$? A natural idea is to require the inner objective of (deterministic) BGD to be a global majorant of the objective function.

Relative smoothness (Bauschke, Bolte, Teboulle 2017)

$f$ is $L$-smooth relative to the reference function $h$ if

$$f(u) \leq f(x) + \nabla f(x)^\top (u - x) + LD_h(u, x) \quad \forall u, x \in C.$$ 

Equivalent to $Lh - f$ convex, or, for twice differentiable functions, that

$$\nabla^2 f(x) \preceq L \nabla^2 h(x).$$

Similarly, relative strong convexity is defined as (Lu, Freund, Nesterov 2018):

$$\mu \nabla^2 h(x) \preceq \nabla^2 f(x)$$

Reduces to the usual notions of smoothness and strong convexity for $h(x) = \frac{1}{2} \|x\|^2$.

We denote $\kappa = \frac{L}{\mu}$ the relative condition number.
Linear inverse problems with Poisson noise (Bauschke et al., 2017): let $b \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times d}_+$, 

$$
\min_{x \in \mathbb{R}^d_+} D_{KL}(b, Ax) = \sum_{j=1}^n b_j \log \left( \frac{b_j}{A_j x} \right) - A_j x + b_j
$$
Example 1: problems with unbounded curvature

Linear inverse problems with Poisson noise (Bauschke et al., 2017): let $b \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times d}_+$,

$$\min_{x \in \mathbb{R}^d_+} D_{KL}(b, Ax) = \sum_{j=1}^{n} b_j \log \left( \frac{b_j}{A_j x} \right) - A_j x + b_j$$

Standard smoothness does not hold as the Hessian is singular when $A_j x \to 0$, but relative smoothness holds with $L = \sum_i b_i$ and the log barrier

$$h(x) = \sum_{i=1}^{d} - \log(x^i).$$
Example 2: Bregman preconditioning

Statistical preconditioning for distributed optimization (Hendrikx et al., 2020):

\[
\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x)
\]
Example 2: Bregman preconditioning

Statistical preconditioning for distributed optimization (Hendrikx et al., 2020):

\[
\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x)
\]

Even if \( f \) is smooth, better performance can be achieved by choosing

\[
h(x) = f_1(x) + \frac{\lambda}{2} \|x\|^2
\]

Typically, \( f_1 \) is the loss function on a part of a dataset of size \( n_{\text{prec}} \).
Statistical preconditioning for distributed optimization (Hendrikx et al., 2020):

\[
\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x)
\]

Even if \( f \) is smooth, better performance can be achieved by choosing

\[
h(x) = f_1(x) + \frac{\lambda}{2} \|x\|^2
\]

Typically, \( f_1 \) is the loss function on a part of a dataset of size \( n_{\text{prec}} \). Relative smoothness and strong convexity hold with high probability, and allows to improve conditioning as

\[
\kappa_{\text{rel}} = 1 + \mathcal{O} \left( \frac{\kappa_{\text{eucl}}}{n_{\text{prec}}} \right).
\]

**Tradeoff:** solving the Bregman subproblem becomes harder as \( n_{\text{prec}} \) grows.
Dual Bregman divergence

Introduce the convex conjugate of \( h \) as

\[
h^*(y) = \sup_{x \in \mathbb{R}^d} x^\top y - h(x).
\]

Then (under some regularity properties) we have that

\[
D_h(x, y) = D_{h^*}(\nabla h(y), \nabla h(x)).
\]

Typically, the quantity

\[
D_{h^*}(\nabla h(x) + v, \nabla h(x))
\]

represents the “squared length relative to \( h \)” of a vector \( v \in \mathbb{R}^d \) at \( x \in C \), and is the analogous of \( \|v\|^2 \) in the Euclidean setting.
Bregman Stochastic Gradient Descent
Variance assumption

Recall the problem

\[
\min_{x \in \mathcal{C}} f(x) := \mathbb{E}_\xi[f_\xi(x)],
\]

(P)

Let \( \eta > 0 \) be the step size.

Assumption on stochastic gradients

The stochastic gradients \( \{g_t\}_{t \geq 0} \) satisfy the following conditions:

- **Sampling**: \( g_t = \nabla f_\xi_t(x_t) \), with \( \mathbb{E}_{\xi_t}[f_\xi] = f \),
Variance assumption

Recall the problem

\[
\min_{x \in \mathcal{C}} f(x) := \mathbb{E}_\xi [f_\xi(x)],
\]  \hspace{1cm} \text{(P)}

Let \( \eta > 0 \) be the step size.

**Assumption on stochastic gradients**

The stochastic gradients \( \{g_t\}_{t \geq 0} \) satisfy the following conditions:

- **Sampling:** \( g_t = \nabla f_\xi_t(x_t) \), with \( \mathbb{E}_{\xi_t} [f_{\xi_t}] = f \),

- **Variance:** there exists a constant \( \sigma^2 > 0 \) such that

\[
\frac{1}{2\eta^2} \mathbb{E}_{\xi_t} \left[ D_{h^*} \left( \nabla h(x_t) - 2\eta \nabla f_{\xi_t}(x^*) , \nabla h(x_t) \right) \right] \leq \sigma^2 \hspace{1cm} \text{(1)}
\]
Variance assumption

Recall the problem
\[
\min_{x \in C} f(x) := \mathbb{E}_\xi [f_\xi(x)] ,
\tag{P}
\]

Let \( \eta > 0 \) be the step size.

Assumption on stochastic gradients

The stochastic gradients \( \{g_t\}_{t \geq 0} \) satisfy the following conditions:

- **Sampling:** \( g_t = \nabla f_{\xi_t}(x_t) \), with \( \mathbb{E}_{\xi_t}[f_{\xi_t}] = f \),

- **Variance:** there exists a constant \( \sigma^2 > 0 \) such that
  \[
  \frac{1}{2\eta^2} \mathbb{E}_{\xi_t} \left[ D_{h^*}(\nabla h(x_t) - 2\eta \nabla f_{\xi_t}(x^*), \nabla h(x_t)) \right] \leq \sigma^2
  \tag{1}
  \]

If \( h \) is \( \mu_{\text{eucl}} \)-strongly convex, then (1) holds for instance if

\[
\mathbb{E}_{\xi_t} \left[ \|\nabla f_{\xi_t}(x^*)\|^2 \right] \leq \mu_{\text{eucl}} \cdot \sigma^2
\]
Convergence analysis of B-SGD

\[ x_{t+1} = \arg \min_{x \in C} \left\{ f(x_t) + g_t^\top (x - x_t) + \frac{1}{\eta} D_h(x, x_t) \right\} \]  

(B-SGD)

Convergence rate, relatively strongly convex case

In addition to the previous assumption, assume that

- \( f_\xi \) is \( L \)-smooth relative to \( h \) for every \( \xi \),
- \( f \) is \( \mu \)-strongly convex relative to \( h \),
- \( \eta \leq 1/(2L) \),
- Generalizes the Euclidean result for SGD
- Interpolation setting: if \( \sigma^2 = 0 \), i.e., \( \nabla f_\xi(x^*) = 0 \) for all \( \xi \), linear convergence rate of Bregman gradient descent (Lu et al, 2018) is recovered.
Convergence analysis of B-SGD

\[ x_{t+1} = \arg \min_{x \in C} \left\{ f(x_t) + g_t^\top (x - x_t) + \frac{1}{\eta} D_h(x, x_t) \right\} \] (B-SGD)

Convergence rate, relatively strongly convex case

In addition to the previous assumption, assume that

- \( f_\xi \) is \( L \)-smooth relative to \( h \) for every \( \xi \),
- \( f \) is \( \mu \)-strongly convex relative to \( h \),
- \( \eta \leq 1/(2L) \),

then the iterates of B-SGD satisfy

\[
\mathbb{E} [D_h(x^*, x_t)] \leq (1 - \eta L)^t D_h(x^*, x_0) + \eta \frac{\sigma^2}{\mu}. \tag{2}
\]
Convergence analysis of B-SGD

\[ x_{t+1} = \arg \min_{x \in C} \left\{ f(x_t) + g_t^\top (x - x_t) + \frac{1}{\eta} D_h(x, x_t) \right\} \]  \hspace{1cm} \text{(B-SGD)}

Convergence rate, relatively strongly convex case

In addition to the previous assumption, assume that

- \( f_\xi \) is \( L \)-smooth relative to \( h \) for every \( \xi \),
- \( f \) is \( \mu \)-strongly convex relative to \( h \),
- \( \eta \leq \frac{1}{2L} \),

then the iterates of B-SGD satisfy

\[ \mathbb{E} [D_h(x^*, x_t)] \leq (1 - \eta L)^t D_h(x^*, x_0) + \eta \frac{\sigma^2}{\mu}. \] \hspace{1cm} (2)

- Generalizes the Euclidean result for SGD
- **Interpolation setting:** if \( \sigma^2 = 0 \), i.e., \( \nabla f_\xi(x^*) = 0 \) for all \( \xi \), linear convergence rate of Bregman gradient descent (Lu et al, 2018) is recovered.
Convergence analysis of B-SGD

\[ x_{t+1} = \arg \min_{x \in C} \left\{ f(x_t) + g_t^\top (x - x_t) + \frac{1}{\eta} D_h(x, x_t) \right\} \]  \hspace{1cm} \text{(B-SGD)}

Convergence rate, convex case

With the same assumptions than before, we have, if \( \mu = 0 \),

\[
\mathbb{E} \left[ \frac{1}{T} \sum_{t=0}^{T} D_f(x^*, x_t) \right] \leq \frac{D_h(x^*, x_0)}{\eta T} + \eta \sigma^2
\]  \hspace{1cm} (3)
Variance reduction
We now assume that the problem is a finite sum:

\[
\min_{x \in C} f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x),
\]

where \( f_i \) are \( L \)-smooth and \( \mu \)-strongly convex relative to \( h \).

In the Euclidean setting, variance reduction can be used to obtain fast linear convergence rates: SAG (Schmidt et al., 2013), SVRG (Johnson and Zhang, 2013), SAGA (Defazio et al., 2014).

**Objective:** combine information used by gradients of previous iterates to reduce the variance of \( g_t \).
Algorithm 1 Bregman-SAGA\((\eta_t)_{t \geq 0}, x_0\)

1: \( \phi_i = x_0 \) for \( i = 1, \ldots, n \)

2: for \( t = 0, 1, 2, \ldots \) do

3: Pick \( i_t \in \{1, \ldots, n\} \) 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: \( \phi_{i_t}^{t+1} = x_t \), 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: gain function

There exists a gain function $G$ such that for any $x, y, v \in \mathbb{R}^d$ and $\lambda \in [-1, 1],

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

### Assumption: gain function

There exists a gain function $G$ such that for any $x, y, v \in \mathbb{R}^d$ and $\lambda \in [-1, 1]$,

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

- Models lack of homogeneity of Bregman divergence for nonquadratic functions
- $G$ will determine the theoretical step size needed for convergence of Bregman-SAGA
- Same issue as for accelerated Bregman algorithms: additional assumptions are unavoidable (Dragomir et al., 2021)
Bregman-SAGA convergence analysis

**Quadratic case:** if $h$ is quadratic, then $G$ can be chosen equal to 1 and the rate in expected function values is

$$\mathbb{E}[\psi_t] \leq \left(1 - \min \left(\frac{1}{8\kappa}, \frac{1}{2n}\right)\right)^t \psi_0.$$
Bregman-SAGA convergence analysis

**Quadratic case:** if $h$ is quadratic, then $G$ can be chosen equal to 1 and the rate in expected function values is

$$
\mathbb{E} [\psi_t] \leq \left(1 - \min \left(\frac{1}{8\kappa}, \frac{1}{2n}\right)\right)^t \psi_0.
$$

**“Mirror descent” setting:** if $h$ is $\mu_{\text{eucl}}$-strongly convex and $f$ is $L_{\text{eucl}}$-smooth w.r.t the Euclidean norm, then

$$
\mathbb{E} [\psi_t] \leq \left(1 - \min \left(\frac{\mu_{\text{eucl}} \cdot \mu}{8L_{\text{eucl}}}, \frac{1}{2n}\right)\right)^t \psi_0.
$$
**Bregman-SAGA convergence analysis**

**Quadratic case:** if $h$ is quadratic, then $G$ can be chosen equal to 1 and the rate in expected function values is

$$
\mathbb{E} [\psi_t] \leq \left( 1 - \min \left( \frac{1}{8\kappa}, \frac{1}{2n} \right) \right)^t \psi_0.
$$

**“Mirror descent” setting:** if $h$ is $\mu_{\text{eucl}}$-strongly convex and $f$ is $L_{\text{eucl}}$-smooth w.r.t the Euclidean norm, then

$$
\mathbb{E} [\psi_t] \leq \left( 1 - \min \left( \frac{\mu_{\text{eucl}} \cdot \mu}{8L_{\text{eucl}}}, \frac{1}{2n} \right) \right)^t \psi_0.
$$

**Issue:** $\frac{L_{\text{eucl}}}{\mu_{\text{eucl}}}$ can be very large. How to get a rate that depends only on the relative condition number $\kappa$ for nonquadratic $h$?
Bregman-SAGA convergence analysis

**Lipschitz-Hessian setting:** if \( h \) is locally smooth and \( \nabla^2 h^* \) is \( M \)-Lipschitz,

\[
\mathbb{E} [\psi_{t+1}] \leq \left( 1 - \min \left( \frac{1}{8G_t \kappa}, \frac{1}{2n} \right) \right) \psi_t, \tag{4}
\]

with \( G_t \to 1 \) as \( t \to +\infty \), for well-chosen step sizes \( \{\eta_t\}_{t \geq 0} \).

The “good” convergence rate is reached asymptotically: same result as for accelerated Bregman gradient descent (Hendrikx et al., 2020).
Numerical experiments
Poisson inverse problems

\[
\min_{x \in \mathbb{R}^d_+} \sum_{j=1}^n \left( b_j \log \left( \frac{b_j}{A_j x} \right) - A_j x + b_j \right) \quad \text{with} \quad h(x) = -\sum_{i=1}^d \log x^i
\]

MU: standard baseline algorithm (a.k.a Lucy-Richardson/Expectation-Maximization)

(a) Toy problem, interpolation setting, \( n = 10000 \), \( d = 1000 \)

(b) Tomographic reconstruction problem, \( n = 360 \), \( d = 10000 \)
Logistic regression, RCV1 dataset. \( n = 100 \) nodes with \( N = 10000 \) samples each.

\( h \) is the loss function on a smaller part of the dataset, with \( n_{\text{prec}} = 1000 \) samples.

**Figure 1:** Logistic regression, \( n = 100, \; d = 47236 \)
Conclusion

- Bregman SGD: tight convergence rate, adapted notion of variance,

- Bregman SAGA: full theory in the quadratic setting, asymptotical rate for nonquadratic $h$.

Open question: understanding the transient regime, with additional regularity assumptions (self-concordance ?)
References


