Fast Stochastic Bregman Gradient Methods

Sharp Analysis and Variance Reduction under Relative Smoothness

Hadrien Hendrikx[2], joint work with Radu-Alexandru Dragomir[1,2] and Mathieu Even[2]

[1] Université Toulouse Capitole, [2] INRIA Paris

Problem setup

Consider the problem

$$\min_{x \in \mathbb{R}^d} f(x), \tag{P}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a convex function.

Problem setup

Consider the problem

$$\min_{x \in \mathbb{R}^d} f(x), \tag{P}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a convex function.

Standard method: Stochastic Gradient Descent

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

where

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

is an unbiased gradient estimate. An equivalent form is

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

Bregman stochastic gradient descent

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

$$x_{t+1} = \arg\min_{x \in \mathbb{R}^d} \left\{ g_t^\top x + \frac{1}{\eta_t} D_h(x, x_t) \right\} \tag{B-SGD}$$

where

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

is the **Bregman divergence** induced by function h.

Bregman stochastic gradient descent

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

$$x_{t+1} = \arg\min_{x \in \mathbb{R}^d} \left\{ g_t^\top x + \frac{1}{\eta_t} D_h(x, x_t) \right\} \tag{B-SGD}$$

where

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

is the **Bregman divergence** induced by function h.

When is this a good idea ? When f is **smooth relative** to h [Bauschke et al., 2017]:

$$f(x) \leq f(x_t) + \nabla f(x_t)^\top (x - x_t) + L D_h(x, x_t).$$

Note: also known as stochastic *Mirror Descent*.

Convergence analysis of B-SGD

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

Convergence rate, relatively strongly convex case

- $g_t = \nabla f_\xi(x_t)$ and f_ξ is L-smooth relative to h for every ξ,
- f is μ-strongly convex relative to h,
- there exists a constant $\sigma^2 > 0$ (**variance**) such that for some z_t,

$$\mathbb{E}_{\xi_t} \left[\| \nabla f_{\xi_t}(x^\star) \|^2_{\nabla^2 h(z_t)^{-1}} \right] \leq \sigma^2. \qquad (1)$$

Convergence analysis of B-SGD

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

Convergence rate, relatively strongly convex case

- $g_t = \nabla f_\xi(x_t)$ and f_ξ is L-smooth relative to h for every ξ,
- f is μ-strongly convex relative to h,
- there exists a constant $\sigma^2 > 0$ (**variance**) such that for some z_t,

$$\mathbb{E}_{\xi_t} \left[\| \nabla f_{\xi_t}(x^\star) \|^2_{\nabla^2 h(z_t)^{-1}} \right] \leq \sigma^2. \tag{1}$$

Then if $\eta \leq 1/(2L)$, the iterates of B-SGD satisfy

$$\mathbb{E}\left[D_h(x^\star, x_t) \right] \leq (1 - \eta L)^t D_h(x^\star, x_0) + \eta \frac{\sigma^2}{\mu}. \tag{2}$$

Convergence analysis of B-SGD

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

Convergence rate, relatively strongly convex case

- $g_t = \nabla f_\xi(x_t)$ and f_ξ is L-smooth relative to h for every ξ,
- f is μ-strongly convex relative to h,
- there exists a constant $\sigma^2 > 0$ (**variance**) such that for some z_t,

$$\mathbb{E}_{\xi_t} \left[\| \nabla f_{\xi_t}(x^\star) \|^2_{\nabla^2 h(z_t)^{-1}} \right] \leq \sigma^2. \tag{1}$$

Then if $\eta \leq 1/(2L)$, the iterates of B-SGD satisfy

$$\mathbb{E}\left[D_h(x^\star, x_t) \right] \leq (1 - \eta L)^t D_h(x^\star, x_0) + \eta \frac{\sigma^2}{\mu}. \tag{2}$$

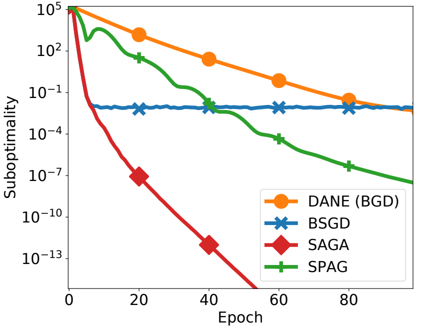
- Generalizes the Euclidean result for SGD
- **Interpolation setting:** if $\sigma^2 = 0$, i.e., $\nabla f_\xi(x^*) = 0$ for all ξ, linear convergence rate of Bregman gradient descent (Lu et al, 2018) is recovered.

Bregman Variance Reduction

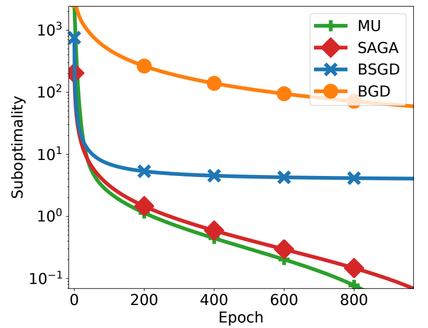
Similarly to B-SGD, a Bregman-SAGA algorithm can be obtained by replacing g_t by a SAGA-style variance-reduced gradient in the finite sum case.

(Informal) For well-chosen step-sizes, Bregman-SAGA converges linearly with rate $n + \kappa G_t$, where $G_t \to 1$ as $t \to +\infty$ and $\kappa = L/\mu$.

The "good" convergence rate is reached asymptotically: same result as for accelerated Bregman gradient descent (Hendrikx et al., 2020).



(a) Distributed logistic regression problem



(b) Tomographic reconstruction problem

Stochasticity can be leveraged to speed up Bregman methods.

References

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.