Informs, Denver
March 23, 2018

Stochastic and Accelerated Algorithms for Minimizing Relatively Smooth Functions

Filip Hanzely[1,2], Lin Xiao[3] and Peter Richtárik[1,2]
[1]KAUST, [2]University of Edinburgh, [3]Microsoft Research

Outline

- Relative Smoothness
  - Problem setting
  - Baseline Algorithm – Relative Gradient Descent
- Relative Stochastic Gradient Descent
  - Algorithm and Convergence rate
- Relative Randomized Coordinate Descent
  - ESO Assumption
  - Algorithm and Convergence rate
- Accelerated Relative Gradient Descent
  - Triangle scaling equality
  - General Algorithm
  - Examples

Relative Smoothness

Problem Setting

- Optimization problem

closed, convex subset of $\mathbb{R}^n$

$$\text{minimize } f(x) \text{ subject to } x \in Q$$

convex and differentiable function

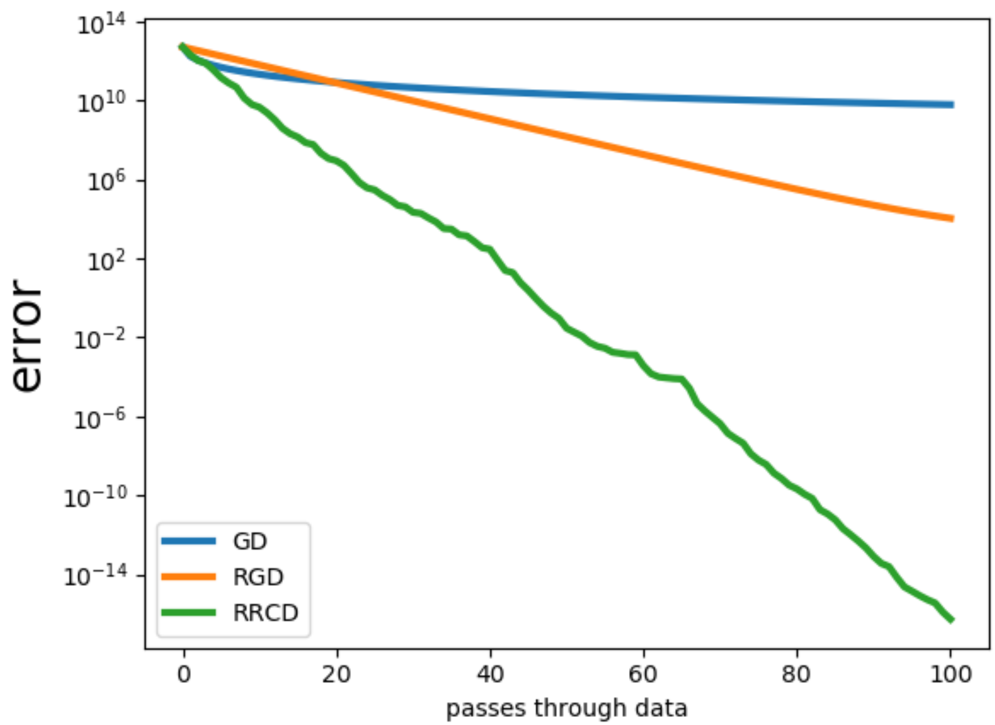
- Gradient oracle

$$|\nabla f(x) - \nabla f(y)| \leq L|x - y| \qquad \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$$

- Smoothness and Strong convexity (standard assumptions) does not necessarily hold
- Approach – Relative smoothness [Birnbaum et. al. 2011], [Bauschke et. al. 2016], [Lu et. al. 2016]

Experiment

$$f(x) = x^\top \left(1_{100}1_{100}^\top + I_{100}\right) x + \frac{1}{100}\sum_{i=1}^{100} x_i^4$$



Relative Smoothness

- L - smoothness of f relative to h

reference function

$$\boxed{Lh(x) - f(x) \text{ is convex}}$$

- Equivalent formulations

Bregman Divergence
$D_h(x,y) = h(x) - h(y) - \langle \nabla h(y), x - y \rangle$

$$D_f(x,y) \leq L D_h(x,y)$$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L \langle \nabla h(x) - \nabla h(y), x - y \rangle$$

$$\nabla^2 f(x) \preccurlyeq L \nabla^2 h(x)$$

- Standard smoothness - special case for $h(x) = \frac{1}{2}\|x\|^2$

Relative Strong Convexity

- μ- strong convexity of f relative to h

$$f(x) - \mu h(x) \text{ is convex}$$

reference function

- Equivalent formulations

Bregman Divergence

$$D_h(x,y) = h(x) - h(y) - \langle \nabla h(y), x - y \rangle$$

$$D_f(x,y) \geq \mu D_h(x,y)$$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \langle \nabla h(x) - \nabla h(y), x - y \rangle$$

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

- Standard strong convexity for $h(x) = \frac{1}{2}\|x\|^2$

Relative Gradient Descent [Lu et. al. 2016]

- f is L-smooth and μ- strongly convex relative to h
- Idea – minimize local upper bound from Relative smoothness:

$$\boxed{D_f(x,y) \le L D_h(x,y)}$$

minimize $f(x)$ subject to $x \in Q$

$$\boxed{f(x) \le f(y) + \langle \nabla f(y), x - y\rangle + L D_h(x,y)}$$

$$x^{t+1} \leftarrow \mathrm{argmin}_{x\in Q} \langle \nabla f(x^t), x\rangle + L D_h(x, x^t)$$

- $h(x) = \frac{1}{2}\|x\|^2$- Gradient Descent

Mirror descent

Relative Gradient Descent [Lu et. al. 2016]

- f is L-smooth and μ- strongly convex relative to h
- Algorithm:

$$x^{t+1} \leftarrow \text{argmin}_{x \in Q} \langle \nabla f(x^t), x \rangle + L D_h(x, x^t)$$

- Complexity result:

Optimal solution

$$f(x^k) - f(x^*) \leq \frac{\mu D_h(x^*, x^0)}{\left(1 + \frac{\mu}{L-\mu}\right)^k - 1}$$

Linear Convergence rate

Relative Stochastic Gradient Descent

Relative Stochastic Gradient Descent

- f is L- smooth and μ- strongly convex relative to h
- Update made via unbiased gradient estimator

stepsize controlling parameter

$$x^{t+1} \leftarrow \text{argmin}_{x \in Q} \langle g^t, x \rangle + L^t D_h(x, x^t)$$

unbiased estimator of $\nabla f(x^t)$

- $h(x) = \frac{1}{2}\|x\|^2$ - Stochastic Gradient Descent
- Useful when access to g^t is much cheaper than $\nabla f(x^t)$

Convergence rate

$$x^{t+1} \leftarrow \text{argmin}_{x\in Q} \langle g^t, x\rangle + L^t D_h(x, x^t)$$

- $L^t = L + \alpha t$ for $\alpha < \mu$:

c is k dimensional positive vector

$$\boxed{\frac{1}{\sum_{t=1}^k c_t} \sum_{t=1}^k c_t E\left[f(x^t) - f(x^*)\right] \le \mathcal{O}(1/k)}$$

- $\mu = 0,\ L^t = \mathcal{O}\left(\sqrt{k}\right)$:

$$\boxed{\frac{1}{k} \sum_{t=1}^k E\left[f(x^t) - f(x^*)\right] \le \mathcal{O}\left(1/\sqrt{k}\right)}$$

- Recovered rates from smooth setting (except constant)

Relative Randomized Coordinate Descent

Relative Randomized Coordinate Descent

- Update a random subset of coordinates
- Separable h:

$$\boxed{h(x) = \sum_{i=1}^{n} h_i(x_i)}$$

- h-Expected Separable Overapproximation (ESO)
- Separable relative strong convexity

$$D_h(y,z)_u = \sum_{i=1}^{n} u_i \left(h_i(y_i) - h_i(z_i) - \langle \nabla h_i(z_i), y_i - z_i \rangle \right)$$

$$D_f(x,y) \geq D_h(x,y)_w$$

- $h(x) = \frac{1}{2}\|x\|^2$ - Randomized Coordinate Descent

Smoothness vs. ESO

$$D_f(x,y) \le L D_h(x,y)$$

- h - smoothness

i–th column of $n \times n$ identity matrix

$$f\left(x + \sum_{i \in \hat{S}} q_i e_i\right) \le f(x) + \left\langle \nabla f(x), \sum_{i \in \hat{S}} q_i e_i \right\rangle + \cancel{L} D_h\left(x + \sum_{i \in \hat{S}} q_i e_i, x\right)_v$$

random subset of $\{1, 2, \ldots, n\}$

$D_h(y,z)_u = \sum_{i=1}^n u_i \left(h_i(y_i) - h_i(z_i) - \langle \nabla h_i(z_i), y_i - z_i \rangle\right)$

- h -ESO

$$E\left[f\left(x + \sum_{i \in \hat{S}} q_i e_i\right)\right] \le f(x) + \langle \nabla f(x), q \rangle_p + L D_h(x+q, x)_{p \circ v}$$

$P(i \in \hat{S}) = p_i;\ p = (p_1, \ldots, p_n)^\top$

vector of parameters v for h–ESO,
potentially smaller than L

Convergence rate

subset od $\mathbb{R}$ guaranteeing $x_j^{t+1} \in Q$

$$x_j^{t+1} \leftarrow \operatorname{argmin}_{x \in Q_j^t} \nabla f(x^t)_j \cdot x + v_j D_{h_j}(x, x_j^t)$$

j belong to random subset of $\{1, 2, \ldots, n\}$

positive vector with entries summing to 1

$$\Delta = \min_i \frac{w_i}{v_i}$$

assume that $p_0 = P(1 \in \hat{S}) = \, = P(2 \in \hat{S}) = \cdots = P(n \in \hat{S})$

$$\sum_{t=1}^{k} c_t \big(E\left[f(x^t) \right] - f(x^*) \big) \leq \frac{(1 - p_0 \Delta) D_h(x^*, x^0)_v + (1 - p_0)(f(x^0) - f(x^*))}{1 - \Delta^{-1} + \Delta^{-1} \left(\frac{1}{1 - p_0 \Delta} \right)^{k-1}}$$

linear rate

Faster than Relative Gradient Descent

Captures the result in smooth case (except constant)

Accelerated Relative Gradient Descent

Accelerated Relative Gradient Descent

- Assume only relative smoothness and convexity
- Complexity of RGD:

$$\boxed{\mathcal{O}(k^{-1})}$$

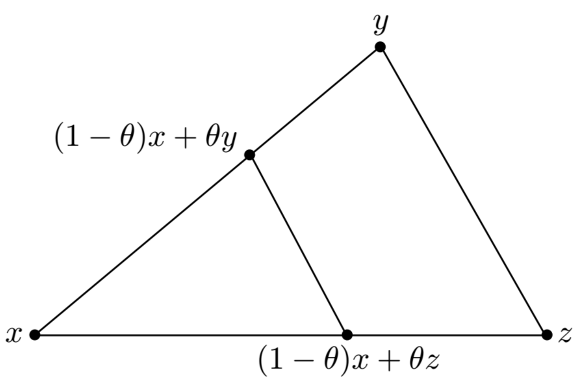
- Can we get $\mathcal{O}(k^{-2})$?

optimal rate for general
smooth convex functions

Triangle Scaling Equality

- Decomposing ratio of Bregman divergences

$$\frac{D_h((1-\theta)x+\theta y, (1-\theta)x+\theta z)}{D_h(y,z)} = g(x,y,z,\theta)\nu(\theta)$$



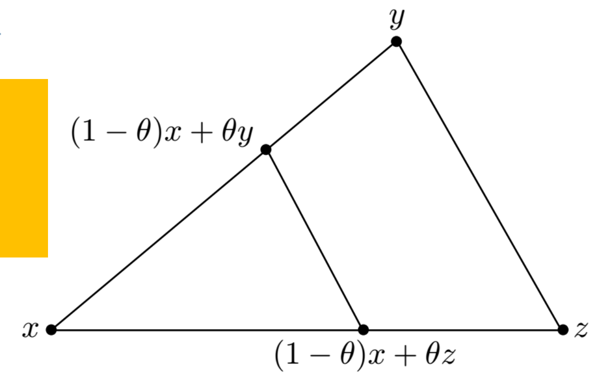
iterates of the algorithm

specific stepsize parameter

$$\mathcal{O}(\max_{t<k} g(x^t, z^{t+1}, z^t, \theta^{t+1})\nu(\theta^k))$$

Triangle Scaling Equality

$$\frac{d((1-\theta)x+\theta y,(1-\theta)x+\theta z)}{d(y,z)} = g(x,y,z,\theta)\nu(\theta)$$



Upper bounded by constant

Fast decay for $\theta \to 0$

$$\mathcal{O}(\max_{t<k} g(x^t, z^{t+1}, z^t, \theta^{t+1})\nu(\theta^k))$$

$$h(x) = \frac{1}{2}\|x\|^2 \;\to\; \nu(\theta^k) = \frac{4}{(k+1)^2},\; g(x,y,z,\theta)=1$$

$$\mathcal{O}(k^{-2})$$

Algorithm

$y^{k+1} \leftarrow (1-\theta^{k+1})x^k + \theta^{k+1} z^k$

Find some $G^{k+1} \in \mathbb{R}$ s. t. $G^{k+1} \geq g(x^k, z^{k+1}, z^k, \theta^{k+1})$ and $G^{k+1} \geq G^k$

$z^{k+1} \leftarrow \mathrm{argmin}_{z \in Q} \left\{ \langle \nabla f(y^{k+1}), z \rangle + \frac{\nu(\theta^{k+1})}{\theta^{k+1}} G^{k+1} L D_h(z, z^k) \right\}$

$x^{k+1} \leftarrow (1-\theta^{k+1})x^k + \theta^{k+1} z^{k+1}$

Choose G^{k+1} as $G^{k+1} \leftarrow \max\left(\sup_z g(x^k, z, z^k, \theta^{k+1}), G^k\right)$

Determine G^{k+1} with linesearch

Bounded $\max_{t \leq k}\left(\sup_z g(x^t, z, z^t, \theta^{t+1})\right)$
→ constant number of linesearch iterations

Smoothness with Relative Entropy

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

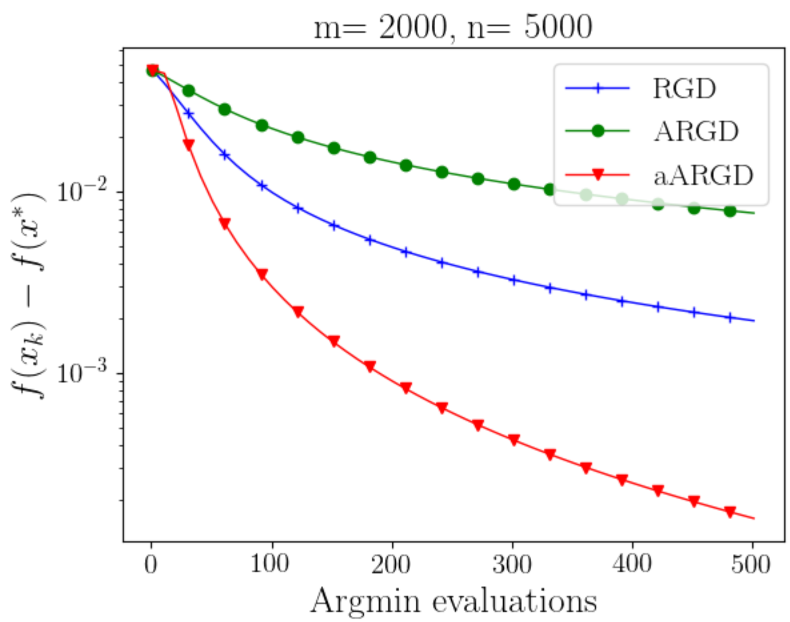
$$D_h(x, y) = \sum_{i=1}^{n} x_i \log(x_i / y_i)$$

$$Q = \left\{ x \mid \sum_{i=1}^{n} x_i = 1,\ \forall i : \ x_i \geq 0 \right\}$$

$$\boxed{\nu(\theta^k) = \mathcal{O}(k^{-2})}$$

Bounded
$\max_{t \leq k} \left(\sup_z g(x^t, z, z^t, \theta^{t+1}) \right)$

Fisher market equilibrium problem



Smoothness with Burg's Entropy

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

$$D_h(x, y) = \sum_{i=1}^{n} \left(\log\left(\frac{y_i}{x_i}\right) + \frac{x_i - y_i}{y_i} \right)$$

$$Q = \left\{ x \mid \sum_{i=1}^{n} x_i = 1,\ \forall i:\ x_i \geq 0 \right\}$$

$$\boxed{\nu(\theta^k) = \mathcal{O}(k^{-2})}$$

Bounded
$\max_{t \leq k} \left(\sup_z g(x^t, z, z^t, \theta^{t+1}) \right)$

D-optimal design