Universal Intermediate Gradient Method with One Proxmap

D. Kamzolov

Moscow Institute of Physics and Technology
Department of Control and Applied Mathematics

19 June 2017

Outline

Problem statement

- Convex optimization problem

$$f(x) \longrightarrow \min_{x \in Q}$$

- $Q \subset \mathbb{R}^n$ – simple closed convex set
- $f(x)$ – convex function with Lipschitz-continuous

Bregman distance

Definition 1.

$d(x)$ – prox-function, *if* $d(x)$ – *is a continuousky differentiable strongly convex function with convexity parameter equal to one* $\|\cdot\|$

$$d(y) \geq d(x) + \langle \nabla d(x), y - x \rangle + \frac{1}{2}\|x - y\|^2, \quad \forall x, y \in Q$$

and $d(x) \geq 0,\ d(x_0) = 0,\ d(x) \geq \frac{1}{2}\|x - x_0\|^2, \quad x \in Q.$

Definition 2.

$V(x, y)$ – Bregman distance *for* $d(x)$, *if*

$$V(x, y) = d(y) - d(x) - \langle \nabla d(x), y - x \rangle.$$

and $V(x, x) = 0,\ V(x, y) \geq \frac{1}{2}\|x - y\|^2, \quad x, y \in Q.$

Basic methods

Gradient descent

$$x_{k+1} = \operatorname*{argmin}_{x \in Q} \left\{ \frac{L}{2} \|x - x_k\|^2 + \langle \nabla f(x_k), x - x_k \rangle \right\}$$

Dual averaging method

$$x_{k+1} = \operatorname*{argmin}_{y \in Q} \left\{ Ld(x) + \sum_{i=0}^{k} \alpha_i \langle \nabla f(x_i), x - x_i \rangle \right\}$$

$$\boxed{f(x^N) - f_* \leq \Theta\left(\frac{LR^2}{N}\right)}$$

Fast Gradient Method

1: $x_0 = \text{argmin}_{x \in Q} d(x)$
2: **for** $k := 0, \ldots$ **do**
3: Compute

$$y_k := \operatorname*{argmin}_{y \in Q} \left\{ \frac{L}{2} \|y - x_k\|^2 + \langle \nabla f(x_k), y - x_k \rangle \right\}$$

4: Compute

$$z_k := \operatorname*{argmin}_{y \in Q} \left\{ Ld(x) + \sum_{i=0}^{k} \alpha_i \langle \nabla f(x_i), x - x_i \rangle \right\}$$

5: Compute

$$x_{k+1} := \tau_k z_k + (1 - \tau_k) y_k$$

$$\boxed{f(x^N) - f_* \leq \Theta\left(\frac{LR^2}{N^2}\right)}$$

Inexact oracle

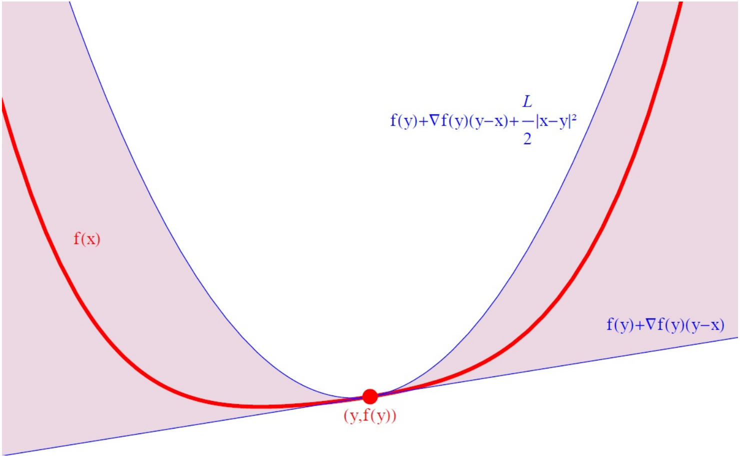
Definition 3.

Let function f be convex on convex set Q. We say that it is equipped with *a first-order* (δ, L)*-oracle* if for any $y \in Q$ we can compute a pair $(f_{\delta,L}(y), g_{\delta,L}(y))$ such that

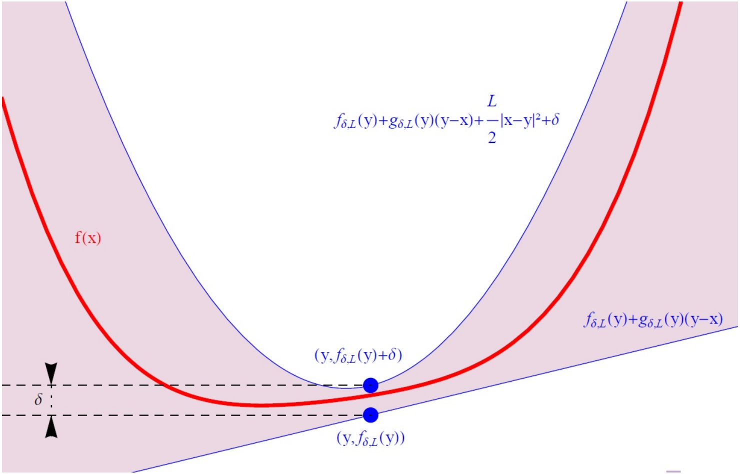
$$0 \le f(x) - f(y) - \langle \nabla f(y), x - y \rangle \le \frac{L}{2}\|x - y\|_2^2 + \delta, \quad \forall x, y \in Q$$

Constant δ is called the *accuracy* of the oracle.

Exact oracle



Inexact oracle



Intermediate Gradient Method (Devolder, Glineur, Nesterov)

1: Compute $x_0 = \text{argmin}_{x\in Q} d(x)$,
$y_0 := \underset{x\in Q}{\text{argmin}} \{Ld(x_0) + \langle g_{\delta,L}(x), x - x_0\rangle\}$

2: **for** $k := 0, \dots$ **do**

3: Compute

$$z_k := \text{argmin}_{x\in Q} \left\{ Ld(x) + \sum_{i=0}^{k} \alpha_i \langle g_{\delta,L}(x_i), x - x_i\rangle \right\}$$

4: Compute $x_{k+1} := \tau_k z_k + (1-\tau_k) y_k$

5: Obtain $(f_{\delta,L}(x_{k+1}), g_{\delta,L}(x_{k+1}))$

6: Compute

$$\tilde{x}_{k+1} = \underset{x\in Q}{\text{argmin}} \{LV(x, z_k) + \alpha_{k+1}\langle g_{\delta,L}(x_{k+1}), x - z_k\rangle\}$$

7: Compute $w_{k+1} = \tau \tilde{x}_{k+1} + (1-\tau_k) y_k$

8: Compute

$$y_{k+1} = \frac{A_{k+1} - B_{k+1}}{A_{k+1}} y_k + \frac{B_{k+1}}{A_{k+1}} w_{k+1}$$

Intermediate Gradient Method

Parameters

Let $\{\alpha_i\}_{i\geq 0}$ and $\{B_i\}_{i\geq 0}$ be two sequences of coefficients satisfying for all $i \geq 0$, $\quad \alpha_i^2 \leq B_i \leq \sum\limits_{j=0}^{i} \alpha_j, \quad 0 \leq \alpha_i \leq B_i$. We define also $A_i = \sum\limits_{j=0}^{i} \alpha_j$ and $\tau_i = \frac{\alpha_{i+1}}{B_{i+1}}$.

Statement 1.

For $p \in [1,2]$, $\alpha_i = \left(\frac{i+p}{p}\right)^{p-1}$, $B_i = a_i^2$ IGM converge with rate:

$$f(y^N) - f_* \leq \Theta\left(\frac{LR^2}{N^p}\right) + \Theta\left(N^{p-1}\delta\right)$$

$$f(y^N) - f_* \leq \varepsilon, \quad N = O\left(\left(\frac{LR^2}{\varepsilon}\right)^{\frac{1}{p}}\right), \quad \delta \leq O\left(\frac{\varepsilon}{N^{p-1}}\right)$$

Composite problem

- Convex composite optimization problem

$$F(x) = f(x) + h(x) \rightarrow \min_{x \in Q}$$

- Q – is a simple closed convex set
- $h(x)$ – is a simple convex function
- $f(x)$ – is a convex function with Lipschitz-continuous constant L with norm $\|\cdot\|$

Hólder continuous gradient

Defenition 4.

Hólder continuous gradient constant L_ν for function $f(x)$ defined as:

$$L_\nu = \sup_{x \neq y \in Q} \frac{\|\nabla f(x) - \nabla f(y)\|_*}{\|x - y\|^\nu}$$

Statement 2.

If $L_\nu < \infty$, then

$$\|\nabla f(y) - \nabla f(x)\|_* \leq L_\nu \|y - x\|^\nu, \quad \forall x, y \in Q$$

And

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_\nu}{1 + \nu} \|x - y\|^{1+\nu}, \quad \forall x, y \in Q$$

FGM with Hólder continuous gradient

Statement 3.

Let $\|\nabla f(y) - \nabla f(x)\|_* \leq L_\nu \|y - x\|^\nu, \quad \forall x, y \in Q$ for $\nu \in [0, 1]$. Then

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2}\|y - x\|^2 + \delta, L = L_\nu \left[\frac{L_\nu}{2\delta} \frac{1-\nu}{1+\nu} \right]^{\frac{1-\nu}{1+\nu}}$$

Statement 4.

If $f(x)$ has Hólder continuous gradient with constant L_ν, then Composite FGM converges with optimal rate:

$$F(y^N) - F(x_*) \leq \varepsilon, \quad N = O\left(\left(\frac{L_\nu R^{1+\nu}}{\varepsilon} \right)^{\frac{2}{1+3\nu}} \right), \delta \leq O\left(\frac{\varepsilon}{N} \right)$$

1: Choose $L_0 > 0, \alpha_0 = A_0 = B_0 = 1/L_0$
2: $x_0 = \underset{x\in Q}{\operatorname{argmin}}\, d(x)$
3: $y_0 := \underset{x\in Q}{\operatorname{argmin}} \{d(x) + \alpha_0\langle g_{\delta,L}(x), x - x_0\rangle + \alpha_0 h(x)\}$
4: **for** $k := 0, \ldots$ **do**
5: Find the smallest $i_k \geq 0$, such that coefficient $\alpha_{k+1,i_k} > 0$, computed from equation $\alpha^2_{k+1,i_k} = \frac{B_{k+1,i_k}}{2^{i_k} L_k}$, ensures the following relation:

$$f(w_{k+1,i_k}) \leq f(x_{k+1,i_k}) + \langle g_{\delta,L_{k+1}}(x_{k+1,i_k}), w_{k+1,i_k} - x_{k+1,i_k}\rangle + \frac{2^{i_k} L_k}{2}||w_{k+1,i_k} - x_{k+1,i_k}||^2 + \frac{\alpha_{k+1,i_k}}{B_{k+1,i_k}}\frac{\varepsilon}{2} + \delta,$$

Where $(f_{\delta,2^{i_k}L_k}(x_{k+1}), g_{\delta,2^{i_k}L_k}(x_{k+1}))$ is obtained

$$A_{k+1,i_k} = A_k + \alpha_{k+1,i_k}, \tau_{k,i_k} = \frac{\alpha_{k+1,i_k}}{B_{k+1,i_k}},$$

$$x_{k+1,i_k} := \tau_{k,i_k} z_k + (1-\tau_{k,i_k}) y_k$$

$$z_{k+1,i_k} := \operatorname*{argmin}_{x\in Q} \Big\{ A_{k+1,i_k} h(x) + d(x) + \sum_{i=0}^{k} \alpha_i \langle g_{\delta,L_i}(x_i), x - x_i\rangle + \alpha_{k+1}\langle g_{\delta,2^{i_k}L_k}(x_{k+1,i_k}), x - x_{k+1,i_k}\rangle \Big\},$$

$$w_{k+1,i_k} = \tau_{k,i_k} z_{k+1,i_k} + (1-\tau_{k,i_k}) y_k$$

$$z_{k+1} = z_{k+1,i_k},\ w_{k+1} = w_{k+1,i_k},\ x_{k+1} = x_{k+1,i_k},\ \tau_k = \tau_{k,i_k},$$
$$\alpha_{k+1} = \alpha_{k+1,i_k},\ A_{k+1} = A_{k+1,i_k},\ B_{k+1} = B_{k+1,i_k}$$

7: $y_{k+1} = \frac{A_{k+1} - B_{k+1}}{A_{k+1}} y_k + \frac{B_{k+1}}{A_{k+1}} w_{k+1}$

8: $L_{k+1} = 2^{i_k - 1} L_k$ **end for**; **return** y_k

Theorem 1

Let f satisfies condition (Hölder) with certain $M_\nu < +\infty$ with (δ, M)-oracle. Then all iterations of UIGM are well defined. Moreover, for all $k \geq 0$ we have

$$A_k F(y_k) - E_k \leq \Psi_k^*$$

where $E_k = \left(\sum\limits_{i=0}^{k} B_i\right) \delta + A_k \frac{\varepsilon}{2}$

$$\Psi_k^* = \min_{x \in Q} \Big\{ \Psi_k(x) = d(x) + \\ + \sum_{i=0}^{k} \alpha_i \left[f_{\delta, L_i}(x_i) + \langle g_{\delta, L_i}(x_i), x - x_i \rangle + h(x) \right] \Big\}$$

UIGM Convergence Rate

Theorem 2

Assume that the UIGM is applied to a function F with a (δ, M)–oracle with the sequences $\{\alpha_i\}_{i\geq 0}$ and $\{B_i\}_{i\geq 0}$. Then for all $k \geq 0$, we have

$$F(y_k) - F^* \leq \frac{d(x^*)}{A_k} + \frac{\sum\limits_{i=0}^{k} B_i}{A_k}\delta + \frac{\varepsilon}{2}$$

UIGM Convergence Rate

Statement 6.

Adaptive composite UIGM with Hólder continuous gradient and inexact oracle converges with rates:

$$F(y^N) - F_* \leq \varepsilon, N = O\left(\inf_{\nu\in[0,1]}\left(\frac{L_\nu R^{1+\nu}}{\varepsilon}\right)^{\frac{2}{1+2p\nu-\nu}}\right),$$

$$\delta \leq O\left(\frac{\varepsilon}{N^{p-1}}\right)$$

Strong convex optimization problem

Restart policy

We restart when $\frac{4}{\mu A_k}\omega_n \leq 1$
A_k and ω_n is easy computed. μ is known.

Statement 7.

Strong convexity version of UIGM with restarts converges with rates:

$$F(y^N) - F_* \leq \varepsilon,$$

$$N = O\left(\inf_{\nu\in[0,1]}\left(\frac{L_\nu^{\frac{2}{1+\nu}}}{\mu\varepsilon^{\frac{1-\nu}{1+\nu}}}\omega_n\right)^{\frac{1+\nu}{1+2p\nu-\nu}} \cdot \left\lceil \ln\left(\frac{\mu R^2}{\varepsilon}\right)\right\rceil\right)$$

$$\delta \leq O\left(\frac{\varepsilon}{N^{p-1}}\right)$$

Refernces

- *Devolder O., Glineur F., Nesterov Yu.* Intermediate gradient methods for smooth convex problems with inexact oracle // CORE Discussion Paper 2013/17. 2013.
- *Nesterov Yu.* Universal gradient methods for convex optimization problems // Math. Prog. 2015.
- *A.Gasnikov, Yu.Nesterov* Universal fast gradient method for stochastic composit optimization problems. 2017

Thank you for your attention!