First-order Methods for Convex Optimization with Inexact Oracle

O. Devolder (F.R.S.-FNRS Research Fellow),
F. Glineur and Y. Nesterov

Center for Operations Research and Econometrics (CORE),
Université catholique de Louvain (UCL)

OPTEC Workshop on Large Scale Convex Quadratic Programming, Leuven, November 25, 2010

Outline

1. First-order methods in smooth convex optimization
2. Definition of inexact oracle
3. Examples of inexact oracles
4. Effect of inexact oracle on GM/FGM
5. Applications in Non-smooth Convex Optimization

Outline

Smooth convex optimization

$$f^* = \min_{x \in Q} f(x)$$

where

- $Q \subset \mathbb{R}^n$ is a closed convex set
- $f : Q \to \mathbb{R}$ is
  1. convex:
  $$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle \quad \forall x, y \in Q$$
  2. smooth with Lipschitz-continuous gradient:
  $$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L(f)}{2} \|x - y\|_2^2 \quad \forall x, y \in Q.$$

Notation: $f \in F^{1,1}_{L(f)}(Q)$

First-order Methods

- Numerical methods using only values of the function and of the gradient at some points.
  This first-order information is given by an **Oracle** $\mathcal{O}$.

- Oracle = Unit (Black-box) that computes $f(x_k)$ and $\nabla f(x_k)$ for the numerical method at each point x_k :

$$(f(x_k), \nabla f(x_k)) = \mathcal{O}(x_k).$$

First-order Methods

- Why FOM ?
  Methods of choice for large-scale problems due to their cheap iteration cost.
  Obtaining an ϵ-solution $\tilde{x}$ i.e.:

$$f(\tilde{x}) - f^* \leq \epsilon$$

  can take large number of iterations but each iteration is very easy.

- In Smooth Convex Optimization, two main FOM:
  1. Gradient method (GM)
  2. Fast gradient method (FGM)

Gradient Method

Very simple algorithm:

Initialization

Choose $x_0 \in Q$

Iteration $k \geq 0$

- $(f(x_k), \nabla f(x_k)) = \mathcal{O}(x_k)$
- $x_{k+1} = \arg\min_{x \in Q} [f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{L(f)}{2} \|x - x_k\|_2^2]$

Remark: When $Q = \mathbb{R}^n$: $x_{k+1} = x_k - \frac{1}{L(f)} \nabla f(x_k)$.

GM: Convergence rate

Convergence rate proportional to $\frac{1}{k}$:

$$f(x_k) - f^* \leq \frac{L(f)\,\|x_0 - x^*\|_2^2}{2k} = \Theta\left(\frac{L(f)R^2}{k}\right)$$

where $R = \|x_0 - x^*\|_2$.

Complexity: ϵ-solution obtained after $O\left(\frac{L(f)R^2}{\epsilon}\right)$ iterations.

Fast Gradient Method

Accelerated version of the gradient method due to Nesterov:
Let $\{\alpha_k\}_{k=0}^{\infty}$ satisfying $\alpha_0 \in]0,1]$, $\quad \alpha_k^2 \leq \sum_{i=0}^{k} \alpha_i$.

Initialization

Choose $x_0 \in Q$

Iteration $k \geq 0$

- $(f(x_k), \nabla f(x_k)) = \mathcal{O}(x_k)$
- $y_k = \arg\min_{x \in Q}\{f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{L(f)}{2} \|y - x_k\|_2^2\}$
- $z_k = \arg\min_{x \in Q}\{\sum_{i=0}^{k} \alpha_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] + \frac{L(f)}{2} \|x - x_0\|_2^2\}$
- $\tau_k = \frac{\alpha_{k+1}}{\sum_{i=0}^{k+1} \alpha_i}$
- $x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k$

FGM: Convergence rate

Convergence rate proportional to $\frac{1}{k^2}$:

$$f(y_k) - f^* \leq \frac{4L(f)\left\|x_0 - x^*\right\|_2^2}{(k+1)(k+2)} = \Theta\left(\frac{L(f)R^2}{k^2}\right)$$

This rate of convergence is optimal for FOM on $F^{1,1}_{L(f)}(Q)$.

Complexity: ϵ-solution can be obtained after $O\left(\sqrt{\frac{L(f)}{\epsilon}}R\right)$ iterations.

Outline

Why inexact oracle ?

- Sometimes: impossible/costly to compute exact first-order information (function and gradient value).
- Possible reasons:
  1. Numerical errors
  2. $f(x)$ is defined by another (simple) optimization problem that can be solved only approximately.
  3. f is not as smooth as we want
- Our goal: study the effect of inexact first-order informations on GM and FGM.

Previous definitions of inexact oracle

1. ϵ-**subgradient** (Rockafellar, Shor,...)

$$g_{y,\epsilon} \text{ s.t. } f(x) \geq f(y) + \langle g_{y,\epsilon} x - y \rangle - \epsilon \quad \forall x \in Q$$

Weak condition. Easy to satisfy but good only for non-smooth convex function.

2. **Comparison with exact gradient/subgradient** (Baes, D'Aspremont,...)
Various possible conditions, $g_{y,\mu}$ such that:
   - $\|\nabla f(y) - g_{y,\mu}\| \leq \mu$
   - $\|g(y) - g_{y,\mu}\| \leq \mu, g(y) \in \partial f(y)$
   - $|\langle \nabla f(y) - g_{y,\mu}, x - z \rangle| \leq \mu \quad \forall x, z \in Q$

Good results can be obtained **but**
Strong conditions: Difficult to guarantee in practice.
Restrictive assumptions: Sometime $\nabla f(y)$ must exist, sometime Q must be bounded.

Definition of inexact oracle

Exact Oracle:
If $f \in F^{1,1}_{L(f)}(Q)$ then the output of the oracle $(f(y), \nabla f(y)) = \mathcal{O}(y)$ is characterized by:

$$f(y)+\langle \nabla f(y), x-y\rangle \leq f(x) \leq f(y)+\langle \nabla f(y), x-y\rangle+\frac{L(f)}{2}\|x-y\|_2^2$$

for all $x \in Q$.

Inexact Oracle:
f is equipped with a first-order (δ, L) oracle if for all $y \in Q$, we can compute $(f_{y,\delta}, g_{y,\delta}) = \mathcal{O}_{\delta,L}(y)$:

$$f_{y,\delta}+\langle g_{y,\delta}, x-y\rangle \leq f(x) \leq f_{y,\delta}+\langle g_{y,\delta}, x-y\rangle+\frac{L}{2}\|x-y\|_2^2+\delta \quad \forall x \in Q.$$

Definition of inexact oracle

Consequences:

- $f_{y,\delta}$ is a δ-lower approximation of $f(y)$:

$$f_{y,\delta} \leq f(y) \leq f_{y,\delta} + \delta.$$

- $g_{y,\delta}$ is a δ-subgradient of f at y:

$$f(x) \geq f(y) + \langle g_{y,\delta}, x - y \rangle - \delta.$$

- In general, L is not the original Lipschitz constant $L(f)$.

Outline

1) Exact computation at shifted point

Let $f \in F^{1,1}_{L(f)}(Q)$.

Inexact oracle: At each point $y \in Q$, the oracle provides exact value of f and ∇f but at a different point y_δ such that

$$\|y - y_\delta\|_2^2 \leq \frac{\delta}{L(f)}.$$

If we define:

$$f_{y,\delta} = f(y_\delta) + \langle \nabla f(y_\delta), y - y_\delta \rangle, \quad g_{y,\delta} = \nabla f(y_\delta)$$

$\Rightarrow$ (δ, L)-oracle with $L = 2L(f)$.

2) Inexact oracle for saddle-point problems

Assume that $f \in F^{1,1}_{L(f)}(Q)$ is defined by another optimization problem:

$$f(x) = \max_{u \in U} \Psi(x, u)$$

where Ψ is concave in u, convex in x and U is closed and convex.

Computations of $f(x)$ and $\nabla f(x)$ require

$$u_x \in \operatorname{Arg}\max_{u \in \overline{U}} \Psi(x, u)$$

since:

$$f(x) = \Psi(x, u_x) \quad \nabla f(x) = \nabla_x \Psi(x, u_x).$$

But in practice, we are only able to solve this subproblem approximatively, computing $\overline{u}_x$, an approximate solution.

Consequences?
Which quality of $\overline{u}_x$ ensures a (δ, L)-oracle ?

2a) Function obtained by the smoothing technique

When applying smoothing technique, we need to solve saddle-point problem with:

$$\Psi(x,u) = G(u) + \langle Au, x\rangle$$

where G is strongly concave with parameter κ.
We know that:

- $f(x) = \max_{u\in U} \Psi(x,u) \in F^{1,1}_{L(f)}(Q)$ with $L(f) = \frac{\|A\|_2^2}{\kappa}$
- $f(x) = \Psi(x, u_x)$ and $\nabla f(x) = Au_x$.

<u>Inexact oracle:</u> If $\overline{u}_x$ satisfies

$$\Psi(x,u_x) - \Psi(x,\overline{u}_x) \le \frac{\delta}{2}$$

and

$$f_{x,\delta} = \Psi(x,\overline{u}_x) \quad g_{x,\delta} = A\overline{u}_x$$

$\Rightarrow$ (δ, L)-oracle with $L = 2L(f)$.

2b) Function obtained in the Augmented Lagrangian Approach

When solving the convex problem $\min_{u \in U}\{H(u) \text{ s.t. } Au = 0\}$ using augmented Lagrangian approach, we need to solve saddle-point problem with:

$$\Psi(x,u) = -H(u) + \langle Au, x\rangle - \frac{\kappa}{2}\|Au\|_2^2 .$$

We know that:

- $f(x) = \max_{u \in U} \Psi(x,u) \in F^{1,1}_{L(f)}(Q)$ with $L(f) = \frac{1}{\kappa}$
- $$f(x) = \Psi(x, u_x) \quad \nabla f(x) = Au_x.$$

<u>Inexact oracle:</u> If $\overline{u}_x$ satisfies

$$\max_{u \in U}\langle \nabla_u \Psi(x, \overline{u}_x), u - \overline{u}_x\rangle \leq \delta$$

and

$$f_{x,\delta} = \Psi(x, \overline{u}_x) \quad g_{x,\delta} = A\overline{u}_x$$

$\Rightarrow$ (δ, L)-oracle with $L = L(f)$.

3) Applications in Non-smooth Convex Optimization

The condition on $(f_{y,\delta}, g_{y,\delta})$:

$$f_{y,\delta} + \langle g_{y,\delta}, x - y\rangle \le f(x) \le f_{y,\delta} + \langle g_{y,\delta}, x - y\rangle + \frac{L}{2}\|x - y\|_2^2 + \delta \quad (1)$$

does not imply differentiability.

Consider the case of a non-smooth convex function f with bounded variation of the subgradients:

$$\|g(x) - g(y)\|_* \le M(f) \quad \forall g(x) \in \partial f(x), g(y) \in \partial f(y), \forall x, y \in Q.$$

Then $(f(y), g(y))$ provides a (δ, L)-oracle with arbitrary δ and $L = \frac{M(f)^2}{2\delta}$.

3) Application in Non-smooth convex Optimization

Remarks:

- The first-order informations $(f(y), g(y))$ are exact, we have a exact oracle but of non-smooth optimization.
- This non-smooth exact oracle can be seen as a inexact (δ, L) smooth oracle.
- δ is not really a given accuracy, it is a parameter that we can choose but there is a tradeoff with $L = \frac{M(f)^2}{2\delta}$.

Outline

Effect of inexact oracle on FOM ?

Effect on gradient method (GM) and on fast gradient method (FGM) if we use an (δ, L)-oracle instead of a exact one by replacing:

$$(f(y), \nabla f(y)) \text{ by } (f_{y,\delta}, g_{y,\delta})$$

and

$$L(f) \text{ by } L?$$

Important Issues:

- Link between desired solution accuracy (SA) and accuracy needed for the oracle (OA).
- Does the FGM still outperform GM when a inexact oracle is used ?

Gradient Method with Inexact Oracle

Exact oracle:

$$f(x_k) - f^* \leq \frac{L(f)R^2}{2k}$$

(δ, L)-oracle:

$$f(x_k) - f^* \leq \frac{LR^2}{2k} + \delta.$$

- No accumulation of errors
  Error asymptotically tends to δ (OA).
- Complexity: ϵ-solution if $k \geq O\left(\frac{LR^2}{\epsilon-\delta}\right)$
- Let ϵ be the desired accuracy for the solution (SA). We can take OA of same order than SA: $\delta = \Theta(\epsilon)$.

Fast Gradient Method with Inexact Oracle

Exact oracle:

$$f(y_k) - f^* \leq \frac{4L(f)R^2}{(k+1)(k+2)}$$

(δ, L)-oracle:

$$f(y_k) - f^* \leq \frac{4LR^2}{(k+1)(k+2)} + \frac{1}{6}(2k+6)\delta.$$

- Accumulation of errors
  Divergence: Error asymptotically tends to ∞ (Decreases fast at first then increases).
- Complexity: ϵ-solution if $\Theta\left(\sqrt{\frac{L}{\epsilon}}R\right) \leq k \leq \Theta\left(\frac{\epsilon}{\delta}\right)$
- OA must be smaller than SA: $\delta = \Theta(\epsilon^{3/2})$.

Which method should we choose?

We have to consider three cases depending on the available oracle:

1. Exact oracle
2. Inexact oracle with a fixed accuracy δ
3. Inexact oracle but the accuracy δ can be chosen.

Case 1: Exact oracle

In order to have a SA of ϵ:

$$\text{GM} : O\left(\frac{L(f)R^2}{\epsilon}\right) \text{ iterations}$$

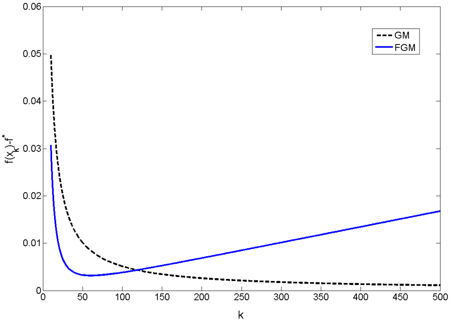
$$\text{FGM} : O\left(\sqrt{\frac{L(f)}{\epsilon}}R\right) \text{ iterations}$$

FGM outperforms GM in all cases.

Case 2: Inexact oracle with fixed OA δ

GM : $f(x_k) - f^* \leq \frac{LR^2}{2k} + \delta$

FGM : $f(y_k) - f^* \leq \frac{4LR^2}{(k+1)(k+2)} + \frac{1}{6}(2k+6)\delta$



We need to stop the FGM after $k^* = \Theta\left(\sqrt[3]{\frac{LR^2}{\delta}}\right)$ iterations:

best SA reachable by the FGM $\epsilon^* = \Theta(\delta^{2/3})$.

Case 2: Inexact oracle with fixed OA ϵ

We need to stop the FGM after $k^* = \Theta\left(\sqrt[3]{\frac{LR^2}{\delta}}\right)$ iterations:

best SA reachable by the FGM $\epsilon^* = \Theta(\delta^{2/3})$.

- If such accuracy is sufficient for the solution: FGM
- If not, the only possibility: GM.

Case 3: Inexact oracle but the OA δ can be chosen

In order to have a SA of ϵ:

$$\text{GM} : O\left(\frac{LR^2}{\epsilon}\right) \text{ iterations but with } \delta = \Theta(\epsilon)$$

$$\text{FGM} : O\left(\sqrt{\frac{L}{\epsilon}}R\right) \text{ iterations but with } \delta = \Theta(\epsilon^{3/2})$$

Choice depends on the complexity of inexact oracle.
Let $C(\delta)$= number of operations needed by the inexact oracle to compute $(f_{x,\delta}, g_{x,\delta})$.

- If $C(\delta) = \Omega\left(\frac{1}{\delta}\right)$ (expensive inexact oracle), we have to use GM.
- If $C(\delta) = o\left(\frac{1}{\delta}\right)$ (cheap inexact oracle), we have to use FGM.

Outline

Applications in Non-smooth Convex Optimization

Recall that when f is a non-smooth convex function with bounded variation of the subgradients i.e:

$$\|g(x) - g(y)\|_* \leq M(f) \quad \forall g(x) \in \partial f(x), g(y) \in \partial f(y), \forall x, y \in Q$$

The non-smooth exact oracle can be seen as a inexact (δ, L) smooth oracle:

$$f_{y,\delta} = f(y) \quad g_{y,\delta} = g(y) \in \partial f(y)$$

where δ is arbitrary and $L = \frac{M(f)^2}{2\delta}$.

Applications in Non-smooth Convex Optimization

These observations gives us the possibility to apply any FOM of smooth convex-optimization to a non-smooth function:

1. We can apply GM with inexact oracle to the non-smooth function f. With a optimal choice of δ:
   Optimal rate of convergence $\Theta\left(\frac{M(f)R}{\sqrt{k}}\right)$ for the non-smooth problem.
2. We can apply FGM with inexact oracle to the non-smooth function f. With a optimal choice of δ:
   Optimal rate of convergence $\Theta\left(\frac{M(f)R}{\sqrt{k}}\right)$ for the non-smooth problem.

Intrinsic accumulation of errors for fast FOM

The applicability of our definition of inexact oracle to non-smooth function gives us also the possibility to prove that:

Accumulation of errors = Intrinsic and unavoidable property of any fast FOM using inexact oracle.

If there exists FOM of smooth convex optimization with:

- optimal rate $\Theta\left(\frac{L(f)R^2}{k^2}\right)$ in the exact case
- without accumulation of errors in the inexact case

then we could solve the non-smooth problem $\min_{x \in Q} f(x)$ with a strictly better convergence rate than $\Theta\left(\frac{M(f)R}{\sqrt{k}}\right)$.

Impossible!

Intrinsic accumulation of errors for fast FOM

More generaly, we can prove the following result:

Theorem

Consider a FOM using a (δ, L)-oracle with convergence rate:

$$f(x_k) - f^* \leq \frac{C_1 L R^2}{k^p} + C_2 k^q \delta$$

then necessarily $q \geq p - 1$.

In particular:

- $q = 0 \Rightarrow p \leq 1$: GM is the fastest FOM without error accumulation
- $p = 2 \Rightarrow q \geq 1$: Any FOM with convergence rate $\frac{1}{k^2}$ must suffer from error accumulation and FGM has the lowest possible error accumulation for such a method: $\Theta(k\delta)$.

Conclusion

- Introduction of a new definition of inexact oracle: (δ, L)-oracle.
- Important examples fit with this definition: computation at shifted point, approximative resolution of subproblems for saddle-point functions, function not as smooth as we want...
- GM is slow but robust with respect to oracle error. It is the fastest FOM without error accumulation.
- FGM is faster but sensitive with respect to oracle error. Like any FOM with optimal convergence rate, it suffers from accumulation of errors.

Further Research

- Using (δ, L)-oracle for saddle-point problems, find the exact total complexity of Augmented Lagrangian approach. Better to use GM or FGM ?
- Development of intermediate FOM between GM and FGM (kind of interpolation) with intermediate accumulation of errors
- Generalization of our results to non-smooth functions when the non-smooth oracle is also inexact
- Study of random inexact oracle
- ...

Thanks for your attention !

UCL

Université catholique de Louvain