First-Order Methods of Smooth Convex Optimization with Inexact Oracle

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

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

2011 SIAM Conference on Optimization, Darmstadt, May 17

Outline

1. First-order methods in smooth convex optimization
2. Definition of inexact oracle
3. FOM of smooth convex optimization with inexact oracle
4. Applications
   - Lack of accuracy in the first-order information
   - Lack of smoothness for the function

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 :

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

- Why FOM ?
  Methods of choice for large-scale problems due to their cheap iteration cost.
- 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 $k \geq 0$

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

Convergence rate in $O(\frac{1}{k}) \Rightarrow$ **Non-optimal FOM for** $F^{1,1}_{L(f)}(Q)$

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 $k \geq 0$

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

Convergence rate in $O(\frac{1}{k^2}) \Rightarrow$ **Optimal FOM for** $F_{L(f)}^{1,1}(Q)$

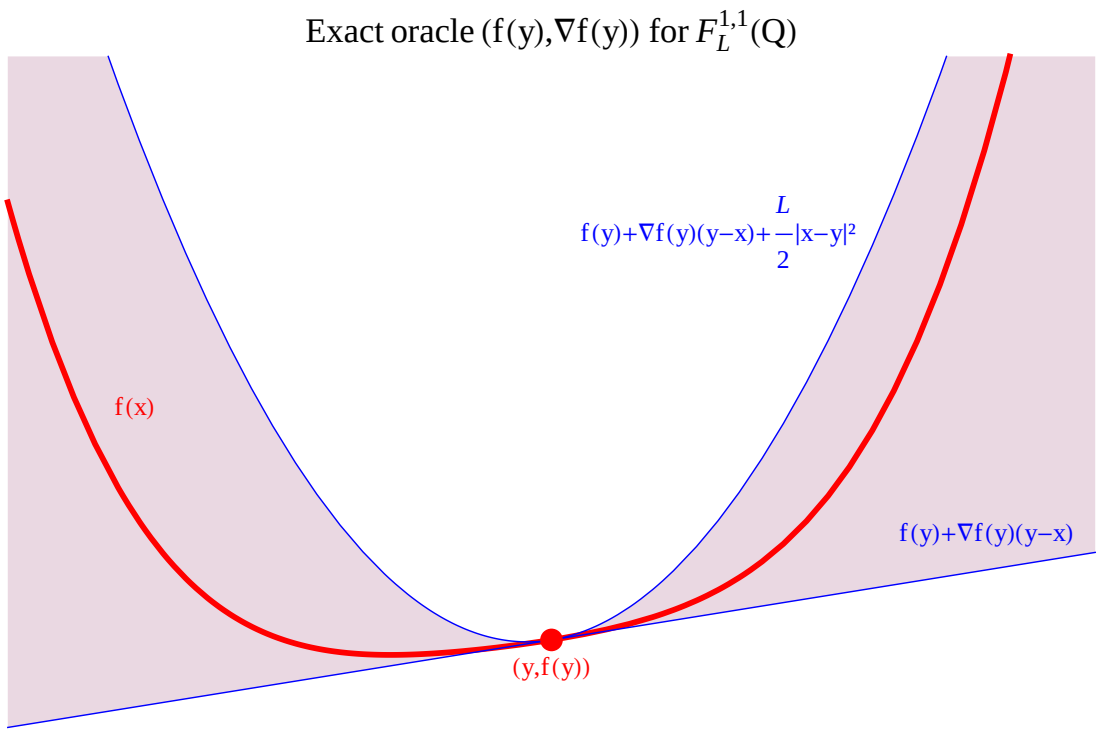
Outline

Exact Oracle for $F^{1,1}_{L(f)}(Q)$

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:

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

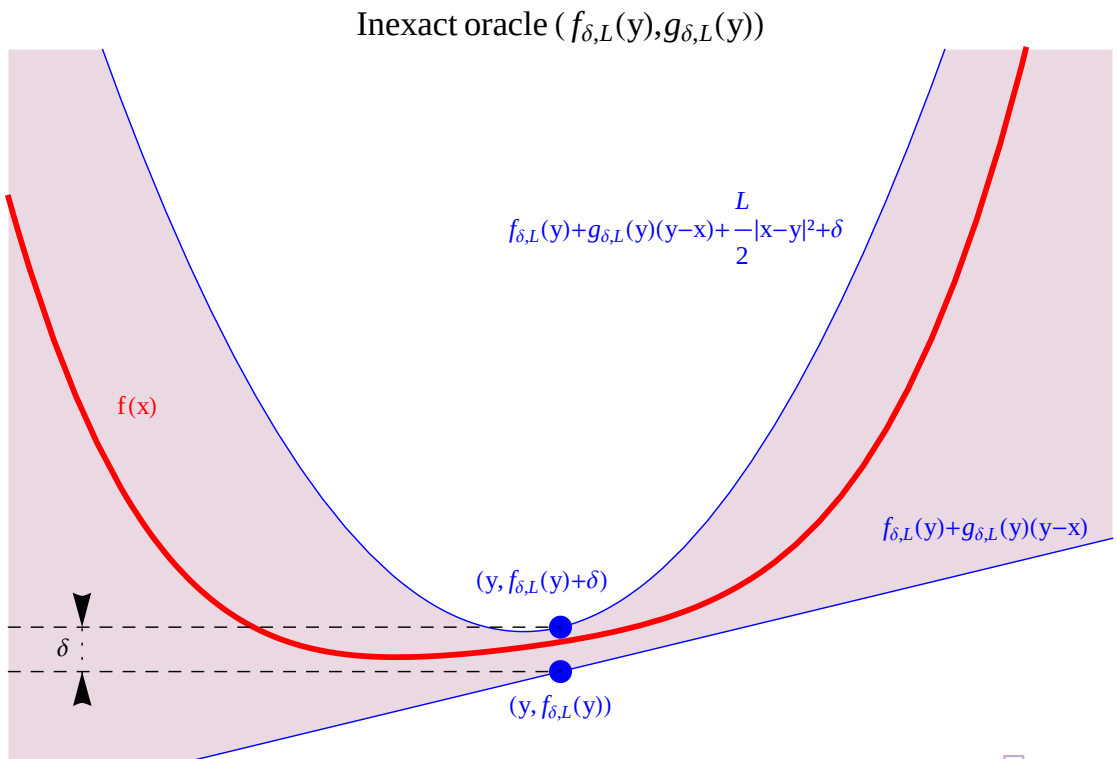
for all $x \in Q$.



(δ, L)-oracle

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

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



Outline

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: $\delta = \Theta(\epsilon)$ e.g. $\delta = \frac{\epsilon}{2}$

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:

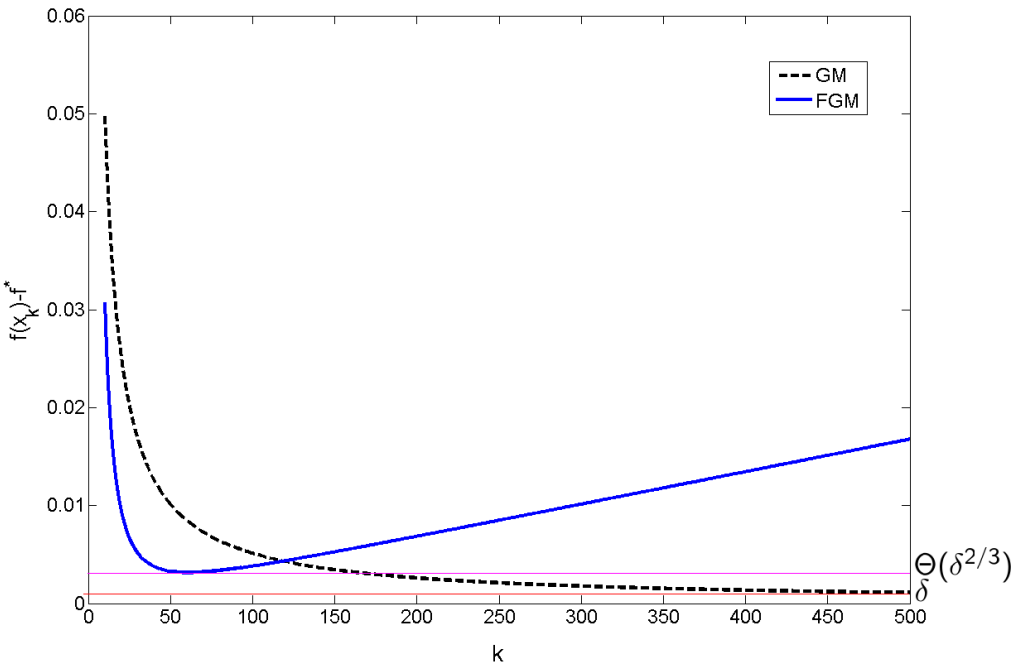
$$f(y_k) - f^* \leq \frac{4LR^2}{(k+1)(k+2)} + \frac{1}{6}(2k+6)\delta.$$

- 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?
Case 1: 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 $\epsilon^* = \Theta(\delta^{2/3})$.

If such accuracy is sufficient: FGM. If not, only possibility: GM.

Which method should we choose?
Case 2: 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 $C(\delta) = \Theta\left(\frac{1}{\delta}\right)$, the two methods are equivalent.
- If $C(\delta) = o\left(\frac{1}{\delta}\right)$ (cheap inexact oracle), we have to use FGM.

Intrinsic accumulation of errors for fast FOM

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

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: $\Theta(k\delta)$.

Outline

Applications

Two kind of situations where a (δ, L) oracle can be available:

1. **Lack of accuracy in the first-order information**
   Smooth function (i.e. in $F^{1,1}_{L(f)}(Q)$) when the first-order information is computed approximately.
   In this case, δ represent the accuracy of the first-order information.

2. **Lack of smoothness for the function**
   Function with weaker level of smoothness (but typically with exact first-order information).
   In this case, δ can be chosen but there is a trade-off with L.

Outline

(δ, L)-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

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

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 approximately, computing $\overline{u}_x$, an approximate solution.

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

(δ, L)-oracle for saddle-point problems: Examples

- **Smoothing technique**:
  $f(x) = \max_{u \in U}\{\Psi(x,u) = G(u) + \langle Au, x\rangle\}$ where G is strongly concave with parameter κ.
  If $V_1(\overline{u}_x) = \Psi(x, u_x) - \Psi(x, \overline{u}_x) \leq \frac{\delta}{2}$,
  $f_{x,\delta} = \Psi(x, \overline{u}_x) \quad g_{x,\delta} = A\overline{u}_x \Rightarrow (\delta, 2L(f))$-oracle
- **Moreau-Yosida Regularization**:
  $f(x) = \min_{u \in U}\{\mathcal{L}(x,u) = h(u) + \frac{\kappa}{2}\|u - x\|_2^2\}$.
  If $V_2(\overline{u}_x) = \max\limits_{u \in U}\left\{\mathcal{L}(x, \overline{u}_x) - \mathcal{L}(x,u) + \frac{\kappa}{2}\|u - \overline{u}_x\|_2^2\right\} \leq \delta$,
  $f_{x,\delta} = \mathcal{L}(x, \overline{u}_x) - \delta \quad g_{x,\delta} = \kappa(x - \overline{u}_x) \Rightarrow (\delta, L(f))$-oracle.
- **Augmented Lagrangian Approach**
  $f(x) = \max_{u \in U}\{\Psi(x,u) = -H(u) + \langle Au, x\rangle - \frac{\kappa}{2}\|Au\|_2^2\}$.
  If $V_3(\overline{u}_x) = \max_{u \in U}\langle \nabla_u \Psi(x, \overline{u}_x), u - \overline{u}_x\rangle \leq \delta$
  $f_{x,\delta} = \Psi(x, \overline{u}_x) \quad g_{x,\delta} = A\overline{u}_x \Rightarrow (\delta, L(f))$-oracle.

Outline

1. First-order methods in smooth convex optimization
2. Definition of inexact oracle
3. FOM of smooth convex optimization with inexact oracle
4. Applications
   - Lack of accuracy in the first-order information
   - Lack of smoothness for the function

(δ, L) oracle for non-smooth or weakly-smooth functions

Assume that f satisfies the following smoothness condition:

$$\|g(x) - g(y)\|_* \leq L_\nu \|x - y\|^\nu, \forall x, y \in Q, \forall g(x) \in \partial f(x), g(y) \in \partial f(y).$$

When:

1. $\nu = 1$: f is smooth with a Lipschitz-continuous gradient
2. $\nu = 0$: f is non-smooth with bounded variation of the subgradients
3. $0 < \nu < 1$: f is weakly-smooth i.e. with a Hölder-continuous gradient.

Important Observation: The exact oracle $(f(y), g(y))$ can be seen as a inexact (δ, L) smooth oracle where δ is arbitrary and

$$L = L_\nu \left[\frac{L_\nu}{2\delta} \cdot \frac{1-\nu}{1+\nu}\right]^{\frac{1-\nu}{1+\nu}}.$$

The FGM as a Universal Optimal FOM

This observation gives us the possibility to apply any FOM of smooth convex-optimization to a function with weaker level of smoothness:

1. We can apply GM with inexact oracle to a non- or weakly-smooth function. With a optimal choice of δ:
   Non-optimal rate of convergence $\Theta\left(\frac{L_\nu R^{1+\nu}}{k^{\frac{1+\nu}{2}}}\right)$.
2. We can apply FGM with inexact oracle to a non- or weakly-smooth function. With a optimal choice of δ:
   Optimal rate of convergence $\Theta\left(\frac{L_\nu R^{1+\nu}}{k^{\frac{1+3\nu}{2}}}\right)$.

The FGM can reach optimal convergence rate for various classes of convex problems characterized by different levels of smoothness.
$\Rightarrow$ **FGM= Universal Optimal FOM**.

Conclusion

- Introduction of a new definition of inexact oracle: (δ, L)-oracle.
- Important examples where the first-order information is computed with numerical errors or using approximative solution of subproblems fit with this definition
- The GM is slow but robust with respect to oracle error. It is the fastest FOM without error accumulation.
- The FGM is faster but sensitive to oracle error. Like any FOM with optimal convergence rate, it suffers from accumulation of errors.
- As exact non-smooth oracle= inexact smooth oracle
  We can apply FOM of smooth convex opt. to non-smooth (and weakly-smooth) convex problems.
  $\Rightarrow$ FGM= Universal Optimal FOM.

Thanks for your attention !

Paper: O. Devolder, F. Glineur and Y. Nesterov
First-order Methods of Smooth Convex Optimization with Inexact Oracle
Available on Optimization Online.
Submitted to Mathematical Programming.