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)

CORE Mathematical Programming seminar, March 15, 2011.

Convex Optimization Problems

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

where:

1. Q is
   - closed
   - convex: $\alpha x + (1-\alpha)y \in Q \quad \forall x, y \in Q, \alpha \in [0,1]$
2. $f : Q \to \mathbb{R}$ is
   - closed i.e that epi$f$ is closed
   - convex:
     $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y) \quad \forall x, y \in Q, \alpha \in [0,1].$

Outline

Outline

First-order Methods

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

- First-order Oracle = Unit (Black-box) that computes $f(x_k)$ and $g(x_k) \in \partial f(x_k)$ for the numerical method at each search point x_k :
$$(f(x_k), g(x_k)) = \mathcal{O}(x_k).$$
- Why FOM ?
  Methods of choice for large-scale problems due to their cheap iteration cost.
  Obtaining an ϵ-solution $\overline{x}$ i.e.:
$$f(\overline{x}) - f^* \leq \epsilon$$
  can take large number of iterations but each iteration is very easy.

What we can expect from a FOM ? Information-based Complexity

Let $\mathcal{F}(Q)$ be a family/class of convex problems of the form: $\min_{x \in Q} f(x)$.
Let $\mathcal{P}$ be an instance in $\mathcal{F}(Q)$
Let $\mathcal{M}$ be a first-order method that satisfies the following properties:

1. **Initial Knowledge**
   When starting to solve $\mathcal{P}$ using $\mathcal{M}$:
   - An accuracy $\epsilon > 0$ to which the problem $\mathcal{P}$ should be solved is given
   - $\mathcal{M}$ does not know what is $\mathcal{P}$ but only that $\mathcal{P} \in \mathcal{F}(Q)$.

Information-based Complexity (2)

2. **Accumulation of Information**
   - In course of solving $\mathcal{P}$, the only information that can obtain $\mathcal{M}$ about $\mathcal{P}$ comes from a **first-order oracle**. The method has no access to the problem structure (Black-box approach).
   - $\mathcal{M}$ generates a sequence of search points $x_1, x_2, \ldots$ using the information given by the oracle.
     Rules for building x_t must be **non-anticipating**: x_t can depend only on the information already accumulated at the first $t-1$ steps i.e. $f(x_1), g(x_1), \ldots, f(x_{t-1}), g(x_{t-1})$.
3. **Termination and construction of the output**
   After $N = N_{\mathcal{M}}(\mathcal{P}, \epsilon)$ calls to the oracle, $\mathcal{M}$ terminates and outputs the result $\hat{x}$ such that:
   - $\hat{x}$ depends only on the information on $\mathcal{P}$ already accumulated
   - $\hat{x} \in Q$
   - $f(\hat{x}) - f^* \leq \epsilon$

Information-based Complexity (3)

- **Complexity of the method $\mathcal{M}$ on $\mathcal{F}(Q)$:**

$$\mathrm{Compl}_{\mathcal{M}}(\epsilon) = \max_{\mathcal{P}\in\mathcal{F}(Q)} N_{\mathcal{M}}(\mathcal{P},\epsilon)$$

  $=$ Minimal number of steps in which $\mathcal{M}$ is capable to solve with accuracy ϵ every problem $\mathcal{P}$ in $\mathcal{F}(Q)$

- **Information-based complexity of the family $\mathcal{F}(Q)$:**

$$\mathrm{Compl}(\epsilon) = \min_{\mathcal{M}} \ \mathrm{Compl}_{\mathcal{M}}(\epsilon)$$

  $=$ Optimal complexity of a first-order method for $\mathcal{F}(Q)$

- $\mathcal{M}$ is an **Optimal Method** for $\mathcal{F}(Q)$ if:

$$\mathrm{Compl}_{\mathcal{M}}(\epsilon) = \Theta\left(\ \mathrm{Compl}(\epsilon)\right).$$

Outline

Convexity versus Strong Convexity

- $f : Q \to \mathbb{R}$ is **convex** if:

  $f(\alpha x+(1-\alpha)y) \leq \alpha f(x)+(1-\alpha)f(y) \quad \forall x, y \in Q, \forall \alpha \in [0,1]$

  First-order information $(f(x), g(x))$ with $g(x) \in \partial f(x)$ satisfies:

$$f(y) \geq f(x) + \langle g(x), y - x \rangle \quad \forall y \in Q$$

- $f : Q \to \mathbb{R}$ is **strongly convex with parameter** $\mu > 0$ if:

  $f(\alpha x+(1-\alpha)y) \leq \alpha f(x)+(1-\alpha)f(y)-\alpha(1-\alpha)\frac{\mu}{2}\|x-y\|^2$

  $\forall x, y \in Q, \forall \alpha \in [0,1]$.

  First-order information $(f(x), g(x))$ with $g(x) \in \partial f(x)$ satisfies:

$$f(y) \geq f(x) + \langle g(x), y - x \rangle + \frac{\mu}{2}\|x - y\|^2 \quad \forall y \in Q$$

Convexity assumptions : a way to obtain lower bounds on f.

Lipschitz-continuity of f versus Lipschitz-continuity of ∇f

- $f : Q \to \mathbb{R}$ is **Lipschitz-continuous with constant** M if:

$$|f(x) - f(y)| \leq M \|x - y\| \quad \forall x, y \in Q.$$

First-order information $(f(x), g(x))$ with $g(x) \in \partial f(x)$ satisfies:

$$f(y) \leq f(x) + \langle g(x), y - x \rangle + M \|x - y\| \quad \forall y \in Q.$$

- $f : Q \to \mathbb{R}$ has a **Lipschitz-continuous gradient with constant** L if:

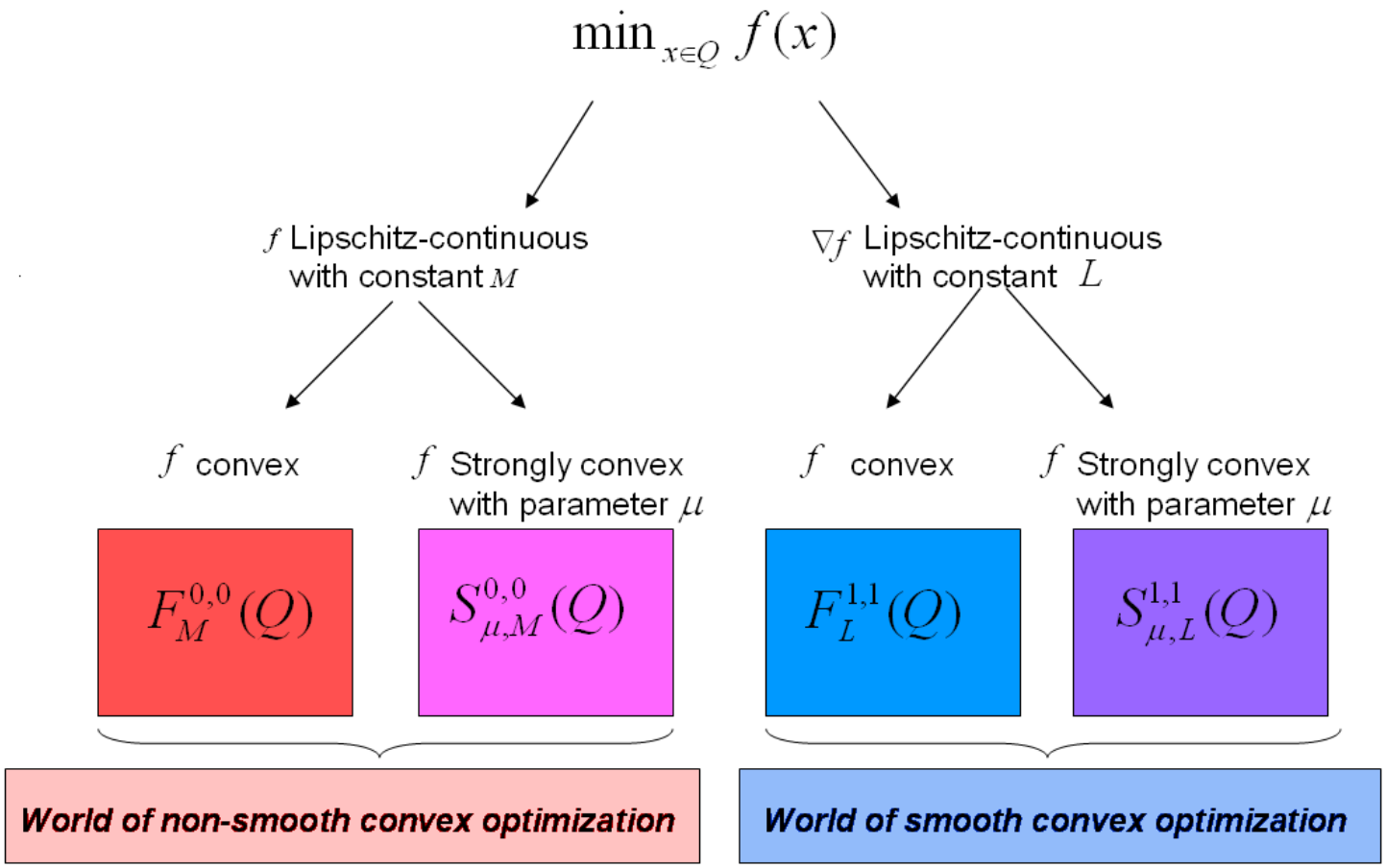
$$\|\nabla f(x) - \nabla f(y)\|_* \leq L \|x - y\| \quad \forall x, y \in Q.$$

First-order information $(f(x), \nabla f(x))$ satisfies:

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

Lipschitz assumptions : a way to obtain upper bounds on f.

Classes of Convex Functions



Optimal Complexity of Classes of Convex Functions

Class	Optimal Complexity	Optimal Methods.
$F_M^{0,0}(Q)$: f conv. f Lipscht-cont.	$\Theta\left(\frac{M^2R^2}{\epsilon^2}\right)$	Subgradient Methods, Mirror descent Methods
$S_{\mu,M}^{0,0}(Q)$: f S. conv. f Lipscht-cont.	$\Theta\left(\frac{M^2}{\mu\epsilon}\ln\left(\frac{\mu R^2}{\epsilon}\right)\right)$	Subgradient Methods, Mirror descent Methods
$F_L^{1,1}(Q)$: f conv. ∇f Lipscht-cont.	$\Theta\left(\sqrt{\frac{LR^2}{\epsilon}}\right)$	~~Gradient Method~~ Fast Gradient Method
$S_{\mu,L}^{1,1}(Q)$: f S. conv. ∇f Lipscht-cont.	$\Theta\left(\sqrt{\frac{L}{\mu}}\ln\left(\frac{\mu R^2}{\epsilon}\right)\right)$	~~Gradient Method~~ Fast Gradient Method

where $R = \|x_0 - x^*\| \leq diam(Q)$.

Optimal Complexity of Classes of Convex Functions (2)

Two important observations:

1. Regarding to the (optimal) numerical methods, the world of non-smooth convex optimization and the world of smooth convex optimization are separated
2. For smooth convex optimization, the most classical and known FOM, the gradient method is not optimal.
   There exist optimal methods for these classes, the fast gradient methods (FGM) that outperform theoretically, and often in practice, the classical gradient method (GM).

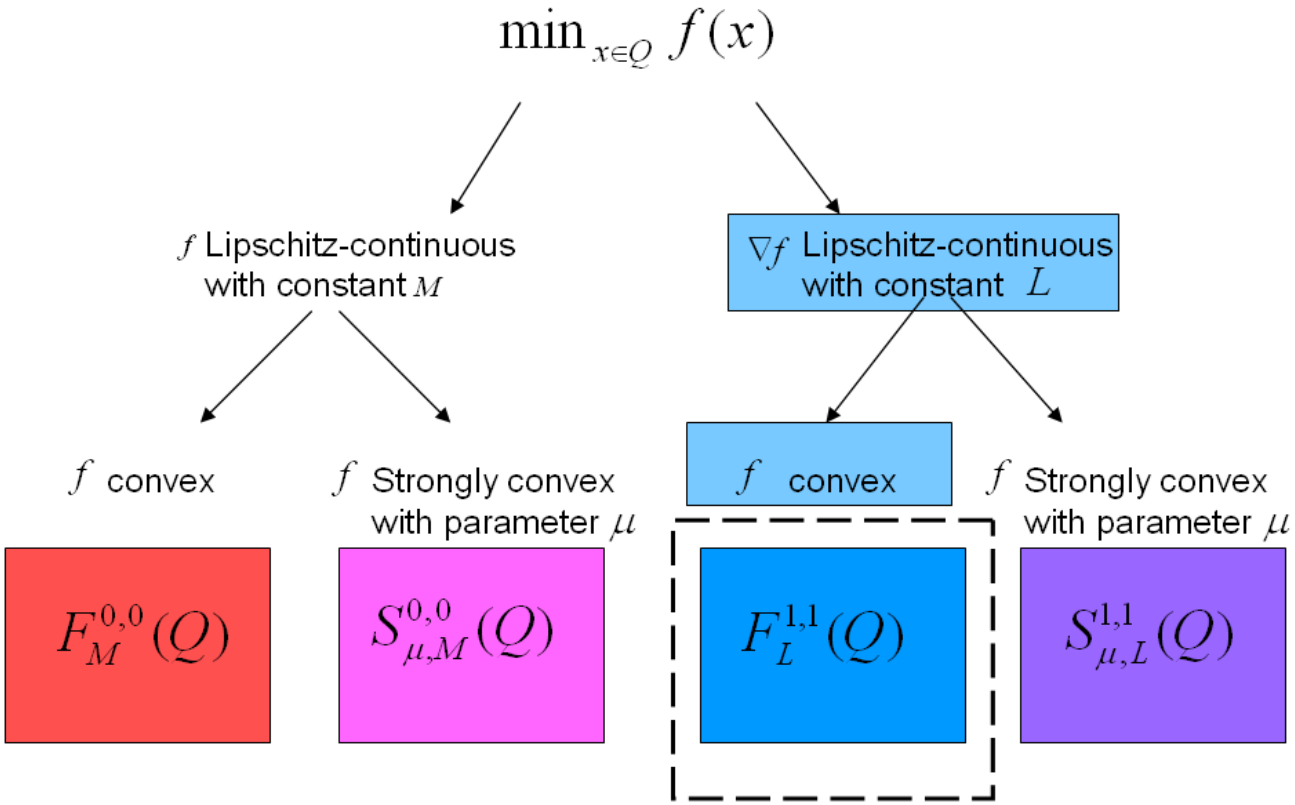
In this work:

We

- Study the effect on the methods of smooth convex optimization when **inexact first-order information** is used.
  $\Rightarrow$ Superiority of the Fast gradient method (FGM) on the classical gradient method (GM) is not anymore absolute (Problem of accumulation of errors).
- Prove that the first-order **methods of smooth convex optimization** (classical and fast gradient methods) can be also **applied to non-smooth convex problems** provided that inexact first-order information is allowed in these schemes.
  $\Rightarrow$The two worlds are not completely separated.

Outline

Our starting point: the class $F^{1,1}_{L(f)}(Q)$



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

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.

$\Rightarrow$ **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 $k \geq 0$

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

FGM: Convergence rate

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

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

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

$\Rightarrow$ **Optimal FOM for** $F^{1,1}_{L(f)}(Q)$

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 information 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** (Mordukhovich, Lemaréchal, Baes, D'Aspremont,...)
   Various possible conditions, $g_{y,\eta}$ such that:
   - $\|\nabla f(y) - g_{y,\eta}\| \leq \eta$
   - $\|g(y) - g_{y,\eta}\| \leq \eta, g(y) \in \partial f(y)$
   - $|\langle \nabla f(y) - g_{y,\eta}, x - z \rangle| \leq \eta \quad \forall x, z \in Q$

   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:

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

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 $(f_{y,\delta}, g_{y,\delta}) = \mathcal{O}_{\delta,L}(y)$:

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

Definition of inexact oracle

Remarks

- FOM for $F^{1,1}_{L(f)}(Q)$ are based on the lower and upper bounds on f.
  Principal motivation of this definition of inexact oracle.
- In general, L is not the original Lipschitz constant $L(f)$
- In some case, $L = L(\delta)$ is a function of the oracle accuracy, which can be chosen arbitrarily. We have a one-parametric family of $(\delta, L(\delta))$ oracle.
- In some case, the oracle accuracy δ is fixed, cannot be chosen. We have a (δ, L) oracle.

Important properties of (δ, L) oracle

- $f_{y,\delta}$ 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:

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

- If f_i has (δ_i, L_i) oracle, $i = 1, 2$, then $f_1 + f_2$ has $(\delta_1 + \delta_2, L_1 + L_2)$ oracle.
- If $Q = \mathbb{R}^n$, for any $g_y \in \partial f(y)$, we have:

$$\|g_y - g_{y,\delta}\|_* \leq [2\delta L]^{1/2}.$$

Outline

1) Exact computation at shifted point

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

<u>Inexact oracle:</u> 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

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

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

$$V_1(\overline{u}_x) = \Psi(x, u_x) - \Psi(x, \overline{u}_x) \leq \frac{\delta}{2}$$

and

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

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

2b) Moreau-Yosida Regularization

Let h be a smooth convex function on a convex set $U \subset \mathbb{R}^n$. The Moreau-Yosida regularization of h is defined by:

$$f(x) = \min_{u \in U} \{\mathcal{L}(x, u) = h(u) + \frac{\kappa}{2} \|u - x\|_2^2\}.$$

We know that:

- $f(x) = \min_{u \in U} \mathcal{L}(x, u) \in F^{1,1}_{L(f)}(Q)$ with $L(f) = \kappa$
- $f(x) = \mathcal{L}(x, u_x)$ and $\nabla f(x) = \kappa(x - u_x)$.

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

$$V_2(\overline{u}_x) = \max_{u \in U} \left\{ \mathcal{L}(x, \overline{u}_x) - \mathcal{L}(x, u) + \frac{\kappa}{2} \|u - \overline{u}_x\|_2^2 \right\} \leq \delta$$

and

$$f_{x,\delta} = \mathcal{L}(x, \overline{u}_x) - \delta \quad g_{x,\delta} = \kappa(x - \overline{u}_x)$$

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

2c) 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

$$V_3(\overline{u}_x) = \max_{u \in U}\langle \nabla_u \Psi(x, \overline{u}_x), u - \overline{u}_x\rangle \leq \delta$$

and

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

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

3) Applications in Non-smooth Convex Optimization

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

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

does not imply differentiability.

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

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

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

3) Applications in Non-smooth convex Optimization

Remarks:

- The first-order information $(f(y), g(y))$ is 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 trade-off with $L = \frac{M(f)^2}{2\delta}$.

More details in the last part of this talk!

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 outperforms GM when a inexact oracle is used ?

Gradient Method with Inexact Oracle

<u>Exact oracle:</u>

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

<u>(δ, L)-oracle:</u>

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

- 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?

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:

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

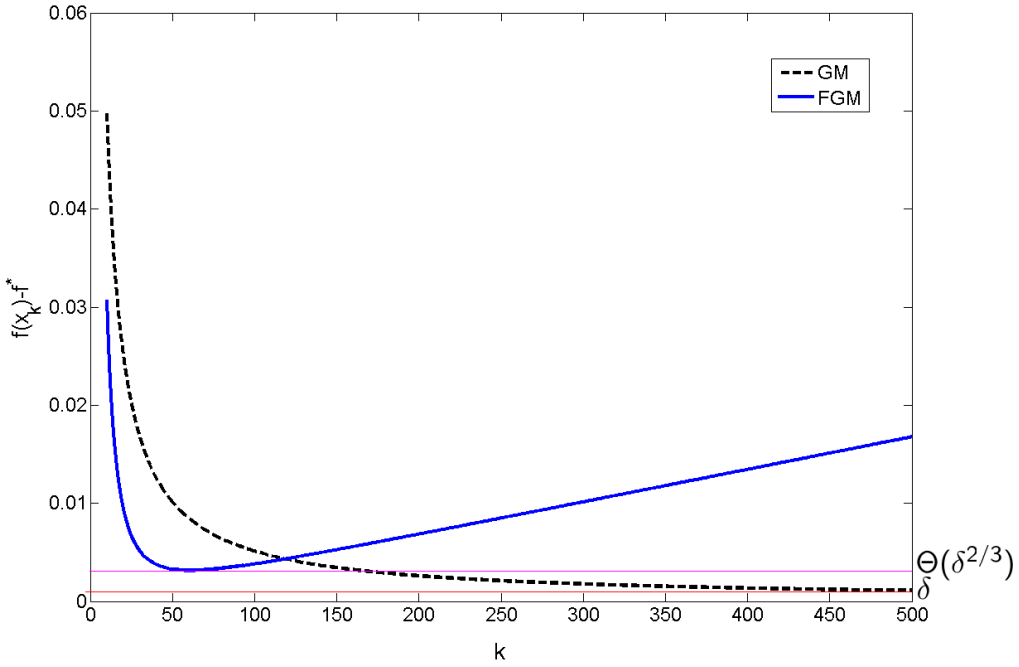
$$\text{FGM} : O\left(\sqrt{\frac{L(f)}{\epsilon}}R\right) \text{ iterations}$$

FGM outperforms GM in all cases.

Case 2: Inexact oracle with fixed OA δ

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

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



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 δ

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

Outline

Applications in Non-smooth Convex Optimization

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

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

This conditions implies:

$$f(x) \leq f(y) + \langle g(y), x - y \rangle + M(f) \|x - y\|, \quad \forall x, y \in Q.$$

But $M(f)t \leq \frac{M(f)^2}{4\delta} t^2 + \delta \quad \forall t \geq 0, \forall \delta > 0$ and therefore:

$$f(x) \leq f(y) + \langle g(y), x - y \rangle + \frac{M(f)^2}{4\delta} \|x - y\|^2 + \delta, \quad \forall x, y \in Q, \forall \delta > 0.$$

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 $L = \frac{M(f)^2}{2\delta}$.

Gradient Method for non-smooth problem

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. We have:

$$f(\hat{x}_k) - f^* \leq \frac{M(f)^2 R^2}{4k\delta} + \delta.$$

With a optimal choice of δ:

$$f(\hat{x}_k) - f^* \leq M(f)R\left(\frac{2}{k}\right)^{1/2}.$$

$\Rightarrow$Optimal rate of convergence $\Theta\left(\frac{M(f)R}{\sqrt{k}}\right)$ for the non-smooth problem (i.e. optimal complexity of $\Theta\left(\frac{M(f)^2R^2}{\epsilon^2}\right)$)..

Fast Gradient Method for non-smooth problem

2. We can apply FGM with inexact oracle to the non-smooth function f. We have:

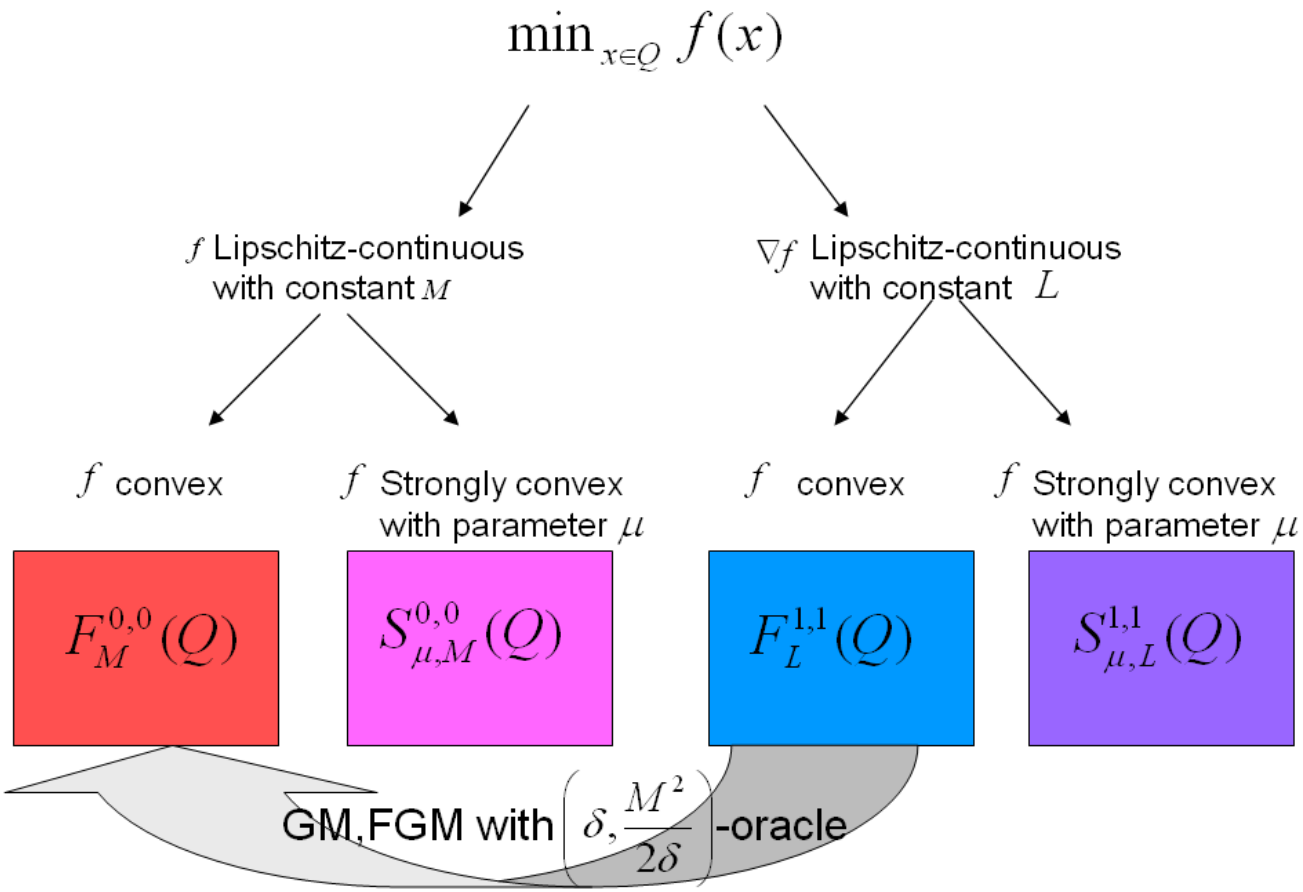
$$f(\hat{x}_k) - f^* \leq \frac{2M(f)^2 R^2}{(k+1)^2 \delta} + \delta(k+1).$$

With a optimal choice of δ:

$$f(\hat{x}_k) - f^* \leq 2M(f)R\left(\frac{2}{k+1}\right)^{1/2}.$$

$\Rightarrow$ Optimal rate of convergence $\Theta\left(\frac{M(f)R}{\sqrt{k}}\right)$ for the non-smooth problem (i.e. optimal complexity of $\Theta\left(\frac{M(f)^2R^2}{\epsilon^2}\right)$).

FOM of smooth convex optimization for non-smooth problems



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 generally, 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: $\Theta(k\delta)$.

Outline

What between Smooth and Non-smooth Problems ?

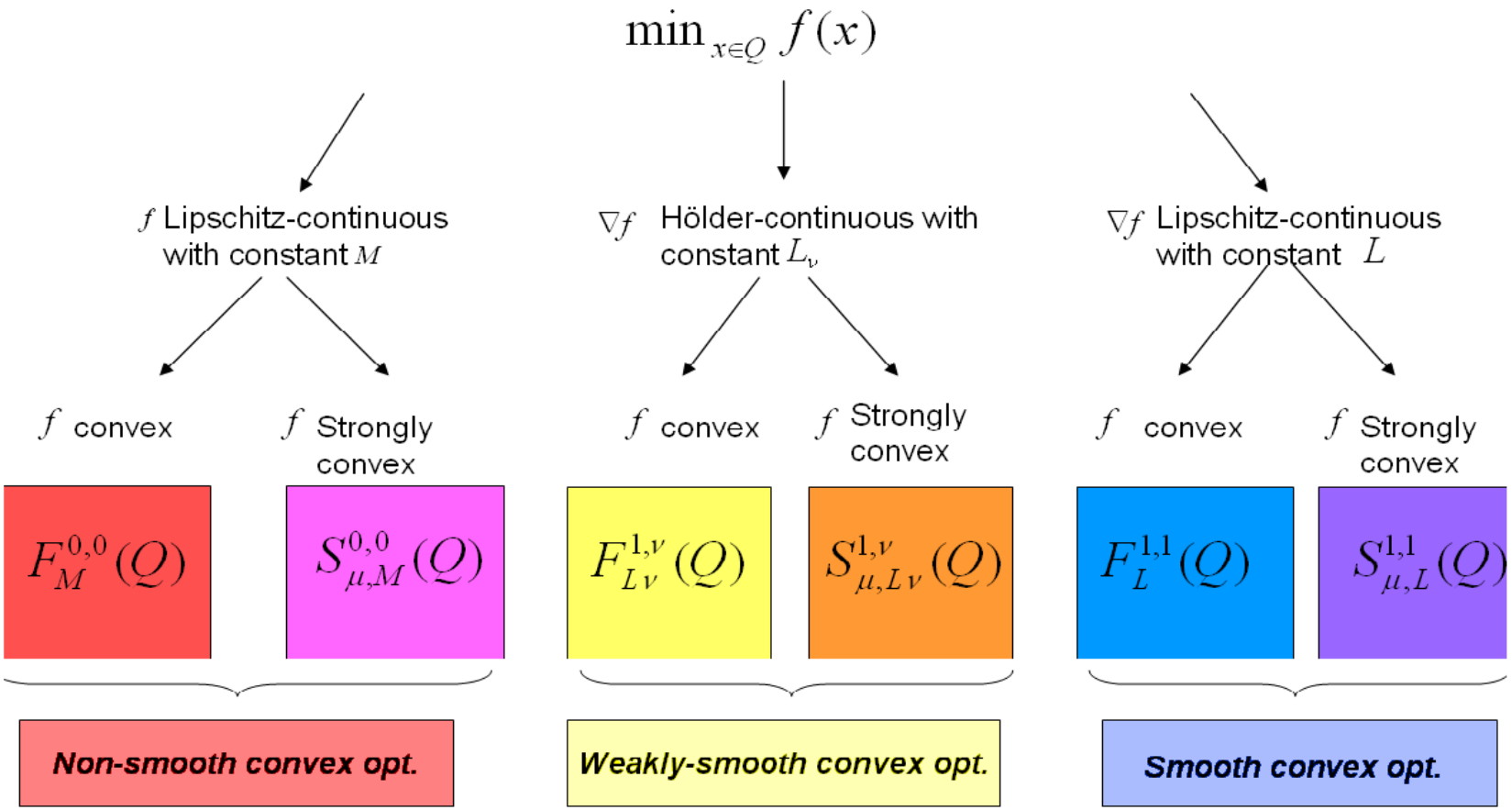
Assume that f satisfies the following smoothness condition:

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

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.

Intermediate case: Weakly-smooth Convex Optimization



A (δ, L)-oracle for weakly-smooth function

Assume that f has a Hölder-continuous gradient:

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

This conditions implies:

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

But $\frac{L_\nu}{1+\nu} t^{1+\nu} \leq \frac{A(\delta,\nu)}{2} t^2 + \delta \quad \forall t \geq 0, \forall \delta > 0$ where

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

A (δ, L)-oracle for weakly-smooth function (2)

We conclude that:

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

$\forall x, y \in Q, \forall \delta > 0$. The weakly-smooth exact oracle can be seen as a inexact (δ, L) smooth oracle:

$$f_{y,\delta} = f(y) \quad g_{y,\delta} = \nabla f(y)$$

where δ is arbitrary and

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

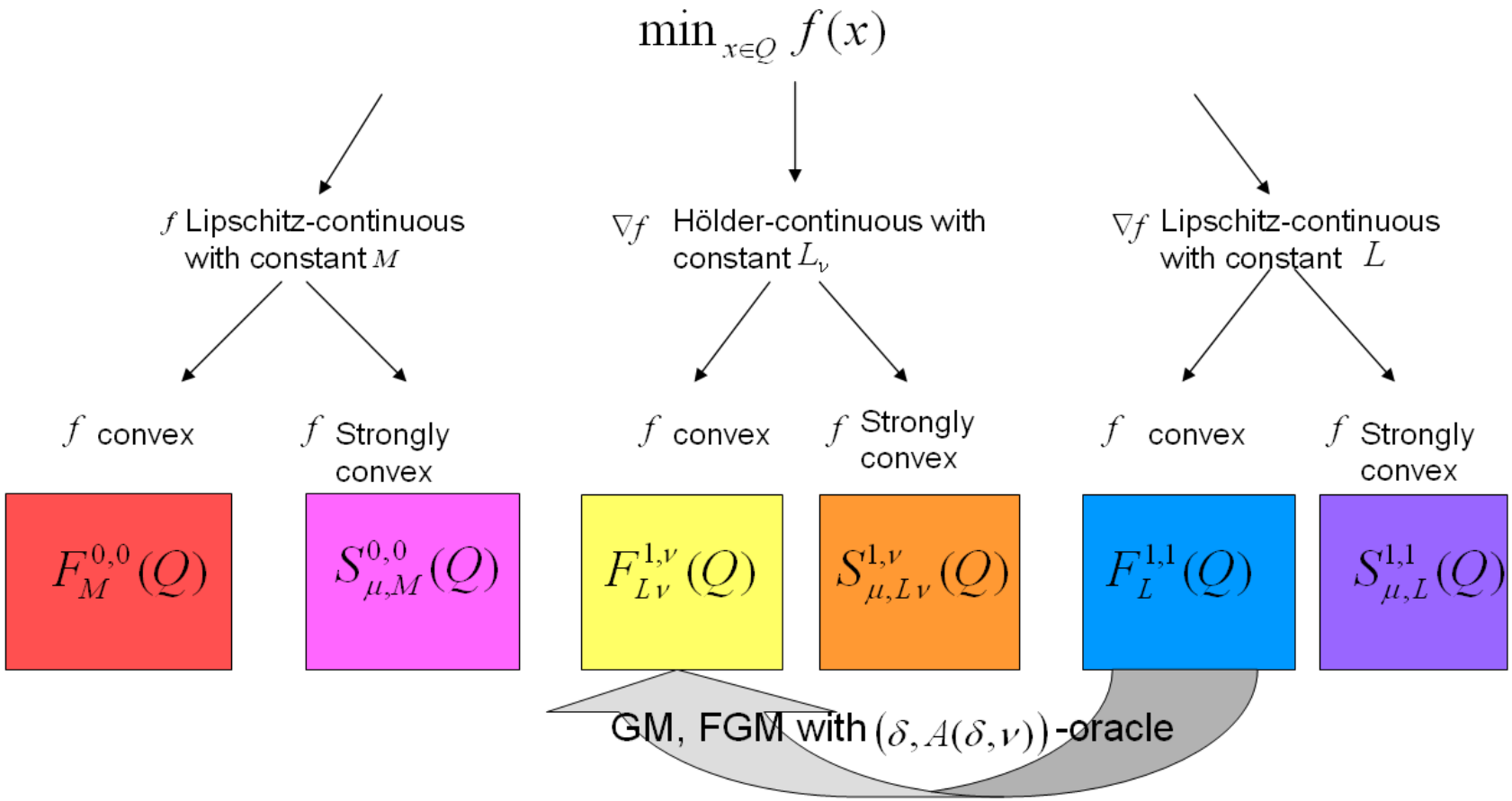
FOM for weakly-smooth problems

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

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

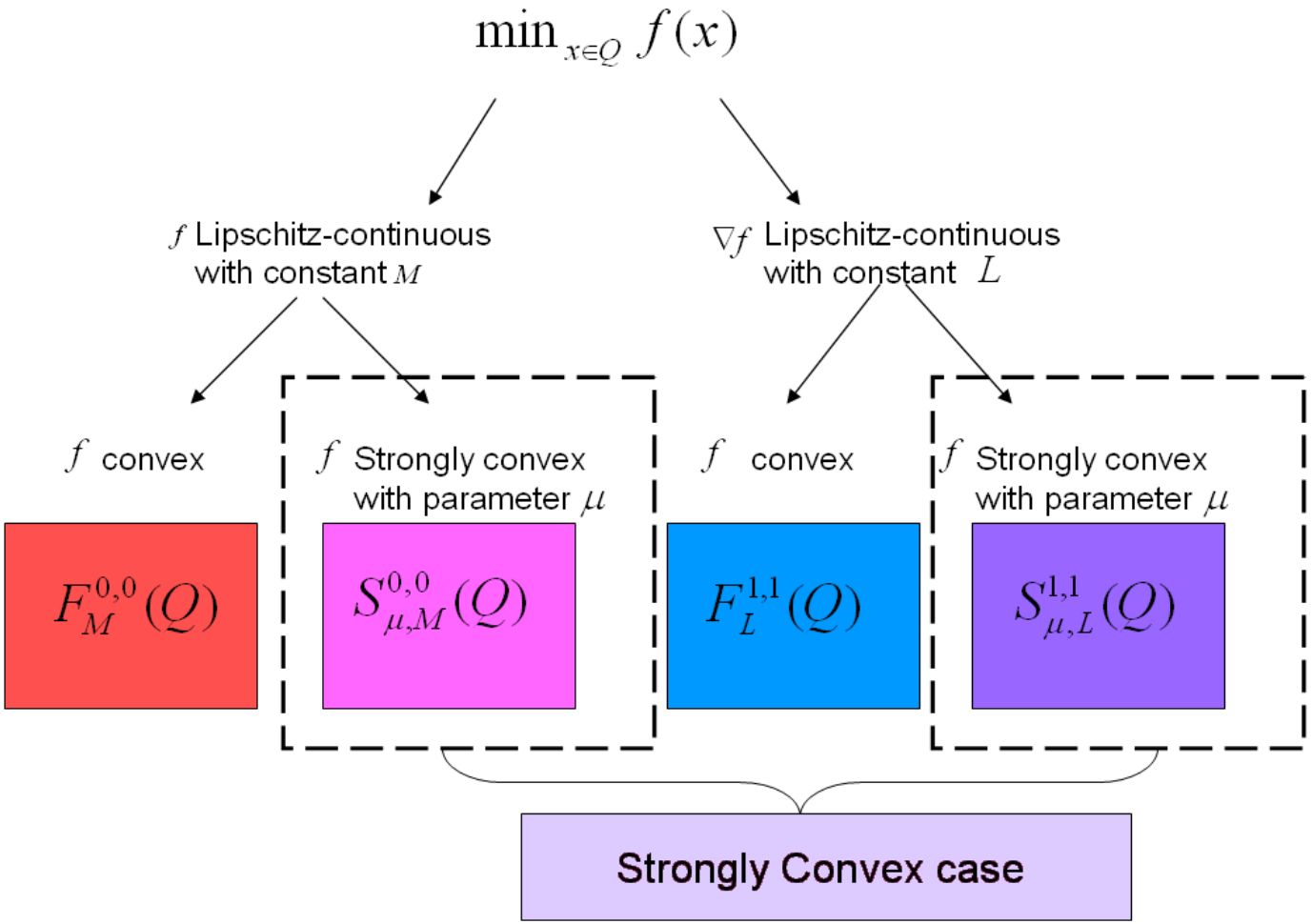
Observation: The FGM can reach optimal convergence rate for various classes of convex problems characterized by different levels of smoothness. $\Rightarrow$ **FGM= Universal Optimal FOM**.

FOM for weakly-smooth problems



Outline

What about the strongly convex case ?



What about strongly convex functions ?

For the moment, we have only considered FOM for convex functions, never assuming strong convexity.
Now, we assume that f is endowed with a (δ, L)-oracle but satisfies also:

$$f(y) \geq f^* + \frac{\mu}{2} \|y - x^*\|^2 \quad \forall y \in Q$$

where $\mu > 0$ and $x^* =$ unique optimal solution of $\min_{x \in Q} f(x)$.
Example: f is strongly convex with parameter $\mu > 0$:

$$f(y) \geq f(x) + \langle g(x), y - x \rangle + \frac{\mu}{2} \|x - y\|^2 \quad \forall x, y \in Q.$$

Goal: Study the effect of the inexact oracle on the first-order methods initially devoted for $S^{1,1}_{\mu,L}(Q)$ (Smooth and strongly convex functions).

GM for strongly convex function using inexact oracle

GM for $S^{1,1}_{\mu,L}(Q)$= GM for $F^{1,1}_{L}(Q)$ restarted each $N = \frac{2L}{\mu}$ iterations.

Inexact case:
If we want an accuracy of $\epsilon > 0$ for the objective function

- Needed number of iterations: $k = \Theta\left(\frac{L}{\mu} \ln\left(\frac{f(x_0)-f^*}{\epsilon}\right)\right)$
  Non-optimal
- Needed oracle accuracy: $\delta = \Theta(\epsilon)$
  No error accumulation

FGM for strongly convex function using inexact oracle

FGM for $S^{1,1}_{\mu,L}(Q)$= FGM for $F^{1,1}_{L}(Q)$ restarted each $N = 4\sqrt{\frac{L}{\mu}} - 1$ iterations.

Inexact case:
If we want an accuracy of $\epsilon > 0$ for the objective function

- Needed number of iterations: $k = \Theta\left(\sqrt{\frac{L}{\mu}} \ln\left(\frac{f(x_0)-f^*}{\epsilon}\right)\right)$
  Optimal
- Needed oracle accuracy: $\delta = \Theta(\sqrt{\frac{\mu}{L}}\epsilon)$
  For bad conditioned problem (i.e $L >> \mu$): $\delta << \epsilon$

Like in the convex case, a faster method need more accurate first-order information !

Application to non-smooth strongly convex function

Assume that f is strongly convex but non-smooth, having only bounded variation of subgradients:

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

As the exact non-smooth oracle = inexact $(\delta, \frac{M^2}{2\delta})$ smooth oracle,

- We can apply the GM to non-smooth strongly convex functions.
  Choosing $\delta = \frac{1}{4}\epsilon$ and restarting the GM each $N = \frac{4M^2}{\mu\epsilon}$ iterations , we can obtain a solution $\hat{x} \in Q$ such that $f(\hat{x}) - f^* \leq \epsilon$ in

$$O\left(\frac{M^2}{\mu\epsilon} \ln\left(\frac{f(x_0) - f^*}{\epsilon}\right)\right) \textit{iterations}.$$

$\Rightarrow$ Optimal (up to a logarithmic term)

Application to non-smooth strongly convex function (2)

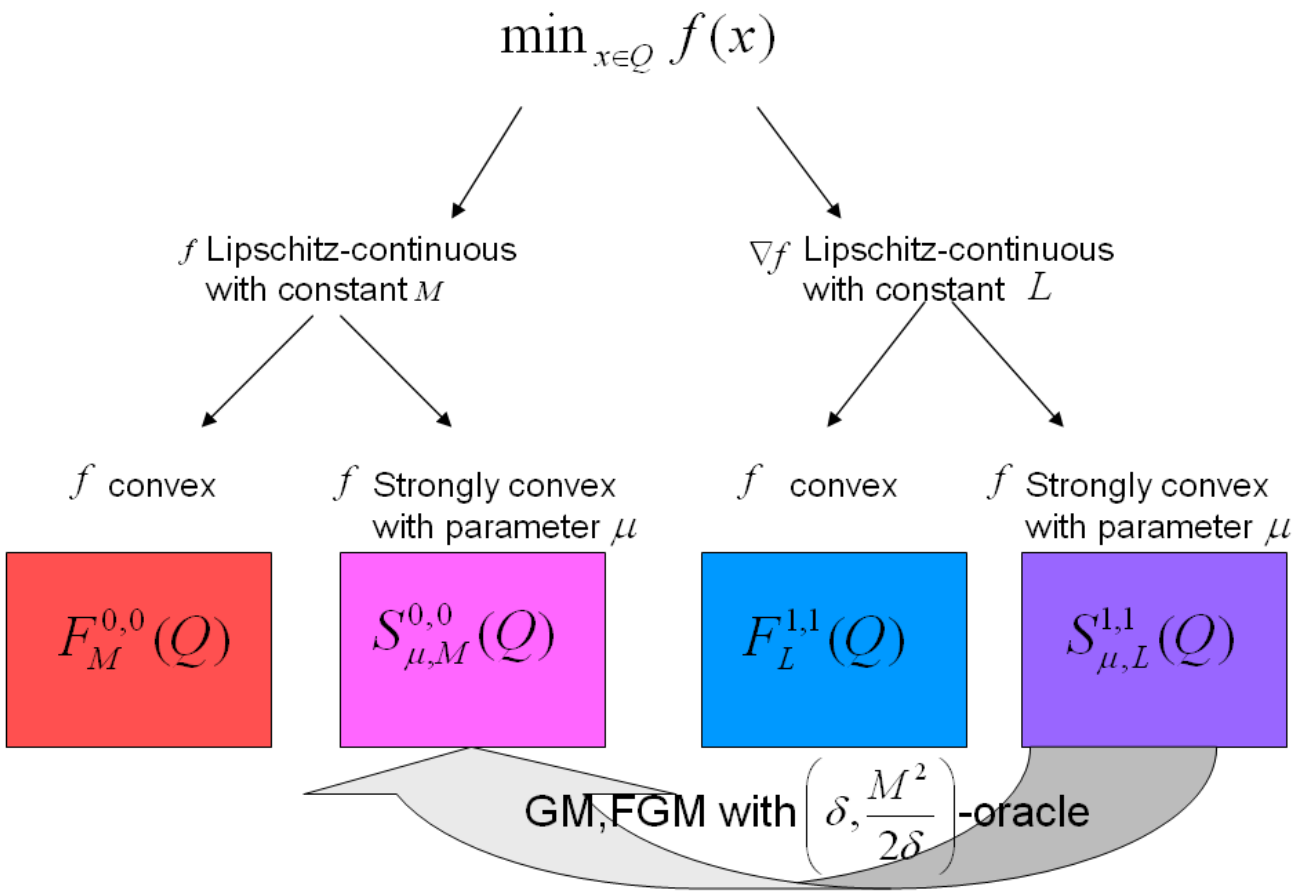
- We can apply the FGM to non-smooth strongly convex function.
  Choosing $\delta = \frac{9\mu}{128M^2}\epsilon^2$ and restarting the GM each $N = \frac{64M^2}{9\mu\epsilon} - 1$ iterations, we can obtain a solution $\hat{x} \in Q$ such that $f(\hat{x}) - f^* \leq \epsilon$ in

$$O\left(\frac{M^2}{\mu\epsilon}\ln\left(\frac{f(x_0) - f^*}{\epsilon}\right)\right) \textit{iterations}.$$

  $\Rightarrow$ Optimal (up to a logarithmic term)

Like in the convex case, we can apply FOM of smooth optimization to non-smooth problems and obtain an optimal method.

Application to non-smooth strongly convex function



Conclusion

- Introduction of a new definition of inexact oracle: (δ, L)-oracle.
- Important examples where the first-order information are 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.
- Same kind of results in the strongly convex case.

Thanks for your attention !

UCL

Université catholique de Louvain