First-order Methods for Smooth Convex Optimization with Inexact Oracle: Classical, Fast and Intermediate Gradient Methods

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

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

ISyE Seminar, Georgia Institute of Technology
Atlanta, November 1 2011

Outline

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:

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

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 that computes $f(x_k)$ and $\nabla f(x_k)$ for the numerical method at each search 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 $k \geq 0$

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

Remark: When $Q = \mathbb{R}^n$ and $\|.\| = \|.\|_2$: $x_{k+1} = x_k - \frac{1}{L(f)} \nabla f(x_k)$.

GM: Convergence rate

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

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

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

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 $k \geq 0$

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

FGM: Convergence rate

Choosing $\alpha_i = \frac{i+1}{2}$ for all $i \geq 0$,
convergence rate proportional to $\frac{1}{k^2}$:

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

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: to study the effect of inexact first-order information on GM and FGM.

Previous definitions of inexact oracle

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

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

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, $g_\eta(y)$ such that:
   - $\|\nabla f(y) - g_\eta(y)\| \leq \eta$
   - $\|g(y) - g_\eta(y)\| \leq \eta, g(y) \in \partial f(y)$
   - $|\langle \nabla f(y) - g_\eta(y), x - z \rangle| \leq \eta \quad \forall x, z \in Q$

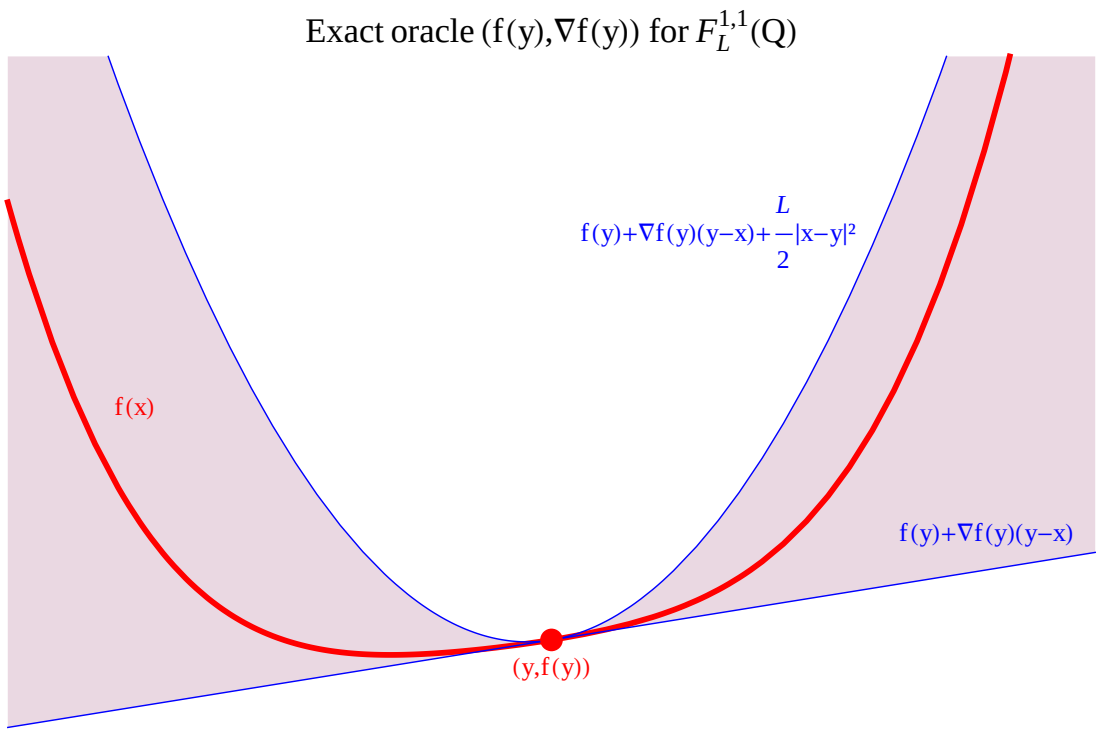
Good results can be obtained **but**
Strong conditions: Difficult to guarantee in practice.
Restrictive assumptions: Sometimes $\nabla f(y)$ must exist, sometimes Q must be bounded.

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:

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

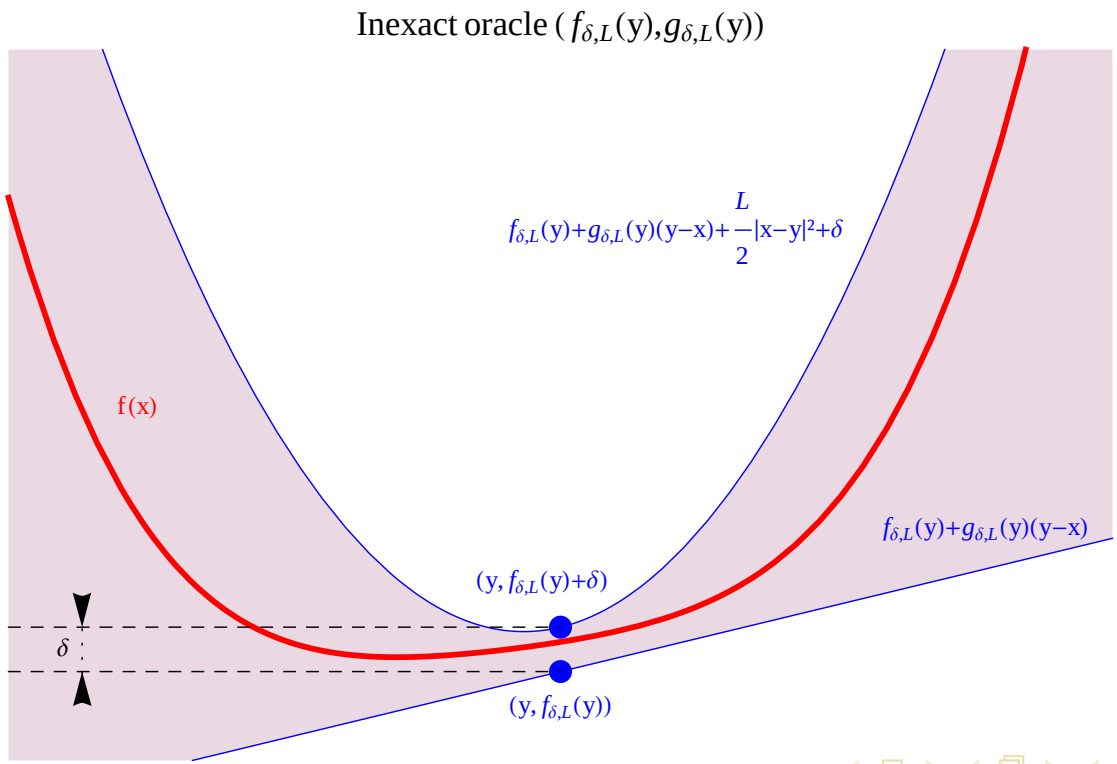
for all $x \in Q$.



(δ, L)-oracle

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

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



Some remarks about this definition

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 $L(f)$
- The existence of a (δ, L) oracle does not imply differentiability.
- $f_{\delta,L}(y)$ is a δ-lower approximation of $f(y)$:

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

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

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

Outline

Examples of inexact oracles

Two kinds 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 represents the accuracy of the first-order information.
   Main subject of this talk

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

Outline

1) Computation at shifted point

Assume that

1. $f \in F^{1,1}_{L(f)}(Q)$.
2. At each point $y \in Q$, the oracle provides exact value of f and ∇f but at a different point y_δ such that

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

then

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

is a (δ, L)-oracle with $L = 2L(f)$.

2) Approximate Gradient

Assume that:

1. $f \in F^{1,1}_{L(f)}(Q)$
2. Q is bounded with diameter $D = \max_{x \in Q, z \in Q} \|x - z\|$
3. $\left\|\nabla f(x) - \tilde{\nabla} f(x)\right\|_* \leq \Delta$

Then

$$f_{\delta,L}(x) = f(x) - \Delta D$$

$$g_{\delta,L}(x) = \tilde{\nabla} f(x)$$

is a (δ, L)-oracle with $\delta = 2\Delta D$ and $L = L(f)$.

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

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

3a) 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 $\nabla f(x) = Au_x$.

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

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

then

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

is a (δ, L)-oracle with $L = 2L(f)$.

3b) 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 $\nabla f(x) = \kappa(x - u_x)$.

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

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

then

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

is a (δ, L)-oracle with $L = L(f)$.

3c) 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 $L(f) = \frac{1}{\kappa}$
-

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

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

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

then

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

is a (δ, L)-oracle with $L = L(f)$.

Outline

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

Assume that f (convex) 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.

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

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

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

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

But that's another story: not the subject of this talk.
For more details: talk Glasgow July 2011, slides available on my webpage.

Outline

Effect of inexact oracle on FOM ?

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

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

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 an inexact oracle is used ?

Gradient Method with Inexact Oracle

Using averaging of the search points i.e. $y_k = \frac{1}{k}\sum_{i=1}^{k} x_i$, we obtain:

Exact oracle:

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

(δ, L)-oracle:

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

- 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:

$$f(y_k) - f^* \leq \frac{4LR^2}{(k+1)(k+2)} + \frac{1}{3}(k+3)\delta.$$

- 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:

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

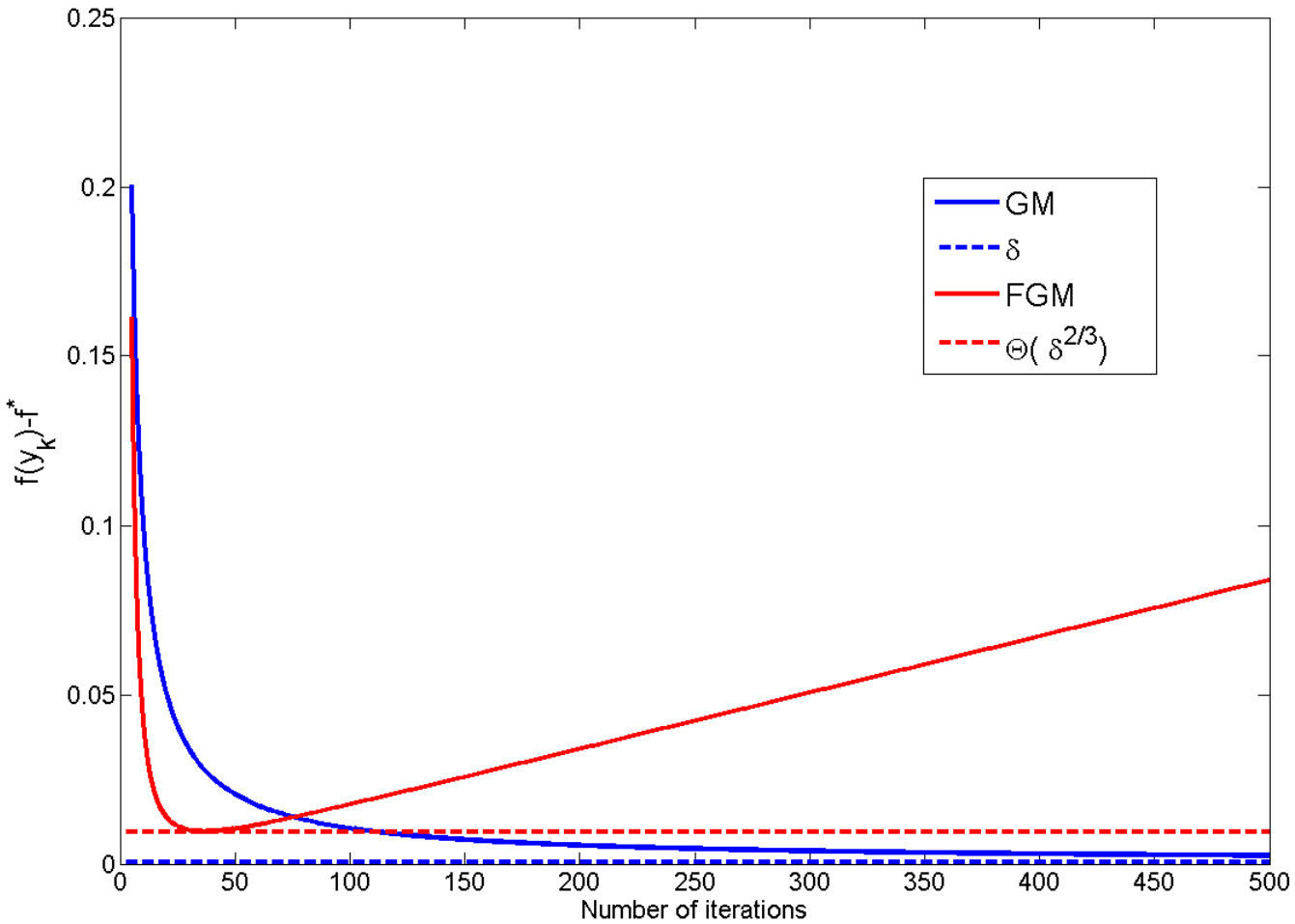
$$\text{FGM} : O\left(\sqrt{\frac{L(f)}{\epsilon}}R\right) \text{ iterations}$$

FGM outperforms GM in all cases.

Case 2: Inexact oracle with fixed OA δ

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

FGM : $f(y_k) - f^* \leq \frac{4LR^2}{(k+1)(k+2)} + \frac{1}{3}(k+3)\delta$



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 $(f_{\delta,L}(x), g_{\delta,L}(x))$.

- 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.

First-order methods with inexact oracle: Summary

Gradient method: Slow but Robust to errors

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

Non-optimal rate of convergence *but* No accumulation of errors.

Fast gradient method: Fast but Sensitive to errors

$$f(y_k) - f^* \leq \frac{4LR^2}{(k+1)(k+2)} + \frac{1}{3}(k+3)\delta.$$

Optimal rate of convergence *but* Accumulation of errors.

Two natural questions

1. In practice, does the FGM really suffer from an higher sensitivity to oracle errors ?

2. Is it possible to modify the FGM, keeping the optimal convergence rate and avoiding accumulation of errors ?

Two natural questions

1. **In practice, does the FGM really suffer from an higher sensitivity to oracle errors ? YES!**
2. Is it possible to modify the FGM, keeping the optimal convergence rate and avoiding accumulation of errors ?

Numerical Experiment 1

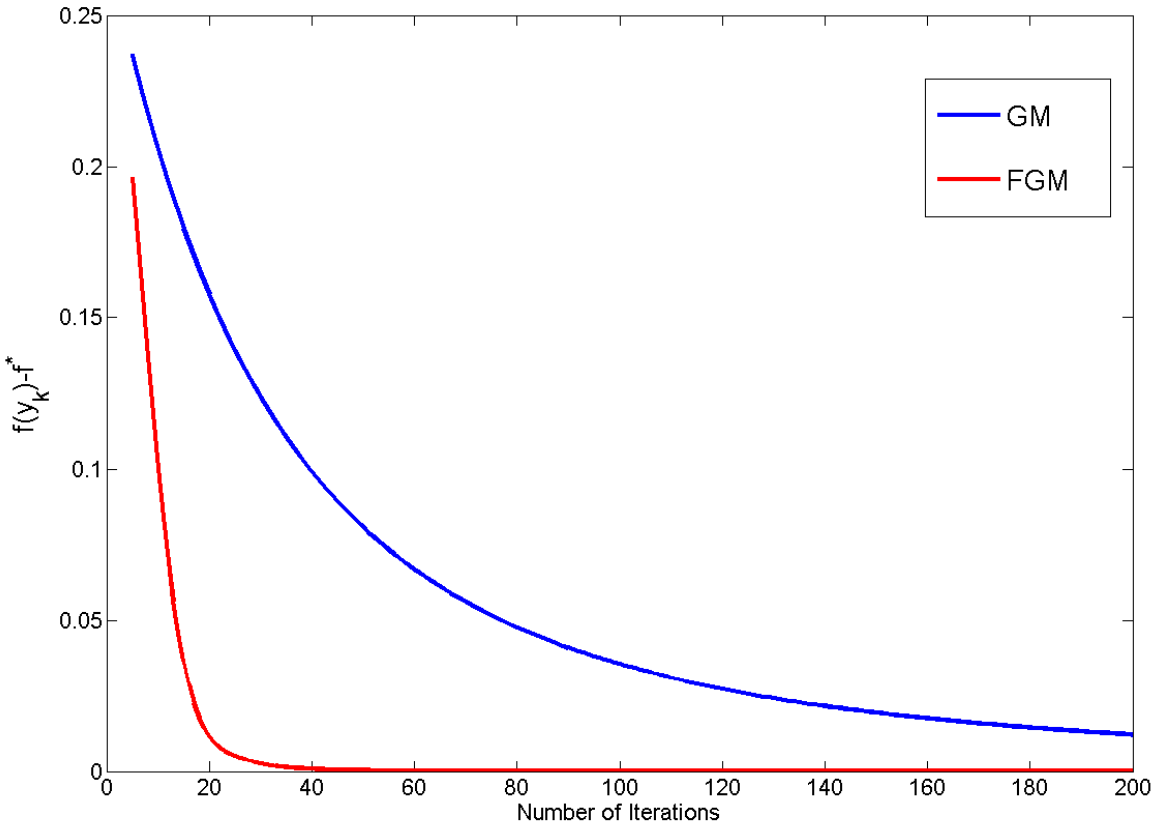
$$\min_{\|x\|_2 \leq 1} \frac{1}{2} x^T A x + \frac{1}{2} x^T B x$$

where:

1. $\|.\| = \|.\|_2$
2. $A \succeq 0, B \succeq 0$
3. $\|A\|_2 = 100 \|B\|_2 .$

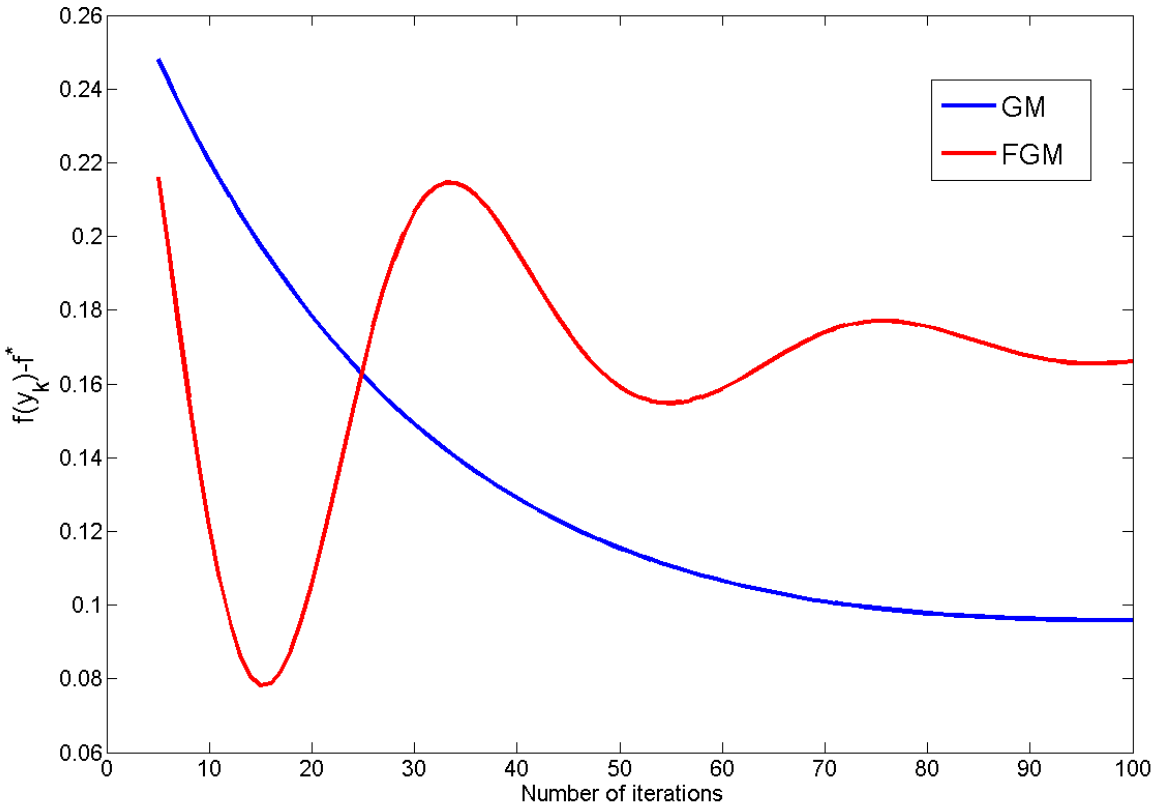
Exact gradient: $\nabla f(x) = Ax + Bx$
Inexact Gradient: $\tilde{\nabla} f(x) = Ax - Bx.$
$\rightarrow (\delta, L)$ oracle with $\delta = 2 \|B\|_2$ and $L = \|A + B\|_2 .$

Numerical Experiment 1: Exact Case



In the exact case: FGM significantly faster than GM.

Numerical Experiment 1: Inexact Case



In the inexact case: FGM faster at first but suffers from accumulation of errors !

Two natural questions

1. In practice, does the FGM really suffer from an higher sensitivity to oracle errors ? YES!

2. **Is it possible to modify the FGM, keeping the optimal convergence rate and avoiding accumulation of errors ? NO !**

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: $\Theta(k\delta)$.

It is a bad news but...

The previous theorem is not a good news: there is no hope to develop a first-order method which is at the same time as Fast as the FGM and as Robust as the GM.

Faster the method is, higher the sensitivity to errors is.
There is no free lunch !

but...

between the two extreme choices of

1. the robust but slow GM
2. the fast but sensitive FGM

it could be preferable to use methods with

- intermediate speed
- intermediate sensitivity to errors.

(Belgian Compromise!)

Between GM and FGM ? Intermediate Gradient Methods

Outline

Developement of Intermediate Gradient Methods (IGM)

Our goal: we want to develop first-order methods with intermediate rate of convergence $\Theta(\frac{1}{k^p})$ $(1 < p < 2)$ and corresponding optimal rate of error accumulation $\Theta(k^{p-1}\delta)$.
We will obtain a whole family of FOM interpolating between GM and FGM.

The Approach: Modify the FGM such that we slow down the rate of error accumulation and, unavoidably, also the rate of convergence.

First Try: Modification of the weights α_i

A natural idea is to modify the sequence of weights α_i in the FGM (keeping however the condition $\alpha_k^2 \leq A_k = \sum_{i=0}^{k} \alpha_i$).

Convergence rate with an exact oracle:

$$f(y_k) - f^* \leq \frac{LR^2}{A_k}.$$

$\Rightarrow$ We choose α_k such that $A_k = \Theta(k^p)$.

Convergence rate with an inexact oracle:

$$f(y_k) - f^* \leq \frac{LR^2}{A_k} + \frac{\sum_{i=0}^{k} A_i}{A_k}\delta$$

$\Rightarrow$ error accumulation of order $\frac{\sum_{i=0}^{k} A_i}{A_k}\delta = \Theta(k\delta)$.

Conclusion: We slow down the method without reducing the rate of error accumulation. Bad Approach ! We need to do more !

Second Try: A new degree of freedom

Idea:
Introduce a new sequence B_k, and therefore a new degree of freedom in the method in order to obtain a convergence rate of the form:

$$f(y_k) - f^* \leq \frac{LR^2}{A_k} + \left(\frac{\sum_{i=0}^{k} B_i}{A_k} \right) \delta$$

with

- $A_k = \Theta(k^p)$ i.e. a rate of convergence of order $\Theta(\frac{1}{k^p})$
- $\frac{\sum_{i=0}^{k} B_i}{A_k} = \Theta(k^{p-1})$ i.e. a rate of error accumulation of order $\Theta(k^{p-1}\delta)$

Intermediate Gradient Method (IGM)

Let $\{\alpha_k\}_{k=0}^{\infty}$ and $\{B_k\}_{k=0}^{\infty}$ satisfying $\alpha_0 = B_0 = 1, \quad \alpha_k^2 \leq B_k$ and $B_k \leq \sum_{i=0}^{k} \alpha_i$

Define $A_k = \sum_{i=0}^{k} \alpha_i$.

Initialization

Choose $x_0 \in Q$

Iteration $k \geq 0$

- $(f_{\delta,L}(x_k), g_{\delta,L}(x_k)) = \mathcal{O}_{\delta,L}(x_k)$
- $w_k = \arg\min_{x \in Q}\{f_{\delta,L}(x_k) + \langle g_{\delta,L}(x_k), y - x_k\rangle + \frac{L}{2}\|y - x_k\|_2^2\}$
- $z_k =$ $\arg \min_{x \in Q}\{\sum_{i=0}^{k} \alpha_i[f_{\delta,L}(x_i) + \langle g_{\delta,L}(x_i), x - x_i\rangle] + \frac{L}{2}\|x - x_0\|_2^2\}$
- $y_k = \frac{A_k - B_k}{A_k} y_{k-1} + \frac{B_k}{A_k} w_k$
- $\tau_k = \frac{\alpha_{k+1}}{B_{k+1}}$
- $x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k$

Intermediate Gradient Method (IGM)

When $B_k = A_k$, we retrieve the FGM. But we have a new degree of freedom, we can choose B_k smaller than A_k.
In fact, we replace $y_k = w_k$ by the more conservative rule:

$$y_k = \frac{A_k - B_k}{A_k} y_{k-1} + \frac{B_k}{A_k} w_k.$$

Two consequences:

1. We slow down the rate of error accumulation : $\frac{\sum_{i=0}^{k} B_i}{A_k} \leq \frac{\sum_{i=0}^{k} A_i}{A_k}$
2. We slow down the rate of convergence (unavoidable) due to the condition $\alpha_k^2 \leq B_k$ (instead of $\alpha_k^2 \leq A_k$).

Choice of the sequences α_k and B_k

Choice of B_k:
Assume $A_k = \Theta(k^p)$ and $B_k = A_k^\beta$.
Then the condition $\frac{\sum_{i=0}^{k} B_i}{A_k} = \Theta(k^{p-1})$ gives us $\beta = \frac{2p-2}{p}$ and therefore

$$B_k = A_k^{\frac{2p-2}{p}}.$$

Choice of α_k:
Consider the choice $\alpha_k = Ck^{p-1}$.
Then the condition $\alpha_k^2 \leq B_k$ gives us $C = \frac{1}{p^{p-1}}$ and therefore

$$\alpha_k = \left(\frac{k}{p}\right)^{p-1}.$$

Convergence rate of the IGM

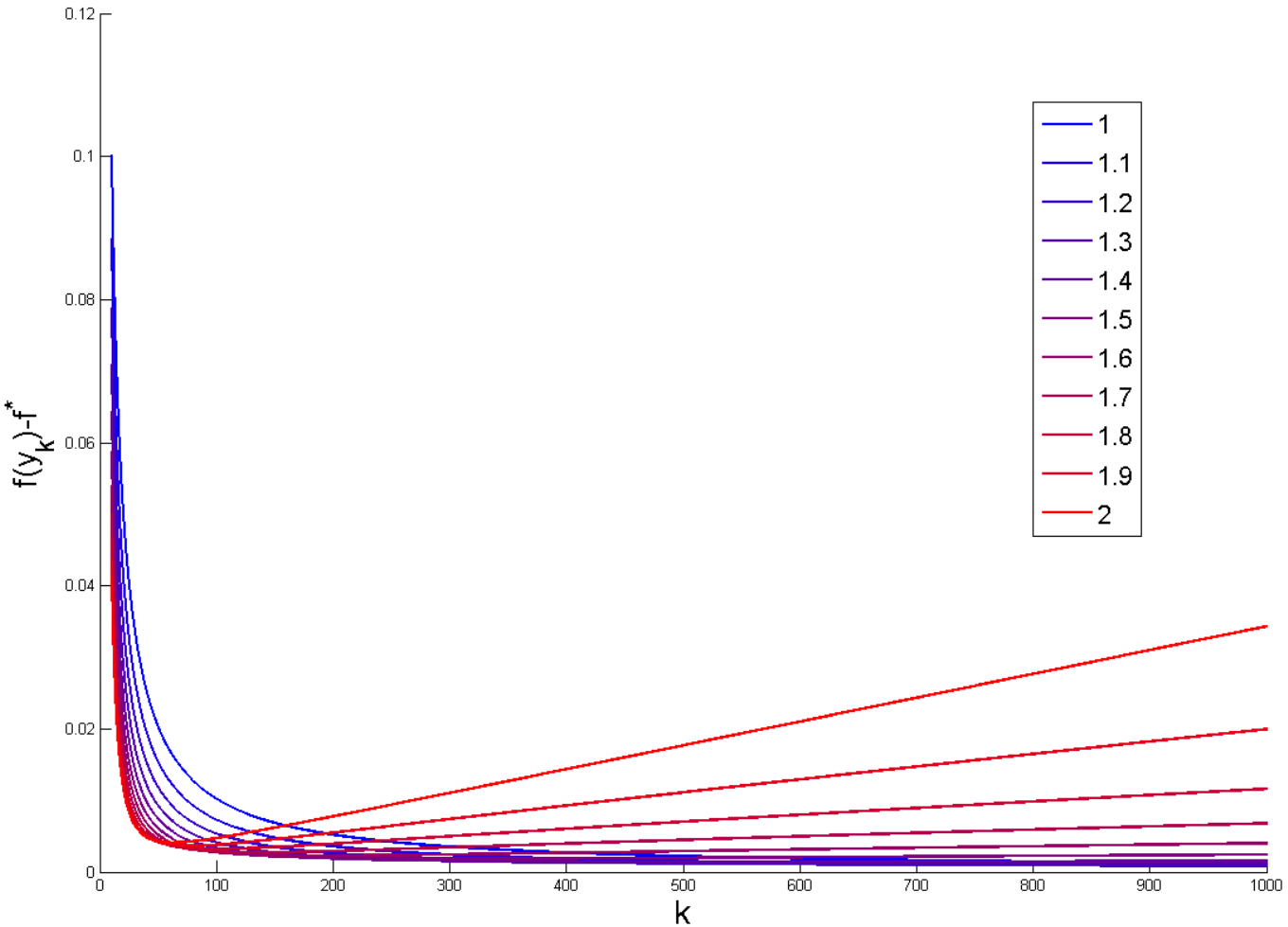
The sequence $\{y_k\}_{k\geq 1}$ generated by the IGM with parameter $1 \leq p \leq 2$ satisfies:

$$\begin{aligned} f(y_k) - f^* &\leq \frac{LR^2}{A_k} + \frac{\sum_{i=0}^{k} B_i}{A_k}\delta \\ &\leq \frac{C_1 LR^2 + C_2\delta}{k^p} + C_3\delta + C_4 k^{p-1}\delta \\ &:= Acc(k, p, \delta). \end{aligned}$$

Conclusion: We have developed a whole family of FOM with intermediate rates of convergence $\Theta\left(\frac{1}{k^p}\right)$ between $\Theta\left(\frac{1}{k}\right)$ (GM) and $\Theta\left(\frac{1}{k^2}\right)$ (FGM) and with intermediate (and optimal !) rates of error accumulation $\Theta(k^{p-1}\delta)$.

Convergence rate of the IGM (cont.)

Convergence rates of the IGM family when $\delta = 1e-4$:



IGM as an interpolation between DGM and FGM

We can consider what we obtain in the two extreme cases:

1. $p = 1$
   We have $\alpha_k = B_k = \tau_k = 1$ for all $k \geq 0$.
   Therefore $y_k = \frac{1}{k} \sum_{i=0}^{k} w_i$ and $x_{k+1} = z_k$.
   $\Rightarrow$ We retrieve the Dual Gradient Method (DGM) [Nes07].
2. $p = 2$
   We have $A_k = B_k$ for all $k \geq 0$.
   Therefore $y_k = w_k$ and $x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k$.
   $\Rightarrow$ We retrieve the Fast Gradient Method (FGM) [Nes05].

Conclusion: The family of IGM can be seen as an interpolation between DGM and FGM.

Which method should we choose ? δ and k fixed

Optimal method:

$$\min_{1\leq p\leq 2} Acc(k,p,\delta).$$

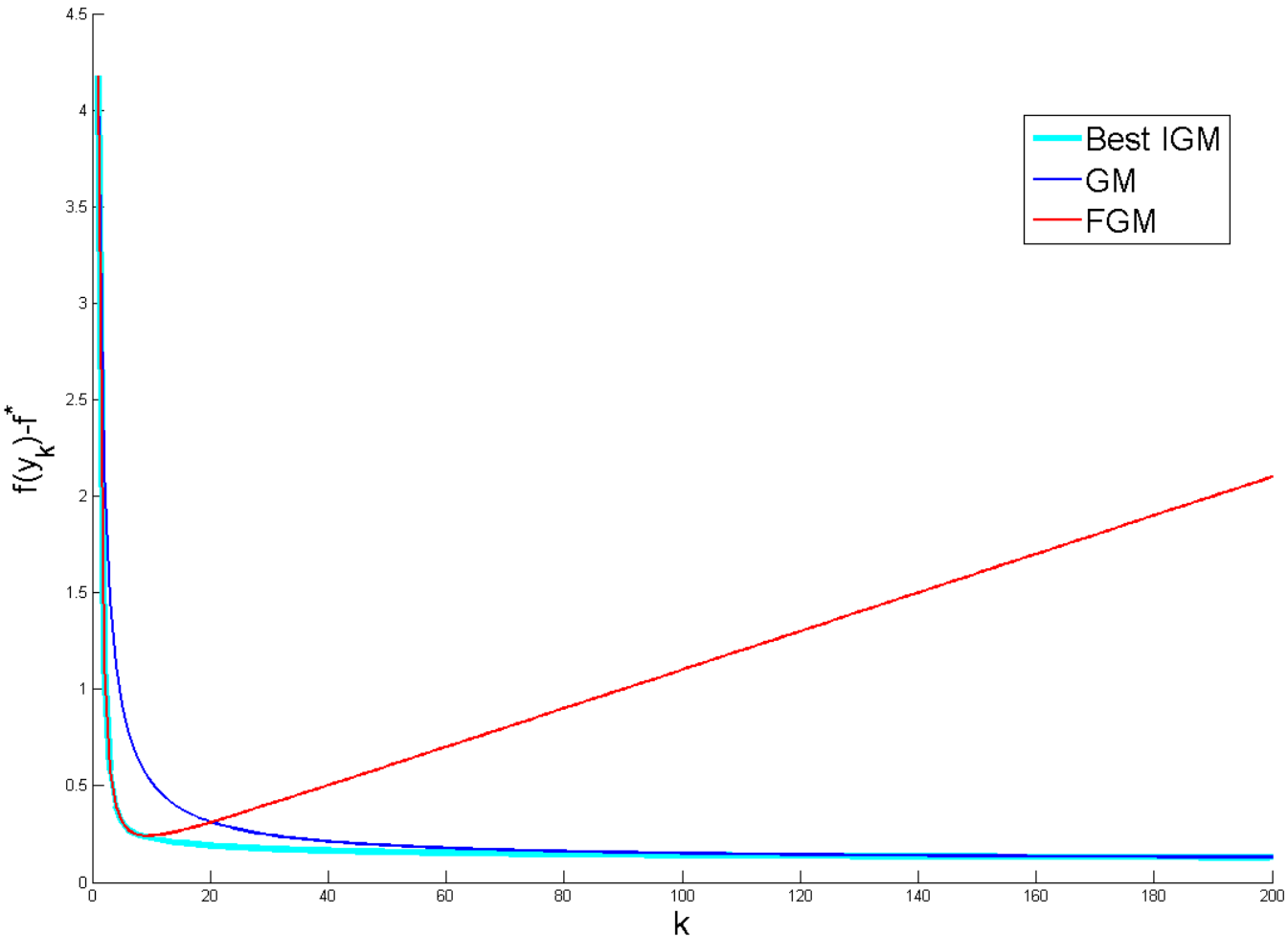
Let $k_1 = \sqrt[3]{\frac{C_1LR^2+C_2\delta}{C_4\delta}}$ and $k_2 = \frac{C_1LR^2+C_2\delta}{C_4\delta}$.

Three different situations:

1. If $0 \leq k \leq k_1$:
   - $p = 2$ (FGM)
   - $BestAcc(k) = \frac{C_1LR^2+C_2\delta}{k^2} + C_3\delta + C_4k\delta.$
2. If $k_1 \leq k \leq k_2$:
   - $p = \frac{1}{2}\left[\frac{\ln\left(\frac{C_1LR^2+C_2\delta}{C_4\delta}\right)}{\ln(k)} + 1\right]$ (IGM)
   - $BestAcc(k) = \frac{2\sqrt{C_1LR^2+C_2\delta}\sqrt{C_4\delta}}{\sqrt{k}} + C_3\delta$
3. If $k \geq k_2$
   - $p = 1$ (GM)
   - $BestAcc(k) = \frac{C_1LR^2+C_2\delta}{k} + (C_3+C_4)\delta$

An improved accuracy using IGM

The function $BestAcc(k)$ is continuous, decreasing and always below the convergence rates of GM and FGM (with $\delta = 1e-2$):



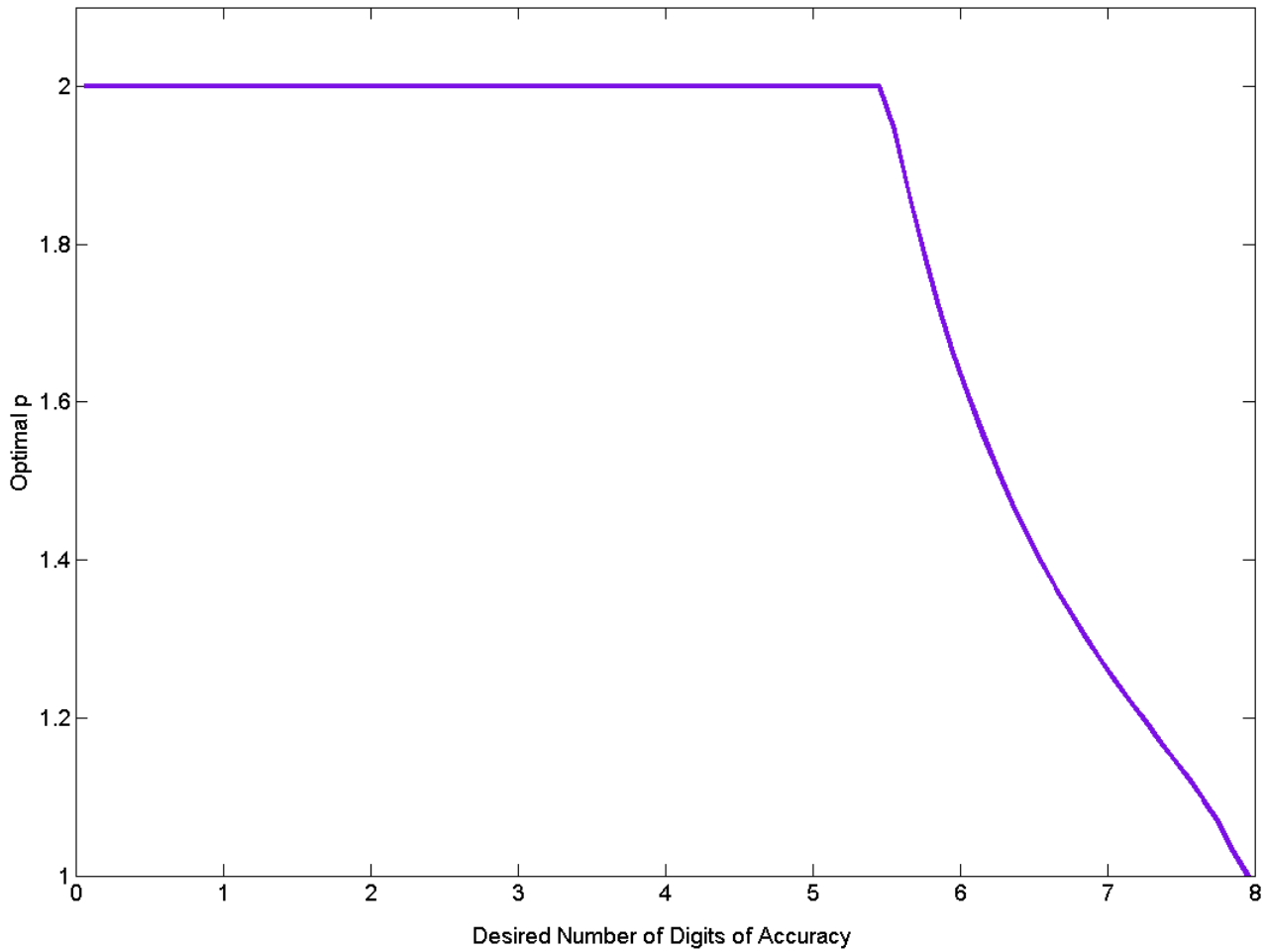
Which method should we choose ? δ and ϵ fixed

Optimal method: $\min_{k\geq 0, 1\leq p\leq 2} k$ s. t. $Acc(k,\delta,p)\leq\epsilon$
Let $\epsilon_1 = 2C_4\delta + C_3\delta$ and $\epsilon_2 = 2(C_1LR^2 + C_2\delta)^{1/3}(C_4\delta)^{2/3} + C_3\delta$.
Three different situations:

1. When $\epsilon \geq \epsilon_2$
   - p=2 (FGM)
   - k= unique root of $P(k) = (C_4\delta)k^3 + (C_3\delta - \epsilon)k^2 + C_1LR^2 + C_2\delta$ on $]0, k_1]$.
2. When $\epsilon_1 \leq \epsilon \leq \epsilon_2$
   - $p = \frac{1}{2}\left[\frac{ln\left(\frac{C_1LR^2+C_2\delta}{C_4\delta}\right)}{\ln\left(\frac{4(C_1LR^2+C_2\delta)C_4\delta}{(\epsilon-C_3\delta)^2}\right)} + 1\right]$ (IGM)
   - $k = \frac{4(C_1LR^2+C_2\delta)C_4\delta}{\epsilon-C_3\delta}$
3. When $C_4\delta + C_3\delta \leq \epsilon \leq \epsilon_1$
   - $p = 1$ (GM)
   - $k = \frac{C_1LR^2+C_2\delta}{\epsilon-(C_3+C_4)\delta}$.

Optimal p depending on the desired accuracy

When $\delta = 1e - 8$, optimal p depending on the desired number of digits of accuracy:



Numerical Experiment 2

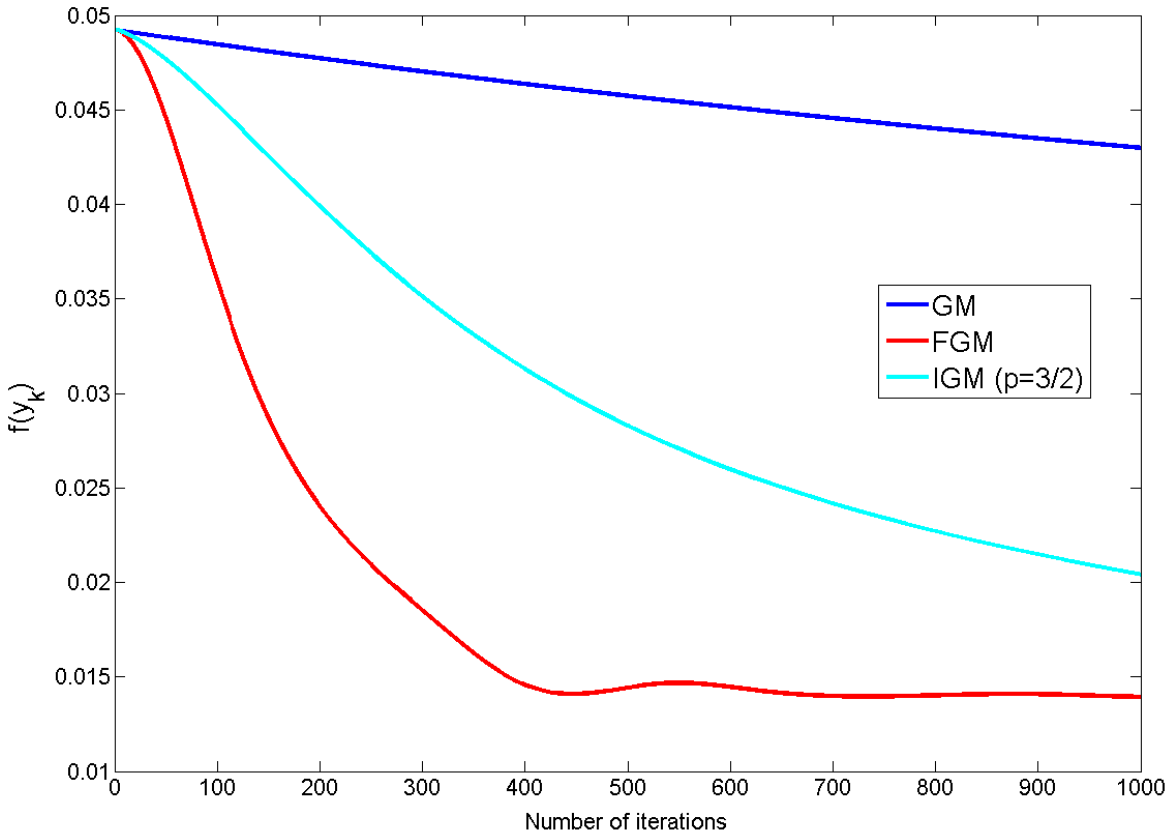
$$\min_{x \in \Delta_n} \frac{1}{2} x^T A x$$

where:

1. $A \succeq 0$
2. $\Delta_n = \{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x^i = 1\}$
3. $\|.\| = \|.\|_1$.

Exact gradient: $\nabla f(x) = Ax$
Inexact gradient: $\tilde{\nabla} f(x) = Ax + \xi$
$\rightarrow (\delta, L)$ oracle with $\delta = 2\|\xi\|_\infty$ and $L = \|A\|_\infty$.

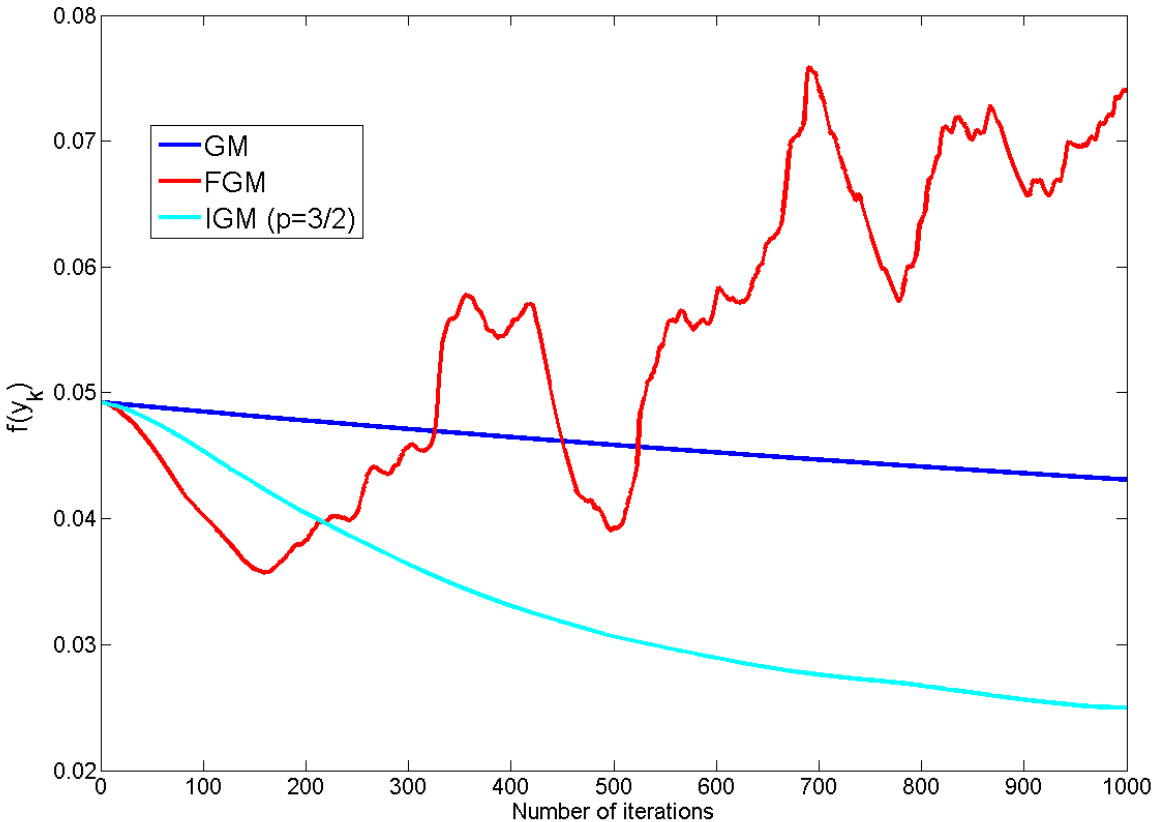
Numerical Experiment 2: Exact Case



FGM significantly faster than GM (which is very slow !)
Intermediate speed for the IGM.

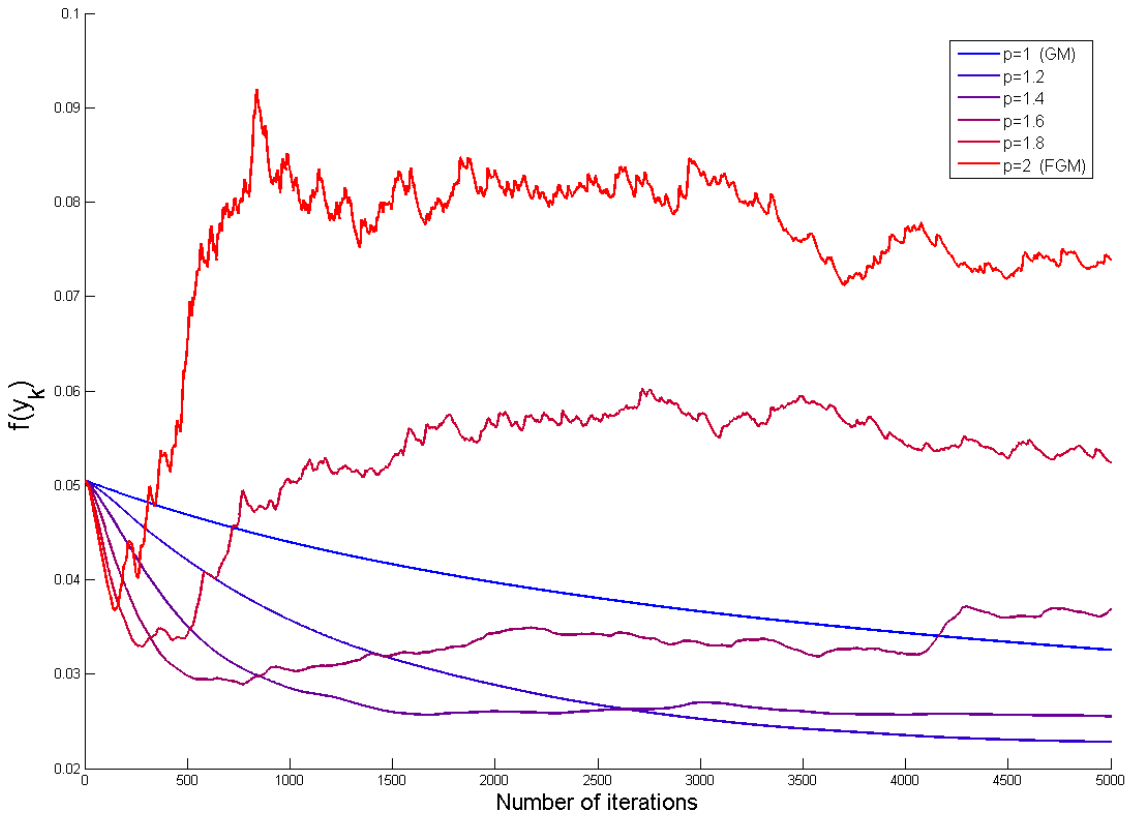
Numerical Experiment 2: Inexact Case

When $\|\xi\|_\infty = 1$ and $\|A\|_\infty = 100$:



GM robust but very slow. FGM highly sensitive to errors.
Method of choice: IGM !

Numerical Experiment 2: Choice of p



The smaller p is, the slower the method is, but the better the reachable accuracy is (confirmation of the theory!).

Numerical Experiment 2: Choice of p

Num. Iter.	p=1	p=1.2	p=1.4	p=1.6	p=1.8	p=2
10	0.0505	0.0504	0.0504	0.0503	0.0503	**0.0503**
50	0.0502	0.0498	0.0493	0.0486	0.0475	**0.0468**
100	0.0498	0.0489	0.0476	0.0454	0.0429	**0.0407**
500	0.0469	0.0421	0.0351	**0.0298**	0.0408	0.0665
1000	0.0440	0.0358	**0.0285**	0.0304	0.0506	0.0824
5000	0.0326	**0.0230**	0.0255	0.0368	0.0541	0.0702
10 000	0.0274	**0.0226**	0.0255	0.0347	0.0459	0.0827

Choice of the method depends on the accuracy needed:

- For obtaining quickly a not so accurate solution: FGM ($p = 2$)
- For obtaining highly accurate solution: GM ($p = 1$)
- For intermediate goals (More Realistic): Use IGM with well-chosen p.

$\Rightarrow$ **The IGM's can effectively accelerate the minimization in the presence of errors.**

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 approximate solution of subproblems
- The GM is slow but robust with respect to oracle error.
- The FGM is faster but sensitive to oracle error. Like any FOM with optimal convergence rate, it suffers from accumulation of errors.
- Developement of new first-order methods with intermediate behavior $\Rightarrow$ Notion of Intermediate Gradient Methods (IGM).
- Choice of the method ? Depend on the needed accuracy ϵ (its relation with the oracle accuracy δ) :
  1. When ϵ is small (close to δ): use GM.
  2. When ϵ is not small at all: use the FGM.
  3. For intermediate accuracy, best choice : use a well-chosen IGM.

Thanks for your attention !

Slides available on my webpage:
http://perso.uclouvain.be/olivier.devolder

Papers:

1. 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.
2. O. Devolder, F. Glineur and Y. Nesterov
   Between Gradient and Fast Gradient Methods: a family of Intermediate First-order Methods.
   In preparation.