Between Gradient and Fast Gradient Methods: a Family of Intermediate First-Order Methods.

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)

OR 2011 International Conference on Operations Research,
Zurich, September 1

Outline

1. First-order methods in smooth convex optimization
2. First-order methods with inexact oracle
3. Intermediate Gradient Methods (IGM)
4. Which method should we choose ?

Outline

1. First-order methods in smooth convex optimization
2. First-order methods with inexact oracle
3. Intermediate Gradient Methods (IGM)
4. Which method should we choose ?

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

In Smooth Convex Optimization, two main FOM:

1. Gradient method (GM)
2. Fast gradient method (FGM)

Gradient Method (GM)

Very simple algorithm:

Initialization

Choose $x_0 \in Q$

Iteration $k \geq 0$

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

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

Fast Gradient Method (FGM)

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 $k \geq 0$

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

If we choose $\alpha_i = \frac{i+1}{2}$:
Convergence rate in $O(\frac{1}{k^2})$ $\Rightarrow$ **Optimal FOM for** $F_{L(f)}^{1,1}(Q)$

Outline

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

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

for all $x \in Q$.

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.
   Examples: Computation at shifted point, saddle-point function with inexact resolution of subproblems...

2. **Lack of smoothness for the function**
   Function with weaker level of smoothness (but typically with exact first-order information).
   Examples: Non-smooth function, Weakly-smooth function...

First-order methods with inexact oracle

Gradient method:

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

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

Fast gradient method:

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

Optimal rate of convergence *but* Accumulation of errors.

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: $\Theta(k\delta)$.

Between GM and FGM ? Intermediate FOM

Outline

1. First-order methods in smooth convex optimization
2. First-order methods with inexact oracle
3. Intermediate Gradient Methods (IGM)
4. Which method should we choose ?

Developement of Intermediate Gradient Methods (IGM)

Our goal: In this work, 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 $\alpha_0 = B_0 = 1, \quad \alpha_k^2 \le B_k$ and $B_k \le \sum_{i=0}^{k} \alpha_i$

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

Initialization

Choose $x_0 \in Q$

Iteration $k \ge 0$

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

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

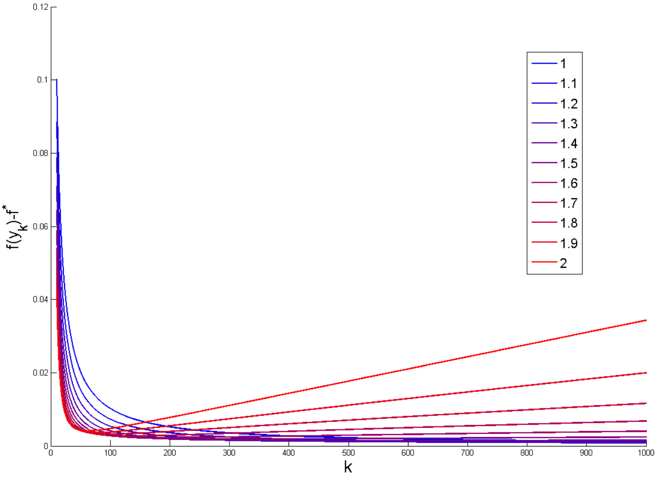
The sequence $\{y_k\}_{k\geq 1}$ generated by the IGM with parameter $1 \leq p \leq 2$ satisfies:

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

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.

Outline

1. First-order methods in smooth convex optimization
2. First-order methods with inexact oracle
3. Intermediate Gradient Methods (IGM)
4. Which method should we choose ?

Which method should we choose ? δ and k fixed

Optimal method:

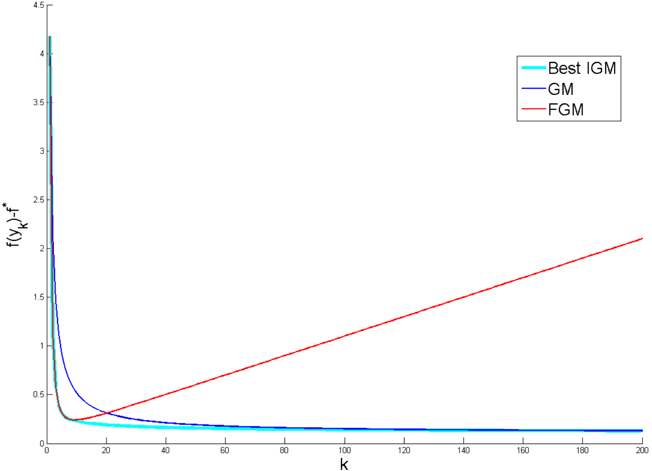
$$\min_{1\leq p\leq 2} C(k,p,\delta).$$

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 $\delta = 1e-2$):



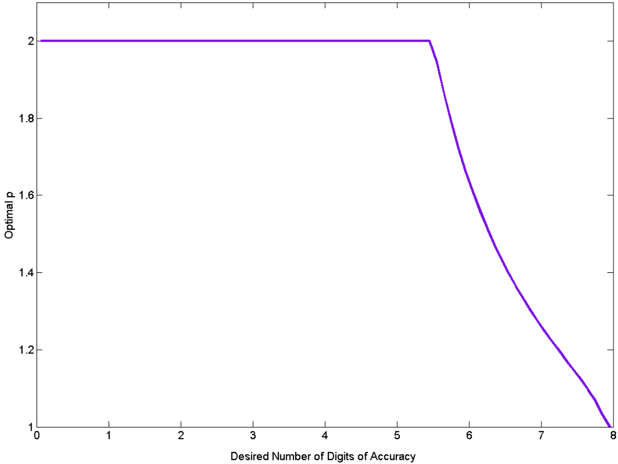
Which method should we choose ? δ and ϵ fixed

Optimal method: $\min_{k\geq 0, 1\leq p\leq 2} k \quad$ s. t. $C(k,\delta,p)\leq\epsilon$
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:



Conclusion

- Developement of new first-order methods with intermediate behavior between
  1. the slow but robust Gradient Method (GM)
  2. the fast but sensitive Fast Gradient Method (FGM).

  $\Rightarrow$ Notion of Intermediate Gradient Methods (IGM).
- For each $1 \leq p \leq 2$, we have developed a method with rate of convergence $\Theta(\frac{1}{k^p})$ and with corresponding optimal rate of error accumulation $\Theta(k^{p-1}\delta)$.
- With availability of IGM, we can minimize a convex function endowed with an inexact oracle more efficiently that just using the GM and FGM.
- 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