The Fast-gradient method as a Universal Optimal First-order Method

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)

Minisymposium Optimization: complexity and applications
Glasgow, July 1

Convex Optimization Problems

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

where:

1. $Q \subset \mathbb{R}^n$ 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

Black-box First-order methods

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 i.e. a numerical method using only values of the function and subgradients at some search points.
Black-box assumption:
In course of solving $\mathcal{P}$, the only information that can obtain $\mathcal{M}$ about $\mathcal{P}$ comes from a
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 :

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

The method has no access to the problem structure.

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

- **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** $\mu(f) > 0$ if:

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

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

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

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

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** $M(f)$ if:

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

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

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

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

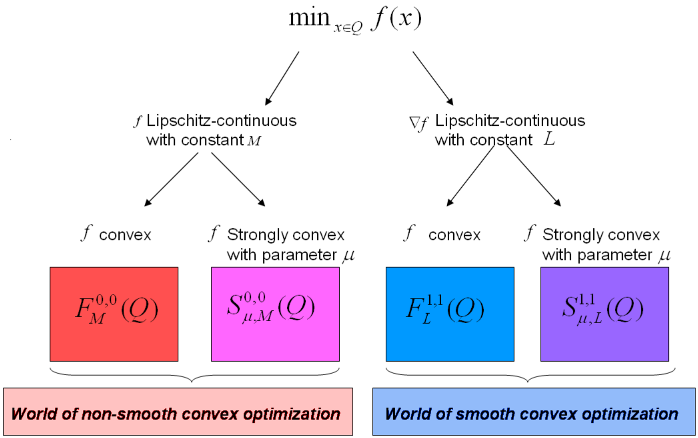
  $$\|\nabla f(x) - \nabla f(y)\|_* \leq L(f) \|x - y\| \quad \forall x, y \in Q.$$

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

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

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

Classes of Convex Functions



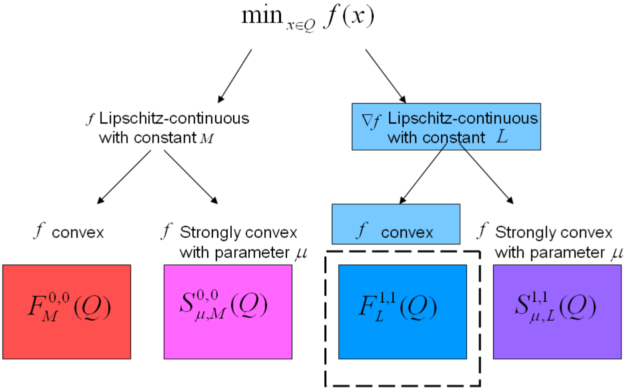
Optimal Complexity of Classes of Convex Functions

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

Outline

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

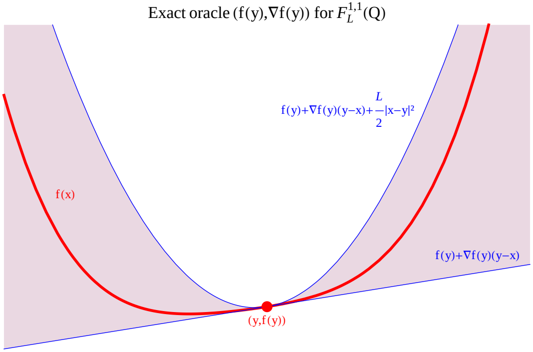


Exact Oracle for $F^{1,1}_{L(f)}(Q)$

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:

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

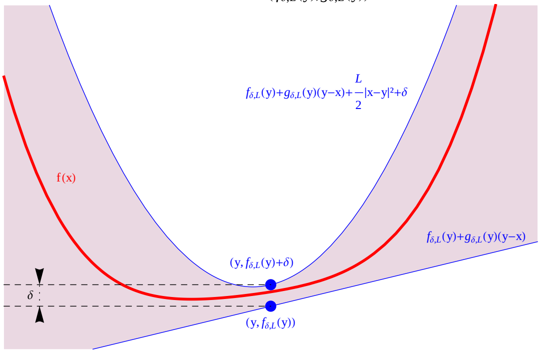
for all $x \in Q$.



(δ, L)-oracle

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

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



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.
   Subject of this talk

Our Goal

Prove, using the notion of (δ, L) oracle, that **the Fast-gradient method**, initially devoted for functions in $F^{1,1}_{L(f)}(Q)$:

- Can be also applied to various other classes of convex problems
- Provides in each case, an optimal method with respect to information-based complexity.

Outline

Fast Gradient Method

First-order method devoted for problems in the class $F^{1,1}_{L(f)}(Q)$.
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 $k \geq 0$

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

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

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

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

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

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

Fast Gradient Method with (δ, L) oracle

Effect 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

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

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

- When $\delta > 0$, the convergence rate is slowed down by an extra term that makes the method asymptotically divergent.
- But allowing $\delta > 0$, we can apply the FGM to functions that are not in $F_{L(f)}^{1,1}(Q)$.

Outline

Applications in Non-smooth Convex Optimization

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

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

does not imply differentiability.

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

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

Applications in Non-smooth Convex Optimization (2)

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

Fast Gradient Method for non-smooth problem

Consequence: We can apply any FOM of smooth convex-optimization to a non-smooth function f.
In particular, we can apply FGM and we have:

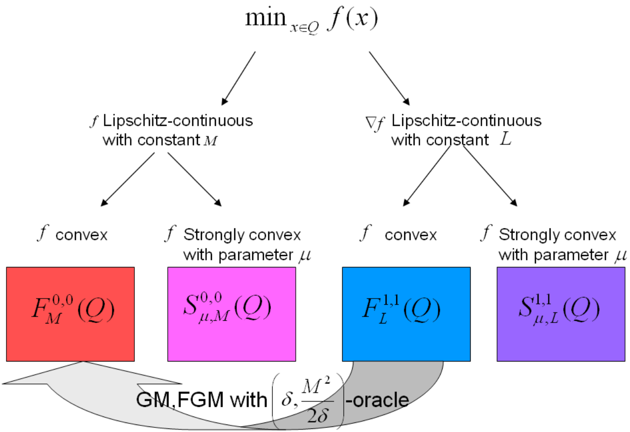
$$f(\hat{x}_k) - f^* \leq \frac{2M(f)^2 R^2}{(k+1)^2 \delta} + \delta(k+1).$$

With a optimal choice of δ:

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

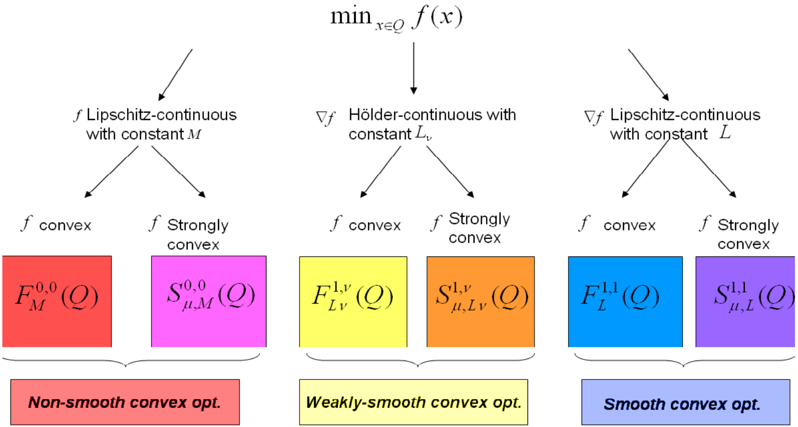
$\Rightarrow$ Optimal rate of convergence $\Theta\left(\frac{M(f)R}{\sqrt{k}}\right)$ for the non-smooth problem (i.e. optimal complexity of $\Theta\left(\frac{M(f)^2R^2}{\epsilon^2}\right)$).

Application to Non-smooth convex problems



Outline

Intermediate case: Weakly-smooth Convex Optimization



(δ, L) oracle for 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

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

FGM for weakly-smooth problems

Consequence: We can apply any FOM of smooth convex-optimization to a weakly-smooth function f.
In particular, we can apply FGM and we have:

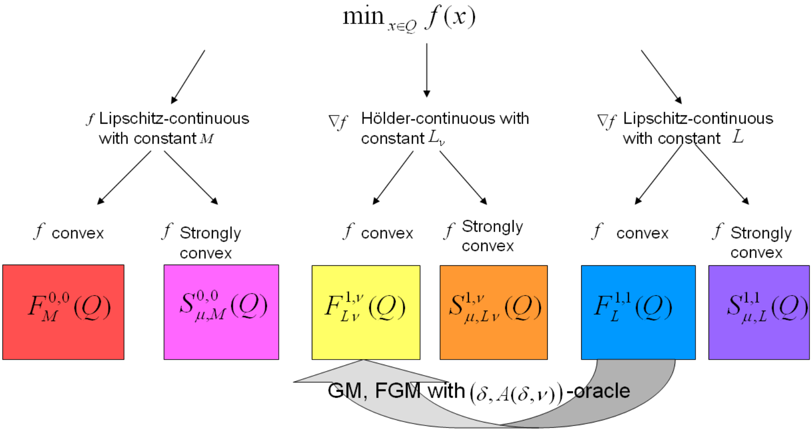
$$f(y_k) - f(x^*) \leq 4L_\nu \left[\frac{L_\nu}{2\delta} \cdot \frac{1-\nu}{1+\nu}\right]^{\frac{1-\nu}{1+\nu}} \frac{R^2}{(k+1)^2} + \delta \cdot (k+1)$$

With a optimal choice of δ:

$$f(y_k) - f(x^*) \leq \frac{2L_\nu R^{1+\nu}}{1+\nu} \left(\frac{2}{k+1}\right)^{\frac{1+3\nu}{2}} .$$

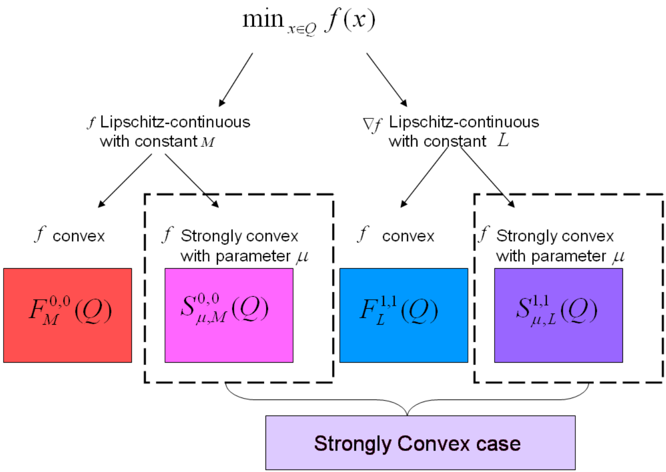
Optimal rate of convergence $\Theta\left(\frac{L_\nu R^{1+\nu}}{k^{\frac{1+3\nu}{2}}}\right)$ for the weakly-smooth problem.

Application to Weakly-smooth convex problems



Outline

What about the strongly convex case ?



Notion of (δ, L, μ) oracle

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

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

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

for all $x \in Q$.

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

$$f_{y,\delta} + \langle g_{y,\delta}, x - y \rangle + \frac{\mu}{2} \|x - y\|^2 \leq$$

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

FGM for strongly convex function using (δ, L, μ) oracle

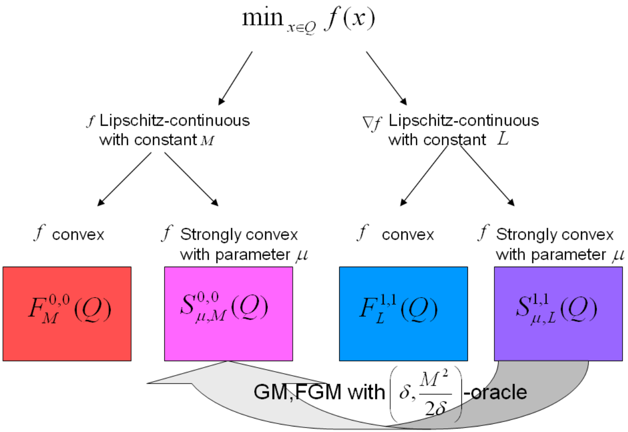
An adapted version of the FGM applied to a function f endowed with a (δ, L, μ) oracle satisfies:

$$f(x_k) - f^* \leq \frac{LR^2}{2} \exp\left(-k\sqrt{\frac{\mu}{L}}\right) + \sqrt{\frac{L}{\mu}}\delta.$$

In particular:

- If $f \in S^{1,1}_{\mu(f),L(f)}(Q)$ (Smooth strongly convex function): a $(0, L(f), \mu(f))$ oracle is available and the FGM reach the optimal complexity $\Theta\left(\sqrt{\frac{L(f)}{\mu(f)}} \ln\left(\frac{f(x_0)-f^*}{\epsilon}\right)\right)$
- If $f \in S^{0,0}_{\mu(f),M(f)}(Q)$ (Non-smooth strongly convex function): a $(\delta, \frac{M(f)^2}{2\delta}, \mu(f))$ oracle is available. With an optimal choice of δ, we obtain the optimal complexity $\Theta\left(\frac{M(f)^2}{\mu(f)\epsilon} \ln\left(\frac{f(x_0)-f^*}{\epsilon}\right)\right)$.

Application to non-smooth strongly convex function



The Fast-gradient method as a Universal Optimal first-order method :

Class	δ	L	Complexity.
$F^{1,1}_{L(f)}(Q)$	0	$L(f)$	$\Theta\left(\sqrt{\frac{L(f)R^2}{\epsilon}}\right)$
$F^{0,0}_{M(f)}(Q)$	δ	$\frac{M(f)^2}{2\delta}$	$\Theta\left(\frac{M(f)^2R^2}{\epsilon^2}\right)$
$F^{1,\nu}_{L_\nu(f)}(Q)$	δ	$L_\nu(f)\left[\frac{L_\nu(f)}{2\delta}\frac{1-\nu}{1+\nu}\right]^{\frac{1-\nu}{1+\nu}}$	$\Theta\left(\left(\frac{L_\nu(f)R^{1+\nu}}{\epsilon}\right)^{\frac{2}{1+3\nu}}\right)$
$S^{1,1}_{\mu,L(f)}(Q)$	0	$L(f)$	$\Theta\left(\sqrt{\frac{L(f)}{\mu}}\ln\left(\frac{\mu R^2}{\epsilon}\right)\right)$
$S^{0,0}_{\mu,M(f)}(Q)$	δ	$\frac{M(f)^2}{2\delta}$	$\Theta\left(\frac{M(f)^2}{\mu\epsilon}\ln\left(\frac{f(x_0)-f^*}{\epsilon}\right)\right)$

Conclusion

- Introduction of the notion of (δ, L)-oracle, a generalization of the first-order oracle in smooth convex optimization.
- With this notion, we can apply the Fast-gradient method, initaly devoted for problems in $F_L^{1,1}(Q)$, to other classes of convex problems with weaker level of smoothness.
- In each case, we obtain an optimal method with respect to information-based complexity.
- Same kind of results for strongly convex problems

$\Rightarrow$ **FGM = Universal Optimal FOM**.

Thanks for your attention !