Recent advances in Structural Convex Optimization

Yurii Nesterov, CORE/INMA (UCL, Belgium)

December 7, 2010 (Fields Institute)

Outline

Black-box optimization

Black-box optimization

Problem formulation: $\min\{f(x) : x \in Q\}$, where

- $Q \subset R^n$ is a closed bounded convex set ($\|x\| \leq R$, $x \in Q$),
- f is a closed convex function.

Black-box optimization

Problem formulation: $\min\{f(x) : x \in Q\}$, where

- $Q \subset R^n$ is a closed bounded convex set ($\|x\| \le R$, $x \in Q$),
- f is a closed convex function.

Main Assumptions: the classes of functional components

Black-box optimization

Problem formulation: $\min\{f(x) : x \in Q\}$, where

- $Q \subset R^n$ is a closed bounded convex set ($\|x\| \leq R$, $x \in Q$),
- f is a closed convex function.

Main Assumptions: the classes of functional components

- $\mathcal{C}_1$: $\|\nabla f(x)\| \leq L \quad \forall x \in Q$,

Black-box optimization

Problem formulation: $\min\{f(x) : x \in Q\}$, where

- $Q \subset R^n$ is a closed bounded convex set ($\|x\| \le R$, $x \in Q$),
- f is a closed convex function.

Main Assumptions: the classes of functional components

- $\mathcal{C}_1$: $\|\nabla f(x)\| \le L \quad \forall x \in Q$,
- $\mathcal{C}_2$: $\|\nabla f(x) - \nabla f(y)\| \le M\|x - y\| \quad \forall x, y \in Q$,

Black-box optimization

Problem formulation: $\min\{f(x) : x \in Q\}$, where

- $Q \subset R^n$ is a closed bounded convex set ($\|x\| \leq R$, $x \in Q$),
- f is a closed convex function.

Main Assumptions: the classes of functional components

- $\mathcal{C}_1 : \quad \|\nabla f(x)\| \leq L \quad \forall x \in Q,$
- $\mathcal{C}_2 : \quad \|\nabla f(x) - \nabla f(y)\| \leq M\|x - y\| \quad \forall x, y \in Q,$
- Convexity: $\quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \quad \forall x, y \in Q.$

Black-box optimization

Problem formulation: $\min\{f(x) : x \in Q\}$, where

- $Q \subset R^n$ is a closed bounded convex set ($\|x\| \leq R$, $x \in Q$),
- f is a closed convex function.

Main Assumptions: the classes of functional components

- $\mathcal{C}_1$: $\|\nabla f(x)\| \leq L \quad \forall x \in Q$,
- $\mathcal{C}_2$: $\|\nabla f(x) - \nabla f(y)\| \leq M\|x - y\| \quad \forall x, y \in Q$,
- Convexity: $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \quad \forall x, y \in Q$.

Black-Box Assumption: *Only* $(f(x_i), \nabla f(x_i))$, $i = 1, \ldots, N$, *are available.*

Complexity results

Model of the objective function: (provided by *oracle*)

Complexity results

Model of the objective function: (provided by *oracle*)

- $\mathcal{L}_k(x) \stackrel{\mathrm{def}}{=} \max\limits_{1 \le i \le k} [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] \;\le\; f(x), \quad \forall x \in Q.$

Complexity results

Model of the objective function: (provided by *oracle*)

- $\mathcal{L}_k(x) \stackrel{\mathrm{def}}{=} \max\limits_{1 \le i \le k} [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] \;\le\; f(x), \quad \forall x \in Q.$
- $\mathcal{C}_2 : \quad f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} M \|y - x\|^2, \; \forall x, y \in Q.$

Complexity results

Model of the objective function: (provided by *oracle*)

- $\mathcal{L}_k(x) \stackrel{\rm def}{=} \max\limits_{1 \le i \le k} [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] \;\le\; f(x), \quad \forall x \in Q.$
- $\mathcal{C}_2 : \quad f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} M \|y - x\|^2, \; \forall x, y \in Q.$

All methods use only this information.

Complexity results

Model of the objective function: (provided by *oracle*)

- $\mathcal{L}_k(x) \stackrel{\mathrm{def}}{=} \max\limits_{1 \le i \le k} [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] \le f(x), \quad \forall x \in Q.$
- $\mathcal{C}_2 : \quad f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} M \|y - x\|^2, \ \forall x, y \in Q.$

All methods use <u>only</u> this information.

Complexity Theory (Nemirovskii, Yudin 1977): # of oracle calls

Complexity results

Model of the objective function: (provided by *oracle*)

- $\mathcal{L}_k(x) \stackrel{\text{def}}{=} \max\limits_{1 \le i \le k} [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] \le f(x), \quad \forall x \in Q.$
- $\mathcal{C}_2 : \quad f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} M \|y - x\|^2, \ \forall x, y \in Q.$

All methods use only this information.

Complexity Theory (Nemirovskii, Yudin 1977): # of oracle calls

Problem class	Limit for calls	Lower bound
$\mathcal{C}_1 : \ \|\nabla f(\cdot)\| \le L$	$\le O(n)$	$O\left(L^2 R^2 / \epsilon^2\right)$

Complexity results

Model of the objective function: (provided by *oracle*)

- $\mathcal{L}_k(x) \stackrel{\text{def}}{=} \max\limits_{1 \le i \le k} [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] \le f(x), \quad \forall x \in Q.$
- $\mathcal{C}_2: \quad f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} M \|y - x\|^2, \ \forall x, y \in Q.$

All methods use only this information.

Complexity Theory (Nemirovskii, Yudin 1977): # of oracle calls

Problem class	Limit for calls	Lower bound
$\mathcal{C}_1: \ \|\nabla f(\cdot)\| \le L$	$\le O(n)$	$O\left(L^2 R^2 / \epsilon^2\right)$
$\mathcal{C}_2: \ \|\nabla^2 f(\cdot)\| \le M$	$\le O(n)$	$O\left(M^{1/2} R / \epsilon^{1/2}\right)$

Complexity results

Model of the objective function: (provided by *oracle*)

- $\mathcal{L}_k(x) \stackrel{\text{def}}{=} \max\limits_{1\leq i\leq k} [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] \leq f(x), \quad \forall x \in Q.$
- $\mathcal{C}_2: \quad f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} M \|y - x\|^2, \ \forall x, y \in Q.$

All methods use <u>only</u> this information.

Complexity Theory (Nemirovskii, Yudin 1977): # of oracle calls

Problem class	Limit for calls	Lower bound
$\mathcal{C}_1: \ \|\nabla f(\cdot)\| \leq L$	$\leq O(n)$	$O\left(L^2 R^2/\epsilon^2\right)$
$\mathcal{C}_2: \ \|\nabla^2 f(\cdot)\| \leq M$	$\leq O(n)$	$O\left(M^{1/2} R/\epsilon^{1/2}\right)$
$\mathcal{C}_3: \ \|\nabla f(\cdot)\| \leq L$	$\geq O(n)$	$O\left(n \ln[LR/\epsilon]\right)$

The bounds are exact!

Worst functions in the world

Worst functions in the world

Class $\mathcal{C}_1$, bounded gradient.

$$f(x) = L \cdot \max_{1 \le i \le n} x^{(i)}, \quad Q = \{x : \|x\| \le R\}.$$

Worst functions in the world

Class $\mathcal{C}_1$, bounded gradient.

$$f(x) = L \cdot \max_{1 \le i \le n} x^{(i)}, \quad Q = \{x : \|x\| \le R\}.$$

Class $\mathcal{C}_2$, bounded Hessian.

$$f(x) = \frac{M}{8}\left[(x^{(1)})^2 + \sum_{i=1}^{n-1} (x^{(i)} - x^{(i+1)})^2 + (x^{(n)})^2 - 2x^{(1)} \right].$$

Worst functions in the world

Class $\mathcal{C}_1$, bounded gradient.

$$f(x) = L \cdot \max_{1 \le i \le n} x^{(i)}, \quad Q = \{x : \|x\| \le R\}.$$

Class $\mathcal{C}_2$, bounded Hessian.

$$f(x) = \frac{M}{8}\left[(x^{(1)})^2 + \sum_{i=1}^{n-1}(x^{(i)} - x^{(i+1)})^2 + (x^{(n)})^2 - 2x^{(1)}\right].$$

- Being in a Black Box, these problems are difficult for <u>all</u> methods.

Worst functions in the world

Class $\mathcal{C}_1$, bounded gradient.

$$f(x) = L \cdot \max_{1 \le i \le n} x^{(i)}, \quad Q = \{x : \|x\| \le R\}.$$

Class $\mathcal{C}_2$, bounded Hessian.

$$f(x) = \frac{M}{8}\left[(x^{(1)})^2 + \sum_{i=1}^{n-1} (x^{(i)} - x^{(i+1)})^2 + (x^{(n)})^2 - 2x^{(1)} \right].$$

- Being in a Black Box, these problems are difficult for all methods.
- They become trivial when we can see their structure.

Worst functions in the world

Class $\mathcal{C}_1$, bounded gradient.

$$f(x) = L \cdot \max_{1 \le i \le n} x^{(i)}, \quad Q = \{x : \|x\| \le R\}.$$

Class $\mathcal{C}_2$, bounded Hessian.

$$f(x) = \frac{M}{8}\left[(x^{(1)})^2 + \sum_{i=1}^{n-1} (x^{(i)} - x^{(i+1)})^2 + (x^{(n)})^2 - 2x^{(1)} \right].$$

- Being in a Black Box, these problems are difficult for all methods.
- They become trivial when we can see their structure.
- The structure is always visible when we code the problem.

Worst functions in the world

Class $\mathcal{C}_1$, bounded gradient.

$$f(x) = L \cdot \max_{1 \le i \le n} x^{(i)}, \quad Q = \{x : \|x\| \le R\}.$$

Class $\mathcal{C}_2$, bounded Hessian.

$$f(x) = \frac{M}{8}\left[(x^{(1)})^2 + \sum_{i=1}^{n-1}(x^{(i)} - x^{(i+1)})^2 + (x^{(n)})^2 - 2x^{(1)}\right].$$

- Being in a Black Box, these problems are difficult for all methods.
- They become trivial when we can see their structure.
- The structure is always visible when we code the problem.

Can we help the methods?

First ring: Theory of self-concordant functions (N. & Nemirovskii, 1987-1990)

First ring: Theory of self-concordant functions (N. & Nemirovskii, 1987-1990)

Model of the problem: $\min\{\langle c, x\rangle : \ x \in Q\}$.

Closed convex set Q is endowed with self-concordant barrier $F(x)$:

First ring: Theory of self-concordant functions (N. & Nemirovskii, 1987-1990)

Model of the problem: $\min\{\langle c, x\rangle : \ x \in Q\}$.

Closed convex set Q is endowed with self-concordant barrier $F(x)$:

- $D^3F(x)[h,h,h] \leq 2\left(D^2F(x)[h,h]\right)^{3/2} \quad \forall x \in \mathrm{int}\, Q, h \in R^n,$

First ring: Theory of self-concordant functions (N. & Nemirovskii, 1987-1990)

Model of the problem: $\min\{\langle c, x\rangle : x \in Q\}$.

Closed convex set Q is endowed with self-concordant barrier $F(x)$:

- $D^3F(x)[h,h,h] \leq 2\left(D^2F(x)[h,h]\right)^{3/2} \quad \forall x \in \mathrm{int}\, Q, h \in R^n$,
- $\langle \nabla F(x), h\rangle^2 \leq \nu \cdot D^2F(x)[h,h] \quad \forall x \in \mathrm{int}\, Q, h \in R^n$.

The value $\nu \geq 1$ is called the *parameter* of the barrier.

First ring: Theory of self-concordant functions (N. & Nemirovskii, 1987-1990)

Model of the problem: $\min\{\langle c, x\rangle : \; x \in Q\}$.

Closed convex set Q is endowed with self-concordant barrier $F(x)$:

- $D^3F(x)[h,h,h] \leq 2\left(D^2F(x)[h,h]\right)^{3/2} \quad \forall x \in \mathrm{int}\, Q, h \in R^n$,
- $\langle \nabla F(x), h\rangle^2 \leq \nu \cdot D^2F(x)[h,h] \quad \forall x \in \mathrm{int}\, Q, h \in R^n$.

The value $\nu \geq 1$ is called the *parameter* of the barrier.

Complexity of finding ϵ-solution:

$$O(\nu^{1/2} \ln \tfrac{\nu}{\epsilon}) \text{ iterations of Newton method.}$$

(This is a Black-Box method.)

First ring: Theory of self-concordant functions (N. & Nemirovskii, 1987-1990)

Model of the problem: $\min\{\langle c, x\rangle : x \in Q\}$.

Closed convex set Q is endowed with self-concordant barrier $F(x)$:

- $D^3 F(x)[h,h,h] \le 2\left(D^2 F(x)[h,h]\right)^{3/2} \quad \forall x \in \mathrm{int}\, Q, h \in R^n$,
- $\langle \nabla F(x), h\rangle^2 \le \nu \cdot D^2 F(x)[h,h] \quad \forall x \in \mathrm{int}\, Q, h \in R^n$.

The value $\nu \ge 1$ is called the *parameter* of the barrier.

Complexity of finding ϵ-solution:

$$O(\nu^{1/2} \ln \tfrac{\nu}{\epsilon}) \text{ iterations of Newton method.}$$

(This is a Black-Box method.)

How wide is the application field?

Applicability of the theory

Universal barrier function

$$f_Q(x) = \kappa \cdot \ln \mathrm{Vol}\, P(x), \quad P(x) = \{s : \langle s, y - x \rangle \leq 1 \; \forall y \in Q\}.$$

For some $\kappa > 0$, this function is $O(n)$-s.c. barrier for Q.

Applicability of the theory

Universal barrier function

$$f_Q(x) = \kappa \cdot \ln \mathrm{Vol}\, P(x), \quad P(x) = \{s : \langle s, y - x\rangle \leq 1 \ \forall y \in Q\}.$$

For some $\kappa > 0$, this function is $O(n)$-s.c. barrier for Q.

Hence, the convex problems can be solved in $O(\sqrt{n} \ln \frac{n}{\epsilon})$ iterations.

Applicability of the theory

Universal barrier function

$$f_Q(x) = \kappa \cdot \ln \mathrm{Vol}\, P(x), \quad P(x) = \{s : \langle s, y - x \rangle \leq 1 \; \forall y \in Q\}.$$

For some $\kappa > 0$, this function is $O(n)$-s.c. barrier for Q.

Hence, the convex problems can be solved in $O(\sqrt{n} \ln \frac{n}{\epsilon})$ iterations. (Impossible in the Black-Box framework!)

Applicability of the theory

Universal barrier function

$$f_Q(x) = \kappa \cdot \ln \mathrm{Vol}\, P(x), \quad P(x) = \{s : \langle s, y - x \rangle \le 1 \; \forall y \in Q\}.$$

For some $\kappa > 0$, this function is $O(n)$-s.c. barrier for Q.

Hence, the convex problems can be solved in $O(\sqrt{n} \ln \frac{n}{\epsilon})$ iterations. (Impossible in the Black-Box framework!)

Conclusion

- It is possible to construct s.c.b. (with appropriate ν) for all convex sets with known structure.

Applicability of the theory

Universal barrier function

$$f_Q(x) = \kappa \cdot \ln \mathrm{Vol}\, P(x), \quad P(x) = \{s : \langle s, y - x \rangle \leq 1 \; \forall y \in Q\}.$$

For some $\kappa > 0$, this function is $O(n)$-s.c. barrier for Q.

Hence, the convex problems can be solved in $O(\sqrt{n} \ln \frac{n}{\epsilon})$ iterations. (Impossible in the Black-Box framework!)

Conclusion

- It is possible to construct s.c.b. (with appropriate ν) for all convex sets with known structure.
- These actions violate the Black-Box assumption for the initial problem.

How to use this theory?

We need to work directly with the elements of the problem.

How to use this theory?

We need to work directly with the elements of the problem.

Example: $\min\limits_x \frac{1}{2}\|Ax-b\|^2 + \|x\|_1$.

How to use this theory?

We need to work directly with the elements of the problem.

Example: $\min\limits_{x} \frac{1}{2}\|Ax - b\|^2 + \|x\|_1.$

1. Rewrite the problem in the standard form:

$$\begin{aligned} \min \quad & \tau_1 + \tau_2 \\ \text{s.t.} \quad & 2\tau_1 \geq \|Ax - b\|^2, \quad \tau_2 \geq \sum_{i=1}^{n} y_i, \\ & y_i \geq |x_i|, \ i = 1, \dots, n. \end{aligned}$$

How to use this theory?

We need to work directly with the elements of the problem.
Example: $\min\limits_x \frac{1}{2}\|Ax - b\|^2 + \|x\|_1$.

1. Rewrite the problem in the standard form:

$$\begin{aligned} \min \quad & \tau_1 + \tau_2 \\ \text{s.t.} \quad & 2\tau_1 \geq \|Ax - b\|^2, \quad \tau_2 \geq \sum_{i=1}^n y_i, \\ & y_i \geq |x_i|, \; i = 1, \dots, n. \end{aligned}$$

2. Construct the s.c.barrier

$$-\ln\left(2\tau_1 - \|Ax - b\|^2\right) - \ln\left(\tau_2 - \sum_{i=1}^n y_i\right) - \sum_{i=1}^n \ln\left(y_i^2 - x_i^2\right)$$

with parameter $\nu = 2n + 2$.

Second ring: Smoothing technique (2005)

Second ring: Smoothing technique (2005)

The most important problem: $\min_{x \in Q} f(x)$ with $f \in \mathcal{C}_1$.

Complexity: $O\left(\frac{L^2 R^2}{\epsilon^2}\right)$ calls of oracle.

Second ring: Smoothing technique (2005)

The most important problem: $\min_{x \in Q} f(x)$ with $f \in \mathcal{C}_1$.

Complexity: $O\left(\frac{L^2R^2}{\epsilon^2}\right)$ calls of oracle.

For a simpler problem ($f \in \mathcal{C}_2$) we have $O\left(\frac{M^{1/2}R}{\epsilon^{1/2}}\right)$ calls.

Second ring: Smoothing technique (2005)

The most important problem: $\min_{x \in Q} f(x)$ with $f \in \mathcal{C}_1$.

Complexity: $O\left(\frac{L^2 R^2}{\epsilon^2}\right)$ calls of oracle.

For a simpler problem ($f \in \mathcal{C}_2$) we have $O\left(\frac{M^{1/2} R}{\epsilon^{1/2}}\right)$ calls.

For $\epsilon = 10^{-2}$ the factor varies from 10 to 10000.

Can we decrease the gap?

Second ring: Smoothing technique (2005)

The most important problem: $\min_{x \in Q} f(x)$ with $f \in \mathcal{C}_1$.

Complexity: $O\left(\frac{L^2 R^2}{\epsilon^2}\right)$ calls of oracle.

For a simpler problem ($f \in \mathcal{C}_2$) we have $O\left(\frac{M^{1/2} R}{\epsilon^{1/2}}\right)$ calls.

For $\epsilon = 10^{-2}$ the factor varies from 10 to 10000.

Can we decrease the gap? (No way in BB-framework!)

Second ring: Smoothing technique (2005)

The most important problem: $\min_{x \in Q} f(x)$ with $f \in \mathcal{C}_1$.

Complexity: $O\left(\frac{L^2 R^2}{\epsilon^2}\right)$ calls of oracle.

For a simpler problem ($f \in \mathcal{C}_2$) we have $O\left(\frac{M^{1/2} R}{\epsilon^{1/2}}\right)$ calls.

For $\epsilon = 10^{-2}$ the factor varies from 10 to 10000.

Can we decrease the gap? (No way in BB-framework!)

Simple Theorem. For $f \in \mathcal{C}_1(R^n)$ there exists $f_\epsilon \in \mathcal{C}_2(R^n)$:

- $f(x) \leq f_\epsilon(x) \leq f(x) + \epsilon$ for any $x \in R^n$,
- $M(f_\epsilon) = \frac{1}{2\epsilon} L^2(f)$.

Second ring: Smoothing technique (2005)

The most important problem: $\min_{x \in Q} f(x)$ with $f \in \mathcal{C}_1$.

Complexity: $O\left(\frac{L^2 R^2}{\epsilon^2}\right)$ calls of oracle.

For a simpler problem ($f \in \mathcal{C}_2$) we have $O\left(\frac{M^{1/2} R}{\epsilon^{1/2}}\right)$ calls.

For $\epsilon = 10^{-2}$ the factor varies from 10 to 10000.
Can we decrease the gap? (No way in BB-framework!)

Simple Theorem. For $f \in \mathcal{C}_1(R^n)$ there exists $f_\epsilon \in \mathcal{C}_2(R^n)$:

- $f(x) \leq f_\epsilon(x) \leq f(x) + \epsilon$ for any $x \in R^n$,
- $M(f_\epsilon) = \frac{1}{2\epsilon} L^2(f)$.

Can we do this in a systematic way?

Second ring: Smoothing technique (2005)

The most important problem: $\min_{x \in Q} f(x)$ with $f \in \mathcal{C}_1$.

Complexity: $O\left(\frac{L^2 R^2}{\epsilon^2}\right)$ calls of oracle.

For a simpler problem ($f \in \mathcal{C}_2$) we have $O\left(\frac{M^{1/2} R}{\epsilon^{1/2}}\right)$ calls.

For $\epsilon = 10^{-2}$ the factor varies from 10 to 10000.
Can we decrease the gap? (No way in BB-framework!)

Simple Theorem. For $f \in \mathcal{C}_1(R^n)$ there exists $f_\epsilon \in \mathcal{C}_2(R^n)$:

- $f(x) \leq f_\epsilon(x) \leq f(x) + \epsilon$ for any $x \in R^n$,
- $M(f_\epsilon) = \frac{1}{2\epsilon} L^2(f)$.

Can we do this in a systematic way? (Then we pass to $O\left(\frac{LR}{\epsilon}\right)$.)

Max-representation of the objective function

Let $Q_d \subseteq E_d$ be a bounded convex dual feasible set and $\phi(u)$ be a convex function. Consider

$$f(x) \quad = \quad \max_{u \in Q_d} \{ \langle Ax - b, u \rangle - \phi(u) \},$$

Max-representation of the objective function

Let $Q_d \subseteq E_d$ be a bounded convex dual feasible set and $\phi(u)$ be a convex function. Consider

$$f(x) \;=\; \max_{u \in Q_d} \{\langle Ax - b, u\rangle - \phi(u)\},$$

Let us choose prox-function $d(u)$ (strongly convex and positive) and define

$$f_\mu(x) \;=\; \max_{u \in Q_d} \{\langle Ax - b, u\rangle - \phi(u) - \mu \cdot d(u)\}, \quad \mu > 0.$$

Max-representation of the objective function

Let $Q_d \subseteq E_d$ be a bounded convex dual feasible set and $\phi(u)$ be a convex function. Consider

$$f(x) = \max_{u \in Q_d} \{\langle Ax - b, u \rangle - \phi(u)\},$$

Let us choose prox-function $d(u)$ (strongly convex and positive) and define

$$f_\mu(x) = \max_{u \in Q_d} \{\langle Ax - b, u \rangle - \phi(u) - \mu \cdot d(u)\}, \quad \mu > 0.$$

Denoting $D_d = \max\limits_{u \in Q_d} d(u)$, we get $\quad f(x) \geq f_\mu(x) \geq f(x) - \mu D_d$.

Max-representation of the objective function

Let $Q_d \subseteq E_d$ be a bounded convex dual feasible set and $\phi(u)$ be a convex function. Consider

$$f(x) \quad = \quad \max_{u \in Q_d} \{ \langle Ax - b, u \rangle - \phi(u) \},$$

Let us choose prox-function $d(u)$ (strongly convex and positive) and define

$$f_\mu(x) \quad = \quad \max_{u \in Q_d} \{ \langle Ax - b, u \rangle - \phi(u) - \mu \cdot d(u) \}, \quad \mu > 0.$$

Denoting $D_d = \max\limits_{u \in Q_d} d(u)$, we get $\quad f(x) \geq f_\mu(x) \; \geq \; f(x) - \mu D_d$.

Note: $M(f_\mu) = \frac{1}{2\mu} \|A\|^2$, with $\|A\| = \max\limits_{\|x\| \leq 1, \|u\| \leq 1} \langle Ax, u \rangle$.

Max-representation of the objective function

Let $Q_d \subseteq E_d$ be a bounded convex dual feasible set and $\phi(u)$ be a convex function. Consider

$$f(x) \quad = \quad \max_{u \in Q_d} \{ \langle Ax - b, u \rangle - \phi(u) \},$$

Let us choose prox-function $d(u)$ (strongly convex and positive) and define

$$f_\mu(x) \quad = \quad \max_{u \in Q_d} \{ \langle Ax - b, u \rangle - \phi(u) - \mu \cdot d(u) \}, \quad \mu > 0.$$

Denoting $D_d = \max\limits_{u \in Q_d} d(u)$, we get $\quad f(x) \geq f_\mu(x) \; \geq \; f(x) - \mu D_d$.

Note: $M(f_\mu) = \frac{1}{2\mu} \|A\|^2$, with $\|A\| = \max\limits_{\|x\| \leq 1, \|u\| \leq 1} \langle Ax, u \rangle$.

Function f_μ must be <u>computable</u>!

Example

Consider $f(x) = \max\limits_{1 \le j \le m} |\langle a_j, x\rangle - b^{(j)}|$.

Example

Consider $f(x) = \max\limits_{1 \le j \le m} |\langle a_j, x \rangle - b^{(j)}|$.

1. $E_d = R^m$, $\phi(u) = \langle b, u \rangle$,

$$f(x) = \max_{u \in R^m} \left\{ \sum_{j=1}^{m} u^{(j)}[\langle a_j, x \rangle - b^{(j)}] : \sum_{j=1}^{m} |u^{(j)}| \le 1 \right\}.$$

Example

Consider $f(x) = \max\limits_{1 \le j \le m} |\langle a_j, x\rangle - b^{(j)}|$.

1. $E_d = R^m$, $\phi(u) = \langle b, u\rangle$,

$$f(x) = \max_{u \in R^m} \left\{ \sum_{j=1}^m u^{(j)}[\langle a_j, x\rangle - b^{(j)}] : \sum_{j=1}^m |u^{(j)}| \le 1 \right\}.$$

2. $E_d = R^{2m}$, $\phi(u)$ is a linear, Q_d is a simplex:

$$f(x) = \max_{u \in R^{2m}_+} \{ \sum_{j=1}^m (u_1^{(j)} - u_2^{(j)}) \cdot [\langle a_j, x\rangle - b^{(j)}] : \sum_{j=1}^m (u_1^{(j)} + u_2^{(j)}) = 1 \}.$$

Application: Consider $\min_{x\in\Delta_n} \left[f(x) \stackrel{\mathrm{def}}{=} \max_{1\leq j\leq m} \langle a_j, x\rangle \right]$,

where $\Delta_n \in R^n$ is a standard simplex.

Application: Consider $\min\limits_{x\in\Delta_n}\left[f(x) \stackrel{\mathrm{def}}{=} \max\limits_{1\le j\le m}\langle a_j, x\rangle\right]$,

where $\Delta_n \in R^n$ is a standard simplex. For the standard subgradient method, we can guarantee

$$f(x_N) - f^* \quad \le \frac{\sqrt{\ln n}}{\sqrt{N+1}} \cdot \max_{i,j} |a_j^{(i)}|.$$

Application: Consider $\min_{x \in \Delta_n} \left[f(x) \stackrel{\text{def}}{=} \max_{1 \le j \le m} \langle a_j, x \rangle \right]$,

where $\Delta_n \in R^n$ is a standard simplex. For the standard subgradient method, we can guarantee

$$f(x_N) - f^* \quad \le \frac{\sqrt{\ln n}}{\sqrt{N+1}} \cdot \max_{i,j} |a_j^{(i)}|.$$

Note that $f(x) = \max_{u \in \Delta_m} \langle Au, x \rangle$. For the smoothing technique, let us use the *entropy function*:

$$d(u) \quad = \quad \ln m + \sum_{i=1}^{n} u^{(i)} \ln u^{(i)}, \quad u \in \Delta_m.$$

Application: Consider $\min_{x \in \Delta_n} \left[f(x) \stackrel{\mathrm{def}}{=} \max_{1 \le j \le m} \langle a_j, x \rangle \right]$,

where $\Delta_n \in R^n$ is a standard simplex. For the standard subgradient method, we can guarantee

$$f(x_N) - f^* \quad \le \frac{\sqrt{\ln n}}{\sqrt{N+1}} \cdot \max_{i,j} |a_j^{(i)}|.$$

Note that $f(x) = \max_{u \in \Delta_m} \langle Au, x \rangle$. For the smoothing technique, let us use the *entropy function*:

$$d(u) \quad = \quad \ln m + \sum_{i=1}^{n} u^{(i)} \ln u^{(i)}, \quad u \in \Delta_m.$$

Then $f_\mu(x) = \mu \ln \left[\frac{1}{m} \sum_{j=1}^{m} e^{\langle a_j, x \rangle / \mu} \right]$, and we obtain the following rate of convergence:

$$f(x_N) - f^* \quad \le \frac{4\sqrt{\ln n \cdot \ln m}}{N} \cdot \max_{i,j} |a_j^{(i)}|.$$

Modern Gradient Methods: primal scheme

Modern Gradient Methods: primal scheme

Problem: $f(x) \quad \to \quad \min_{x \in Q},$ where f is convex function and $\|\nabla f(x) - \nabla f(y)\|^* \leq M(f)\|x - y\| \ \forall x, y \in Q$ (closed, convex).

Modern Gradient Methods: primal scheme

Problem: $f(x) \quad \to \quad \min\limits_{x \in Q}, \quad$ where f is convex function and $\|\nabla f(x) - \nabla f(y)\|^* \leq M(f)\|x - y\| \;\forall x, y \in Q$ (closed, convex).

Primal Gradient Method (PGM): $\quad x_{k+1} = T(x_k), \quad$ where

$$T(x_k) = \arg\min_{x \in Q} [\underbrace{f(x_k) + \langle f(x_k), x - x_k \rangle + \tfrac{1}{2} M(f)\|x - x_k\|^2}_{\geq f(x)}].$$

Modern Gradient Methods: primal scheme

Problem: $f(x) \quad \rightarrow \quad \min_{x \in Q}$, where f is convex function and $\|\nabla f(x) - \nabla f(y)\|^* \leq M(f)\|x - y\| \ \forall x, y \in Q$ (closed, convex).

Primal Gradient Method (PGM): $x_{k+1} = T(x_k)$, where

$$T(x_k) = \arg\min_{x \in Q} [\underbrace{f(x_k) + \langle f(x_k), x - x_k \rangle + \frac{1}{2} M(f) \|x - x_k\|^2}_{\geq f(x)}].$$

Therefore $f(T(x_k)) + \frac{1}{2} M(f) \|x^* - T(x_k)\|^2$
$\leq f(x_k) + \langle f(x_k), x^* - x_k \rangle + \frac{M(f)}{2} \|x^* - x_k\|^2 \leq f^* + \frac{M(f)}{2} \|x^* - x_k\|^2$.

Modern Gradient Methods: primal scheme

Problem: $f(x) \quad \to \quad \min_{x \in Q}, \quad$ where f is convex function and $\|\nabla f(x) - \nabla f(y)\|^* \leq M(f)\|x - y\| \ \forall x, y \in Q$ (closed, convex).

Primal Gradient Method (PGM): $\quad x_{k+1} = T(x_k), \quad$ where

$$T(x_k) = \arg\min_{x \in Q} [\underbrace{f(x_k) + \langle f(x_k), x - x_k \rangle + \tfrac{1}{2} M(f) \|x - x_k\|^2}_{\geq f(x)}].$$

Therefore $\quad f(T(x_k)) + \frac{1}{2} M(f) \|x^* - T(x_k)\|^2$

$$\leq f(x_k) + \langle f(x_k), x^* - x_k \rangle + \frac{M(f)}{2} \|x^* - x_k\|^2 \leq f^* + \frac{M(f)}{2} \|x^* - x_k\|^2.$$

Rate of convergence:

$$\sum_{i=0}^{k} [f(x_i) - f^*] \leq \frac{M(f)}{2} \|x^* - x_0\|^2 \ \Rightarrow \ f(x_k^*) - f^* \leq \frac{M(f)\|x^* - x_0\|^2}{2(k+1)}.$$

Modern Gradient Methods: primal scheme

Problem: $f(x) \to \min\limits_{x \in Q}$, where f is convex function and $\|\nabla f(x) - \nabla f(y)\|^* \le M(f)\|x - y\| \; \forall x, y \in Q$ (closed, convex).

Primal Gradient Method (PGM): $x_{k+1} = T(x_k)$, where

$$T(x_k) = \arg\min_{x \in Q} [\underbrace{f(x_k) + \langle f(x_k), x - x_k \rangle + \tfrac{1}{2} M(f) \|x - x_k\|^2}_{\ge f(x)}].$$

Therefore $f(T(x_k)) + \frac{1}{2}M(f)\|x^* - T(x_k)\|^2$

$$\le f(x_k) + \langle f(x_k), x^* - x_k \rangle + \frac{M(f)}{2}\|x^* - x_k\|^2 \le f^* + \frac{M(f)}{2}\|x^* - x_k\|^2.$$

Rate of convergence:

$$\sum_{i=0}^{k} [f(x_i) - f^*] \le \frac{M(f)}{2}\|x^* - x_0\|^2 \;\Rightarrow\; f(x_k^*) - f^* \le \frac{M(f)\|x^* - x_0\|^2}{2(k+1)}.$$

Main feature: moderate local improvement.

Modern Gradient Methods: primal scheme

Problem: $f(x) \quad \to \quad \min\limits_{x \in Q}, \quad$ where f is convex function and $\|\nabla f(x) - \nabla f(y)\|^* \leq M(f)\|x - y\| \; \forall x, y \in Q$ (closed, convex).

Primal Gradient Method (PGM): $\quad x_{k+1} = T(x_k), \quad$ where

$$T(x_k) = \arg\min_{x \in Q} [\underbrace{f(x_k) + \langle f(x_k), x - x_k \rangle + \tfrac{1}{2} M(f) \|x - x_k\|^2}_{\geq f(x)}].$$

Therefore $\quad f(T(x_k)) + \frac{1}{2} M(f) \|x^* - T(x_k)\|^2$

$$\leq f(x_k) + \langle f(x_k), x^* - x_k \rangle + \frac{M(f)}{2} \|x^* - x_k\|^2 \leq f^* + \frac{M(f)}{2} \|x^* - x_k\|^2.$$

Rate of convergence:

$$\sum_{i=0}^{k} [f(x_i) - f^*] \leq \frac{M(f)}{2} \|x^* - x_0\|^2 \;\Rightarrow\; f(x_k^*) - f^* \leq \frac{M(f)\|x^* - x_0\|^2}{2(k+1)}.$$

Main feature: moderate local improvement.

Interpretation:

Modern Gradient Methods: primal scheme

Problem: $f(x) \;\to\; \min\limits_{x\in Q}$, where f is convex function and $\|\nabla f(x) - \nabla f(y)\|^* \leq M(f)\|x-y\|$ $\forall x, y \in Q$ (closed, convex).

Primal Gradient Method (PGM): $x_{k+1} = T(x_k)$, where

$$T(x_k) = \arg\min_{x\in Q} [\underbrace{f(x_k) + \langle f(x_k), x - x_k\rangle + \tfrac{1}{2}M(f)\|x-x_k\|^2}_{\geq f(x)}].$$

Therefore $f(T(x_k)) + \frac{1}{2}M(f)\|x^* - T(x_k)\|^2$

$$\leq f(x_k) + \langle f(x_k), x^* - x_k\rangle + \frac{M(f)}{2}\|x^*-x_k\|^2 \leq f^* + \frac{M(f)}{2}\|x^*-x_k\|^2.$$

Rate of convergence:

$$\sum_{i=0}^{k}[f(x_i) - f^*] \leq \frac{M(f)}{2}\|x^*-x_0\|^2 \;\Rightarrow\; f(x_k^*) - f^* \leq \frac{M(f)\|x^*-x_0\|^2}{2(k+1)}.$$

Main feature: moderate local improvement.

Interpretation: Practitioners, Industry, etc.

Dual Gradient Method (DGM)

Dual Gradient Method (DGM)

$$v_{k+1} = \arg\min_{x \in Q} \left\{ \psi_k(x) \equiv \sum_{i=0}^{k} [f(v_i) + \langle f(v_i), x - v_i \rangle] + \frac{M(f)}{2} \|x - x_0\|^2 \right\}$$

Dual Gradient Method (DGM)

$$v_{k+1} = \arg\min_{x \in Q} \left\{ \psi_k(x) \equiv \sum_{i=0}^{k} [f(v_i) + \langle f(v_i), x - v_i \rangle] + \frac{M(f)}{2} \|x - x_0\|^2 \right\}$$

Theorem: Let $x_i = T(v_i)$. Then $\sum_{i=0}^{k} [f(x_i) - f^*] \leq \frac{M(f)}{2} \|x^* - x_0\|^2$.

Dual Gradient Method (DGM)

$$v_{k+1} = \arg\min_{x \in Q} \left\{ \psi_k(x) \equiv \sum_{i=0}^{k} [f(v_i) + \langle f(v_i), x - v_i \rangle] + \frac{M(f)}{2} \|x - x_0\|^2 \right\}$$

Theorem: Let $x_i = T(v_i)$. Then $\sum\limits_{i=0}^{k} [f(x_i) - f^*] \leq \frac{M(f)}{2} \|x^* - x_0\|^2$.

Proof: 1. Let us prove by induction that $\sum\limits_{i=0}^{k} f(x_i) \leq \psi_k^*$.

Dual Gradient Method (DGM)

$$v_{k+1} = \arg\min_{x \in Q} \left\{ \psi_k(x) \equiv \sum_{i=0}^{k} [f(v_i) + \langle f(v_i), x - v_i \rangle] + \frac{M(f)}{2} \|x - x_0\|^2 \right\}$$

Theorem: Let $x_i = T(v_i)$. Then $\sum\limits_{i=0}^{k} [f(x_i) - f^*] \le \frac{M(f)}{2} \|x^* - x_0\|^2$.

Proof: 1. Let us prove by induction that $\sum\limits_{i=0}^{k} f(x_i) \le \psi_k^*$. Indeed,

$$\psi_{k+1}(x) = \psi_k(x) + f(v_{k+1}) + \langle f(v_{k+1}), x - v_{k+1} \rangle$$

Dual Gradient Method (DGM)

$$v_{k+1} = \arg\min_{x \in Q} \left\{ \psi_k(x) \equiv \sum_{i=0}^{k} [f(v_i) + \langle f(v_i), x - v_i \rangle] + \frac{M(f)}{2} \|x - x_0\|^2 \right\}$$

Theorem: Let $x_i = T(v_i)$. Then $\sum_{i=0}^{k} [f(x_i) - f^*] \leq \frac{M(f)}{2} \|x^* - x_0\|^2$.

Proof: 1. Let us prove by induction that $\sum_{i=0}^{k} f(x_i) \leq \psi_k^*$. Indeed,

$$\begin{aligned} \psi_{k+1}(x) &= \psi_k(x) + f(v_{k+1}) + \langle f(v_{k+1}), x - v_{k+1} \rangle \\ &\geq \psi_k^* + \frac{M(f)}{2} \|x - v_{k+1}\|^2 + f(v_{k+1}) + \langle f(v_{k+1}), x - v_{k+1} \rangle \end{aligned}$$

Dual Gradient Method (DGM)

$$v_{k+1} = \arg\min_{x \in Q} \left\{ \psi_k(x) \equiv \sum_{i=0}^{k} [f(v_i) + \langle f(v_i), x - v_i \rangle] + \frac{M(f)}{2} \|x - x_0\|^2 \right\}$$

Theorem: Let $x_i = T(v_i)$. Then $\sum\limits_{i=0}^{k} [f(x_i) - f^*] \leq \frac{M(f)}{2} \|x^* - x_0\|^2$.

Proof: 1. Let us prove by induction that $\sum\limits_{i=0}^{k} f(x_i) \leq \psi_k^*$. Indeed,

$$\begin{aligned}
\psi_{k+1}(x) &= \psi_k(x) + f(v_{k+1}) + \langle f(v_{k+1}), x - v_{k+1} \rangle \\
&\geq \psi_k^* + \frac{M(f)}{2} \|x - v_{k+1}\|^2 + f(v_{k+1}) + \langle f(v_{k+1}), x - v_{k+1} \rangle \\
&\geq \sum_{i=0}^{k} f(x_i) + f(x_{k+1}).
\end{aligned}$$

Dual Gradient Method (DGM)

$$v_{k+1} = \arg\min_{x \in Q} \left\{ \psi_k(x) \equiv \sum_{i=0}^{k} [f(v_i) + \langle f(v_i), x - v_i \rangle] + \frac{M(f)}{2} \|x - x_0\|^2 \right\}$$

Theorem: Let $x_i = T(v_i)$. Then $\sum_{i=0}^{k} [f(x_i) - f^*] \leq \frac{M(f)}{2} \|x^* - x_0\|^2$.

Proof: 1. Let us prove by induction that $\sum_{i=0}^{k} f(x_i) \leq \psi_k^*$. Indeed,

$$\begin{aligned} \psi_{k+1}(x) &= \psi_k(x) + f(v_{k+1}) + \langle f(v_{k+1}), x - v_{k+1} \rangle \\ &\geq \psi_k^* + \frac{M(f)}{2} \|x - v_{k+1}\|^2 + f(v_{k+1}) + \langle f(v_{k+1}), x - v_{k+1} \rangle \\ &\geq \sum_{i=0}^{k} f(x_i) + f(x_{k+1}). \end{aligned}$$

2. Note that $\psi_k^* \leq (k+1) f^* + \frac{M(f)}{2} \|x^* - x_0\|^2$. □

Dual Gradient Method (DGM)

$$v_{k+1} = \arg\min_{x \in Q} \left\{ \psi_k(x) \equiv \sum_{i=0}^{k} [f(v_i) + \langle f(v_i), x - v_i \rangle] + \frac{M(f)}{2} \|x - x_0\|^2 \right\}$$

Theorem: Let $x_i = T(v_i)$. Then $\sum_{i=0}^{k} [f(x_i) - f^*] \le \frac{M(f)}{2} \|x^* - x_0\|^2$.

Proof: 1. Let us prove by induction that $\sum_{i=0}^{k} f(x_i) \le \psi_k^*$. Indeed,

$$\begin{aligned} \psi_{k+1}(x) &= \psi_k(x) + f(v_{k+1}) + \langle f(v_{k+1}), x - v_{k+1} \rangle \\ &\ge \psi_k^* + \frac{M(f)}{2} \|x - v_{k+1}\|^2 + f(v_{k+1}) + \langle f(v_{k+1}), x - v_{k+1} \rangle \\ &\ge \sum_{i=0}^{k} f(x_i) + f(x_{k+1}). \end{aligned}$$

2. Note that $\psi_k^* \le (k+1) f^* + \frac{M(f)}{2} \|x^* - x_0\|^2$. □

This method: 1. Updates the *model*.

Dual Gradient Method (DGM)

$$v_{k+1} = \arg\min_{x \in Q} \left\{ \psi_k(x) \equiv \sum_{i=0}^{k} [f(v_i) + \langle f(v_i), x - v_i \rangle] + \frac{M(f)}{2} \|x - x_0\|^2 \right\}$$

Theorem: Let $x_i = T(v_i)$. Then $\sum_{i=0}^{k} [f(x_i) - f^*] \leq \frac{M(f)}{2} \|x^* - x_0\|^2$.

Proof: 1. Let us prove by induction that $\sum_{i=0}^{k} f(x_i) \leq \psi_k^*$. Indeed,

$$\begin{aligned} \psi_{k+1}(x) &= \psi_k(x) + f(v_{k+1}) + \langle f(v_{k+1}), x - v_{k+1} \rangle \\ &\geq \psi_k^* + \frac{M(f)}{2} \|x - v_{k+1}\|^2 + f(v_{k+1}) + \langle f(v_{k+1}), x - v_{k+1} \rangle \\ &\geq \sum_{i=0}^{k} f(x_i) + f(x_{k+1}). \end{aligned}$$

2. Note that $\psi_k^* \leq (k+1) f^* + \frac{M(f)}{2} \|x^* - x_0\|^2$. □

This method: 1. Updates the *model*. 2. Is not monotone.

Dual Gradient Method (DGM)

$$v_{k+1} = \arg\min_{x \in Q} \left\{ \psi_k(x) \equiv \sum_{i=0}^{k} [f(v_i) + \langle f(v_i), x - v_i \rangle] + \frac{M(f)}{2} \|x - x_0\|^2 \right\}$$

Theorem: Let $x_i = T(v_i)$. Then $\sum_{i=0}^{k} [f(x_i) - f^*] \leq \frac{M(f)}{2} \|x^* - x_0\|^2$.

Proof: 1. Let us prove by induction that $\sum_{i=0}^{k} f(x_i) \leq \psi_k^*$. Indeed,

$$\begin{aligned} \psi_{k+1}(x) &= \psi_k(x) + f(v_{k+1}) + \langle f(v_{k+1}), x - v_{k+1} \rangle \\ &\geq \psi_k^* + \frac{M(f)}{2} \|x - v_{k+1}\|^2 + f(v_{k+1}) + \langle f(v_{k+1}), x - v_{k+1} \rangle \\ &\geq \sum_{i=0}^{k} f(x_i) + f(x_{k+1}). \end{aligned}$$

2. Note that $\psi_k^* \leq (k+1) f^* + \frac{M(f)}{2} \|x^* - x_0\|^2$. □

This method: 1. Updates the *model*. 2. Is not monotone.
3. Does not need x_i.

Dual Gradient Method (DGM)

$$v_{k+1} = \arg\min_{x \in Q} \left\{ \psi_k(x) \equiv \sum_{i=0}^{k} [f(v_i) + \langle f(v_i), x - v_i \rangle] + \frac{M(f)}{2} \|x - x_0\|^2 \right\}$$

Theorem: Let $x_i = T(v_i)$. Then $\sum_{i=0}^{k} [f(x_i) - f^*] \leq \frac{M(f)}{2} \|x^* - x_0\|^2$.

Proof: 1. Let us prove by induction that $\sum_{i=0}^{k} f(x_i) \leq \psi_k^*$. Indeed,

$$\begin{aligned}
\psi_{k+1}(x) &= \psi_k(x) + f(v_{k+1}) + \langle f(v_{k+1}), x - v_{k+1} \rangle \\
&\geq \psi_k^* + \frac{M(f)}{2} \|x - v_{k+1}\|^2 + f(v_{k+1}) + \langle f(v_{k+1}), x - v_{k+1} \rangle \\
&\geq \sum_{i=0}^{k} f(x_i) + f(x_{k+1}).
\end{aligned}$$

2. Note that $\psi_k^* \leq (k+1) f^* + \frac{M(f)}{2} \|x^* - x_0\|^2$. □

This method: 1. Updates the *model*. 2. Is not monotone.
3. Does not need x_i. 4. Has the *same* efficiency as PGM.

Dual Gradient Method (DGM)

$$v_{k+1} = \arg\min_{x \in Q} \left\{ \psi_k(x) \equiv \sum_{i=0}^{k} [f(v_i) + \langle f(v_i), x - v_i \rangle] + \frac{M(f)}{2} \|x - x_0\|^2 \right\}$$

Theorem: Let $x_i = T(v_i)$. Then $\sum_{i=0}^{k} [f(x_i) - f^*] \leq \frac{M(f)}{2} \|x^* - x_0\|^2$.

Proof: 1. Let us prove by induction that $\sum_{i=0}^{k} f(x_i) \leq \psi_k^*$. Indeed,

$$\begin{aligned} \psi_{k+1}(x) &= \psi_k(x) + f(v_{k+1}) + \langle f(v_{k+1}), x - v_{k+1} \rangle \\ &\geq \psi_k^* + \frac{M(f)}{2} \|x - v_{k+1}\|^2 + f(v_{k+1}) + \langle f(v_{k+1}), x - v_{k+1} \rangle \\ &\geq \sum_{i=0}^{k} f(x_i) + f(x_{k+1}). \end{aligned}$$

2. Note that $\psi_k^* \leq (k+1) f^* + \frac{M(f)}{2} \|x^* - x_0\|^2$. □

This method: 1. Updates the *model*. 2. Is not monotone.
3. Does not need x_i. 4. Has the *same* efficiency as PGM.

Interpretation:

Dual Gradient Method (DGM)

$$v_{k+1} = \arg\min_{x \in Q} \left\{ \psi_k(x) \equiv \sum_{i=0}^{k} [f(v_i) + \langle f(v_i), x - v_i \rangle] + \frac{M(f)}{2} \|x - x_0\|^2 \right\}$$

Theorem: Let $x_i = T(v_i)$. Then $\sum_{i=0}^{k} [f(x_i) - f^*] \le \frac{M(f)}{2} \|x^* - x_0\|^2$.

Proof: 1. Let us prove by induction that $\sum_{i=0}^{k} f(x_i) \le \psi_k^*$. Indeed,

$$\begin{aligned} \psi_{k+1}(x) &= \psi_k(x) + f(v_{k+1}) + \langle f(v_{k+1}), x - v_{k+1} \rangle \\ &\ge \psi_k^* + \frac{M(f)}{2} \|x - v_{k+1}\|^2 + f(v_{k+1}) + \langle f(v_{k+1}), x - v_{k+1} \rangle \\ &\ge \sum_{i=0}^{k} f(x_i) + f(x_{k+1}). \end{aligned}$$

2. Note that $\psi_k^* \le (k+1) f^* + \frac{M(f)}{2} \|x^* - x_0\|^2$. $\square$

This method: 1. Updates the *model*. 2. Is not monotone.
3. Does not need x_i. 4. Has the *same* efficiency as PGM.

Interpretation: Academic Science.

Dual Gradient Method (DGM)

$$v_{k+1} = \arg\min_{x \in Q} \left\{ \psi_k(x) \equiv \sum_{i=0}^{k} [f(v_i) + \langle f(v_i), x - v_i \rangle] + \frac{M(f)}{2} \|x - x_0\|^2 \right\}$$

Theorem: Let $x_i = T(v_i)$. Then $\sum_{i=0}^{k} [f(x_i) - f^*] \leq \frac{M(f)}{2} \|x^* - x_0\|^2$.

Proof: 1. Let us prove by induction that $\sum_{i=0}^{k} f(x_i) \leq \psi_k^*$. Indeed,

$$\begin{aligned} \psi_{k+1}(x) &= \psi_k(x) + f(v_{k+1}) + \langle f(v_{k+1}), x - v_{k+1} \rangle \\ &\geq \psi_k^* + \frac{M(f)}{2} \|x - v_{k+1}\|^2 + f(v_{k+1}) + \langle f(v_{k+1}), x - v_{k+1} \rangle \\ &\geq \sum_{i=0}^{k} f(x_i) + f(x_{k+1}). \end{aligned}$$

2. Note that $\psi_k^* \leq (k+1) f^* + \frac{M(f)}{2} \|x^* - x_0\|^2$. □

This method: 1. Updates the *model*. 2. Is not monotone.
3. Does not need x_i. 4. Has the *same* efficiency as PGM.

Interpretation: Academic Science.

Can we combine the primal and dual strategy?

Fast Gradient Methods (N. 1984-2005) (Assume $Q \equiv R^n$)

Fast Gradient Methods (N. 1984-2005) (Assume $Q \equiv R^n$)

Estimate sequences: $\{\psi_k(x)\}$, $\{a_k\}$, $\{x_k\}$ such that

Fast Gradient Methods (N. 1984-2005) (Assume $Q \equiv R^n$)

Estimate sequences: $\{\psi_k(x)\}$, $\{a_k\}$, $\{x_k\}$ such that

$$\psi_k(x) \;=\; \sum_{i=0}^{k} a_i [f(y_i) + \langle \nabla f(y_i), x - y_i \rangle] + \frac{1}{2}\|x - x_0\|^2.$$

Fast Gradient Methods (N. 1984-2005) (Assume $Q \equiv R^n$)

Estimate sequences: $\{\psi_k(x)\}$, $\{a_k\}$, $\{x_k\}$ such that

$$\begin{aligned} \psi_k(x) &= \sum_{i=0}^{k} a_i[f(y_i) + \langle \nabla f(y_i), x - y_i \rangle] + \tfrac{1}{2}\|x - x_0\|^2. \\ A_k f(x_k) &\leq \psi_k^* \stackrel{\text{def}}{=} \min_{x \in R^n} \psi_k(x). \quad \left(A_k = \sum_{i=0}^{k} a_i. \right) \end{aligned}$$

Fast Gradient Methods (N. 1984-2005) (Assume $Q \equiv R^n$)

Estimate sequences: $\{\psi_k(x)\}$, $\{a_k\}$, $\{x_k\}$ such that

$$\begin{aligned} \psi_k(x) &= \sum_{i=0}^{k} a_i[f(y_i) + \langle \nabla f(y_i), x - y_i \rangle] + \tfrac{1}{2}\|x - x_0\|^2. \\ A_k f(x_k) &\leq \psi_k^* \stackrel{\text{def}}{=} \min_{x \in R^n} \psi_k(x). \quad \left(A_k = \sum_{i=0}^{k} a_i. \right) \end{aligned}$$

Note: $\psi_k(x^*) \leq A_k f^* + \frac{1}{2}\|x^* - x_0\|^2 \Rightarrow f(x_k) - f^* \leq \frac{\|x^* - x_0\|^2}{2A_k}$.

Fast Gradient Methods (N. 1984-2005) (Assume $Q \equiv R^n$)

Estimate sequences: $\{\psi_k(x)\}$, $\{a_k\}$, $\{x_k\}$ such that

$$\begin{aligned} \psi_k(x) &= \sum_{i=0}^{k} a_i[f(y_i) + \langle \nabla f(y_i), x - y_i \rangle] + \tfrac{1}{2}\|x - x_0\|^2. \\ A_k f(x_k) &\leq \psi_k^* \stackrel{\text{def}}{=} \min_{x \in R^n} \psi_k(x). \quad \left(A_k = \sum_{i=0}^{k} a_i. \right) \end{aligned}$$

Note: $\psi_k(x^*) \leq A_k f^* + \frac{1}{2}\|x^* - x_0\|^2 \Rightarrow f(x_k) - f^* \leq \frac{\|x^* - x_0\|^2}{2A_k}$.

Main properties: Let $v_k = \arg\min_{x \in R^n} \psi_k(x)$.

Fast Gradient Methods (N. 1984-2005) (Assume $Q \equiv R^n$)

Estimate sequences: $\{\psi_k(x)\}$, $\{a_k\}$, $\{x_k\}$ such that

$$\begin{aligned} \psi_k(x) &= \sum_{i=0}^{k} a_i[f(y_i) + \langle \nabla f(y_i), x - y_i \rangle] + \tfrac{1}{2}\|x - x_0\|^2. \\ A_k f(x_k) &\leq \psi_k^* \stackrel{\text{def}}{=} \min_{x \in R^n} \psi_k(x). \quad \left(A_k = \sum_{i=0}^{k} a_i. \right) \end{aligned}$$

Note: $\psi_k(x^*) \leq A_k f^* + \frac{1}{2}\|x^* - x_0\|^2 \Rightarrow f(x_k) - f^* \leq \frac{\|x^* - x_0\|^2}{2A_k}$.

Main properties: Let $v_k = \arg\min_{x \in R^n} \psi_k(x)$.

If $y_{k+1} = \frac{a_{k+1} v_k + A_k x_k}{a_{k+1} + A_k}$, then $\psi_{k+1}^* \geq A_{k+1} f(y_{k+1}) - \frac{a_{k+1}^2 \|\nabla f(y_{k+1})\|_*^2}{2}$.

Fast Gradient Methods (N. 1984-2005) (Assume $Q \equiv R^n$)

Estimate sequences: $\{\psi_k(x)\}$, $\{a_k\}$, $\{x_k\}$ such that

$$\psi_k(x) = \sum_{i=0}^{k} a_i[f(y_i) + \langle \nabla f(y_i), x - y_i \rangle] + \tfrac{1}{2}\|x - x_0\|^2.$$

$$A_k f(x_k) \leq \psi_k^* \stackrel{\mathrm{def}}{=} \min_{x \in R^n} \psi_k(x). \qquad \left(A_k = \sum_{i=0}^{k} a_i. \right)$$

Note: $\psi_k(x^*) \leq A_k f^* + \frac{1}{2}\|x^* - x_0\|^2 \Rightarrow f(x_k) - f^* \leq \frac{\|x^* - x_0\|^2}{2A_k}$.

Main properties: Let $v_k = \arg\min_{x \in R^n} \psi_k(x)$.

If $y_{k+1} = \frac{a_{k+1} v_k + A_k x_k}{a_{k+1} + A_k}$, then $\psi_{k+1}^* \geq A_{k+1} f(y_{k+1}) - \frac{a_{k+1}^2 \|\nabla f(y_{k+1})\|_*^2}{2}$.

If $x_{k+1} = y_{k+1} - \frac{\nabla f(y_{k+1})}{M(f)}$, then $f(y_{k+1}) \geq f(x_{k+1}) + \frac{\|\nabla f(y_{k+1})\|_*^2}{2M(f)}$.

Algorithmic Scheme

1. Compute $v_k = \arg\min\limits_{x \in R^n} \psi_k(x)$.
2. Assume that $f(x_k) \leq \frac{1}{A_k}\psi_k(v_k)$.
3. Find a_{k+1} : $\frac{a_{k+1}^2}{a_{k+1}+A_k} = \frac{1}{M(f)}$.
4. Define $y_{k+1} = \frac{a_{k+1}v_k + A_k x_k}{a_{k+1}+A_k}$.
5. Compute $x_{k+1} = y_{k+1} - \frac{\nabla f(y_{k+1})}{M(f)}$.

$$\Rightarrow \quad f(x_{k+1}) \leq \frac{1}{A_{k+1}}\psi_{k+1}^*.$$

Algorithmic Scheme

1. Compute $v_k = \arg\min\limits_{x \in R^n} \psi_k(x)$.
2. Assume that $f(x_k) \leq \frac{1}{A_k}\psi_k(v_k)$.
3. Find a_{k+1} : $\frac{a_{k+1}^2}{a_{k+1}+A_k} = \frac{1}{M(f)}$.
4. Define $y_{k+1} = \frac{a_{k+1}v_k + A_k x_k}{a_{k+1}+A_k}$.
5. Compute $x_{k+1} = y_{k+1} - \frac{\nabla f(y_{k+1})}{M(f)}$.

$\Rightarrow \quad f(x_{k+1}) \leq \frac{1}{A_{k+1}}\psi_{k+1}^*$.

Note: $a(t) \approx A'(t)$. Hence, $A'(t) = \left(\frac{A(t)}{M(f)}\right)^{1/2} \Rightarrow A(t) \approx \frac{t^2}{4M(f)}$.

Algorithmic Scheme

1. Compute $v_k = \arg\min\limits_{x \in R^n} \psi_k(x)$.
2. Assume that $f(x_k) \leq \frac{1}{A_k}\psi_k(v_k)$.
3. Find a_{k+1} : $\frac{a_{k+1}^2}{a_{k+1}+A_k} = \frac{1}{M(f)}$.
4. Define $y_{k+1} = \frac{a_{k+1}v_k + A_k x_k}{a_{k+1}+A_k}$.
5. Compute $x_{k+1} = y_{k+1} - \frac{\nabla f(y_{k+1})}{M(f)}$.

$\Rightarrow \quad f(x_{k+1}) \leq \frac{1}{A_{k+1}}\psi^*_{k+1}$.

Note: $a(t) \approx A'(t)$. Hence, $A'(t) = \left(\frac{A(t)}{M(f)}\right)^{1/2} \Rightarrow A(t) \approx \frac{t^2}{4M(f)}$.

Interpretation:

Algorithmic Scheme

1. Compute $v_k = \arg\min_{x \in R^n} \psi_k(x)$.
2. Assume that $f(x_k) \leq \frac{1}{A_k}\psi_k(v_k)$.
3. Find $a_{k+1}: \frac{a_{k+1}^2}{a_{k+1}+A_k} = \frac{1}{M(f)}$.
4. Define $y_{k+1} = \frac{a_{k+1}v_k + A_k x_k}{a_{k+1}+A_k}$.
5. Compute $x_{k+1} = y_{k+1} - \frac{\nabla f(y_{k+1})}{M(f)}$.

$\Rightarrow \quad f(x_{k+1}) \leq \frac{1}{A_{k+1}}\psi_{k+1}^*$.

Note: $a(t) \approx A'(t)$. Hence, $A'(t) = \left(\frac{A(t)}{M(f)}\right)^{1/2} \Rightarrow A(t) \approx \frac{t^2}{4M(f)}$.

Interpretation: Efficient collaboration of Theory and Practice organized by the wise government.

Algorithmic Scheme

1. Compute $v_k = \arg\min_{x \in R^n} \psi_k(x)$.
2. Assume that $f(x_k) \leq \frac{1}{A_k}\psi_k(v_k)$.
3. Find a_{k+1} : $\frac{a_{k+1}^2}{a_{k+1}+A_k} = \frac{1}{M(f)}$.
4. Define $y_{k+1} = \frac{a_{k+1}v_k + A_k x_k}{a_{k+1}+A_k}$.
5. Compute $x_{k+1} = y_{k+1} - \frac{\nabla f(y_{k+1})}{M(f)}$.

$\Rightarrow \quad f(x_{k+1}) \leq \frac{1}{A_{k+1}}\psi_{k+1}^*$.

Note: $a(t) \approx A'(t)$. Hence, $A'(t) = \left(\frac{A(t)}{M(f)}\right)^{1/2} \Rightarrow A(t) \approx \frac{t^2}{4M(f)}$.

Interpretation: Efficient collaboration of Theory and Practice organized by the wise government.

Expected outcome: achieve the maximal performance in 10 years instead of 100.

Third ring: Minimization of Composite Functions (2007)

Third ring: Minimization of Composite Functions (2007)

Problem formulation: $\min\{\phi(x) \stackrel{\mathrm{def}}{=} f(x) + \Psi(x) : x \in R^n\}$,

- function f is differentiable ($f \in \mathcal{C}_2$),
- function Ψ is closed and convex on R^n.

Third ring: Minimization of Composite Functions (2007)

Problem formulation: $\min\{\phi(x) \stackrel{\mathrm{def}}{=} f(x) + \Psi(x) : x \in R^n\}$,

- function f is differentiable ($f \in \mathcal{C}_2$),
- function Ψ is closed and convex on R^n.

Note: in general $f + \Psi \notin \mathcal{C}_1$.
(No complexity bounds in BB-framework!)

Third ring: Minimization of Composite Functions (2007)

Problem formulation: $\min\{\phi(x) \stackrel{\text{def}}{=} f(x) + \Psi(x) : x \in R^n\}$,

- function f is differentiable ($f \in \mathcal{C}_2$),
- function Ψ is closed and convex on R^n.

Note: in general $f + \Psi \notin \mathcal{C}_1$.
(No complexity bounds in BB-framework!)

Examples:

- $\Psi(x) = \begin{cases} 0, & \text{if } x \in Q, \\ +\infty, & \text{otherwise.} \end{cases}$

Third ring: Minimization of Composite Functions (2007)

Problem formulation: $\min\{\phi(x) \stackrel{\mathrm{def}}{=} f(x) + \Psi(x) : x \in R^n\}$,

- function f is differentiable ($f \in \mathcal{C}_2$),
- function Ψ is closed and convex on R^n.

Note: in general $f + \Psi \notin \mathcal{C}_1$.
(No complexity bounds in BB-framework!)

Examples:

- $\Psi(x) = \begin{cases} 0, & \text{if } x \in Q, \\ +\infty, & \text{otherwise.} \end{cases}$
- Ψ is a barrier function for Q.

Third ring: Minimization of Composite Functions (2007)

Problem formulation: $\min\{\phi(x) \stackrel{\text{def}}{=} f(x) + \Psi(x) : x \in R^n\}$,

- function f is differentiable ($f \in \mathcal{C}_2$),
- function Ψ is closed and convex on R^n.

Note: in general $f + \Psi \notin \mathcal{C}_1$.
(No complexity bounds in BB-framework!)

Examples:

- $\Psi(x) = \begin{cases} 0, & \text{if } x \in Q, \\ +\infty, & \text{otherwise.} \end{cases}$
- Ψ is a barrier function for Q.
- Ψ is a simple nonsmooth function (e.g. $\|x\|_1$).

Third ring: Minimization of Composite Functions (2007)

Problem formulation: $\min\{\phi(x) \stackrel{\text{def}}{=} f(x) + \Psi(x) : x \in R^n\}$,

- function f is differentiable ($f \in \mathcal{C}_2$),
- function Ψ is closed and convex on R^n.

Note: in general $f + \Psi \notin \mathcal{C}_1$.
(No complexity bounds in BB-framework!)

Examples:

- $\Psi(x) = \begin{cases} 0, & \text{if } x \in Q, \\ +\infty, & \text{otherwise.} \end{cases}$
- Ψ is a barrier function for Q.
- Ψ is a simple nonsmooth function (e.g. $\|x\|_1$).

Main Assumption: The problem $\min_x [q(x) + \Psi(x)]$ is easy.
(q is a "simple" quadratic function.)

Modified tools: Composite Gradient Mapping

For any $y \in \mathrm{dom}\,\Psi$ define

$$
\begin{aligned}
m_L(y;x) &= f(y) + \langle \nabla f(y), x - y \rangle + \tfrac{L}{2}\|x - y\|^2 + \Psi(x), \\
T_L(y) &= \arg\min_{x \in R^n} \; m_L(y;x),
\end{aligned}
$$

where L is a positive constant.

Modified tools: Composite Gradient Mapping

For any $y \in \mathrm{dom}\,\Psi$ define

$$
\begin{aligned}
m_L(y;x) &= f(y) + \langle \nabla f(y), x-y \rangle + \tfrac{L}{2}\|x-y\|^2 + \Psi(x), \\
T_L(y) &= \arg\min_{x \in R^n} \; m_L(y;x),
\end{aligned}
$$

where L is a positive constant. Then the direction

$$
g_L(y) = L \cdot (y - T_L(y))
$$

is a constrained analogue of the gradient of smooth function.

Modified tools: Composite Gradient Mapping

For any $y \in \mathrm{dom}\,\Psi$ define

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

$$T_L(y) = \arg\min_{x \in R^n} m_L(y;x),$$

where L is a positive constant. Then the direction

$$g_L(y) = L \cdot (y - T_L(y))$$

is a constrained analogue of the gradient of smooth function.

Main property: If $L \geq M(f)$ then

$$\phi(y) - \phi(T_L(y)) \geq \frac{1}{2L}\|g_L(y)\|^2.$$

Basic Gradient Methods for Composite Functions

Basic Gradient Methods for Composite Functions

Primal Gradient Method

Consider the method: $x_{k+1} = T_{M(f)}(x_k)$, $k \geq 0$. Then

$$\sum_{i=0}^{k} [\phi(x_i) - \phi^*] \leq 2M(f)\|x^* - x_0\|^2.$$

Basic Gradient Methods for Composite Functions

Primal Gradient Method

Consider the method: $x_{k+1} = T_{M(f)}(x_k)$, $k \geq 0$. Then

$$\sum_{i=0}^{k} [\phi(x_i) - \phi^*] \;\leq\; 2M(f)\|x^* - x_0\|^2.$$

Dual Gradient Method. Consider the method

$$v_{k+1} = \arg\min_{x \in R^n} \{ \hat{\psi}_k(x) \;\equiv\; \sum_{i=0}^{k} [f(v_i) + \langle f(v_i), x - v_i \rangle + \Psi(x)] + \frac{M(f)}{2}\|x - x_0\|^2 \}.$$

Define $x_i = T_{M(f)}(v_i)$. Then $\sum_{i=0}^{k} [\phi(x_i) - \phi^*] \leq 2M(f)\|x^* - x_0\|^2$.

Basic Gradient Methods for Composite Functions

Primal Gradient Method

Consider the method: $x_{k+1} = T_{M(f)}(x_k)$, $k \geq 0$. Then

$$\sum_{i=0}^{k} [\phi(x_i) - \phi^*] \quad \leq \quad 2M(f)\|x^* - x_0\|^2.$$

Dual Gradient Method. Consider the method

$$v_{k+1} = \arg\min_{x \in R^n} \{ \hat{\psi}_k(x) \quad \equiv \quad \sum_{i=0}^{k} [f(v_i) + \langle f(v_i), x - v_i \rangle + \Psi(x)] + \frac{M(f)}{2} \|x - x_0\|^2 \}.$$

Define $x_i = T_{M(f)}(v_i)$. Then $\sum_{i=0}^{k} [\phi(x_i) - \phi^*] \leq 2M(f)\|x^* - x_0\|^2$.

Same as for $\Psi \equiv 0$!

Fast gradient methods for Composite Functions

Fast gradient methods for Composite Functions

Main change: New definition of the model

$$\hat{\psi}_k(x) \quad = \quad \sum_{i=0}^{k} a_i[f(x_i) + \langle f(x_i), x - x_i\rangle + \Psi(x)] + \tfrac{1}{2}\|x - x_0\|^2.$$

Fast gradient methods for Composite Functions

Main change: New definition of the model

$$\hat{\psi}_k(x) \quad = \quad \sum_{i=0}^{k} a_i[f(x_i) + \langle f(x_i), x - x_i \rangle + \Psi(x)] + \frac{1}{2}\|x - x_0\|^2.$$

The scheme becomes as follows:

1. Compute $v_k = \arg\min\limits_{x \in R^n} \hat{\psi}_k(x)$.
2. Compute a_k from equation $\frac{a_k^2}{A_k + a_k} = \frac{2}{M(f)}$.
3. Define $y_k = \frac{A_k x_k + a_k v_k}{A_k + a_k}$ and compute $x_{k+1} = T_{M(f)}(y_k)$.

Fast gradient methods for Composite Functions

Main change: New definition of the model

$$\hat{\psi}_k(x) = \sum_{i=0}^{k} a_i[f(x_i) + \langle f(x_i), x - x_i\rangle + \Psi(x)] + \frac{1}{2}\|x - x_0\|^2.$$

The scheme becomes as follows:

1. Compute $v_k = \arg\min_{x \in R^n} \hat{\psi}_k(x)$.
2. Compute a_k from equation $\frac{a_k^2}{A_k + a_k} = \frac{2}{M(f)}$.
3. Define $y_k = \frac{A_k x_k + a_k v_k}{A_k + a_k}$ and compute $x_{k+1} = T_{M(f)}(y_k)$.

Rate of convergence: $\phi(x_k) - \phi^* \leq \frac{2M(f)\|x_0 - x^*\|^2}{(k+1)^2}$.

Fast gradient methods for Composite Functions

Main change: New definition of the model

$$\hat{\psi}_k(x) \;=\; \sum_{i=0}^{k} a_i[f(x_i) + \langle f(x_i), x - x_i\rangle + \Psi(x)] + \frac{1}{2}\|x - x_0\|^2.$$

The scheme becomes as follows:

1. Compute $v_k = \arg\min_{x \in R^n} \hat{\psi}_k(x)$.
2. Compute a_k from equation $\frac{a_k^2}{A_k + a_k} = \frac{2}{M(f)}$.
3. Define $y_k = \frac{A_k x_k + a_k v_k}{A_k + a_k}$ and compute $x_{k+1} = T_{M(f)}(y_k)$.

Rate of convergence: $\phi(x_k) - \phi^* \le \frac{2M(f)\|x_0 - x^*\|^2}{(k+1)^2}$.

Example: $\phi(x) = \frac{1}{2}\|Ax - b\|^2 + \sum_{i=1}^{n} |x^{(i)}|$.

Conclusion

Conclusion

For breaking the BB-limitations, optimization methods need help!

Conclusion

For breaking the BB-limitations, optimization methods need help!

Possible approaches:

- **Interior-point methods.** Rewrite the problem in a standard form. Construct the s.c.barrier. **Complexity:** $O\left(\sqrt{\nu}\ln\frac{1}{\epsilon}\right)$.

Conclusion

For breaking the BB-limitations, optimization methods need help!

Possible approaches:

- **Interior-point methods.** Rewrite the problem in a standard form. Construct the s.c.barrier. **Complexity:** $O\left(\sqrt{\nu}\ln\frac{1}{\epsilon}\right)$.
- **Smoothing technique.** Find a reasonable max-representation of the objective with computable smooth approximation. **Complexity:** $O\left(\frac{1}{\epsilon}\right)$.

Conclusion

For breaking the BB-limitations, optimization methods need help!

Possible approaches:

- **Interior-point methods.** Rewrite the problem in a standard form. Construct the s.c.barrier. **Complexity:** $O\left(\sqrt{\nu}\ln\frac{1}{\epsilon}\right)$.
- **Smoothing technique.** Find a reasonable max-representation of the objective with computable smooth approximation. **Complexity:** $O\left(\frac{1}{\epsilon}\right)$.
- **Composite function.** Find a possibility to minimize a bad part of the objective. **Complexity:** $O\left(\frac{1}{\epsilon^{1/2}}\right)$.

Conclusion

For breaking the BB-limitations, optimization methods need help!

Possible approaches:

- **Interior-point methods.** Rewrite the problem in a standard form. Construct the s.c.barrier. **Complexity:** $O\left(\sqrt{\nu}\ln\frac{1}{\epsilon}\right)$.
- **Smoothing technique.** Find a reasonable max-representation of the objective with computable smooth approximation. **Complexity:** $O\left(\frac{1}{\epsilon}\right)$.
- **Composite function.** Find a possibility to minimize a bad part of the objective. **Complexity:** $O\left(\frac{1}{\epsilon^{1/2}}\right)$.

And some others!

How to use the structure?

How to use the structure?

1. Fix analytical form of functional components. (Too fragile!)

How to use the structure?

1. Fix analytical form of functional components. (Too fragile!)

2. Reformulations.

$$\mathcal{P} \longrightarrow \ldots \longrightarrow (f^*, x^*).$$

How to use the structure?

1. Fix analytical form of functional components. (Too fragile!)

2. Reformulations.

$$\mathcal{P} \longrightarrow \ldots \longrightarrow (f^*, x^*).$$

Classical example: Cholesky decomposition

For solving the linear system $Ax = b$, we proceed as follows:

How to use the structure?

1. Fix analytical form of functional components. (Too fragile!)

2. Reformulations.

$$\mathcal{P} \longrightarrow \dots \longrightarrow (f^*, x^*).$$

Classical example: Cholesky decomposition

For solving the linear system $Ax = b$, we proceed as follows:

1. Check if A is symmetric and positive definite.

How to use the structure?

1. Fix analytical form of functional components. (Too fragile!)

2. Reformulations.

$$\mathcal{P} \longrightarrow \ldots \longrightarrow (f^*, x^*).$$

Classical example: Cholesky decomposition

For solving the linear system $Ax = b$, we proceed as follows:

1. Check if A is symmetric and positive definite.
2. Compute Cholesky factorization of this matrix: $A = LL^T$, where L is a lower-triangular matrix.

How to use the structure?

1. Fix analytical form of functional components. (Too fragile!)

2. Reformulations.

$$\mathcal{P} \longrightarrow \ldots \longrightarrow (f^*, x^*).$$

Classical example: Cholesky decomposition

For solving the linear system $Ax = b$, we proceed as follows:

1. Check if A is symmetric and positive definite.
2. Compute Cholesky factorization of this matrix: $A = LL^T$, where L is a lower-triangular matrix.
3. Solve the systems $Ly = b$, $L^T x = y$ by sequential elimination of variables.

How to discover a new method?

How to discover a new method?

Golden Rules

How to discover a new method?

Golden Rules

- Find a class of problems which can be solved very efficiently. (e.g. the class of linear systems with triangular matrices.)

How to discover a new method?

Golden Rules

- Find a class of problems which can be solved very efficiently. (e.g. the class of linear systems with triangular matrices.)
- Describe the transformation rules for converting the initial problem into desired form.

How to discover a new method?

Golden Rules

- Find a class of problems which can be solved very efficiently. (e.g. the class of linear systems with triangular matrices.)
- Describe the transformation rules for converting the initial problem into desired form.
- Describe the class of problems for which these transformation rules are applicable.

How to discover a new method?

Golden Rules

- Find a class of problems which can be solved very efficiently. (e.g. the class of linear systems with triangular matrices.)
- Describe the transformation rules for converting the initial problem into desired form.
- Describe the class of problems for which these transformation rules are applicable.

We have seen how it works for

- **IPM:** Newton for s.c.functions + rules for constructing s.c.b.

How to discover a new method?

Golden Rules

- Find a class of problems which can be solved very efficiently. (e.g. the class of linear systems with triangular matrices.)
- Describe the transformation rules for converting the initial problem into desired form.
- Describe the class of problems for which these transformation rules are applicable.

We have seen how it works for

- **IPM:** Newton for s.c.functions + rules for constructing s.c.b.
- **Smoothing:** Fast GM + max representation.

How to discover a new method?

Golden Rules

- Find a class of problems which can be solved very efficiently. (e.g. the class of linear systems with triangular matrices.)
- Describe the transformation rules for converting the initial problem into desired form.
- Describe the class of problems for which these transformation rules are applicable.

We have seen how it works for

- **IPM:** Newton for s.c.functions + rules for constructing s.c.b.
- **Smoothing:** Fast GM + max representation.
- **Composite functions:** Fast GM + exact minimization of difficult parts of the objective.